Challenges of Forecasting and Measuring a Complex Networked World

Bruno Ribeiro
Carnegie Mellon University
Pittsburgh, PA, USA
ribeiro@cs.cmu.edu

ABSTRACT

A new era of data analytics of online social networks promises tremendous high-impact societal, business, and healthcare applications. As more users join online social networks, the data available for analysis and forecast of human social and collective behavior grows at an incredible pace.

The first part of this talk introduces an apparent paradox, where larger online social networks entail more user data but also less analytic and forecasting capabilities [7]. More specifically, the paradox applies to forecasting properties of network processes such as network cascades, showing that in some scenarios unbiased long term forecasting becomes increasingly inaccurate as the network grows but, paradoxically, short term forecasting – such as the predictions in Cheng et al. [2] and Ribeiro et al. [7] – improves with network size. We discuss the theoretic foundations of this paradox and its connections with known information theoretic measures such as Shannon capacity. We also discuss the implications of this paradox on the scalability of big data applications and show how information theory tools – such as Fisher information [3, 8] – can be used to design more accurate and scalable methods for network analytics [6, 8, 10]. The second part of the talk focuses on how these results impact our ability to perform network analytics when network data is only available through crawlers and the complete network topology is unknown [1, 4, 5, 9].

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

Keywords

Social networks; complex networks; network process forecast; information diffusion

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2744720.

Acknowledgements

This work was partially supported by NSF grants CNS-1065133 and ARL Cooperative Agreement W911NF-09-2-0053. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied of the NSF, ARL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

1. REFERENCES

[1] Konstantin Avrachenkov, Bruno Ribeiro, and Don Towsley. Improving random walk estimation accuracy with uniform restarts. In *Algorithms and Models for the Web-Graph*, pages 98–109. Springer, 2010.

[2] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proc. WWW*, 2014.

[3] Thomas M Cover and Joy A Thomas. *Elements of information theory.* John Wiley & Sons, 2012.

[4] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Practical recommendations on crawling online social networks. *Selected Areas in Communications, IEEE Journal on*, 29(9):1872–1892, 2011.

[5] Fabricio Murai, Bruno Ribeiro, Don Towsley, and Krista Gile. Characterizing Branching Processes from Sampled Data. In *Proc. WWW Companion*, pages 805–811, 2013.

[6] Bruno Ribeiro. *On the Design of Methods to Estimate Network Characteristics.* PhD thesis, University of Massachusetts Amherst, 2010.

[7] Bruno Ribeiro, Minh X. Hoang, and Ambuj K. Singh. Beyond models: Forecasting complex network processes directly from data. In *Proc. WWW*, 2015.

[8] Bruno Ribeiro, Don Towsley, Tao Ye, and Jean C Bolot. Fisher information of sampled packets: an application to flow size estimation. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 15–26. ACM, 2006.

[9] Bruno Ribeiro, Pinghui Wang, Fabricio Murai, and Don Towsley. Sampling directed graphs with random walks. In *INFOCOM, 2012 Proceedings IEEE*, pages 1692–1700. IEEE, 2012.

[10] Paul Tune and Darryl Veitch. Fisher information in flow size distribution estimation. *Information Theory, IEEE Transactions on*, 57(10):7011–7035, 2011.