Science Bots:
a Model for the Future of Scientific Computation?

Tobias Kuhn
Department of Humanities, Social and Political Sciences, ETH Zurich, Switzerland
tokuhn@ethz.ch

ABSTRACT

As a response to the trends of the increasing importance of computational approaches and the accelerating pace in science, I propose in this position paper to establish the concept of "science bots" that autonomously perform programmed tasks on input data they encounter and immediately publish the results. We can let such bots participate in a reputation system together with human users, meaning that bots and humans get positive or negative feedback by other participants. Positive reputation given to these bots would also shine on their owners, motivating them to contribute to this system, while negative reputation will allow us to filter out low-quality data, which is inevitable in an open and decentralized system.

Categories and Subject Descriptors

K.4.2 [**Computers and Society**]: Organizational Impacts—*Automation*

Keywords

autonomous agent, bot, nanopublication, scholarly communication, semantic publishing

1. INTRODUCTION

As datasets become increasingly important in all branches of science, many have proposed methods and tools to publish data [3] [12] [1]. Nanopublications [5] are an approach to bundle atomic data snippets (in RDF) in small packages together with their provenance and metadata. Such nanopublications can be manually created by scientists and linked to their articles, but they can also be automatically extracted from existing datasets or be directly created by programs that implement scientific methods.

In general, computer programs form a third kind of scientific contribution, besides narrative articles and datasets. While many such programs are openly available, there are no conventions or standards of how to reliably link data to the software that produced it, including the version of the software and the input it received. Moreover, due to the focus of the scientific life cycle on the publication of articles, scientific software is typically applied only to the data available at the time of writing a paper. It is often not the case that new output data is made public when new input data becomes available. To tackle these problems, I argue that we can encapsulate certain types of scientific algorithms as small independent agents that take inputs of a given type and produce, for example, nanopublications, and they could do this in a real-time and automatic manner as new input data becomes available.

2. BACKGROUND

I borrow the term "bot" from Wikipedia, where bots are applied, for example, to revert edits that are the results of vandalism [7] [14]. A prominent example is a bot that has created around 454 000 articles for the Swedish Wikipedia [6]. The fact that bots can be powerful also in a negative sense has become apparent with the rise of botnets [13], and with the increasing problem of ”social bots” that pretend to be humans [4]. I argue here that the power of bots could also be harnessed in a positive way for scientific computation. In contrast to the agents in the original Semantic Web paper [2], such bots would not propose or make decisions as a kind of personal assistant, but they would only publish data snippets and they would do that without any further interaction with humans.

In previous work, I showed how the concept of nanopublications can be extended and I mentioned the possible use of bots to create them [9]. I also presented an approach to attach cryptographic hash values to nanopublication identifiers to make them verifiable and immutable [11]. Based on that work, I have started to establish a nanopublication server network, with which nanopublications can be published, retrieved, and archived in a reliable, trustworthy, and decentralized manner [10], which could serve as the basis for the communication for bots.

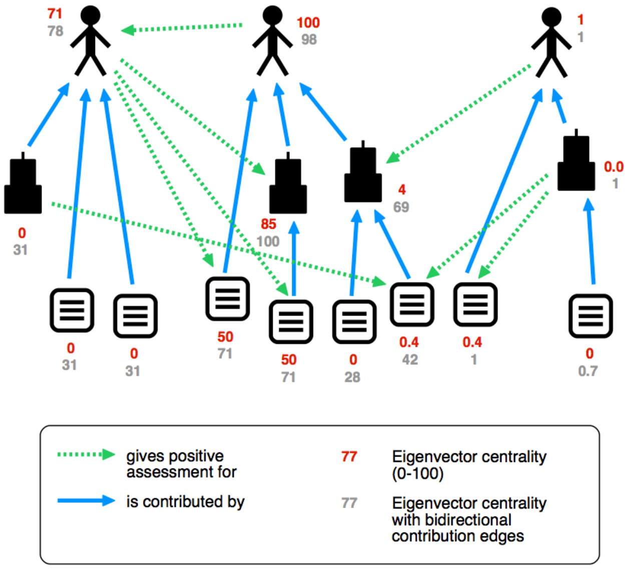
3. APPROACH

In this position paper, I propose bots as a general concept for scientific computation. For example, a bot could apply text mining to extract relations from the abstracts of the constantly growing PubMed database, another bot could regularly measure the temperature at a given location and publish the results, and yet another one could infer new facts from existing nanopublications by applying specified rules or heuristics (e.g. if disease X is related to gene Y, which is targeted by drug Z then Z might help to treat X). Importantly, these bots can automatically publish the obtained data without double-checking or direct supervision by their creators, and these data can be made immediately accessible to everybody (including other bots).

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742014.

In a system that treats bots as first-class citizens, we have to expect that some bots (and humans for that matter) will produce low-quality contributions, and we have to make sure that this does not affect the reliability and trustworthiness of the system. I argue that we can achieve that without introducing a central authority, without making concessions with respect to the openness of the system, and without delaying the publishing of results. We simply need a sufficiently accurate automatic method to discern good contributions from bad ones, which can be achieved by a reputation system. We can let scientists and bots participate in the same reputation system, where they would increase their reputation by receiving positive feedback by other participants on the usefulness and quality of their contributions. Positive reputation of a bot, in turn, would give credit and reputation to the scientist who created it.

To arrive at a simple exemplary model to explain the approach, we can define a relation ”is contributed by", where bots can occur on either side: They are contributions, as they were programmed and created by somebody, but they are also contributors, as they can create new digital entities on their own. We can define a second type of relation to represent assessments. For the sake of simplicity, we model here only positive assessments and strip them of all granularity and detail, and we can call the resulting relation “gives positive assessment for”.



71
78
100
98
1
1
0
31
85
100
4
69
0.0
1
0
31
0
31
50
71
50
71
0
28
0.4
42
0.4
1
0
0.7
gives positive assessment for
is contributed by
77
Eigenvector centrality (0-100)
77
Eigenvector centrality with bidirectional contribution edges

Figure 1: A simple example of a graph of contributors and contributions with edges for creatorship and assessments.

Fig. 1 shows a simple example of such a graph with two kinds of edges, representing creatorship and assessments. To determine the reputation or importance of the nodes, we can in the simplest case treat the two types of edges identically and rank the nodes by applying a network measure such as Eigenvector centrality (which is closely related to Google's PageRank algorithm to rank websites), as shown by the red numbers. The person at the top-left has a high reputation because he is endorsed by the person in the middle. The latter has a high reputation because her direct and indirect contributions were positively assessed by others (even though she has not received a direct assessment herself). The third person to the right, however, has not contributed anything that was positively assessed by others (only by his own bot), and therefore his reputation is low. Of course, there are many possible variations and extensions, such as bidirectional contribution edges for the Eigenvector calculation, as indicated by the gray numbers. In general, as one cannot influence *incoming* links from the part of the network that is not under one's control, there is no way to efficiently game the system. The scalability of such algorithms in open and decentralized systems is demonstrated by their successful application by search engines and peer-to-peer systems [8].

Bots could free scientists from routine tasks and therefore allow them to focus on the interesting questions. Furthermore, this approach could increase the value and appreciation of datasets and software as research products, and give due credit to their creators. With appropriate reputation mechanisms, this can be achieved in a fully open and decentralized environment.

4. REFERENCES

[1] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, I. Buchan. Research objects: Towards exchange and reuse of digital knowledge. *FWCS 2010*.

[2] T. Berners-Lee, J. Hendler, O. Lassila. The Semantic Web . *Scientific American* 284:5.

[3] J. Brase. DataCite - A global registration agency for research data. *COINFO 2009*.

[4] E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini. The Rise of Social Bots. arXiv preprint 1407.5225, 2014.

[5] P. Groth, A. Gibson, J. Velterop. The anatomy of a nanopublication. *Information Services and Use* 30:1, 2010.

[6] L. Guldbrandsson. Swedish Wikipedia surpasses 1 million articles with aid of article creation bot. *Wikimedia blog*, 17 June 2013.

[7] A. Halfaker, J. Riedl. Bots and Cyborgs: Wikipedia's Immune System. *Social Computing* 45:3. 2012.

[8] S. D. Kamvar, M. T. Schlosser, H. Garcia-Molina. The EigenTrust algorithm for reputation management in P2P networks. *WWW 2003*.

[9] T. Kuhn, P. E. Barbano, M. Levente Nagy, M. Krauthammer. Broadening the Scope of Nanopublications. *ESWC 2013*.

[10] T. Kuhn, C. Chichester, M. Dumontier, M. Krauthammer. Publishing without Publishers: a Decentralized Approach to Dissemination, Retrieval, and Archiving of Data. arXiv preprint 1411.2749, 2014.

[11] T. Kuhn, M. Dumontier. Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data. *ESWC 2014*.

[12] M. A. Parsons, R. Duerr, J.-B. Minster. Data Citation and Peer Review. *Eos, Transactions American Geophysical Union* 91:34, 2010.

[13] M. A. Rajab, J. Zarfoss, F. Monrose, A. Terzis. A multifaceted approach to understanding the botnet phenomenon. *IMC 2006*.

[14] T. Steiner. Bots vs. Wikipedians, Anons vs. Logged-Ins. *WWW Companion 2014*.