LETTER

Identification of vital nodes in the fake news propagation

To cite this article: Zilong Zhao 2020 *EPL* **131** 16001

A Letters Journal Exploring the Frontiers of Physics

July 2020

EPL, **131** (2020) 16001 www.epljournal.org
doi: 10.1209/0295-5075/131/16001

Identification of vital nodes in the fake news propagation

Zilong Zhao[(a)]

School of Reliability and Systems Engineering, Beihang University - Beijing 100191, China and Science and Technology on Reliability and Environmental Engineering Laboratory - Beijing 100191, China

received 20 April 2020; accepted in final form 24 June 2020
published online 29 July 2020

PACS 64.60.aq – General studies of phase transitions: Networks
PACS 64.60.ah – General studies of phase transitions: Percolation
PACS 89.75.-k – Complex systems

Abstract – Fake news causes an adverse effect on the regular public order and has become easier to propagate with the popularity of online social networks. The threat of fake news propagation makes it important to explore the vital nodes, which are defined as nodes with large branch sizes and hence generating a wider influence than others in this work. Previous studies about identifying vital nodes are mainly from single propagation of fake news networks, which do not consider that users may participate in different propagation networks. Here we identify vital nodes with the feature named the C_k-*value* that combines structural feature out-degree in a single network and multi-network user activeness. The C_k-*value* could reflect the branch size with a strong correlation, even at the early stage of propagation, and percolation based on C_k-*value* is more efficient than other indicators such as node activeness and out-degree. Thus, this research may provide a better understanding of vital nodes in the fake news propagation from topology properties, and further inspires innovative ways to identify vital nodes of fake news propagation.

Introduction. – With the rapid growth of online social network services, billions of Internet users worldwide exchange information conveniently. Unfortunately, along with the popularization of online social networks, harmful and misleading information also propagates among network citizens, especially bringing panic and social losses during emergency incidents [1–3]. For example, among the 30 million tweets exchanged by 2.2 million users, 29% of these tweets on Twitter are fake news, conspiracy theories, or extremely biased news [4]. The fast circulation of fake news can largely undermine the basic value of modern society by reshaping public opinion [5,6]. However, the relevant detection of fake news in the realistic online social network service is insufficient and time-consuming [7,8]. The mismatching of fake news prevalence and detection inspires the analysis of fake news propagation, especially the study of vital nodes during propagation. Some vital nodes could have a much larger harmful influence on the fake news than others. For example, one spreader of fake news is found to participate in eleven networks in our dataset. This super spreader propagates fake news to around two thousand other users before the identification. Hence, the identification of vital nodes is crucial in fake news research.

In the network research, many studies estimate the importance of nodes by using the topological properties including various centrality metrics. For example, the degree is the most basic and simplest measurement yet it is misleading when viewing the global importance in the network [9]. For the networks with bottlenecks, the nodes with high betweenness [10] are usually considered as vital nodes. For the networks where the ranking of the nodes needs to be updated frequently, the vital nodes could have high closeness centrality [11]. For example, in the movie stars cooperation network, the ranking of movie stars usually uses the closeness centrality because it is very sensitive when a new movie is released [12]. Apart from centrality measurement, many studies focus on node importance ranking methods. For instance, the PageRank method [13,14], developed from webpage ranking in Google search engine, considers the ranking of the neighbours of one node. However, PageRank is less efficient in the social network because it does not utilize the leadership topology [15]. With regard to this, the LeaderRank [15] algorithm could find vital nodes more effectively and reliably. Moreover, the HITS

[(a)]E-mail: zhaozilong@buaa.edu.cn

(Hyperlink-Induced Topic Search) algorithm calculates the hub scores and authority scores, which also measures the importance of nodes [16,17]. It is used as the basic webpage ranking algorithm in Ask.com. Additionally, inspired by the idea of the gravity formula, Ma *et al.* propose an algorithm based on gravity to identify influential spreaders [18]. They also test the effect on a realistic network.

Specifically, as for fake news research, studies on vital nodes are also quite critical. Doerr *et al.* find that the nodes with smaller degrees play an important role in the fake news spreading because they could quickly forward the information to their neighbouring nodes [17]. Recently, Indu *et al.* mapped the propagation of fake news in social networks to the spread of a forest fire, and they identified the major nodes during the fake news dissemination process [19]. Based on the hypothesis that the information propagates among friend relationships, Nam *et al.* excavate the smallest group of vital nodes in order to control the propagation of fake news [20]. More specifically, Leskovec *et al.* study a set of vital nodes during the out-break situation [21]. Wang *et al.* study the immunity of some nodes and hence they stop the spreading between them and their neighbours [22]. These studies above identify vital nodes in the single propagation network. However, single propagation networks are not independent of each other because of the repetitive users and organization between them [23]. From the multi-network perspective, the active users who participate in many fake news networks play crucial roles in the global fake news propagation.

In this work, we find that the C_k-*value*, considering both the node importance in a single propagation network (out-degree) and the node activeness among different networks (R-*value*), could effectively describe the branch size, thus identifying vital nodes during the fake news propagation. Additionally, removing nodes based on C_k-*value* is shown to be more destructive compared with other methods based on out-degree or R-*value*, respectively, let alone random removal. Importantly, these findings could emerge at the early stage of fake news spreading, which demonstrates the advantage of identifying vital nodes.

Methods. –

Definition of fake news. Fake news is defined as fabricated information. For the 1862 fake news networks, the dataset we collect is from the topic of the officially certified fake news.

Branch size. For a given node, its branch contains the following re-postings of it, which is a measurement representing the propagation influence of this node. The re-posting nodes of the given node are divided into two types: some of these nodes repost the given node's tweets directly (DP) and others repost the given node's tweets via other nodes (IP). The node itself and its following reposters (both direct and indirect) form the branch:

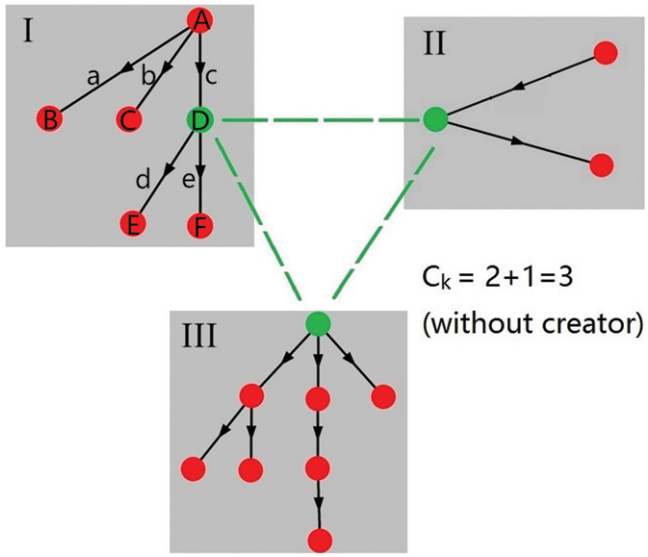

Fig. 1: A schematic diagram of the propagation network and C_k-*value*. The node marked by the green colour participates in three networks in this figure. For example, in the network I, node A stands for the creator who creates the initial message of the propagation network I. Node B reposts node A's message, hence producing the edge a, whose direction is from the information source A to the information receiver B. If more than one re-posting exists between two nodes in the same propagation network, we will only consider the first re-posting by chronological order because this re-posting represents that the information is spread between these two nodes for the first time. We apply the schematic diagram by using Pajek software here.

$$B = DP + IP + 1, \tag{1}$$

where B is the branch size of this node in a network, DP is the number of direct re-posting nodes of this node (edge a, b and c in fig. 1), IP is the number of indirect re-posting nodes of this node (edge d and e in fig. 1).

Note that the average branch size does not count the number of networks that the given node participates as a creator. For example, if this given node takes part in R-*value* networks as a creator for Rc-*value* networks, we only consider the average branch size for this node not as a creator:

$$\langle B \rangle = \frac{\sum_{j=1}^{(R-R_c)} B_j}{(R - R_c)}, \tag{2}$$

where B_j is the branch size of the given node in the j-th network, R is the number of networks that the given node participates in, R_c is the number of networks that the given node participates in as a creator.

For example, in fig. 1, the branch sizes of the green node are two, one and seven, respectively, in network I, II and III. The average branch size is 1.5, which does not consider network III.

Active nodes. The active nodes represent users frequently appearing in different propagation networks [23]. The *R-value* is the number of networks that one node participates in, which reflects user activeness. Specifically, in this work, we consider nodes with a more than 7 *R-value* as fake news active nodes, which are about the highest 0.1% *R-value* nodes.

C_k-value. For a given node in one network, the out-degree is the number of its outgoing edges. The *C_k-value* combines the structural property out-degree of this node in a single network and the number of networks that this node takes part in,

$$C_k = \sum_{j=1}^{(R-R_c)} k_j, \tag{3}$$

where k_j is the out-degree of the given node in the j-th network.

Similarly, if this node participates in some networks as a creator, we will remove these networks and keep the networks that this node participates in as an ordinary re-poster. For example, in fig. 1, the given node acts as a creator in network III but a re-poster in network I and II. The out-degrees of this node are two and one respectively. As a result, the *C_k-value* is the total value three.

S-value. This parameter is used to select nodes which have large average out-degree and small *R-value* or nodes with small out-degree and large *R-value*.

$$S = \frac{\overline{k} + 1}{(R - R_c)}, \tag{4}$$

where $\overline{k}$ is the average out-degree of the given node.

Nodes with large *S-value* have large average out-degree and small *R-value*. Moreover, we find that many nodes have a zero average out-degree. Hence we add one to the numerator to give nodes with larger *R-value* but zero average out-degree an even smaller *S-value* compared with nodes with small *R-value* and zero average out-degree. As a result, nodes with small *S-value* have small average out-degree and large *R-value*. We also do not consider the creator here.

Percolation process. We respectively remove nodes purposefully and randomly and q is the proportion of removed nodes. The giant component G from percolation theory is a property that describes the largest connected groups of nodes after a certain removal, which is used to measure the function of the remaining network. Here we consider a weakly connected component in directed propagation networks. More specifically, for the purposeful removal, we first rank the importance of a node by a property, for example, the *C_k-value* in each propagation network. Then we delete the vital nodes in each propagation network from the largest to the smallest *C_k-value*, and calculate the G of each propagation network. Finally, we calculate the average value of G among all the networks. As for the random removal, we also calculate the G in each network and then study the average value of G.

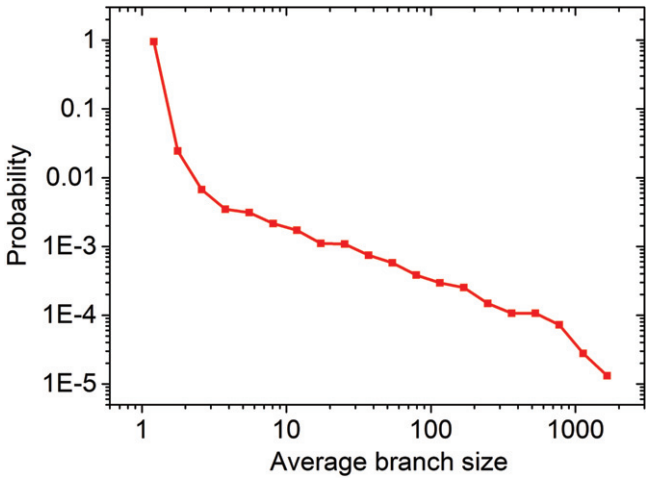

Fig. 2: The distribution of the average branch size of all the nodes.

Results. – Among all the fake news research, the propagation size which reflects the infectious ability of fake news attracts the attention of many scholars. A study based on random network finds that the propagation size of fake news could be at most 79.7% of the entire network [24]. As for the small-world network, the study considers the mean-field equation to propose that the propagation size is less than 80% [25]. Liu *et al.* study both heterogeneous and homogeneous networks and find that the propagation size decreases with larger network heterogeneity [26]. Moreover, a study finds that the scale-free network has a smaller propagation size compared with random network [27]. Lu *et al.* study the epidemic propagation by proposing a dynamic infection rate [28]. From the global perspective, the study of the propagation size is essential because it means the proportion of nodes which hear or believe the fake news. Meanwhile, from the local perspective, the branch size of a node has rarely been investigated.

This work focuses on the identification of vital nodes among the propagation networks. The branch size indicates all the nodes that one node could affect, as a result, it is considered as the measurement of node importance. One node may appear in several networks, and the average branch size is the mean value of the branch size for this node in different networks. Inevitably, the most vital node has the maximum branch size. However, we study the average branch size here to measure the node importance rather than studying the maximization of spreading. As for the creator, it is obvious that it acts as a vital node in the propagation network, hence we do not consider this situation. From the definition of branch size, we know that the branch size of a creator (the size of the network) is very large. As a result, when we calculate the average branch size of a given node, we do not consider networks where this node acts as a creator to avoid possible error. As shown in fig. 2, most nodes have a small average branch size, while a few nodes have a very large average branch

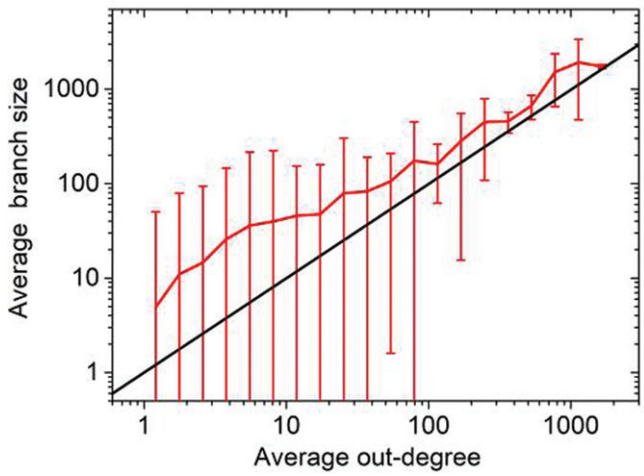

Fig. 3: The relationship between the average out-degree and the average branch size. We first set a group of windows by the average out-degree. For all the nodes in each window, we further calculate the average value and the standard deviation value of branch sizes.

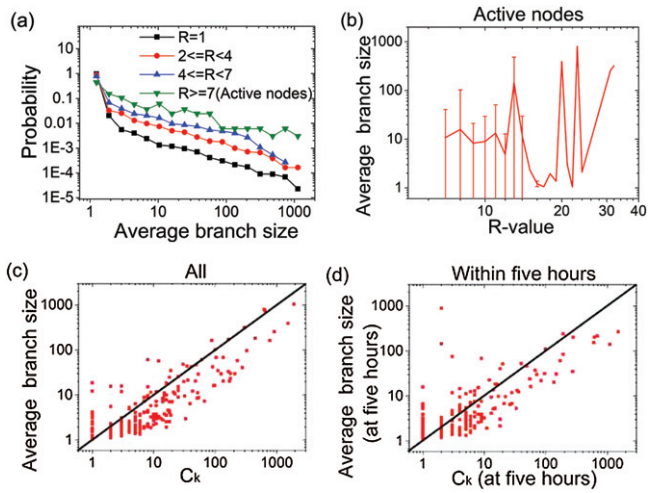

Fig. 4: For the properties of multi-networks, the C_k-*value* could measure the average branch size better than the *R-value*. (a) The distribution of the average branch size for different *R-values*. (b) More specifically, analysing inside the active nodes (green curve) in (a), we plot the average branch size line and its error bar for different *R-values*. The Pearson correlation coefficient between the *R-value* and the average branch size is 0.32. (c) The scatter plot for the C_k-*value* and the average branch size for active nodes for the whole lifespan. (d) The scatter plot for the C_k-*value* and the average branch size for active nodes at an early stage. The Pearson correlation coefficients between the C_k-*value* and the average branch size are respectively 0.85 (c) and 0.81 (d).

size even above one thousand. Thus, we further concentrate on identifying these vital nodes which have a large branch propagation size.

As a typical indicator, we first study the average out-degree which describes the number of direct re-posting nodes. One node may take part in more than one network, and in each network, it has different values of out-degree. As a result, we calculate the mean value of these out-degrees among networks in which this node participates not as a creator. As shown in fig. 3, a node who has a larger average out-degree is more likely to have a larger average branch size. Additionally, the average branch size curve (red) is above the average out-degree line (black), because the direct re-posting nodes are part of the branch. However, the branch size has a large fluctuation as a function of degree. The out-degree used here may ignore the user coupling in multi-networks and its activeness.

Considering multi-networks rather than the single network, we find that some users repeatedly participate in more than one network [23]. The *R-value* is defined as the number of networks that this node participates in, which shows the frequency of node appearing. From the global perspective, as shown in fig. 4(a), the nodes tend to have a relatively large average branch size when the range of *R-value* increases. For example, the red curve ($2 \leq R < 4$) is above the black curve ($R = 1$) when the range of *R-value* rises. Particularly, here we select the nodes with the highest 0.1% *R-value* nodes as active nodes (see the "Methods" section). In fig. 4(a), the green curve shows that active nodes tend to have large average branch sizes with higher probability. Therefore, these active nodes are much more infectious than others according to the branch size.

From fig. 4(a), the positive relationship between *R-value* and the average branch size is prominent in the global perspective. When it comes to individual nodes, the positive relationship remains yet a large fluctuation. In fig. 4(b), when we look inside the detailed information in the green curve of fig. 4(a), we find that the relevance between *R-value* and the average branch size of active nodes shows a positive relationship with the Pearson correlation coefficient 0.32, with the large error bar showing large fluctuation. As a result, although the *R-value* is related to the average branch size in the general perspective, it could not fully explain the branch size. A deeper explanation for the branch size with less fluctuation is required.

Both the average out-degree and the *R-value* could reflect the importance of nodes to some extent. However, the out-degree is a topology property in a single network. The *R-value* means user activeness which considers networks coupling from the user perspective, which neglects the topological effort. Therefore, we propose to use the C_k-*value* (see the "Methods" section) to combine these two properties: the average out-degree and the *R-value*. From above we know that active nodes are more vital, hence we concentrate on active nodes in fig. 4(c). The C_k-*value* and the average branch size of active nodes are found to have a strong positive correlation. On the one hand, the C_k-*value* measures the total interest on this node across different information propagation networks. On the other hand, the average branch size of a given node

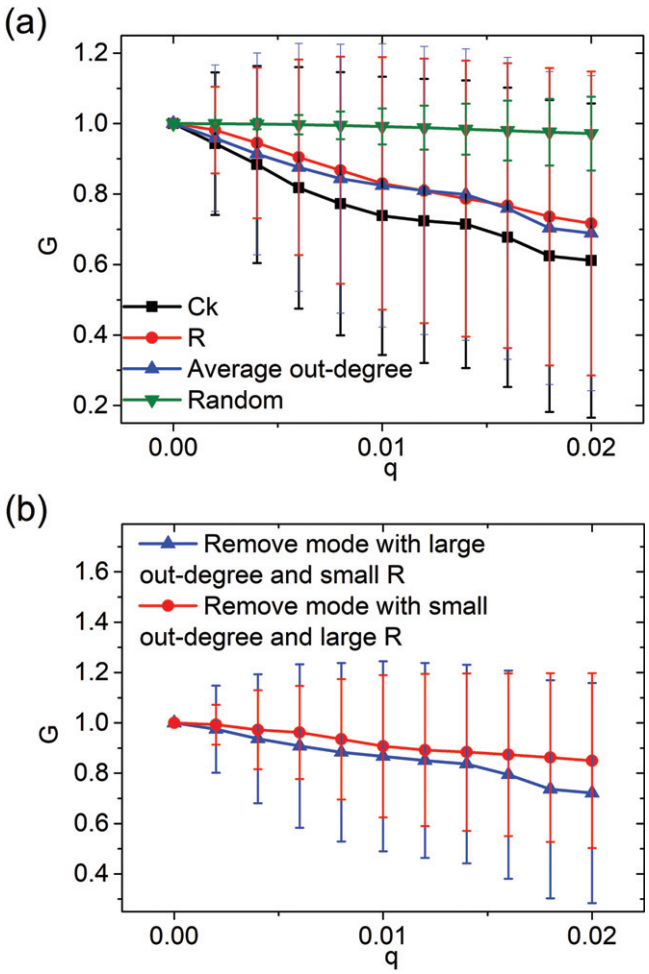

Fig. 5: (a) The average and the standard deviation of G after removing vital nodes for fake news. The average G decreases to 0.72 and 0.69 by removing 2% nodes with high *R-value* or high average out-degree, while the average G is 0.97 with the same amount of nodes removed randomly. The average G even decreases more to 0.61 with the same amount of nodes removed considering C_k-*value*. (b) The average and the standard deviation of G after removing nodes with large average out-degree and small *R-value* or opposite nodes with small out-degree and large *R-value*. The nodes with large average out-degree and small *R-value* could have a larger effect on the network efficiency: the average G decreases to 0.72 by removing 2% nodes compared with 0.85 for nodes with small out-degree and large *R-value*.

in one network could be well predicted by this C_k-*value*. The relatively small fluctuation between the C_k-*value* and the branch size indicates that it is a better measurement of vital nodes rather than only *R-value*. Furthermore, in order to study the relationship between the C_k-*value* and the average branch size at the early stage, we only consider the re-postings within several hours after the first re-posting of the network. We find that the relationships between the C_k-*value* and the average branch size for different time lengths are similar at the early stage, hence we choose five hours as an example time and plot the scatter figure as shown in fig. 4(d). The C_k-*value* and the average branch size still have a strong positive correlation with the Pearson correlation coefficient 0.81. Therefore, the C_k-*value* is stable to predict the branch size timely in the propagation network.

Since the C_k-*value* is a useful indicator of the nodes' information propagation ability, we further analyse the giant component of propagation after removing nodes in descending order of the C_k-*value*. As shown in fig. 5(a), although there is a fluctuation, the average giant component decreases faster by deleting high C_k-*value* nodes than by deleting other removal methods, which indicates that removing nodes with a higher C_k-*value* could prevent the fake news spread better. Both the *R-value* and the average branch size could contribute to the C_k-*value*, hence we further study whether nodes with large average out-degree and small *R-value* or opposite nodes with small out-degree and large *R-value* have a larger influence on the average (G). We first rank all nodes in the propagation network by *S-value* (see the "Methods" section). In the percolation process, we remove nodes in a descending or ascending order of the *S-value*. As shown in fig. 5(b), the nodes with large average out-degree and small *R-value* have a larger effect on network efficiency with relatively smaller average G and larger fluctuation.

Discussion. – In this era of information explosion, various kinds of fake news are easily created and spread. The current situation in social networks indicates the significance and necessity of an analysis for fake news propagation, based on the detection of fake news among news [29], the next important research is the identification of influential spreaders (vital nodes) in fake news propagation.

This paper applies C_k-*value* as a symbol of branch size (also the importance of a node), which considers both the propagation ability in a single propagation network and the user activeness among all the networks. Moreover, the silence of only a few vital users selected by the C_k-*value* could interrupt fake news propagation effectively. This suggested method is simple and understandable: fake news propagation counts on vital spreaders since these users could express novel negative information to attract human attention and encourage human interaction [30]. This method considers properties which form both single networks and multi-networks, and could predict node influence with little fluctuation. More importantly, this method is extremely time-saving: we test the relationship between the C_k-*value* and the branch size, finding that they are highly related at the beginning of fake news spreading. One could perform a further test with more fake news datasets in the future.

Further analysis of the topological features for vital nodes could be a promising research direction. Our finding, combining local structural features and global user activeness, may inspire creative machine-learning methods for detection of fake news in future research. Specifically,

the authorities could pay more attention to vital nodes for the detection and containment of fake news, and the vital users could also be considered as a propagation catalyst in the advertising industry. Therefore, further research in these fields based on vital nodes is well worth carrying out.

REFERENCES

[1] Ruan Z., Tang M. and Liu Z., *Phys. Rev. E*, **86** (2012) 036117.
[2] Kosfeld M., *J. Math. Econ.*, **41** (2005) 646.
[3] Zollo F., Bessi A., Del Vicario M. *et al.*, *PLoS ONE*, **12** (2017) e0181821.
[4] Bovet A. and Makse H. A., *Nat. Commun.*, **10** (2019) 7.
[5] Quattrociocchi W., *How does misinformation spread online?*, World Economic Forum (2015).
[6] Del Vicario M., Bessi A., Zollo F. *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, **2016** (113) 554.
[7] *Fact-checking fake news on Facebook works - just too slowly*, `https://phys.org/news/2017-10-fact-checking-fake-news-facebook-.html#jCp`.
[8] The report of Weibo refuting rumors in 2018, `http://www.piyao.org.cn/2019-02/03/c_1210053804.htm`.
[9] Xia Y. and Fan J., *Efficient attack strategy to communication networks with partial degree information*, in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)* (IEEE) 2011, pp. 1588–1591.
[10] Kim H., Tang J., Anderson R. *et al.*, *Comput. Netw.*, **56** (2012) 983.
[11] Okamoto K., Chen W. and Li X.-Y., *Ranking of closeness centrality for large-scale social networks*, in *International Workshop on Frontiers in Algorithmics* (Springer) 2008, pp. 186–195.
[12] Newman M., *Networks: An Introduction* (Oxford University Press) 2010.
[13] Wang J. and Wang S., *A method of discovering key nodes for online social network based on coritivity theory*, in *Proceedings of the 2017 6th International Conference on Measurement, Instrumentation and Automation (ICMIA 2017)* (Atlantis Press) 2017.
[14] Page L., Brin S., Motwani R. *et al.*, *The PageRank Citation Ranking: Bringing Order to the Web*, Stanford InfoLab (1999).
[15] Lü L., Zhang Y.-C., Yeung C. H. *et al.*, *PLoS ONE*, **6** (2011) e21202.
[16] Kleinberg J. M., *J. ACM*, **1999** (46) 604.
[17] Doerr B., Fouz M. and Friedrich T., *Commun. ACM*, **55** (2012) 70.
[18] Ma L. L., Ma C., Zhang H.-F. *et al.*, *Phys. A: Stat. Mech. Appl.*, **451** (2016) 205.
[19] Indu V. and Thampi S. M., *J. Netw. Comput. Appl.*, **125** (2019) 28.
[20] Nguyen N. P., Yan G., Thai M. T., *et al.*, *Containment of misinformation spread in online social networks*, in *Proceedings of the 4th Annual ACM Web Science Conference* (ACM) 2012, pp. 213–222.
[21] Leskovec J., Krause A., Guestrin C. *et al.*, *Cost-effective outbreak detection in networks*, in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM) 2007, pp. 420–429.
[22] Wang H., Chen C., Qu B. *et al.*, *New J. Phys.*, **19** (2017) 073039.
[23] Li D., Gao J., Zhao J. *et al.*, *Repetitive users network emerges from multiple rumor cascades*, arXiv preprint, arXiv:1804.05711 (2018).
[24] Sudbury A., *J. Appl. Probab.*, **22** (1985) 443.
[25] Zanette D. H., *Phys. Rev. E*, **65** (2002) 041908.
[26] Liu Z., Lai Y.-C. and Ye N., *Phys. Rev. E*, **67** (2003) 031911.
[27] Zhou J., Liu Z. and Li B., *Phys. Lett. A*, **368** (2007) 458.
[28] Lu D., Yang S., Zhang J. *et al.*, *Chaos*, **27** (2017) 083105.
[29] Zhao Z., Zhao J., Sano Y. *et al.*, *EPJ Data Sci.*, **9** (2020) 7.
[30] Itti L. and Baldi P., *Vis. Res.*, **49** (2009) 1295.