Identifying Spreading Sources and Influential Nodes of Hot Events on Social Networks

Nan Zhou[1], Xiu-Xiu Zhan[2], Qiang Ma[1], Song Lin[1], Jun Zhang[3], and Zi-Ke Zhang[1(✉)]

[1] Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 311121, People's Republic of China
Zhangzike@gmail.com
[2] Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft 2628 CD, The Netherlands
[3] Shanghai Surfing City Information S&T Co. Ltd., Shanghai 201315, China

Abstract. The rapid development of World Wide Web accelerates information spreading in various ways. Thanks to the emergence of multiple social platforms, some events which are not much attractive in the past can become social hot spots nowadays. In this paper, we study the information diffusion process of "*IP MAN3* box office fraud", which is widely diffused in the largest Chinese microblogging system, namely *Sina Weibo*, in March 2016. Based on the temporal metric we have proposed, we succeed in finding out the sources of the information, and constructing the panorama of the diffusion process. In addition, a portion of nodes that promote the diffusion are identified by using the node importance algorithms. Finally, the users with abnormal behaviors in the process of event development are identified.

Keywords: Critical nodes identification · Source tracing · Information diffusion

1 Introductions

The rapid development of the Internet brought great convenience for our daily life. The creation of social platforms makes people much closer to each other, which shows a small-world effect [1]. In addition, the information spreading speed are much faster than ever before, which makes the outbreak size of the information even larger. Recently, much attention has been paid on information diffusion via social networks. Most social networks studied have shown the property of small-world and scale-free effect [2]. Researches on social networks includes information spreading, user behavior analysis, as well as social recommendation [3].

In the past, most studies of the information spreading process are based on the compartment models, which are originated from epidemic spreading models. Usually, researchers use mean field approach to solve these models [4], thus ignoring the network structure in the real world. In fact, the network structure

C. Cherifi et al. (eds.), *Complex Networks & Their Applications VI*,
Studies in Computational Intelligence 689, https://doi.org/10.1007/978-3-319-72150-7_76

can affect the diffusion of information, and in turn the diffusion of information can also change the network structure. Therefore, nodes which play an important role in the information spreading process has become a new hot topic recently. At present, there are four main sorting methods for node importance [5]: (1) ranking methods based on node neighbor information [6]; (2) path-based ranking methods [7]; (3) eigenvector based sorting methods [8]; (4) ranking methods based on node removal [9,10]. These methods depict the importance of nodes in the process of information diffusion from different perspectives.

However, we not only need to find the important nodes in the social network, but also need to track the information sources. In this paper, we study the diffusion process of "IP MAN3 box office fraud" on Sina microblogging system. We first crawl the data from the platform and use the complex network theory to further identify the important nodes and information sources in the diffusion process [3,11–14]. Section 2 gives some background and data collection process; In Sect. 3, we give the main methods used in the work. In Sect. 4, the main results are shown. We conclude our work and give some future direction of this work in Sect. 5.

2 Data Collection and Analysis

2.1 Background

The film "*IP MAN3*" was released on March 4th, 2016 in China, but soon the news about its box office fraud was exposed by mainstream media. Three days later, on March 7th, *SARFT* interviewed with the chief person responsible for the movie and confirmed the truth of fake box office, which means 32 million RMB is faked at the box office. As punishment, the film distributor was forbidden to distribute new movies in next month, resulting in huge economic losses. The main purpose of this paper is to examine the corresponding information diffusion process, and find out the sources and promoters of the very event.

2.2 Data Collection

There are 212 million monthly active users on the Sina microblogging platform, which brings a large number of blogging behaviors, including posting, forwarding, comments and likes of microblogs, which could easily be built as a complex information flow network. In order to reflect the diffusion process of the particular event "IP MAN 3" in the network, we firstly extracted all the keywords associated with the event and then collected the information of microblogs containing relevant keywords with the web crawler technology [15–17].

2.3 Data Processing

We construct the information diffusion network based on the collected sdata of posting and forwarding behaviors. The steps are as follows:

1. Forwarding path extraction: Take out each instance of microblog forwarding behavior in the local database and analyse the data by the regular expression "//@nickname:" to extract the previous forwarding user of the microblog, then preserve the data of "repost_user_id, from_user_id, repost_time" in the local database;
2. Forwarding network construction: Use each forwarding path from the local database to construct the directed network, in which each directed edge has a time attribute.
3. Network undirected: The constructed directed network is transformed into an undirected network without regard to multiple edge.

3 Identification of Information Sources and Internet Promoter

3.1 Identification of Information Sources

Current methods of information sources identification are mostly based on text similarity without considering the timeliness. The calculation process is of low computational efficiency and requires many iterations, resulting in lack of capacity in processing hot events of large-scale diffusion. This paper presents an improved tracing method based on the temporal information of text.

We sort all microblogs according to their posting time and measure the possible source sequentially, which simplifies the calculation. Suppose the result of sorting is $\{d_1, d_2, d_3, \ldots d_i, \ldots d_j, d_n\}$, where $j > i$ means that the posting time of d_j is later than d_i. Then we have $H(d_j) = 1/j$, which means that the earlier the posting time, the greater the value of H. However, we cannot determine whether the mircroblog is the information source just by H. For example, some blogs may contain related keywords but the actual content is not related. Another example is that some microblogs have been posted very early, but has not been forwarded. In this paper, we set a time threshold to identify the information sources [18,19].

3.2 Node Importance Identification Metrics

In social network, there are two ways of formation of hot events: content-driven and form-driven. Content-driven makes an event to be popular by making the topic attractive to people. Form-driven is the employment of a large number of spammers who post a large number of related articles or posts. In order to find out those Internet marketers, we can consider them as influential nodes in the network, and then use the complex network theory to find them. Important nodes in complex network are that can have stronger influence on the network structure and function compared to the ordinary nodes, e.g., gross-roots.

3.2.1 Identification of Important Nodes

In the social network analysis, the importance of nodes is also called "centrality". We use three kinds of node importance indices that are defined as follows:

1. Degree centrality: the degree is defined as the number of neighbors of the node. In this method, larger degree implies the node is more important [20].
2. Eigenvector centrality: the centrality of eigenvectors indicates that the importance of a node depends on both the number of its neighbors (i.e., the degree of the node) and the importance of each neighboring node. Let x_i be the eigenvector centrality of node V_i, the importance of all nodes can be expressed as a vector $X = [x1, x2, x3, \ldots, xn]^T$. Giving initial value $X(0)$ for X, the iterative algorithm is $X(t) = cAX(t-1)$, where A is the adjacency matrix of the network. In general, the parameter c is taken as the largest eigenvalue of A, which can speed up the convergence rate [8]. The eigenvector centrality considers more about the topologic structure. From the perspective of diffusion process, the eigenvector centrality is suitable to describe the long-term influence. Such as in the spread of diseases and rumors, the larger eigenvector a node, the greater probability the node close to the source.
3. K-shell centrality: recent studies have shown that nodes' location in the network is also a crucial factor in characterizing node importance. Nodes which are in the core of the network usually have higher influence, even with small degree. Kitsak et al. [21] applied k-shell decomposition in determining the location of the nodes in the network. The specific decomposition process is as follows: we first remove the nodes with degree 1 until there is no nodes with degree 1 in the network, and these removed nodes are defined as 1-shell. Then we remove the nodes with degree 2 until all nodes with this degree are removed, this is 2-shell of the network. We repeat this process as all the nodes in the network are given a shell number [22].

For each of the aforementioned method, we can get a score vector for all the nodes in the network. We then rank the scores in a descending order and the nodes which are located in the top L position of those three vectors are the most important nodes in the network.

3.2.2 Key Neighborhood Identification

Normally, nodes influence can be affected by their neighbors in the process of diffusion [23, 24]. Therefore, it is of great importance to find the neighbors which are with large amount of forwarding. The detailed algorithm is as follows:

(1) Mining the information source nodes IS with degree larger than 1;
(2) Mining the first order neighbors of IS, which are with degree larger than k, where k is a constant;

(3) Mining the second-order important neighbor of IS, which are with degree larger than k;
(4) Mining the third-order important neighbor of IS, which are with degree larger than k;

By artificial validation, we found that 1^{st}-, 2^{nd}-, and 3^{rd}-order important neighbors of IS we have identified here play an important role in the information spreading process. In addition, most of them posted a large number of blogs about "IP Man 3 box office fraud". Therefore, we define the 1st, 2nd, and 3rd order neighbors of the source nodes with forwarding amount greater than k (here we set k = 10) are the abnormal users.

4 Results and Analysis

4.1 Information Source Identification

We analyze the data collected in Sect. 3.1 and find that the keyword of "IP MAN 3 Fraud" firstly appeared in the Sina microblogging platform on March 4th. Then the outbreak of this event is on March 7th with an amount of more than 600 forwarding (Fig. 1).

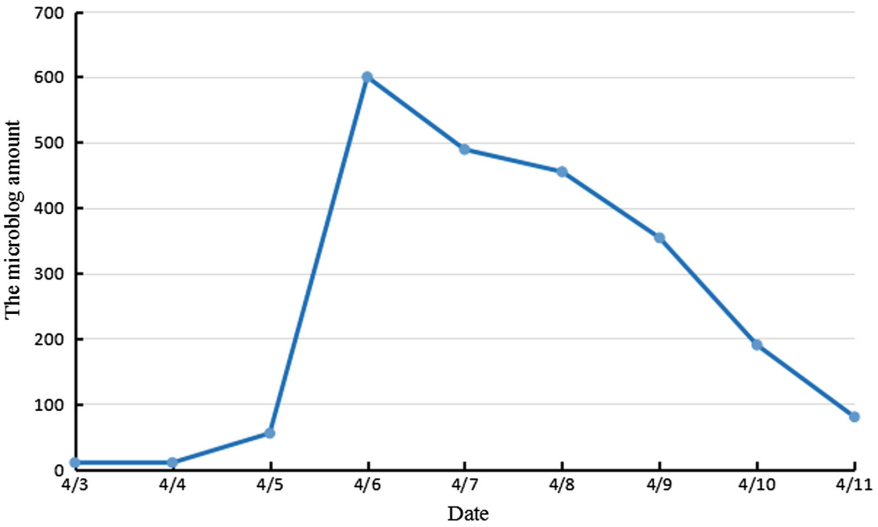

Fig. 1. The trend of microblogging in "IP MAN 3".

According to the method introduced in Sect. 3.1, we identify the source of information about "IP MAN 3 box office fraud" on the microblog platform. We consider the time that those users that have posted the fake news about the movie as movie releasing time. Taking the user with ID 806147150 as an

example, the spreading paths of his blog information is shown in Fig. 2. The size of the nodes shows the forwarding amount of the nodes. It is observed that s/he blog not only contains a large amount of forwarding, but also with long depth of information spreading.

4.2 Information Propagation Path

The company, Kuailu Group, who have invested "IP MAN 3", have suffered a lot of exposure after the spread of "box office fraud". We study users who posted blogs containing all the related keywords of "IP MAN 3", and make a whole network of how the information is diffused in Sina mircroblogging system in Fig. 3.

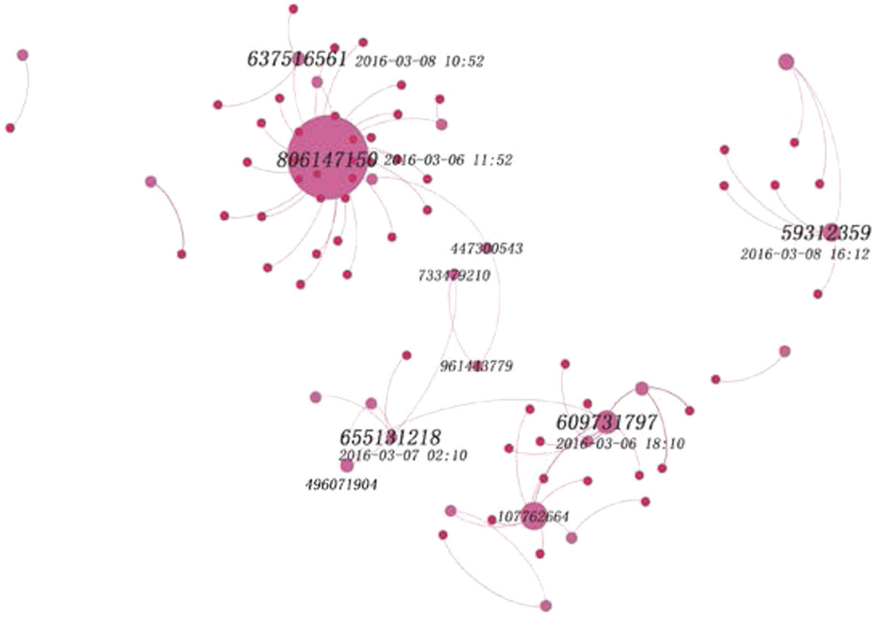

Fig. 2. Constructed forwarding network of an ordinary user #806147150.

The Yellow nodes in the picture are the identified important nodes which have promoted the propagation process. Nodes with larger size indicates larger number of forwarding. At the same time, we also find that some common accounts are showing abnormal behavior, such as releasing a large number of related posts in a short period of time. These abnormal accounts information can be seen in the **Appendix**.

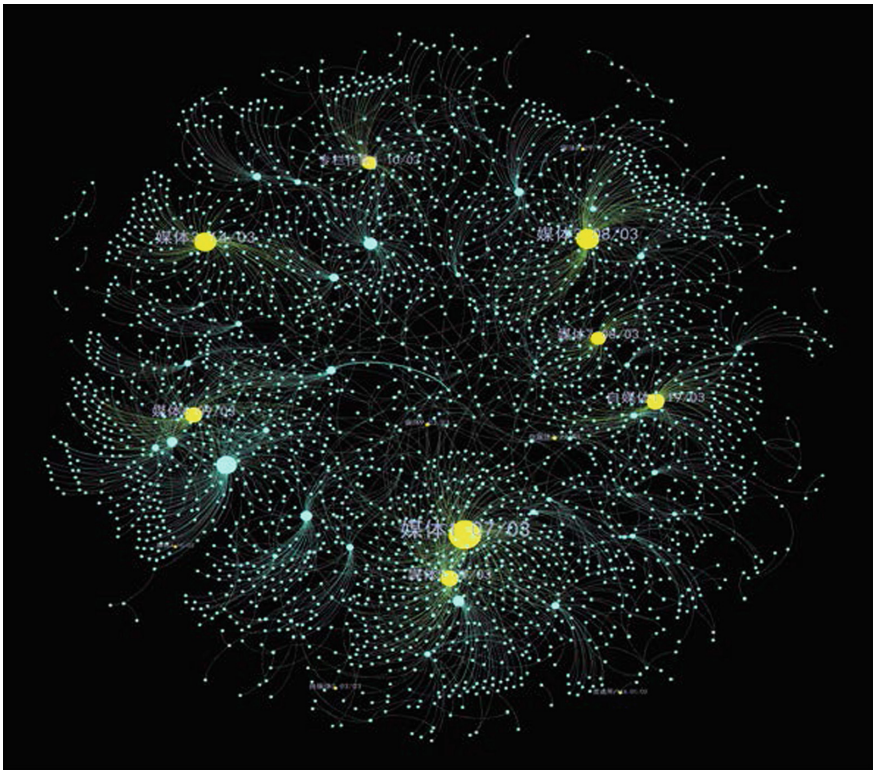

Fig. 3. Information diffusion panorama of "IP MAN 3 fraud" path.

5 Conclusions

This paper studies spreading process of "IP MAN 3 box office fraud" based on one-week data collected from Sina microblog platform from March 4th to 12th, 2016. We obtain the original data through the crawling technology, and construct the information diffusion network. Then, according to node importance theory in complex network, the important nodes in the propagation process are identified. The main contributions of this work are as follows: (1) we found out information sources as well as nodes with large amount of forwarding in the network; (2) users with abnormal behavior in this event are also identified. (3) we construct the information diffusion map of this event. This paper improves the algorithms of finding the source of information and promotes the computational efficiency, then identifies the network promoter by using the methods of node importance. The future work should pay more attention on exploring how to improve the accuracy of finding information sources and abnormal behavioring users with high computation efficiency.

Acknowledgments. This work was partially supported by Zhejiang Provincial Natural Science Foundation of China (Grant Nos. LR18A050001, LY18A050004 and LQ16F030006), Natural Science Foundation of China (Grant Nos. 61673151 and 11671241) and the EUFP7 Grant 611272 (project GROWTHCOM).

Appendix

Abnormal users information		
Anonymous ID	Anonymous name	Type
97394430	Media 5	Official medias
699174196	Media 10	
731638928	Media 11	
853079662	Media 12	
708870469	Media 13	
617383185	Media 14	
612128353	Media 4	
119305779	Media 15	
206755762	Enterprise user 1	Enterprise users
197819660	Enterprise user 2	
124872340	Enterprise user 3	
144799753	Enterprise user 4	
210073340	Enterprise user 5	
735707880	Enterprise user 6	
657976480	Enterprise user 7	
666111015	Enterprise user 8	
200485887	Enterprise user 9	
194044453	Enterprise user10	
207588018	Enterprise user11	
203921855	Enterprise user12	
784250235	Enterprise user13	
209345803	Enterprise user14	
654156459	We Media2	*We* media
609509321	We Media 4	
786955035	Columnist2	Columnist
512500175	Columnist3	
505128349	Ordinary user 17	Ordinary user
205250357	Ordinary user 18	
213261996	Ordinary user 19	
665988504	Ordinary user 20	
441603440	Ordinary user 21	
657288882	Ordinary user 22	
202591342	Ordinary user 23	
803846391	Ordinary user 24	
611645649	Ordinary user 25	
110999748	Ordinary user 26	
772151948	Ordinary user 27	
816024305	Ordinary user 28	
202048903	Ordinary user 29	
806147150	Ordinary user 8	
701165025	Ordinary user 30	
707565922	Ordinary user 14	
116682933	Ordinary user 31	

References

1. Wang, X., Li, X., Chen, G.: Introduction to Network Science, pp. 232–265 (2012)
2. Wu, F., Huberman, B.A., Adamic, L.A., et al.: Information flow in social groups. Phys. A Stat. Mech. Appl. **337**(1), 327–335 (2004)
3. Zanette, D.H.: Dynamics of rumor propagation on small-world networks. Phys. Rev. E **65**(4), 041908 (2002)
4. Moreno, Y., Nekovee, M., Pacheco, A.F.: Dynamics of rumor spreading in complex networks. Phys. Rev. E **69**(6), 066130 (2004)
5. Ren, X., Lü, L.: Review of ranking methods for important node of network. Chin. Sci. Bull. **59**(13), 1175–1197 (2014)
6. Burt, R.S., Minor, M.J.: Applied Network Analysis: A Methodological Introduction. Sage Publications, Inc (1983)
7. Dolev, S., Elovici, Y., Puzis, R.: Routing betweenness centrality. J. ACM (JACM) **57**(4), 25 (2010)
8. Cheng, X.Q., Ren, F.X., Shen, H.W., et al.: Bridgeness: a local index on edge significance in maintaining global connectivity. J. Stat. Mech. Theory Exp. **2010**(10), 10011 (2010)
9. Martin, T., Zhang, X., Newman, M.E.J.: Localization and centrality in networks. Phys. Rev. E **90**(5), 052808 (2014)
10. Yuejin Tan, W., Jun, H.D.: Node contraction method for evaluating node importance in complex network. Syst. Eng. Theory Pract. **26**, 79–83 (2006)
11. Hu, H., Zhu, J.H.: Social networks, mass media and public opinions. J. Econ. Interact. Coord., 1–19 (2015)
12. Xie, J.: Rumor spread in microblogging and its avoidance mechanism. J. Hefei Univ. Soc. Sci. **29**(2), 51–54 (2012)
13. Wang, H., Han, J., Deng, L., et al.: Research on rumor spread dynamics based on mobile social network. Acta Phys. Sin. **62**(11), 110505–110505 (2013)
14. Bao, Y., Yi, C., Xue, Y., et al.: A new rumor propagation model and control strategy on social networks. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1472–1473. ACM (2013)
15. Zhou, D., Li, Z.: High performance web reptile: a summary of research. Comput. Sci. **36**(08), 26–29 (2009)
16. Zhou, Z., Zhang, H., Xie, J.: Sina microblogging data crawler based on Python. J. Comput. Appl. **34**(11), 3131–3134 (2014)
17. Zhou, L., Lin, L.: A summary of the research on crawler technology. J. Comput. Appl. **25**(9), 1965–1969 (2005)
18. Li, G., Gan, T., Guang, Z.: Identifying network pushing hands based on text emotion classification. Libr. Inf. Syst. **54**(8), 77–80 (2010)
19. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. J. Math. Sociol. **2**(1), 113–120 (1972)
20. Chen, D.B., Gao, H., Lü, L., et al.: Identifying influential nodes in large-scale directed networks: the role of clustering. PloS One **8**(10), e77455 (2013)
21. Kitsak, M., Gallos, L.K., Havlin, S., et al.: Identification of influential spreaders in complex networks. Nat. Phys. **6**(11), 888–893 (2010)
22. Freeman, L.C.: Centrality in social networks conceptual clarification. Soc. Netw. **1**(3), 215–239 (1978)
23. Zeng, A., Zhang, C.J.: Ranking spreaders by decomposing complex networks. Phys. Lett. A **377**, 1031–1035 (2013)
24. Liu, J.G., Ren, Z.M., Guo, Q.: Ranking the spreading influence in complex networks. Phys. A **392**, 4154–4159 (2013)