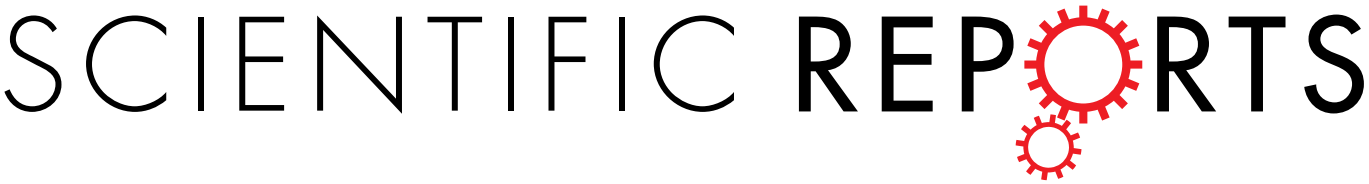

OPEN

Identifying a set of influential spreaders in complex networks

Jian-Xiong Zhang[1,2], Duan-Bing Chen[1,2], Qiang Dong[1,2] & Zhi-Dan Zhao[2]

Received: 09 February 2016
Accepted: 23 May 2016
Published: 14 June 2016

Identifying a set of influential spreaders in complex networks plays a crucial role in effective information spreading. A simple strategy is to choose top-*r* ranked nodes as spreaders according to influence ranking method such as PageRank, ClusterRank and *k*-shell decomposition. Besides, some heuristic methods such as hill-climbing, SPIN, degree discount and independent set based are also proposed. However, these approaches suffer from a possibility that some spreaders are so close together that they overlap sphere of influence or time consuming. In this report, we present a simply yet effectively iterative method named VoteRank to identify a set of decentralized spreaders with the best spreading ability. In this approach, all nodes vote in a spreader in each turn, and the voting ability of neighbors of elected spreader will be decreased in subsequent turn. Experimental results on four real networks show that under Susceptible-Infected-Recovered (SIR) and Susceptible-Infected (SI) models, VoteRank outperforms the traditional benchmark methods on both spreading rate and final affected scale. What's more, VoteRank has superior computational efficiency.

In real world, many complex systems can be represented as complex networks[1–6], in which, Many activities such as advertising over media and word-of-mouth on social networks can be described by information spreading on complex networks[5,7–11]. Maximizing the scale of spreading is a common target. If a market manager want to advertise a new product on Twitter.com, she/he tries to choose a small number of users to provide them with free products in exchange for posting tweets about the product to influence their friends to buy the products. So, the task of market manager is to choose a few users such that the product information can be transmitted to more users and, more products can be sold finally. With the topology unchanged or changed slightly, the location of source spreaders determines the final scale of spreading on large degree. The problem of choosing initial nodes as source spreaders to achieve maximum scale of spreading is defined as *influence maximization problem*[12]. Our research focuses on the strategy of choosing a set of critical nodes as source spreaders in this report.

As influential nodes have strong ability to affect other nodes, selecting top-ranked influential nodes as source spreaders is a common and classical strategy. Up to now, many ranking methods have been proposed, such as degree, closeness[13], betweenness[14] centralities, and other heuristic algorithms[15–18]. Random-walk based methods such as well-known PageRank[19] and LeaderRank[20] have been receiving great attentions and shown significant value in last few years. Pei *et al.*[21] addressed a direct method to search for influential spreaders by following the real spreading dynamics in a wide range of networks. Some other methods such as HITS[22] and TwitterRank[23] are also useful and effective. Recently a local based method ClusterRank[24] has also good performance in some cases. Ref. 25 shows that the crucial factor of node's influence is its location in network measured by *k*-shell value. Under this measuring strategy, nodes with larger *k*-shell values usually have more ability on spreading. Wei *et al.*[26] proposed a weighted *k*-shell decomposition to identify influential nodes. Liu *et al.*[27] introduced a measure based on link diversity of shells to distinguish the true core and core-like group so as to find the real influential spreaders. Based on ref. 27, Liu *et al.*[28] proposed an improved *k*-shell method by removing redundant links. Lü *et al.*[29] unveiled the elegant mathematical relationship among three simple yet important centrality measures of networks, i.e., degree, H-index and coreness. Ref. 25 indicates that top-ranked nodes obtained by *k*-shell decomposition are of significant influences. However, if select them as a group to spread, the result is not so good, even is worse than the result of pure degree centrality. Like *k*-shell method, other ranking methods such as closeness, PageRank, LeaderRank and ClusterRank suffer the similar limitation.

Kempe *et al.*[12] proposed a hill-climbing based greedy algorithm that can find a group of important nodes to affect the widest scope of nodes. Their work can overcome the shortcoming mentioned before. However, it is very time consuming, especial in large scale networks. Based on greedy strategy, Narayanam and Narahari[30] proposed

[1]Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 611731, P.R. China. [2]Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, P.R. China. Correspondence and requests for materials should be addressed to D.-B.C. (email: dbchen@uestc.edu.cn)

a much faster algorithm SPIN approach than greedy algorithm while its quality decreasing a little. Unfortunately, SPIN is also hardly applied to large scale networks. For example, the CPU running time is 28.25 minutes if to find top-30 important nodes in network with 1589 nodes. For this reason, some fast heuristic algorithms are presented in recent years. Chen *et al.*[31] proposed degree discount heuristic algorithm, which nearly matches the performance of the greedy methods for the IC model. Tang *et al.*[32] presented a Two-phase Influence Maximization (TIM) algorithm that aimed to bridge the theory and practice in influence maximization. In theory, TIM runs in $O((r+\ell)(n+m)\log n/\varepsilon^2)$ expected time and returns a $(1-1/e-\varepsilon)$-approximate solution with at least $1-n^{-\ell}$ probability where ℓ and ε are parameters. Zhao *et al.*[33] made an attempt to find a set of important spreaders by generalizing the idea of the coloring problem in graph theory[34] to complex networks. Ji *et al.*[35] proposed an effective multiple leaders identifying method based on percolation theory. The method well utilizes the similarities between the pre-percolated state and the average of information propagation in each social cluster to obtain a set of distributed and coordinated spreaders. Very recently, Morone and Makse presented an effective method to find a set of critical nodes by mapping the problem onto optimal percolation in random networks[36]. He *et al.*[37] proposed a novel method to identify multiple spreaders in complex networks with community structures.

In this report, we propose a simply yet effectively iterative method named VoteRank to choose a set of influential spreaders. In our method, influential spreaders are elected one by one according to their voting scores obtained from their neighbors. At each iteration, the voting ability of elected spreader will be set to zero while that of its neighbors will be decreased by a factor. Our method can be applied to large scale network with millions of nodes since it just updates local information after selecting a spreader. Experimental results on real datasets show that our method outperforms traditional methods on both final affected scale and spreading rate. What's more, VoteRank is also superior to other group-spreader identifying methods on computational time.

Methods and Materials

Spreading Models. In this report, we mainly use SIR epidemic model with limited contact[38,39] to evaluate methods. In SIR model, each node is in one of three statuses, i.e., Susceptible(S), Infected(I) and Recovered(R). Initially, all nodes are susceptible status except for a set of r infected nodes selected as source spreaders. At each time step, infected node tries to infect one of its neighbors with probability μ. At the same time, each infected node will be recovered with a probability β, if success, it won't be infected again and no longer infect other susceptible nodes. The process terminates if there isn't any infected node in network. In this report, we use $\lambda=\mu/\beta$ to represent *infected rate*, which is crucial to infected speed and final affected scale that are often used to indicate the spreading ability of r source spreaders. Besides SIR model with limited contact, the performance of methods can also be evaluated by SIR model with full contact and SI model[40] that is usually used to evaluate the method on spreading rate especially in the early stage.

VoteRank Algorithm. In real world, if a person A has supported person B, the support strength of A to others will fade generally. Under this perspective, a vote based approach for identifying influential spreaders named VoteRank is presented in this report. In VoteRank, the main idea is to choose a set of spreaders one by one according to voting scores of nodes obtained from their neighbors. If we need to select top-r influential spreaders, every node has to vote r turns. The node getting the most votes in each turn is regarded as the most influential node in that turn and will be elected as one of top-r influential spreaders. If a node has been elected as a spreader, it doesn't participate in subsequent voting, and the voting ability of its neighbors also be decreased. Actually, when a node u is elected as spreader, the propagation range has increased a little if the nodes near u are elected as spreaders again since u can transfer information to these nodes. So, it's better to select far apart nodes because they can affect as many nodes as possible. That is to say, after a node is elected as spreader, the selection probability of its neighbors and neighbors' neighbors will decrease. Under this mechanism, the selected nodes are far apart and are important in its local structure. In fact, similar idea has been reported in references. For example, Kitsak *et al.*[25] pointed out that the propagation range would be improved greatly if any two selected spreaders are disconnected comparing with simply selecting nodes with maximum degree or k-shell value one by one.

In VoteRank, each node u is attached with a tuple (s_u, va_u), where voting score s_u denotes the number of votes obtained from u's neighbors and voting ability va_u represents the number of votes that u can give its neighbors. The details of VoteRank are described as following five steps:

step 1: Initialize. Tuples of all nodes are set to (0, 1).

step 2: Vote. Nodes vote for their neighbors, at the same time are voted by their neighbors. After voting step, the voting score of each node will be calculated. It is noted that the voting score of node is set to zero if it has been elected in earlier turn so as to avoid electing it again. For example, node v_0 has three neighbors v_1, v_2 and v_3. Node v_0 will vote for v_1, v_2, v_3 with va_{v_0} votes, and v_1, v_2, v_3 will vote for their corresponding neighbors with va_{v_1}, va_{v_2} and va_{v_3} votes respectively. So, the voting score of node v_0 is $s_{v_0}=va_{v_1}+va_{v_2}+va_{v_3}$. This voting process is different from political voting because some nodes just vote for less one vote in VoteRank.

step 3: Select. According to voting scores calculated in step 2, select the node $v_{\max}$ that gets the most votes. This node will not participate in subsequent voting turns, that is, its voting ability $va_{v_{\max}}$ will be zero from now on.

step 4: Update. Weaken the voting ability of nodes those voted for $v_{\max}$ in step 2. For example, if node u voted for $v_{\max}$, update the voting ability of u with va_u-f unless va_u has been decreased to zero, where f is a decreasing factor being between 0 and 1. For special case of $f=0$, just the degree of newly elected node's neighbors will minus one since only the voting ability of newly elected node turns to zero. In this report, we mainly focus f on a simple form $\frac{1}{<k>}$, where $<k>$ is the average degree of the network.

step 5: Repeat steps 2 to 4 until r spreaders are elected.

In order to give an intuitive explanation, we use VoteRank to choose top-2 nodes on a small toy network with 10 nodes, as shown in Fig. 1. Figure 1(a) represents the first turn of voting. The value of voting score and voting

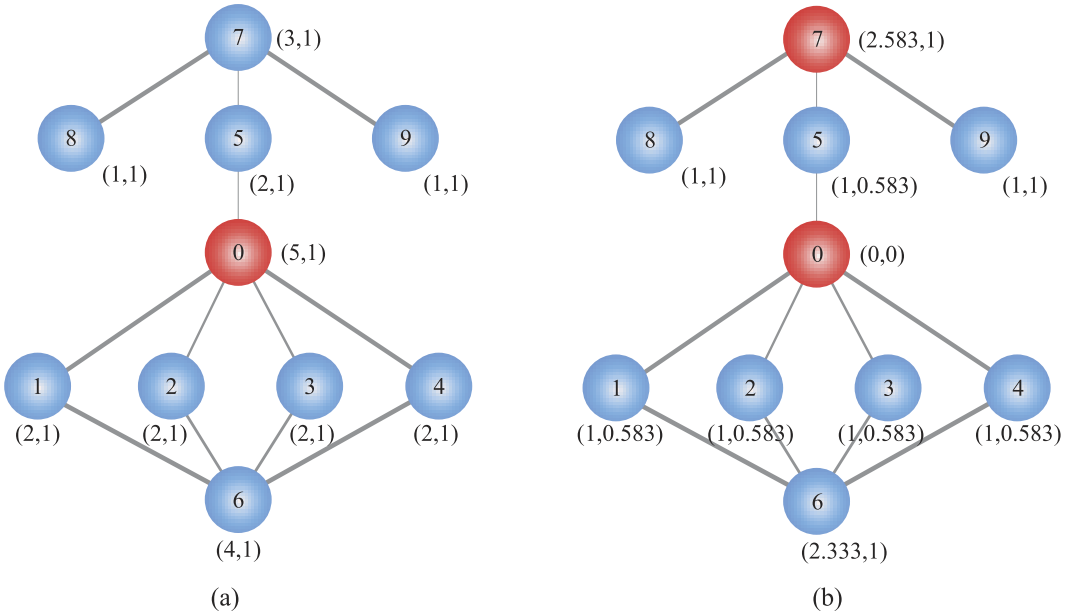

Figure 1. A toy network. In (**a**) node 0 is selected as one of top-2 spreaders, and in (**b**), node 7 is selected as one of top-2 spreaders.

ability for each node is marked as tuple (*s*, *va*) in Fig. 1(a). In this turn, node 0 is chosen and its voting ability is set to 0. The voting abilities of nodes 1, 2, 3, 4 and 5 are reduced by $\frac{1}{2.4} = 0.417$. The updated voting ability of each node is marked in Fig. 1(b). According to new voting abilities, node 7 is chosen since it gets the highest voting score 2.583 at the second voting.

VoteRank algorithm not only can be used to choose top-*r* spreaders in undirected networks, but also can be used in directed networks. In directed network, if there is a link from node *u* to node *v*, *u* is the in-neighbor of *v*, and correspondingly, *v* is the out-neighbor of *u*. In this report, a link from node *u* to *v* indicates that *v* receives information from *u*. The directed version of VoteRank is slightly different from undirected one. Firstly, nodes only vote for their in-neighbors, and secondly, only the voting ability of elected node and its out-neighbors will be updated.

Performance Metrics. In this report, three metrics are used to evaluate the performance of methods. The first two are based on spreading scale under SIR or SI spreading model, and the third is based on structural properties of elected spreaders.

In order to compare the spread speed for different methods, we introduce infected scale $F(t)$ at time t which is defined as:

$$F(t) = \frac{n_{I(t)} + n_{R(t)}}{n}, \tag{1}$$

where n is the number of nodes of network, $n_{I(t)}$ and $n_{R(t)}$ ($n_{R(t)} = 0$ for SI model) are the number of infected and recovered nodes at time t respectively.

In order to investigate the final scale of affected nodes, final affected scale $F(t_c)$ is introduced:

$$F(t_c) = \frac{n_{R(t_c)}}{n}, \tag{2}$$

where $n_{R(t_c)}$ is the number of recovered nodes when spread process achieving steady state.

Besides $F(t)$ and $F(t_c)$, the structural properties among selected spreaders are also used to evaluate the performance of different methods. In this report, the *average shortest path length* L_s between each pair of source spreaders S is used as evaluating metric. It is defined as:

$$L_s = \frac{1}{|S|(|S| - 1)} \sum_{\substack{u,v \in S \\ u \neq v}} l_{u,v}, \tag{3}$$

where $l_{u,v}$ denotes the length of the shortest path from node u to v.

Data Description. Four real networks are used to test the performance of VoteRank in this report. Networks YOUTUBE[41] and COND-MAT[42] are undirected and Networks BERKSTAN[43] and NOTRE DAME[44] are directed. YOUTUBE is a video-sharing web site that includes a social network, in which, nodes represent users and edges represent friendships between two users. COND-MAT is a collaboration network, which generates from the

Networks	n	m	$<k>$	k_{max}	$<c>$	H
YOUTUBE[41]	1134890	2987624	5.2650	28754	0.0808	93.9270
COND-MAT[42]	23133	93497	8.0834	279	0.6334	2.7305
BERKSTAN[43]	685230	7600595	11.0920	249	0.5967	3.1744
NOTRE DAME[44]	325729	1497134	4.5120	3444	0.2346	23.4647

Table 1. The basic topological features of four real networks. n and m are the total number of nodes and edges, respectively. $<k>$ is the average degree for undirected networks or the average out-degree for directed networks. k_{max} is the maximum degree for undirected networks or the maximum out-degree for directed networks. $<c>$ is the average clustering coefficient and H is the degree heterogeneity, defined as $H = \frac{<k^2>}{<k>^2}$[45].

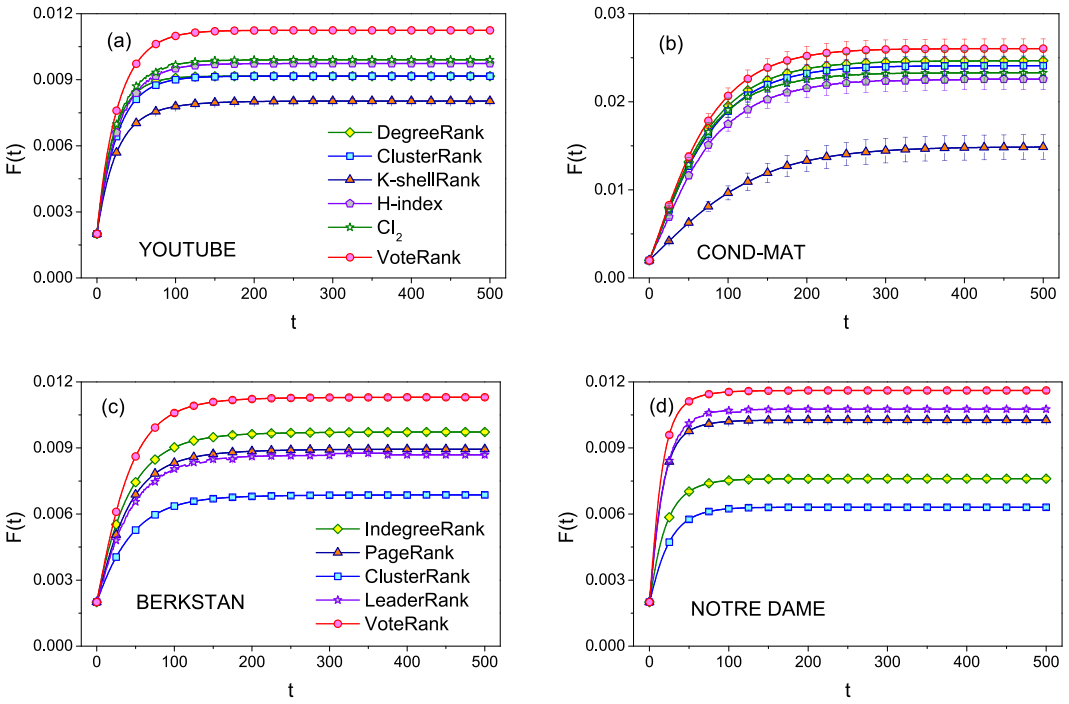

Figure 2. The infected scale $F(t)$ on four networks under different methods, where $\lambda = 1.5$ and $p = 0.002$. The results are averaged over 100 independent runs.

e-print arXiv and covers scientific collaborations between authors who submit papers to Condense Matter category. In BERKSTAN, nodes represent pages from berkely.edu or stanford.edu domains and directed edges represent hyperlinks between them. In NOTRE DAME network, nodes represent pages from University of Notre Dame and directed edges represent hyperlinks between them. Some topological features of these four networks, including the number of nodes n, the number of edges m, the average degree (or average out-degree for directed networks) $<k>$, the maximum degree (or maximum out-degree for directed networks) k_{max}, the average clustering coefficient $<c>$, and the degree heterogeneity H which is defined as $\frac{<k^2>}{<k>^2}$[45], are shown in Table 1.

Results

The performances of VoteRank and other methods are evaluated by three metrics mentioned before on four real networks. Figure 2 shows the infected scale $F(t)$ on four networks under different methods with infected rate $\lambda = 1.5$ and $p = 0.002$ where p is the ratio of the number of source spreaders and that of nodes in network. From Fig. 2, it can be seen that from the source spreaders obtained by VoteRank, information can spread faster and eventually affect larger scale than that by other methods. Moreover, the deviation of $F(t)$ is generally small especial for YOUTUBE, BERKSTAN and NOTRE DAME.

Figure 3 shows the final affected scale $F(t_c)$ with different number of source spreaders. It's obvious that VoteRank can achieve wider final affected scale $F(t_c)$ than other benchmark methods under same number of source spreaders, especially when the number of source spreaders is large.

Figure 4 shows the $F(t_c)$ with different λ for different methods on four networks. From Fig. 4, it can be seen that VoteRank can achieve wider spread scale than other methods under different λ, especial in YOUTUBE, BERKSTAN and NOTRE DAME networks. If λ is too small, information can not be effectively spread no matter how to choose source spreaders. And if λ is too large, information can spread all over the network. For this reason, λ just be ranging from 1 to 2 in this report so as to compare the difference of methods clearly.

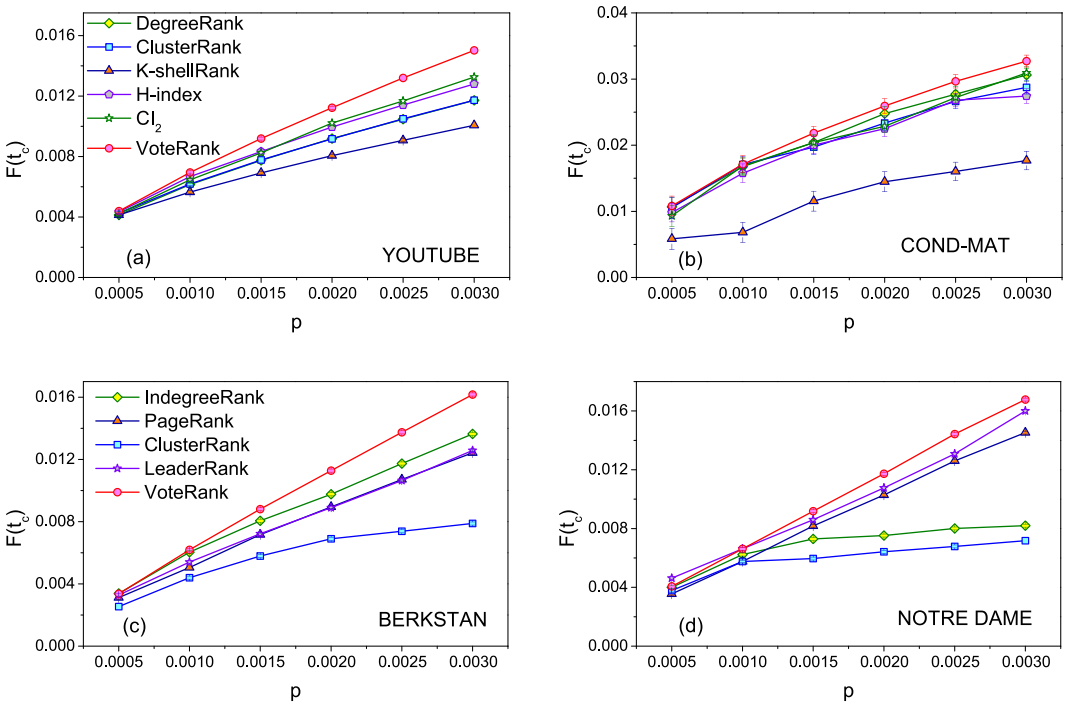

Figure 3. The final affected scale $F(t_f)$ with different number of source spreaders. In (**a–d**), $\lambda = 1.5$, in (**a,b**), $\beta = \frac{1}{< k >}$ and in (**c,d**), $\beta = \frac{1}{< k^{out} >}$. The results are averaged over 100 independent runs.

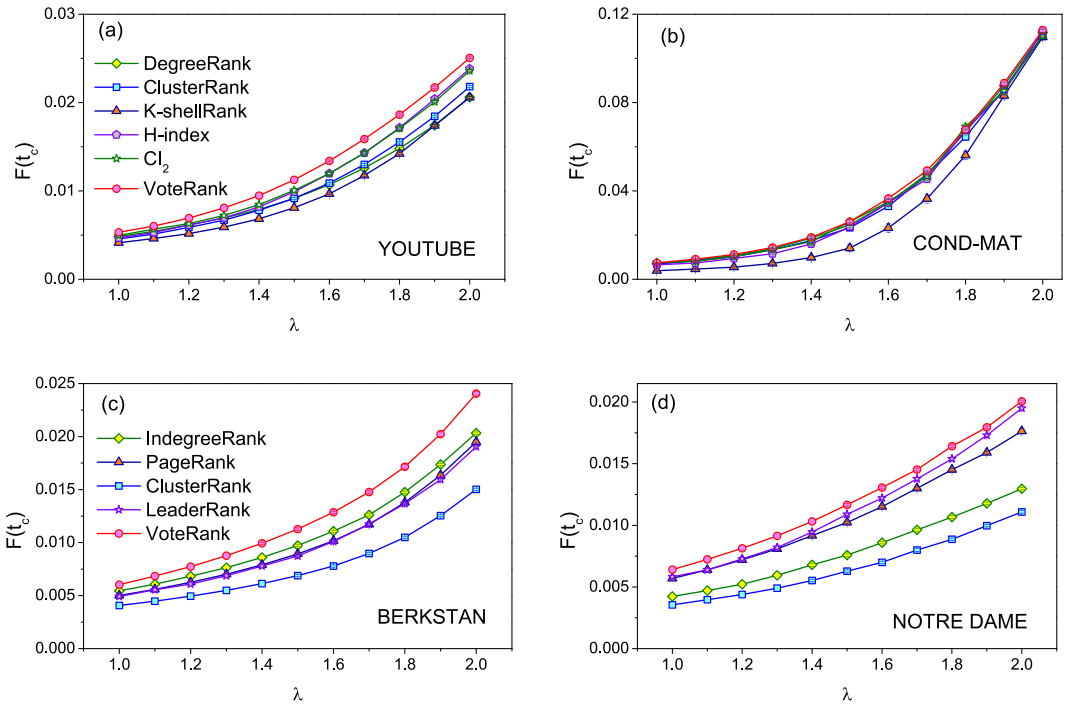

Figure 4. The final affected scale $F(t_c)$ with different infected rate λ where $p = 0.002$. The results are averaged over 100 independent runs.

Actually, final affected scale $F(t_c)$ is not only determined by the influence of source spreaders, but also by their relative location. For this reason, k-shell decomposition can dig out influential single spreader effectively, but perform poorly on selecting group spreaders by simply selecting nodes with the biggest k-shell value. To overcome this limitation in some degree, a reasonably improved strategy is to choose nodes with the highest voting score or k-shell value such that any two of selected spreaders are not directly linked. That is, under current state, if a node with the highest score is the neighbor of any selected spreader, we will skip this node and consider the next

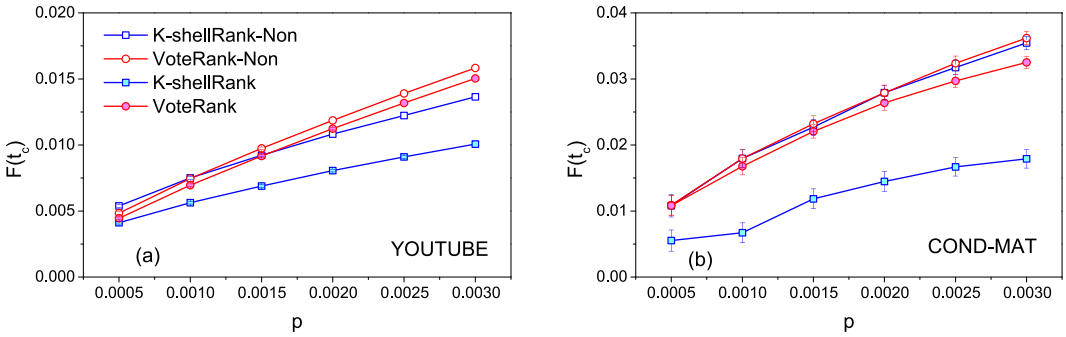

Figure 5. The final affected scale $F(t_c)$ under different initial infected scale p. Both in (**a**,**b**), $\lambda = 1.5$, and K-shellRank-Non and VoteRank-Non are improved versions of K-shellRank and VoteRank, in which, any two spreaders are not directly linked. The results are averaged over 100 independent runs.

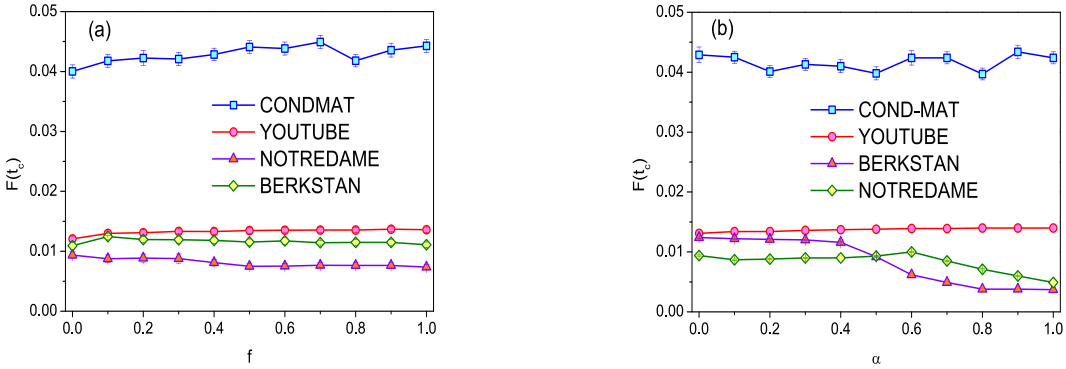

Figure 6. The final affected scale $F(t_c)$ with (**a**) different decreasing factor f and (**b**) different initial voting ability where $\lambda = 1.5$ and $p = 0.002$. The results are averaged over 100 independent runs.

one. Under this improved strategy, VoteRank and k-shellRank can be modified as their improved versions, i.e., VoteRank-Non and K-shellRank-Non, respectively. In order to evaluate the performance of VoteRank with this improved selecting strategy, we compare K-shellRank and VoteRank with K-shellRank-Non and VoteRank-Non.

Figure 5 shows the results of $F(t_c)$ against p ranging from 0.0005 to 0.003 on two undirected networks. Both K-shellRank-Non and VoteRank-Non are improved compared with K-shellRank and VoteRank. Particularly, K-shellRank gets significant improvement. Even though, VoteRank-Non outperforms K-shellRank-Non when the number of source spreaders is large, especial in YOUTUBE. The results of VoteRank-Non and original VoteRank are very close, as shown in Fig. 5. This indicates that the source spreaders selected by VoteRank are more disperse and diverse than K-shellRank. Interestingly, VoteRank even outperforms K-shellRank-Non when p is larger than 0.0015 in YOUTUBE, as shown in Fig. 5(a).

In VoteRank, there are two parameters, i.e., decreasing factor f and initial voting ability. The final affected scale $F(t_c)$ under different f is compared, as shown in Fig. 6(a). From this figure, it can be seen that the final affected scale $F(t_c)$ for $f > 0$ is larger than that for $f = 0$ except for NOTREDAME. The effect of initial voting ability on VoteRank is also analyzed. The initial voting ability of node i is set as k_i^{α} ($(k_i^{out})^{\alpha}$ for directed network) where α is a parameter whose value is from zero to one, correspondingly. For curtain α, the parameter f of node i is set as $\frac{k_i^{\alpha}}{<k>}$ ($\frac{(k_i^{out})^{\alpha}}{<k^{out}>}$ for directed network). The initial voting ability is 1 when $\alpha = 0$, and it equals node degree when $\alpha = 1$. The effect of initial voting ability is shown in Fig. 6(b). Generally, in undirected networks, initial voting ability has little effect on $F(t_c)$. However, in directed networks, the smaller initial voting ability is a relatively better choice.

Besides SIR model with limited contact, the performance of methods are also compared on other spreading models such as SI model and SIR model with full contact process, in which, a node will contact its all neighbors. SI model is usually used to evaluate the method on spreading rate especially in the early stage. In SI model, the infected scale $F(t)$ of early stage of different methods is compared, as shown in Fig. 7. From this figure, it can be seen that from the source spreaders obtained by VoteRank, the information will spread faster than that from other methods. The performance of VoteRank on SIR spreading model with full contact process with $\beta = 1$, $\lambda = 1.5\lambda_c$ is compared with other methods where λ_c is the threshold[46–48]. The results of final affected scale $F(t_c)$ for different methods are shown in Table 2. From this table, it can be seen that in most of cases, VoteRank is rather good.

To verify source spreaders selected by VoteRank are more scattered than that by other methods, the *average shortest path length* L_s obtained by VoteRank and other methods are compared. We just use two small networks, NOTRE DAME(directed) and COND-MAT(undirected) to analyze in this report for calculating the length of

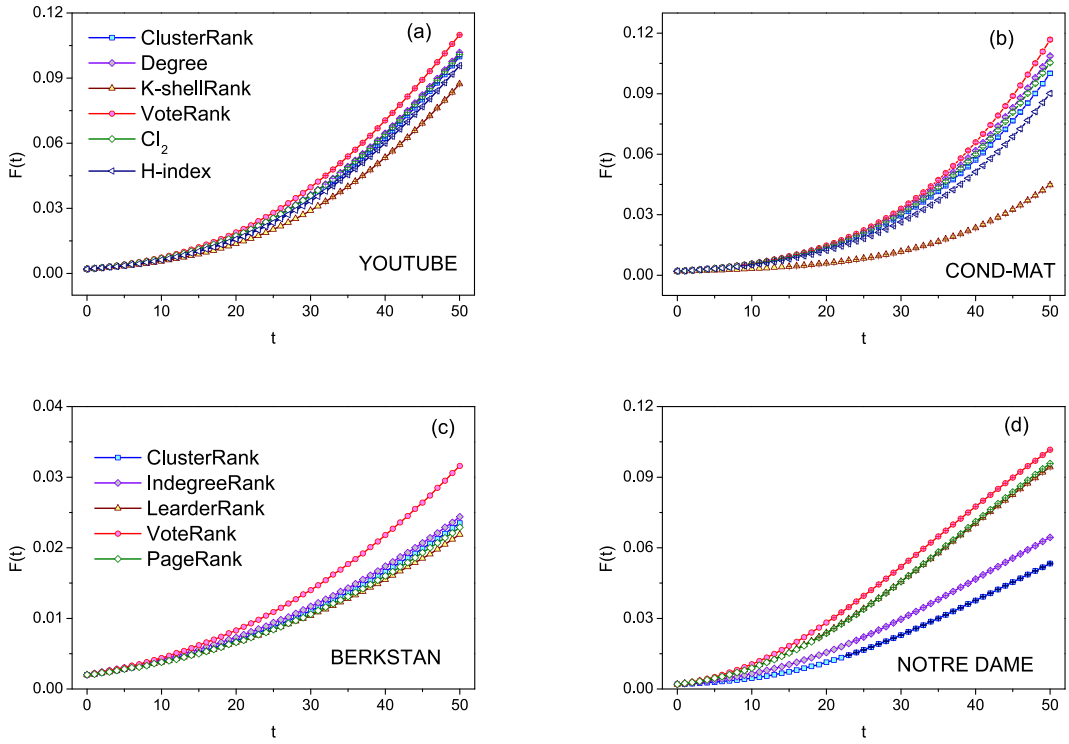

Figure 7. The infected scale $F(t)$ on SI model with $\lambda = 1.5$, $p = 0.002$, and $\beta = \frac{1}{< k >}$ in (**a**,**b**), $\beta = \frac{1}{< k^{out} >}$ in (**c**,**d**). The results are averaged over 100 independent runs.

	YOUTUBE	COND-MAT	BERKSTAN	NOTRE DAME
Degree	0.0065	0.1213	0.2117	0.0365
ClusterRank[24]	0.0055	0.1181	0.2018	0.0326
KshellRank[25]	0.0054	0.1198	/†	/
CI_2[36]	0.0051	0.1183	/	/
H-index[29]	0.0045	0.1196	/	/
PageRank[19]	/	/	0.2045	0.0415
LeaderRank[20]	/	/	0.2037	0.0423
VoteRank	0.0064	0.1239	0.2190	0.0385

Table 2. The final affected scale $F(t_c)$ for different methods on SIR spreading model with full contact process where $\lambda = 1.5\lambda_c$, $\beta = 1$ and $p = 0.003$. The results are averaged over 100 independent runs. †Just undirected versions of KshellRank, CI_2 and H-index and directed versions of PageRank and LeaderRank are considered in this report.

the shortest path in large scale network being time consuming. Figure 8 shows L_s of source spreaders selected by different methods under different scale of source spreaders. From Fig. 8, it can be seen that spreaders selected by VoteRank have larger L_s than that by other methods, especially when p is large. So, compared with other methods, the source spreaders selected by VoteRank are more decentralized in the whole network. Actually, as pointed out in ref. 49, the spreading will be more effective when L_s gets larger.

Computational complexity analysis. The total computational time includes three parts as follows: the time of initializing voting ability and voting score, the time of selecting a node with the highest voting score, and the time of updating the voting ability and voting score. For the first part, the time of initializing voting ability is $O(n)$ and that of initializing voting score is $O(<k> n) = O(m)$ where $<k>$ is the average degree of network and m is the number of edges, so, the computational complexity of this step is $O(n + m) = O(m)$. Particularly, the computational complexity is $O(1)$ if we set initial voting ability as 1. For the second part, the computational complexity is $O(n)$ so as to select a node with the highest voting score. And if we take high efficient data structure such as red-black tree, the computational complexity will decrease to $O(\log n)$. For the third part, just the information of nodes with a distance of 2 from the newly selected spreader needs updating. Hence, the computational complexity is $O(<k>^2) = O(m^2/n^2)$. To select r spreaders with r times in step 2 and 3, the total computational complexity is $O(m) + O(r \log n) + O(r <k>^2) = O(m + r \log n + rm^2/n^2)$. If networks is sparse and $r \ll n$, the computational

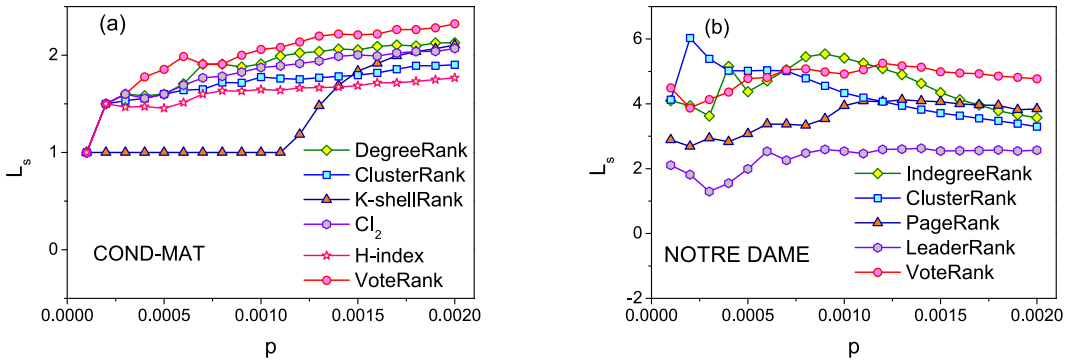

Figure 8. Average shortest path length L_s for different methods under different scale of source spreaders.

complexity of VoteRank can be reaching $O(n)$. Although above analysis is based on undirected network, it is similar for case of directed network.

Discussion

In summary, with utilizing information of $r-1$ ranked nodes to rank the r^{th} node, we get an obvious boost on information spreading in complex networks, especial in large scale networks. However, when r is small, little information can be utilized and the advantage of VoteRank is not significant. When r becomes large, the information accumulated by the $r-1$ previous nodes becomes abundant and can make a significant improvement. VoteRank provides a simple yet effective way to determine the next most influential node based on the selected nodes. It is worth mentioning that VoteRank outperforms K-shellRank on undirected network, and also outperforms other benchmark algorithms such as PageRank, ClusterRank and IndegreeRank on directed network. The results also indicate that performance of VoteRank is fairly stable with different infected rate λ and different scale of initial spreaders p in terms of information spreading. Another interesting question is how to judge the optimal number of r to get the best spreading ability. In fact, this problem has two equal forms. The first is maximizing the spreading ability while fixing the number of initial spreaders. The second is minimizing the number of initial spreaders for giving spreading ability, i.e., fixing the number of final affected nodes. We just take into account the first form in our work, the other form can be analyzed similarly. Besides, some researchers use the robustness R[50], which is defined as $R = \frac{1}{n}\sum_{i=1}^{n}\sigma(i/n)$ where $\sigma(i/n)$ is the fraction of nodes belonging to giant component after removing i/n of nodes from network, to evaluate the attacking ability of a method. The method has higher attacking ability if R is smaller. In recent years, many researchers aimed to the study of temporal networks, including structure and dynamics[51]. To identify a set of influential nodes in temporal networks is an important and interesting topic. How to extend our work to temporal networks is worth further studying.

References

1. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. Complex networks: Structure and dynamics. *Phys. Rep.* **424,** 175–308 (2006).
2. Gao, Z.-K. *et al.* Multi-frequency complex network from time series for uncovering oil-water flow structure. *Sci. Rep.* **5,** 8222 (2015).
3. Gao, Z.-K., Fang, P.-C., Ding, M.-S. & Jin, N.-D. Multivariate weighted complex network analysis for characterizing nonlinear dynamic behavior in two-phase flow. *Exp. Therm. Fluid Sci.* **60,** 157–164 (2015).
4. Gao, Z.-K. & Jin, N.-D. A directed weighted complex network for characterizing chaotic dynamics from time series. *Nonlinear Anal-Real* **13,** 947–952 (2012).
5. Pastor-Satorras, R. & Vespignani, A. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86,** 3200–3203 (2001).
6. Luo, J. & Qi, Y. Identification of essential proteins based on a new combination of local interaction density and protein complexes. *PLoS ONE* **10,** e0131418 (2015).
7. Lü, L., Chen, D.-B. & Zhou, T. The small world yields the most effective information spreading. *New J. Phys.* **13,** 123005 (2011).
8. Myers, A. A., Zhu, C. & Leskovec, J. Information diffusion and external influence in networks. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China. New York: ACM Press (doi: 10.1145/2339530.2339540), August 12–16, 2012, pp. 33–41 (2012).
9. Liu, C. & Zhang, Z.-K. Information spreading on dynamic social networks. *Commun. Nonlinear Sci. Numer. Simul.* **19,** 896–904 (2014).
10. Cinimi, G. *et al.* Enhancing topology adaptation in information-sharing social networks. *Phys. Rev. E* **85,** 046108 (2012).
11. Chen, D.-B., Xiao, R. & Zeng, A. Predicting the evolution of spreading on complex networks. *Sci. Rep.* **4,** 6108 (2014).
12. Kempe, D., Kleinberg, J. & Tardos, E. Maximizing the spread of influence through a social network. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA. New York: ACM Press (doi: 10.1145/956750.956769), August 2003, pp. 137–146 (2003).
13. Sabidussi, G. The centrality index of a graph. *Psychometrika* **31,** 581–603 (1996).
14. Freeman, L. C. Centrality in social networks conceptual clarification. *Social Netw.* **1,** 215–239 (1979).
15. Chen, D.-B., Xiao, R., Zeng, A. & Zhang, Y.-C. Path diversity improves the identification of influential spreaders. *EPL* **104,** 68006 (2013).
16. Ren, Z.-M., Zeng, A., Chen, D.-B., Liao, H. & Liu, J.-G. Iterative resource allocation for ranking spreaders in complex networks. *EPL* **106,** 48005 (2014).
17. Chen, D.-B., Lü, L., Shang, M.-S., Zhang, Y.-C. & Zhou, T. Identifying influential nodes in complex networks. *Physica A* **391,** 1777–1787 (2012).
18. AskariSichani, O. & Jalili, M. Influence maximization of informed agents in social networks. *Appl. Math. Comput.* **254,** 229–239 (2015).
19. Brin, S. & Page, L. The anatomy of a largescale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **30,** 107–117 (1998).

20. Lü, L., Zhang, Y.-C., Yeung, C. H. & Zhou, T. Leaders in social networks, the delicious case. *PLoS ONE* **6,** e21202 (2011).
21. Pei, S., Muchnik, L., Andrade, J. S., Zheng, Z. & Makse, H. A. Searching for superspreaders of information in real-world social media. *Sci. Rep.* **4,** 5547 (2014).
22. Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *J. ACM* **46,** 604–632 (1999).
23. Weng, J., Lim, E.-P., Jiang, J. & He, Q. Twitterrank: finding topic-sensitive influential twitterers. Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, New York, USA. New York: ACM Press (doi: 10.1145/1718487.1718520), February 4–6, 2010, pp. 261–270 (2010).
24. Chen, D.-B., Gao, H., Lü, L. & Zhou, T. Identifying influential nodes in large-scale directed networks: the role of clustering. *PLoS ONE* **8,** e77455 (2013).
25. Kitsak, M. *et al.* Identification of influential spreaders in complex networks. *Nat. Phys.* **6,** 888–893 (2010).
26. Wei, B., Liu, J., Wei, D. J., Gao, C. & Deng, Y. Weighted k-shell decomposition for complex networks based on potential edge weights. *Physica A* **420,** 277–283 (2015).
27. Liu, Y., Tang, M., Zhou, T. & Do, Y. Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition. *Sci. Rep.* **5,** 9602 (2015).
28. Liu, Y., Tang, M., Zhou, T. & Do, Y. Improving the accuracy of the k-shell method by removing redundant links: From a perspective of spreading dynamics. *Sci. Rep.* **5,** 13172 (2015).
29. Lü, L., Zhou, T., Zhang, Q.-M. & Stanley, H. E. The H-index of a network node and its relation to degree and coreness. *Nat. Comm.* **7,** 10168 (2016).
30. Narayanam, R. & Narahari, Y. A shapley value-based approach to discover influential nodes in social networks. *IEEE T. Autom. Sci. Eng.* **8,** 130–147 (2011).
31. Chen, W., Wang, Y. & Yang, S. Efficient influence maximization in social networks. Proceedings of the 15th ACM SIGKDD International Conference on Knowlwdge Discovery and Data Mining, Paris, France. New York: ACM Press (doi: 10.1145/1557019.1557047), June 28-July 1, 2009, pp. 199–208 (2009).
32. Tang, Y., Xiao, X. & Shi, Y. Influence maximization: near-optimal time complexity meets practical efficiency. Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, Snowbird, UT, USA. New York: ACM Press (doi: 10.1145/2588555.2593670), June 22–27, 2014, pp. 75–86 (2014).
33. Zhao, X.-Y., Huang, B., Tang, M., Zhang, H. F. & Chen, D.-B. Identifying effective multiple spreaders by coloring complex networks. *EPL* **108,** 68005 (2014).
34. Welsh, D. J. & Powell, M. B. An upper bound on the chromatic number of a graph and its application to timetabling problems. *Comput. J.* **10,** 85 (1967).
35. Li, S., Lü, L., Yeung, C.-h. & Hu, Y. Effective spreading from multiple leaders identified by percolation in social networks. arXiv, 1508.04294 (2015).
36. Morone, F. & Makse, H. A. Influence maximization in complex networks through optimal percolation. *Nature* **524,** 7563 (2015).
37. He, J.-L., Fu, Y. & Chen, D.-B. A novel top-k strategy for influence maximization in complex networks with community structure. *PLoS ONE* **10,** e0145283 (2015).
38. Zhou, T., Fu, Z. & Wang, B.-H. Epidemic dynamics on complex networks. *Progr. Nat. Sci*. **16,** 452–457 (2006).
39. Hethcote, H. W. The mathematics of infectious diseases. *SIAM Rev.* **42,** 599–653 (2000).
40. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286,** 509–512 (1999).
41. Yang, J. & Leskovec, J. Defining and evaluating network communities based on ground-truth. IEEE 12th International Conference on Data Mining, Brussels, Belgium. Piscataway, New Jersey: IEEE Press, 10–13 December 2012, pp. 745–754 (2012).
42. Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* **98,** 404–409 (2001).
43. Leskovec, J., Lang, K., Dasgupta, A. & Mahoney, M. Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics* **6,** 29–123 (2009).
44. Albert, R., Jeong, H. & Barabsi, A.-L. Internet: diameter of the World-Wide Web. *Nature* **401,** 130–131(1999).
45. Hu, H.-B. & Wang X.-F. Unified index to quantifying heterogeneity of complex networks. *Physica A* **387,** 3769–3780 (2008).
46. Castellano, C. & Pastor-Satorras, R. Thresholds for epidemic spreading in networks. *Phys. Rev. Lett.* **105,** 218701 (2010).
47. Chu S. & Fu, X.-C. Epidemic spreading in directed networks with degree correlation. *Journal of Biomathematics* **30,** 29–37 (2015).
48. Li, C., Wang, H. & Mieghem, P. V. Epidemic threshold in directed networks. *Phys. Rev. E* **88,** 062802 (2013).
49. Hu, Z.-L., Liu, J.-G., Yang, G.-Y. & Ren, Z.-M. Effects of the distance among multiple spreaders on the spreading. *EPL* **106,** 18002 (2014).
50. Schneider, C. M., Moreira, A. A., Andrade, J. S., Havlin, S. & Herrmann, H. J. Mitigation of malicious attacks on networks. *Proc. Natl. Acad. Sci. USA* **108,** 3838–3841(2011).
51. Holme, P. & Saramäki, J. Temporal networks. *Phys. Rep.* **519,** 97–125 (2012).

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China under Grant Nos 61433014 and 61300018, by the National High Technology Research and Development Program under Grant No. 2015AA7115089, by the Fundamental Research Funds for the Central Universities under Grant Nos ZYGX2014Z002, ZYGX2015J152 and ZYGX2015J156 and by the Shanghai Research Institute of Publishing and Meida under Grant No. SAYB1402.

Author Contributions

J.-X.Z. and D.-B.C. designed the research and prepared all figures. J.-X.Z. and Z.-D.Z. performed the experiments. D.-B.C. and Q.D. analyzed the data. All authors wrote and reviewed the manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Zhang, J.-X. *et al.* Identifying a set of influential spreaders in complex networks. *Sci. Rep.* **6**, 27823; doi: 10.1038/srep27823 (2016).