The Hoaxy Misinformation and Fact-Checking Diffusion Network

Pik-Mai Hui,[1*] **Chengcheng Shao,**[1, 2]
Alessandro Flammini[1,] **Filippo Menczer,**[1] **Giovanni Luca Ciampaglia**[3]
[1]School of Informatics, Computing, and Engineering, Indiana University, Bloomington, USA
[2]College of Computer, National University of Defense Technology, China
[3]Indiana University Network Science Institute, Bloomington, USA

Abstract

Massive amounts of misinformation flood social media like Twitter and Facebook. Digital misinformation includes articles about hoaxes, conspiracy theories, fake news, and other misleading claims. This content has been alleged to disrupt the public debate, leading to questions about its impact on the real world. A number of research questions have been formulated around the ways misinformation spreads, who are its main purveyors, and whether fact-checking efforts can be helpful at mitigating its diffusion. Here we release a large longitudinal dataset from Twitter, consisting of retweeted messages with links to misinformation and fact-checking articles. These data have been collected using Hoaxy (hoaxy.iuni.iu.edu), an open social media analytics platform whose goal is to provide a comprehensive picture of how digital misinformation spreads and competes with fact-checking efforts. The released dataset contains over 20 million retweets, spanning the period from May 2016 to the end of 2017. We provide basic statistics about the data and the associated diffusion networks.

Introduction

In little over a decade, social media have come to play a prominent life in our everyday life. Much of our social interactions now occur on social media, and in recent years they have also started replacing more traditional carriers of information, becoming one of the primary news source for a majority of the population worldwide. This phenomenal adoption has unfortunately created incentives for the production and dissemination of digital misinformation, leading to real-world consequences ranging from an increase incidence of risky healthy behavior (Hotez 2016), to stock market manipulation (Ferrara et al. 2016).

Researchers have investigated what contributes to our vulnerability to misinformation. These investigations have come from different perspectives. Vulnerabilities originate from information overload (Qiu et al. 2017), echo chambers in social networks (Pariser 2011; Sunstein 2009), selective exposure (Stroud 2011), and motivated reasoning (Kahan 2012). Social media platforms may exacerbate the problem by introducing algorithmic biases, such as the promotion of popular posts (Nematzadeh et al. 2017). It is unclear, however, which of these factors dominate, and how they interact with each other in deceiving information consumers (Lazer et al. 2018).

Fact-checking initiatives have multiplied to combat the flood of digital misinformation (Graves, Nyhan, and Reifler 2016). Despite these efforts, the abuse of social media has not abated in recent years (Ciampaglia 2018), sometimes facilitated by social bots (Ferrara et al. 2016). The fight between misinformation sources and fact checkers has never been as evident as in the current era. To study this competition, comprehensive data are needed.

To this end, here we release a large longitudinal dataset collected from Twitter, consisting of retweeted messages with links to either fact-checking or misinformation articles (or both). The released dataset is a part of Hoaxy, an open social media analytics system (Shao et al. 2016; 2018). The goal of Hoaxy is to track how misinformation spreads and competes with fact-checking efforts on Twitter. Hoaxy uses the "POST statuses/filter" API endpoint to collect all public tweets that include links to fact-checking and misinformation articles. The domains in the URLs of these articles fall into one of two pre-compiled lists of misinformation sources and fact-checking organizations (Shao et al. 2017). We refer to articles from misinformation sources as "claims." Note that the particular Twitter API endpoint we used for the collection enables Hoaxy to obtain not a sample stream, but a complete set of tweets that link to our target domains. The retweets in this dataset can be used to build directed, weighted diffusion networks of fact-checking and claim articles on Twitter. The dataset can be downloaded at doi:10.5072/FK2/XSEHDL, hosted at dataverse.mpi-sws.org/dataverse/icwsm18.

Data Description

For each tweet in our dataset, we provide the following information in a list of retweets: its numerical identifier (tweet ID), its timestamp, the two numerical identifiers of the retweeting and tweeting users (user IDs), and a label indicating the type of article linked in the tweet (claim or fact-checking). If a tweet included multiple URLs matching our list of domains, we provide one extra entry for each additional URL, with the corresponding label. In total, the dataset includes 20,987,210 retweets, with 19,917,712

*Corresponding author: huip@indiana.edu.

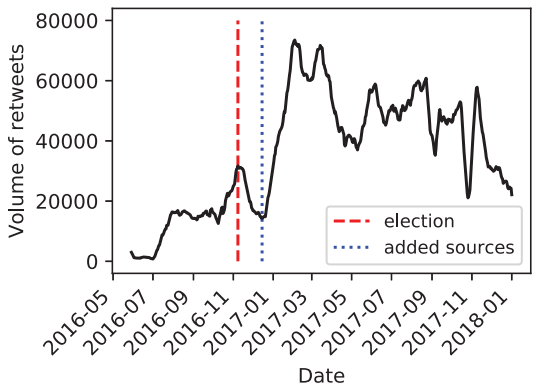

Figure 1: Running average of daily volume of retweets in the Hoaxy dataset, computed over a sliding window of 14 days. The dashed vertical line denotes the date of the 2016 U.S. Presidential Election, whereas the dotted vertical line denotes the date when additional entries were added to the list of domains tracked by Hoaxy.

(95%) linking to claim articles and 1,069,498 (5%) to fact-checking articles.

Hoaxy started collecting data on the 16th of May 2016. The dataset presented here include retweets collected from that date until the 31st of December 2017. Figure 1 shows the rolling average of daily retweet volume. Two special events are marked with vertical lines in the figure. The first is the 2016 U.S. Presidential Election; the second is an addition to the Hoaxy misinformation sources, from 71 to 121 domains. Some of the dips in volume are due to occasional downtimes of Hoaxy.

Let us now present a procedure to construct meaningful diffusion networks of claim and fact-checking articles from the retweet data. We start by defining the constructed diffusion networks. In these networks, two nodes are connected by a directed edge if one account retweeted the other during the observation window. That is, edges are weighted by the amount of retweets observed over the data collection window. We follow the conventional flow of information to determine the direction of a retweet, i.e., an edge is drawn from the retweeted account to the retweeting account. Since the networks are directed, the edge weights account only for the number of retweets in its corresponding direction. In other words, between any two nodes there can be up to two edges in opposing directions, which in general have different weights.

Let us consider an edge e of this network. Because edge weights are the result of aggregating several retweets, in principle the weight $w(e)$ should be split in two parts, one accounting for the retweets of claim articles, and one for retweets of fact-checking articles. Let us call them $w_c(e)$ and $w_f(e)$. Then, by defintion, $w(e) = w_c(e) + w_f(e)$. In practice, we find that $w_c(e) \cdot w_f(e) > 0$ only a small minority of edges $e \in E$. Based on this observation, we apply a simple majority rule and categorize each edge as either a

	# nodes	# edges
fact-checking	395,085	764,831
claim	1,424,733	9,255,428
combined	1,607,628	9,966,326

Table 1: Basic statistics about the diffusion networks constructed from retweet data.

'claim' or a 'fact-checking' edge based on the majority label of its retweets.

In the data repository that comes with the present article we provide both the full diffusion network, computed according to the above aggregation procedures, as well as the original disaggregated list of retweets. Retweet timestamps can be used to slice the diffusion networks into smaller temporal snapshots. Temporal analysis of the diffusion patterns can thus be performed by aggregating the data at different resolutions.

Table 1 shows the statistics of the fact-checking, claim, and combined diffusion networks. Note that there is an overlap between the set of users in the fact-checking network and that in the claim network. Therefore the number of nodes in the combined network is smaller than the sum of the claim and fact-checking network sizes.

Figure 2 shows that the distributions of degree and strength (a.k.a. weighted degree) are heavy-tailed in all three networks. This suggests that these networks are dominated by extremely influential (heavily retweeted) and extremely active (amplifier) accounts, which spread many links to claims and fact-checking articles.

Conclusion

In this paper, we present a large longitudinal dataset from Hoaxy, comprising of retweeted messages with links to fact-checking and misinformation articles. We provide statistics about the dataset, and the associated diffusion networks. Among potential applications, the dataset opens up new possibilities to evaluate algorithms and methods that predict and/or control the spreading of misinformation using real-world records. It covers some important events such as the 2016 U.S. Presidential Election. We hope that the released dataset will be a valuable addition to the research community and in particular that will foster further research into the development of effective countermeasures against digital misinformation.

Acknowledgments

We are grateful to Ben Serrette and Valentin Pentchev of the Indiana University Network Science Institute (iuni.iu.edu) and to Democracy Fund for supporting the development of the Hoaxy platform. We are also indebted to Twitter for providing data through their API. C.S. thanks the Center for Complex Networks and Systems Research (cnets.indiana.edu) for the hospitality during his visit at the Indiana University School of Informatics, Computing, and Engineering. C.S. was supported by the China Scholarship Council. G.L.C. was supported by IUNI. A.F. and F.M. were supported in part by the James S. McDonnell Foundation (grant

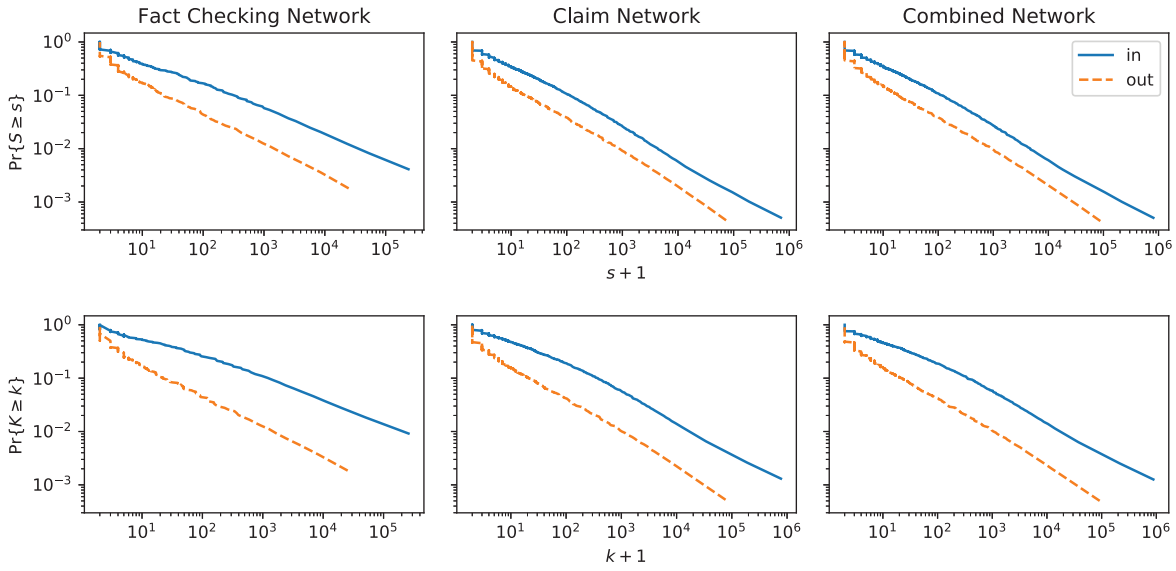

Figure 2: Cumulative distributions (CCDF) of in- and out-strenth (s, top) and in- and out-degree (k, bottom) for the fact-checking, claim, and combined diffusion networks.

220020274) and the National Science Foundation (award CCF-1101743). The funders had no role in data collection and analysis, decision to publish or preparation of the manuscript.

References

Ciampaglia, G. L. 2018. Fighting fake news: a role for computational social science in the fight against digital misinformation. *Journal of Computational Social Science* 1(1):147–153.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM* 59(7):96–104.

Graves, L.; Nyhan, B.; and Reifler, J. 2016. Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking. *Journal of Communication* 66(1):102–138.

Hotez, P. J. 2016. Texas and its measles epidemics. *PLoS medicine* 13(10):e1002153.

Kahan, D. M. 2012. Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision Making* 8(407):24.

Lazer, D.; Baum, M.; Benkler, Y.; Berinsky, A.; Greenhill, K.; Menczer, F.; Metzger, M.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S.; Sunstein, C.; Thorson, E.; Watts, D.; and Zittrain, J. 2018. The science of fake news. *Science* 359(6380):1094–1096.

Nematzadeh, A.; Ciampaglia, G. L.; Menczer, F.; and Flammini, A. 2017. How algorithmic popularity bias hinders or promotes quality. Preprint arXiv:1707.00574, CoRR.

Pariser, E. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. London, UK: Penguin.

Qiu, X.; Oliveira, D. F.; Shirazi, A. S.; Flammini, A.; and Menczer, F. 2017. Limited individual attention and online virality of low-quality information. *Nature Human Behaviour* 1(7):0132.

Shao, C.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, 745–750. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Flammini, A.; and Menczer, F. 2017. The spread of misinformation by social bots. Preprint arXiv:1707.07592, CoRR.

Shao, C.; Hui, P.-M.; Wang, L.; Jiang, X.; Flammini, A.; Menczer, F.; and Ciampaglia, G. L. 2018. Anatomy of an online misinformation network. Preprint arXiv:1801.06122, CoRR.

Stroud, N. J. 2011. *Niche news: The politics of news choice*. Oxford, UK: Oxford University Press.

Sunstein, C. R. 2009. *Going to extremes: How like minds unite and divide*. Oxford, UK: Oxford University Press.