Twitter Trends Manipulation: A First Look Inside the Security of Twitter Trending

Yubao Zhang, *Student Member, IEEE,* Xin Ruan, *Student Member, IEEE,*
Haining Wang, *Senior Member, IEEE,* Hui Wang, and Su He

Abstract—Twitter trends, a timely updated set of top terms in Twitter, have the ability to affect the public agenda of the community and have attracted much attention. Unfortunately, in the wrong hands, Twitter trends can also be abused to mislead people. In this paper, we attempt to investigate whether Twitter trends are secure from the manipulation of malicious users. We collect more than 69 million tweets from 5 million accounts. Using the collected tweets, we first conduct a data analysis and discover evidence of Twitter trend manipulation. Then, we study at the topic level and infer the key factors that can determine whether a topic starts trending due to its popularity, coverage, transmission, potential coverage, or reputation. What we find is that except for transmission, all of factors above are closely related to trending. Finally, we further investigate the trending manipulation from the perspective of compromised and fake accounts and discuss countermeasures.

Index Terms—Twitter trend, Social computing, Security.

1 Introduction

The Internet has subverted the autocratic way of disseminating news by traditional media like newspapers. Online trends are different from traditional media as a method for information propagation. For instance, Google Hot Trends ranks the hottest searches that have recently experienced a sudden surge in popularity [2]. Meanwhile, these trends may attract much more attention than before due to their appearance on Google Hot Trends.

More recently, Online Social Networking (OSN) like Twitter has inaugurated a new era of "We Media." Twitter is a real-time microblogging service. Users broadcast short messages no longer than 140 characters (called *tweets*) to their followers. Users can also discuss with the others on a variety of topics at will. The topics that gain sudden popularity are ranked by Twitter as a list of *trends* (also known as *trending topics*) [3]. Twitter and Google trends have become an important tool for journalists. Twitter in particular is used to develop stories, track breaking news, and assess how public opinion is evolving in the breaking story. Taking election campaigns as an example [5], journalists, campaigns, and pundits have tracked trends in Twitter traffic to determine candidates' popularity and predict likely election outcomes [4].

Previous research have studied trend taxonomy [7], [9], [10], trend detection [14], [17], [19], [20], and real events extraction from Twitter trends [6], [38]. However, researchers have paid little attention to Twitter trend manipulation. It is reported that attackers manipulate Google trends by simply employing large group of people to visit Google and search for a specific keyword phrase [23]. Also, Just *et al.* [4] inspected Twitter manipulation in an election campaign. As reported in *The Wall Street Journal*, robots have been used to undermine the "trending topics" on Twitter [1]. Thus, the focus of this work is on Twitter trend manipulation.

In this paper, the primary questions we attempt to answer are whether the malicious users can manipulate the Twitter trends and how they might be able to do that? Being exposed to real-time trending topics, users are entitled to have insight into how those trends actually go trending. Moreover, this research also cast light on how to enhance a commercial promotion campaign by reasonably using Twitter trends. To investigate the possibility of manipulating Twitter trends, we need to deeply understand how Twitter trending works. Twitter states that trends are determined by an algorithm and are always topics that are immediately popular. However, the detailed trending algorithm of Twitter is unknown to the public, and we have no way to find out what it specifically is. Instead, we study Twitter trending at the topic level and infer the key factors that can determine whether a topic trends from its popularity, coverage, transmission, potential coverage, and reputation. After identifying those key factors that are associated with the

- *Yubao Zhang is with the Department of Electrical and Computer Engineering, University of Delaware.*
  E-mail: ybzhang@udel.edu
- *Xin Ruan is with the Department of Computer Science, College of William and Mary.*
- *Haining Wang is with the Department of Electrical and Computer Engineering, University of Delaware.*
- *Hui Wang and Su He are with College of Information System and Management, National University of Defense Technology.*

by compromised and fake accounts which are in hand of a few malicious users, we can detect the anomaly of tweet source as an indicator of trend manipulation. Moreover, the monitoring of *topological hierarchy* of the accounts in a topic help detect trend anomaly. As figure 6 shows, spamming infrastructure exists in the topological hierarchy of spam accounts and this kind of anomaly indicates trend manipulation.

Detecting Manipulation Using Previously Manipulated Topics. We can classify different topics into two classes: manipulated and normal. There should be some connections among manipulated topics due to similar manipulation strategies. The connections among normal trending topics and the connections among manipulated topics, can be exploited for the early detection of Twitter trends using previously trending topics [3]. One feasible way to trace the connection between two topics with respect to manipulation is to treat one topic as the training set and the other as the testing set. In this regard, an SVM classifier can be employed to train the classification model based on the training set and then perform the classification task based on the testing set. The classification result reflects the connection between the two topics. Thus, the connections among manipulated topics enable us to detect manipulated topics one by one from the very beginning of identifying the first set of manipulated topics. The challenges here include identifying the first set of manipulated topics and verifying the manipulated topics. The influence model that we use to demonstrate the evidence of manipulation can be utilized to identify the first set of manipulated topics. The development of an accurate and practical verification method remains as our future work.

6 Related Works

To the best of our knowledge, this is the first effort to investigate whether Twitter trends could be manipulated.

Research on trending topics in Twitter includes real event recognization [6], [7], realtime trending topic detection [14], [15], [16], [38], the evolution of trending topic characterization [17], [18], and the taxonomy of trending topics [9], [21], [22]. Becker *et al.* [6] analyzed the stream of Twitter messages and distinguished the messages about real events from non-event messages based on a clustering method. Zubiaga *et al.* [7] categorized different triggers that leverage the trending topics by using social features rather than content-based approaches.

In the detection of realtime trending topics, Agarwal *et al.* [38] identified the emerging events before they became trending topics by modeling the detection problem as discovering dense clusters in highly dynamic graphs. Kasiviswanathan *et al.* [14] presented a dictionary-learning-based framework for detecting emerging topics in social media via the user-generated stream. Lu *et al.* [15] used an energy function to model the life activity of news events on Twitter and proposed a news event detection method based on online energy function. Cataldi *et al.* [16] identified emerging terms from user content by measuring user authority and proposing a keyword life cycle model, and then detected the emerging topics by formalizing the keyword-based topic graph.

To address the evolution and taxonomy of trending topics, Altshuler and Pan [17] presented the lower bounds of the probability that emerging trends successfully spread through the scale-free networks. Asur *et al.* [18] studied trending topics on Twitter and theoretically analyzed the formation, persistence, and decay of trends. Naaman *et al.* [9] characterized the trends in multiple dimensions and presented a taxonomy of trends. They also proposed a collection of hypotheses on different kinds of trends and evaluated them. Lehmann *et al.* [21] classified the popular hashtags by the temporal dynamics of hashtags. Irani *et al.* [22] focused on the trend-stuffing issue and developed a classifier to automatically identify the trend-stuffing in tweets.

Whether a topic begins trending is closely related to (1) the influence of users who are involved with the topic and (2) the topic adoption for users who are exposed to the topic. Cha *et al.* [24] performed a comparison of three different measures of influence: indegree, retweet, and mention. Weng *et al.* [25] proposed a topic-sensitive PageRank measure for user influence. Romero *et al.* [26] proposed an algorithm to measure the relative influence and passivity of each user from the viewpoint of a whole network. Bakshy *et al.* [27] measured the influence from the diffusion tree. The studies of topic adoption in Twitter mainly concentrate on hashtag adoption. Lin *et al.* [28] classified the adoption of hashtags into two classes and proposed a framework to capture the dynamics of hashtags based on their topicality, interactivity, diversity, and prominence. Yang *et al.* [29] studied the effect of the dual role of a hashtag on hashtag adoption.

7 Limitation and Future Work

There are some limitations of our work, some of which will be addressed in our future work.

First, we use a linear influence model to capture the network impact on the diffusion of a topic in Twitter, which enables us to find the evidence of manipulation. The application of the model is limited to linear scenarios. We will develop a non-linear model in our future work.

Second, we randomly choose 11 topics and more than 10,000 related tweets to infer the relevance of five key factors over Twitter trending. Although we have tried our best to guarantee the randomness, those 11 sample topics may not be large enough to represent the overall scenario in practice. Besides, we study five comparatively straight-forward factors that may affect trending. In the future work, we will consider more complicated factors and sample more topics to study the factors over trending.

Finally, we propose the countermeasures against Twitter trend manipulation but most of them remain in

the discussion stage. We leave the implementation and evaluation of those countermeasures for our future work. Specifically, we plan to develop a manipulation detection mechanism by using an SVM classifier. We will train the classifier using previously manipulated topics and then classify future trends as manipulated or not.

8 Conclusions

With the datasets we collected via Twitter API, we first evidence the manipulation of Twitter trending and observe a suspect spamming infrastructure. Then, we employ the SVM classifier to explore how accurately five different factors at the topic level (popularity, coverage, transmission, potential coverage, and reputation) could predict the trending. We observe that, except for transmission, the other factors are all closely related to Twitter trending. We further investigate the interacting patterns between authenticated accounts and malicious accounts. Finally, we present the threat posed by compromised and fake accounts to Twitter trending and discuss the corresponding countermeasures against trending manipulation.

References

[1] Wall Street Journal (Inside a Twitter Robot Factory), *http://online.wsj.com*

[2] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. *Detecting influenza epidemics using search engine query data.* Nature, 457(7232), 1012-4.

[3] Nikolov, S. *Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series* (Doctoral dissertation, Massachusetts Institute of Technology).

[4] Just, M., Crigler, A., Metaxas, P., and Mustafaraj, E. *It's Trending on Twitter-An Analysis of the Twitter Manipulations in the Massachusetts 2010 Special Senate Election.* In APSA 2012 Annual Meeting Paper.

[5] Ratkiewicz, J., Conover, M., and Meiss, M. *Detecting and tracking the spread of astroturf memes in microblog streams.* 5th International Conference on Weblogs and Social Media, 2010.

[6] Becker, H., Naaman, M., and Gravano, L. *Beyond trending topics: Real-world event identification on twitter.* ICWSM 2011.

[7] Zubiaga, A., Spina, D., and Martinez, R. *Classifying Trending Topics: A Typology of Conversation Triggers on Twitter.* CIKM 2011.

[8] Agarwal, M. K., Ramamritham, K., and Bhide, M. *Identifying Real World Events in Highly Dynamic Environments.* VLDB 2012.

[9] Naaman, M., Becker, H., and Gravano, L. *Hip and trendy: Characterizing emerging trends on Twitter.* Journal of the American Society for Information Science and Technology, 62(5), 902-918.

[10] Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., and Choudhary, A. *Twitter Trending Topic Classification.* 2011 IEEE 11th International Conference on Data Mining Workshops, 251-258.

[11] Morstatter, F., Ave, S. M., and Carley, K. M., *Is the Sample Good Enough? Comparing Data from Twitters Streaming API with Twitters Firehose,* AAAI 2013.

[12] Lin, J., *Divergence measures based on the Shannon entropy,* IEEE Transactions on Information theory, 37(1), 145-151, 1991.

[13] Cover, T.M. and Thomas, J.A., *Elements of information theory,* John Wiley and Sons, 2012.

[14] Kasiviswanathan, S. P., Melville, P., Banerjee, A., and Sindhwani, V. *Emerging topic detection using dictionary learning.* CIKM 2011.

[15] Lu, R., Xu, Z., Zhang, Y., and Yang, Q. *Life Activity Modeling of News Event.* Advances in Knowledge and Data Discovery 2012.

[16] Cataldi, M., Di Caro, L., and Schifanella, C. *Emerging topic detection on Twitter based on temporal and social terms evaluation.* In Proceedings of the Tenth International Workshop on Multimedia Data Mining 2010.

[17] Altshuler, Y., and Pan, W. *Trends Prediction Using Social Diffusion Models.* Social Networks 2011.

[18] Asur, S., Huberman, B. a., Szabo, G., and Wang, C. *Trends in Social Media: Persistence and Decay.* SSRN Electronic Journal 2011.

[19] Fang, F., Pervin, N., Datta, A., Dutta, K., and VanderMeer, D. *Detecting Twitter Trends in Real-Time.* WITS 2011 (pp. 49-54).

[20] Mathioudakis, M., and Koudas, N. *Twittermonitor: trend detection over the twitter stream.* In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (pp. 1155-1158). ACM.

[21] Lehmann, J., Goncalves, B., Ramasco, J. J., and Cattuto, C. *Dynamical classes of collective attention in twitter.* Proceedings of the 21st International Conference on World Wide Web - WWW 2012.

[22] Irani, D., Webb, S., Pu, C., Drive, F., and Gsrc, B. *Study of Trend-Stuffing on Twitter through Text Classification*, CEAS 2010.

[23] Google trend manipulation. *http://piloseo.com/google/trends-manipulation/*

[24] Cha, M., and Haddadi, H. *Measuring user influence in twitter: The million follower fallacy.* ICWSM 2010.

[25] J. Weng, E. P. Lim, J. Jiang, and Q. He. *TwitterRank: finding topic-sensitive influential twitterers.* In Proceedings of the third ACM International Conference on Web Search and Data Mining (WSDM '10).

[26] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. *Influence and passivity in social media.* In Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases, 18-33.

[27] Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. *Identifying 'influencers' on twitter.* In Fourth ACM International Conference on Web Seach and Data Mining (WSDM).

[28] Lin, Y. R., Margolin, D., Keegan, B., Baronchelli, A., and Lazer, D. *#Bigbirds Never Die: Understanding Social Dynamics of Emergent Hashtag.* ICWSM 2012.

[29] Yang, L., Sun, T., Zhang, M., and Mei, Q. *We know what@ you# tag: does the dual role affect hashtag adoption?* In Proceedings of the 21st International Conference on World Wide Web (pp. 261-270). ACM.

[30] Agrawal, R., Potamias, M., and Terzi, E. *Learning the Nature of Information in Social Networks.* ICWSM 2012.

[31] Yang, J., and Leskovec, J. *Modeling information diffusion in implicit networks.* 2010 IEEE 10th International Conference on Data Mining (ICDM 2010).

[32] Vapnik, V. *Statistical Learning Theory.* Wiley, 1998.

[33] Papageorgiou, C., Oren, M. and Poggio, T. *A general framework for object detection.* In Proceedings of the International Conference on Computer Vision, 1998.

[34] Joachims, T. *Text categorization with support vector machines: Learning with many relevant features.* In Proceedings of European Conference on Machine Learning, 1998.

[35] Tong, S. *Support vector machine active learning for image retrieval.* In Proceedings of the ninth ACM International Conference on Multimedia, 2001.

[36] DTREG. SVM - Support Vector Machines. *http://www.dtreg.com/svm.htm*, Feb 2011.

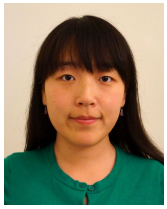
[37] Chang, C. C., and Lin, C. J. *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2011.

[38] Agarwal, M. K., Ramamritham, K., and Bhide, M. *Real time discovery of dense clusters in highly dynamic graphs: Identifying real world events in highly dynamic environments.* Proceedings of the VLDB Endowment 2012, 5(10), 980-991.

[39] Thomas, K., Mccoy, D., Grier, C., and Paxson, V. *Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse.* USENIX Security Symposium 2013.

[40] Egele, M., Kruegel, C., and Vigna, G. *COMPA: Detecting Compromised Accounts on Social Networks.* NDSS 2013.

Yubao Zhang Yubao Zhang is a Ph.D. student in the Department of Electrical and Computer Engineering, University of Delaware. He received his bachelor and master degrees from National University of Defense Technology, China. His research interests lie in online social network and user behavior analysis.



Xin Ruan Xin Ruan is a Ph.D. candidate in Computer Science at the College of William and Mary, Williamsburg VA, USA. She received her bachelor and master degrees from Xidan University, China. Her research interests lie in online social networks and user privacy protection.

Haining Wang Haining Wang received his Ph.D. in Computer Science and Engineering from the University of Michigan at Ann Arbor in 2003. Currently He is a Professor of Electrical and Computer Engineering at the University of Delaware, Newark DE, USA. His research interests lie in the areas of security, networking system, and cloud computing. He is a senior member of IEEE.

Hui Wang Hui Wang received Ph.D. degree in Engineering Systems from National University of Defense Technology, China. He is now a Professor at National University of Defense Technology. His research interests are in the areas of n multimedia intelligence analysis, large-scale P2P systems modeling and dynamic social networks analysis.

Su He Su He is a Ph.D. candidate in National University of Defense Technology, Changsha, China. He received his bachelor and master degrees from National University of Defense Technology, China. His research interests lie in online social network and data mining.