8-2019

Evaluating vulnerability to fake news in social networks: A community health assessment model

Bhavtosh RATH

Wei GAO
Singapore Management University, weigao@smu.edu.sg

Jaideep SRIVASTAVA

Citation

RATH, Bhavtosh; GAO, Wei; and SRIVASTAVA, Jaideep. Evaluating vulnerability to fake news in social networks: A community health assessment model. (2019). *Proceedings of the 2019 IEEE/ACM International Conference on Social Networks Analysis and Mining (ASONAM 2019)*. 432-435. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/4556

Evaluating Vulnerability to Fake News in Social Networks: A Community Health Assessment Model

Bhavtosh Rath[1], Wei Gao[2], and Jaideep Srivastava[1]
[1]Dept. of Computer Science & Engineering, University of Minnesota, Twin Cities, MN, USA
[2]School of Information Management, Victoria University of Wellington, Wellington, New Zealand
rathx082@umn.edu, wei.gao@vuw.ac.nz, srivasta@umn.edu

***Abstract*—Understanding the spread of false information in social networks has gained a lot of recent attention. In this paper, we explore the role community structures play in determining how people get exposed to fake news. Inspired by approaches in epidemiology, we propose a novel *Community Health Assessment* model, whose goal is to understand the vulnerability of communities to fake news spread. We define the concepts of neighbor, boundary and core nodes of a community and propose appropriate metrics to quantify the vulnerability of nodes (individual-level) and communities (group-level) to spreading fake news. We evaluate our model on communities identified using three popular community detection algorithms for twelve real-world news spreading networks collected from Twitter. Experimental results show that the proposed metrics perform significantly better on the fake news spreading networks than on the true news, indicating that our community health assessment model is effective.**

I. Introduction

Use of social media platforms like Facebook and Twitter is ubiquitous in modern times, making them powerful platforms for news propagation and consumption. However, the good inevitably is accompanied by the bad, which can be witnessed by the problem of *fake news spreading* [1]. It spreads when someone propagates it via endorsements such as replying, sharing or re-posting, without validating its authenticity. There is significant interest in understanding the nature of fake news spreading. Our focus is on *assessing the vulnerability of social networks to fake news spreading*. Specifically, we focus on people and the communities they create, with the goal of identifying how vulnerable individuals and communities are to believing and propagating fake news. We propose the Community Health Assessment model that distinguishes between neighbor, boundary and core nodes of a community, and propose novel metrics to quantify the vulnerability of an individual node, as well as the community, to external exposure. We propose methods to estimate the likelihood of a boundary node of a community to believe fake news sent from its immediate neighbors; and also estimate the likelihood of a community's entire boundary node set to believe fake news coming from its neighborhood. It is important to note that the method used to quantify vulnerability of a boundary node can be generalized to any node. Intuitively, if an external node infects a member of a community, the likelihood of the entire community getting infected increases due to high connectivity and trust among its members. Thus, while assessing vulnerability of community, we focus on examining the influence of news propagated from external nodes into the community rather than considering the propagation of the news within the community. We evaluate our model on the propagation networks of twelve real-world news collected from snopes.com[1].

Our contributions are summarized as follows:

- We propose the Community Health Assessment model that introduces the ideas of neighbor, boundary and core nodes for a community structure in a social network.
- We propose metrics to quantify the vulnerability of a node and a community to fake news exposure from outside.
- Using Twitter news item spreading network (a subgraph of Twitter network induced by the news item, *news network* in short) we demonstrate that our proposed metrics can assess the vulnerability of social networks to fake news better than for true news.

II. Related Work

There has been a recent surge in interest among researchers and practitioners to develop approaches to prevent fake news spread. Most approaches in the literature use content-based [2], [3] and propagation-based characteristics [4], [5]. Approaches using neural networks [6], [7] have also shown promising results. Infection spread models inspired from epidemiology [8], [9] have also been used to model rumor spreading. Other models have tried to identify the rumor spreading source [10], [11]. A community perspective to rumor spread has also been attempted. Fan et al. [12] proposed an approach to identify a minimal set of boundary nodes that would prevent spread of rumors from neighboring communities. Nguyen et al. [13] proposed a community-based heuristic method to find the smallest set of highly influential nodes whose decontamination with good information would contain rumor spreading. Vosoughi

ASONAM '19, August 27-30, 2019, Vancouver, Canada

ACM ISBN 978-1-4503-6868-1/19/08?/$15.00
http://dx.doi.org/10.1145/3341161.3342920

[1]https://www.snopes.com/fact-check-ratings/

et al. [14] empirically analyzed the spread of true and false news online, and is close to our research.

III. Community Health Assessment Model

Social networks comprise of communities, which are structures that are modular groups, where within-group members are highly connected, and across-group members are loosely connected. *Modularity* is the ratio of density of edges inside a community to edges outside the community [15]. If such communities are exposed to fake news being propagated from neighboring nodes, the likelihood of the whole community getting infected would be high. Thus it is important to identify vulnerable communities that lie in the path of fake news spread in order to protect them, and thus limit the overall influence of fake news in the network. As part of the Community Health Assessment model, we first propose the ideas of neighbor, boundary and core nodes of a community, and then derive metrics to quantify vulnerability of nodes and communities based on the fundamental measures of trust.

The three types of nodes with respect to a community which are affected during the process of news spreading are explained below:

1. *Neighbor nodes*: These nodes are directly connected to at least one node of the community. The set of neighbor nodes is denoted by $\mathcal{N}$. They are not a part of the community.

2. *Boundary nodes*: These are community nodes that are directly connected to at least one neighbor node. The set of boundary nodes is denoted by $\mathcal{B}$.

3. *Core nodes*: These nodes are only connected to members within the community. The set of core nodes is denoted as $\mathcal{C}$.

A. Preliminaries

1) Trustingness and Trustworthiness: In the context of social media, researchers have used social networks to understand how trust manifests among users. A recent work is the Trust in Social Media (TSM) model which assigns a pair of complementary trust scores to each actor, called *Trustingness* and *Trustworthiness*. *Trustingness* quantifies the propensity of an actor to trust its neighbors and *Trustworthiness* quantifies the willingness of the neighbors to trust the actor. The details of the model are excluded due to space constraints and can be found in [16].

2) Believability: Believability is an edge score derived from Trustingness and Trustworthiness scores. It helps us quantify how likely is the receiver of a message to believe its sender. Believability for a directed edge is naturally computed as a function of the trustworthiness of the sender and the trustingness of the receiver. The idea has been applied in [17] where a classification model was built to identify rumor spreaders in Twitter network.

B. Vulnerability Metrics

Motivation: Fake news generally gets no coverage from mainstream news platforms (such as press or television), so the biggest factor contributing to a user's decision to spread a fake news on social media is its inherent trust on other users endorsing it. On the other hand, a user would most likely endorse a true news since it is typically spread from more credible news sources, such as mainstream media. We hypothesize that *the less credible nature of fake news makes it much more reliant on user trust for spreading than true news does*. Thus, we propose our vulnerability metrics based upon the idea of computational trust, particularly the believability measure, for assessing the health of individuals and communities encountering fake news.

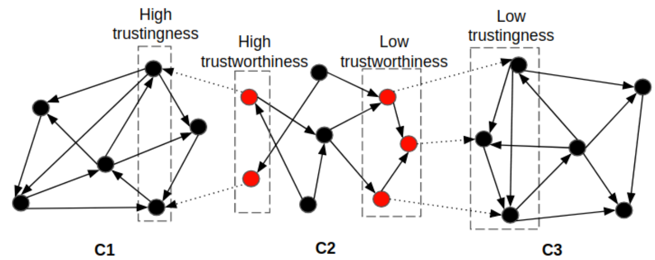

Fig. 1: Illustration of vulnerability to fake news spread.

Illustrative Example: We illustrate the idea of the proposed vulnerability metrics through figure 1. Red nodes in community C2 represent fake news spreaders. C1 and C3 are two other communities having identical structure. C3 and C1 have 3 and 2 boundary nodes, respectively, that are directly connected to the fake news spreaders. Based on edge count one would believe that C3 is more vulnerable to fake news spreading than C1. But low trusting boundary nodes of C3 are connected to low trustworthy spreaders, while high trusting boundary nodes of C1 are connected to high trustworthy spreaders. Therefore, our metric should be able to identify C1 as more vulnerable than C3.

While TSM has also been used to assess news organization's impact on audience engagement [18], we build upon the idea of believability to propose *Vulnerability Metrics* that help us quantify the likelihood of boundary nodes and communities believing a news spread from their neighbors. We assume that the news spreading is widespread outside of the community, i.e., at least some of the neighbor nodes of the community are spreaders. We define the node- and community-level vulnerability metrics as follows:

Vulnerability of boundary node $V(b)$: This metric measures the likelihood of a boundary node b to become a spreader. The metric is derived as follows: The likelihood of node b to believe an immediate neighbor n is a function of the trustworthiness of the neighbor n ($n \in \mathcal{N}_b$, where $\mathcal{N}_b$ is the set of all neighbor nodes of b) and the trustingness of b, and is quantified as $bel_{nb} = tw(n) * ti(b)$, that is, $Believability(n \rightarrow b)$. Thus, the likelihood that b is *not* vulnerable to n can be quantified as $(1 - bel_{nb})$. Generalizing this, the likelihood of b *not* being vulnerable to all of its neighbor nodes is $\prod_{\forall n \in \mathcal{N}_b}(1 - bel_{nb})$. Therefore, the likelihood of b to believe any of its neighbors, i.e., the vulnerability of the boundary node b is computed as:

$$V(b) = 1 - \prod_{\forall n \in \mathcal{N}_b} (1 - bel_{nb}) \quad (1)$$

Vulnerability of community, $\widetilde{V}(C)$: This metric measures likelihood of the boundary node set of a community C ($\mathcal{B}_C$) to believe a news from any of its neighbors. The metric is derived as follows: Going forward with the idea in 1), the likelihood that boundary node b is *not* vulnerable to its neighbors can be quantified as $(1 - V(b))$. Generalizing this to all $b \in \mathcal{B}_C$, the likelihood that none of the boundary nodes of a community are vulnerable to their neighbors can be quantified as $\prod_{\forall b \in \mathcal{B}_C}(1 - V(b))$. Thus, the likelihood of community C being vulnerable to any its neighbors, i.e., the vulnerability of the community, is defined as:

$$\widetilde{V}(C) = 1 - \prod_{\forall b \in \mathcal{B}_C} (1 - V(b)) \tag{2}$$

IV. Experiments and Results

A. Dataset and Setup

We collected twelve different designated news articles and their ground truth ratings through snopes.com. Based on the rating type, we categorized the news into three categories: News M1, M2, M3 and M4 are labelled as *Mixture* which indicates that the news has significant elements of both truth and falsity in it, news F1, F2, F3 and F4 are labelled as *False* which indicates that the primary elements of the news are basically false, and news T1, T2, T3 and T4 are labelled as *True* which indicates that the primary elements of a claim are basically true. The general statistics for the twelve networks are presented in Table I.

We identified the specific source tweet related to each news in question. For evaluation of metrics, we then identified all the spreaders of the source tweet associated with the news, which comprised of the source tweeter (identified using Twitter API) and the list of retweeters (accessible through *twren.ch*). We considered the follower-following network of the spreaders obtained from Twitter API, as the directed social network. We ran the TSM algorithm [16] on this network to compute the trustingness and trustworthiness scores for every node. We then identified disjoint communities using three popular community detection algorithms on large networks: Louvain [19], Infomap [20], and Label Propagation [21]. For each of the communities generated we identified the sets of boundary and neighbor nodes.

B. Evaluation of Metrics

To measure how good the proposed metrics are able to quantify the vulnerability of nodes and communities, we evaluate the quality of ranking on boundary nodes and communities based on vulnerability scores in comparison with the ground-truth ranking of nodes and communities derived from the news spread in the network. We adopt the ranking evaluation measures widely used in Information Retrieval literature [22].

1) Evaluation of $V(b)$: A vulnerable boundary node is highly likely to have strong believability with its neighbors. We thus consider the ground truth of a vulnerable node as a node which retweets. The ground truth vulnerability of boundary nodes is binary as we only have information of whether the node retweets or not. We thus evaluate this metric using *Average Precision@k* (AP@k, where k represents the top-k vulnerable boundary nodes) and *Mean Average Precision* (MAP) over all communities in a network.

2) Evaluation of $\widetilde{V}(C)$: A community with more number of spreader boundary nodes is more vulnerable to news penetration. As most communities of a network have at least few spreader boundary nodes, it is not feasible to use node ranking metrics above for evaluating community vulnerability. We thus rank the communities by their vulnerability scores and compare with the ground-truth ranking given by the relative count of spreader boundary nodes in the community. We use Kendall's tau (τ), which is a correlation measure for ordinal data, as evaluation metric.

C. Results

Table II shows the evaluation results for the proposed metric assessing the vulnerability of boundary nodes. For the twelve networks we show the Average Precision for k = 1, 3, 5, 10 and 15 and compute the MAP for the top-15 results. AP@1 shows how well we are able to identify the first spreader boundary node based on our metric. Our metric is able to identify the most vulnerable boundary node in AP of 0.712, 0.91, 0.471 averaged over mixture, false, true news networks respectively for Louvain; 0.695, 0.923, 0.459 averaged over mixture, false, true news networks respectively for Informap; 0.811, 0.915, 0.74 averaged over mixture, false, true news networks respectively for Label Propagation. As expected, our metrics show better performance particularly for fake news networks, followed by mixture and then true news networks. Average precision for rest of k also shows similar trend. Metrics for Louvain-/Infomap-based communities follow a similar trend for the remaining k values. However, Label Propagation communities for k=3 and 5 show better performance on true news networks compared to mixture news networks. This insensitivity in evaluation could be attributed to the fact that label propagation algorithm tends to generate more number of communities. We also observe that the MAP shows a similar trend, with better performance for false news networks compared to true news networks.

Table III shows the evaluation results for proposed metric to compute the vulnerability of a community. For the twelve networks the table shows Kendall's tau value (τ) for communities generated using the three algorithms. We observe that the τ for mixture and true news networks tend to have a negative correlation with the ground truth community ranking. False news networks on the other hand show a positive correlation.

V. Conclusion

We propose novel metrics based on the concept of believability derived from computational trust measures to compute vulnerability of nodes and communities to news spread and show that the metrics are much more sensitive to fake news than true news. This confirms our hypothesis that fake news have to rely on strong trust among spreaders to propagate while true news do not. Through experiments on large news spread

TABLE I. Statistics for the news propagation networks.

	M1	M2	M3	M4	F1	F2	F3	F4	T1	T2	T3	T4
# of nodes in network	2,385,188	3,669,213	6,462,462	3,512,201	1,883,329	4,981,319	782,209	503,160	10,929,291	953,040	2,155,927	1,530,958
# of edges in network	11,684,879	7,054,734	10,621,364	6,108,311	16,658,841	12,625,672	12,498,122	7,797,449	14,933,611	1,250,463	3,221,985	2,484,553
# of spreaders	2,833	2,296	2,834	2,668	2,879	2,833	465	290	2,788	198	693	1,053

*TABLE II. Evaluation of vulnerability of boundary nodes (**L**: Louvain; **I**: Infomap; **LP**: Label Propagation).*

	AP@1			**AP@3**			**AP@5**			**AP@10**			**AP@15**			**MAP**		
	L	**I**	**LP**	**L**	**I**	**LP**	**L**	**I**	**LP**	**L**	**I**	**LP**	**L**	**I**	**LP**	**L**	**I**	**LP**
M1	0.759	0.676	0.712	0.770	0.567	0.523	0.736	0.548	0.519	0.606	0.543	0.533	0.661	0.505	0.566	0.672	0.546	0.555
M2	0.818	0.749	0.907	0.737	0.888	0.722	0.769	0.733	0.799	0.821	0.699	0.999	0.733	0.666	0.999	0.785	0.733	0.875
M3	0.805	0.642	0.878	0.620	0.595	0.784	0.567	0.509	0.749	0.590	0.512	0.674	0.524	0.586	0.833	0.596	0.577	0.751
M4	0.468	0.714	0.750	0.409	0.619	0.633	0.366	0.674	0.633	0.323	0.523	0.659	0.325	0.454	0.799	0.350	0.569	0.660
M_{avg}	0.712	0.695	0.811	0.634	0.667	0.665	0.609	0.616	0.675	0.585	0.569	0.716	0.560	0.552	0.799	0.600	0.606	0.710
F1	0.892	0.749	0.855	0.793	0.722	0.761	0.824	0.679	0.999	0.922	0.499	0.799	0.899	0.422	0.999	0.876	0.552	0.905
F2	0.819	0.999	0.874	0.657	0.777	0.851	0.727	0.499	0.839	0.741	0.399	0.924	0.706	0.266	0.999	0.714	0.518	0.900
F3	0.933	0.945	0.933	0.999	0.999	0.999	0.955	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.972	0.985	0.995
F4	0.999	0.999	0.999	0.999	0.999	0.999	0.955	0.999	0.999	0.979	0.999	0.999	0.999	0.999	0.999	0.991	0.999	0.999
F_{avg}	0.910	0.923	0.915	0.862	0.874	0.902	0.865	0.794	0.959	0.901	0.724	0.930	0.900	0.671	0.999	0.888	0.763	0.949
T1	0.222	0.531	0.868	0.469	0.466	0.802	0.424	0.492	0.716	0.439	0.349	0.479	0.377	0.344	0.533	0.450	0.424	0.644
T2	0.548	0.374	0.482	0.407	0.238	0.666	0.299	0.399	0.999	0.049	0.299	0.699	0.033	0.033	0.466	0.173	0.264	0.726
T3	0.666	0.470	0.913	0.472	0.499	0.999	0.519	0.499	0.999	0.299	0.499	0.899	0.266	0.433	0.799	0.391	0.479	0.900
T4	0.449	0.464	0.699	0.371	0.666	0.541	0.399	0.000	0.479	0.409	0.000	0.499	0.362	0.000	0.366	0.399	0.106	0.500
T_{avg}	0.471	0.459	0.740	0.429	0.467	0.752	0.410	0.347	0.798	0.299	0.286	0.644	0.259	0.202	0.541	0.353	0.318	0.692

*TABLE III. Evaluation of vulnerability of communities (**L**: Louvain; **I**: Infomap; **LP**: Label Propagation).*

	$\tau_{\mathbf{M1}}$	$\tau_{\mathbf{M2}}$	$\tau_{\mathbf{M3}}$	$\tau_{\mathbf{M4}}$	$\tau_{\mathbf{F1}}$	$\tau_{\mathbf{F2}}$	$\tau_{\mathbf{F3}}$	$\tau_{\mathbf{F4}}$	$\tau_{\mathbf{T1}}$	$\tau_{\mathbf{T2}}$	$\tau_{\mathbf{T3}}$	$\tau_{\mathbf{T4}}$
L	-0.027	0.003	-0.149	-0.035	0.050	0.164	0.457	0.161	-0.045	-0.255	-0.090	-0.030
I	0.072	0.000	0.274	0.138	0.642	0.667	0.117	0.146	-0.037	-0.222	-0.025	-0.031
LP	0.039	-0.014	0.019	0.018	0.039	0.029	0.381	0.714	0.003	0.005	-0.110	-0.036

networks on Twitter we show that our proposed metrics can identify the vulnerable nodes for false news networks with higher precision.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.

[2] Benjamin D Horne and Sibel Adali. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*, 2017.

[3] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

[4] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of AAAI*, pages 2972–2978, 2016.

[5] Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. Rumor cascades. In *Proceedings of ICWSM*, 2014.

[6] Jing Ma, Wei Gao, and Kam-Fai Wong. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of ACL*, pages 1980–1989, 2018.

[7] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of IJCAI*, pages 3818–3824, 2016.

[8] Mark EJ Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

[9] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, page 8. ACM, 2013.

[10] Devavrat Shah and Tauhid Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on information theory*, 57(8):5163–5181, 2011.

[11] Kai Zhu and Lei Ying. Information source detection in the sir model: A sample-path-based approach. *IEEE/ACM Transactions on Networking*, 24(1):408–421, 2016.

[12] Lidan Fan, Zaixin Lu, Weili Wu, Bhavani Thuraisingham, Huan Ma, and Yuanjun Bi. Least cost rumor blocking in social networks. In *Proceedings of IEEE ICDCS*, pages 540–549, 2013.

[13] Nam P Nguyen, Guanhua Yan, My T Thai, and Stephan Eidenbenz. Containment of misinformation spread in online social networks. In *Proceedings of ACM Web Science Conference*, pages 213–222. ACM, 2012.

[14] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[15] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[16] Atanu Roy, Chandrima Sarkar, Jaideep Srivastava, and Jisu Huh. Trustingness & trustworthiness: A pair of complementary trust measures in a social network. In *Proceedings of ASONAM*, pages 549–554, 2016.

[17] Bhavtosh Rath, Wei Gao, Jing Ma, and Jaideep Srivastava. From retweet to believability: Utilizing trust to identify rumor spreaders on twitter. In *Proceedings of ASONAM*, pages 179–186. ACM, 2017.

[18] Bhavtosh Rath, Jisu Kim, Jisu Huh, and Jaideep Srivastava. Impact of news organizations' trustworthiness and social media activity on audience engagement. *arXiv preprint arXiv:1808.09561*, 2018.

[19] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[20] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[21] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3):036106, 2007.

[22] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press, 2008.