Truthy: Enabling the Study of Online Social Networks

Karissa McKelvey
Filippo Menczer
Center for Complex Networks
and Systems Research
Indiana University
Bloomington, IN, USA

Abstract

The broad adoption of online social networking platforms has made it possible to study communication networks at an unprecedented scale. Digital trace data can be compiled into large data sets of online discourse. However, it is a challenge to collect, store, filter, and analyze large amounts of data, even by experts in the computational sciences. Here we describe our recent extensions to Truthy, a system that collects Twitter data to analyze discourse in near real-time. We introduce several interactive visualizations and analytical tools with the goal of enabling citizens, journalists, and researchers to understand and study online social networks at multiple scales.

Author Keywords

Research Data Management; Visualization; HCID; Collective Intelligence; Online Social Networks

ACM Classification Keywords

H.5 [**Human-centered computing**]: Information visualization; Collaborative and social computing devices; H.3.7 [**Information systems**]: Digital libraries and archives; Data stream mining; Collaborative filtering.

Introduction

Reseachers can study commnication networks on a larger scale than has been possible before in human history due to the high availability and use of the Internet as a central communication platform [9]. Recent studies have demonstrated that digital trace data can be combined with sophisticated statistical tools to produce insights into the behavior and interactivity patterns of hundreds of thousands of individual actors [4, 1]. With these new information and communication technologies, researchers are afforded with many opportunities to further human knowledge and understanding of community discourse and deliberation.

Despite the promises of these advances, there are various limitations to using social networks as a primary data source. It is often difficult or expensive for researchers trained in the social sciences to utilize the techonological expertise required to collect, store, filter, and analyze large amounts of data. Spam and misinformation add noise to results, often originating from compromised accounts of otherwise legitimate users [2, 3, 7], and should be flagged for removal. Research is also difficult to reproduce when performed on a wide variety of datasets gathered with custom toolkits. Thus, it is beneficial for researchers to utilize a centralized platform. It is also important that any such platform be free or inexpensive, unlike for-profit social media analytics services such as GNIP (http://gnip.com/), as researchers often operate within limited budgets.

We demonstrate the following contributions: interactive visualizations enabling users to better understand social networks and online trends; tools allowing users to freely download derived data from our large historical repository of online discourse; interfaces that allow users to tag content in order to improve automatic classifiers that detect behavior such as spam and misinformation.

The Truthy System

The *Truthy* system (http://truthy.indiana.edu) was originally designed to analyze and detect the emergence of coordinated misinformation campaigns on Twitter [7]. Now tasked with advancing the study of social networks in general, Truthy monitors a real-time feed of 140-character messages known as tweets and clusters them into groups of related messaged called "memes."

Memes typically correspond to discussion topics, communication channels, or information resources shared among Twitter users, so that one can focus attention to understandable units of information transfer. We define each meme as the set of all tweets containing a common hashtag (e.g., #bahrain), mentioned username (e.g., @BarackObama), hyperlink, or phrase. Memes are extracted from tweets that match lists of hand-picked keywords ("themes"); users can browse memes according to these themes (Fig. 1). We refer the reader to previous work for a detailed description of the algorithms utilized for data collection, storage and filtering [6].

For each meme, the user is presented with an interactive dashboard containing a crowdsourced definition from Tagdef (http://tagdef.com), a high-resolution image of the meme's information diffusion network (Fig. 2), and various interactive visualizations and statistics. Available information includes the number of users and tweets, meme diffusion network statistics such as mean degree and largest connected component size, as well as user-specific statistics such as predicted political partisanship, sentiment score, language, and activity. Users can download the aforementioned derived data,

recent tweets, and the network graphs themselves for use in a spreadsheet or an analysis application such as Network Workbench [8]. We ensure that this data download function abides by the Twitter Terms of Service.

We provide various interactive visualizations allowing users to interrogate the network structure, the characteristics of geography and time, and meme-meme co-occurrence patterns [5]. Another important new contribution of the Truthy system is in enabling active improvement of our algorithms by facilitating the tagging of suspicious content. Through the meme detail interface (Fig. 3), one can tweet about a meme (Fig. 4) or user (Fig. 5) in a syntax that can be automatically parsed by our system. We collect these posts for future studies analyzing the reliability of crowdsourced data in identifying persuasion and spam campaigns.

Future Work

In future work, we would like to provide a broader set of historical data while improving our visualizations via collaboration with the public. These initiatives include: providing a public REST API for derived data about individual tweets, memes, and users; expanding the scope to include all of the tweets we have collected since September 2010; facilitating user-defined themes to expand visualizations to customized content.

Conclusion

Tools like Truthy stimulate the study of online social networks. In particular, we aim to support social scientists who currently find this research difficult, expensive, or impossible to reproduce. We hope that open data and open tools like ours will advance efforts to leverage large data streams from the Internet as a primary source in the social sciences. Furthermore, we hope that our interactive

visualizations facilitate the navigation and understanding of online discourse, whether for research, journalism, or general use cases.

Acknowledgements

References

[1] Conover, M. D., Ratkiewicz, J., Gonçalves, B., Flammini, A., and Menczer, F. Political polarization on twitter. In *International Conference on Weblogs and Social Media 2011* (2011).

[2] Gonçalves, B., Conover, M., and Menczer, F. Abuse of social media and political manipulation. In *The Death of The Internet*, M. Jakobsson, Ed. Wiley, 2012.

[3] Grier, Chris and Thomas, Kurt and Paxson, Vern and Zhang, M. @ spam : The Underground on 140 Characters or Less Categories and Subject Descriptors. In *Proceedings of the 17th ACM conference on Computer and communications security*, ACM (Chicago, IL, USA, 2010), 27–37.

[4] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. Computational social science. *Science 323*, 5915 (2009), 721–723.

[5] McKelvey, K., Rudnick, A., Conover, M. D., and