Physica A 391 (2012) 1777–1787

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Identifying influential nodes in complex networks

Duanbing Chen [a], Linyuan Lü [b,*], Ming-Sheng Shang [a], Yi-Cheng Zhang [a,b], Tao Zhou [a,c]

[a] Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 611731, People's Republic of China

[b] Physics Department, University of Fribourg, Chemin du Musée 3, CH-1700 Fribourg, Switzerland

[c] Department of Modern Physics, University of Science and Technology of China, Hefei 230026, People's Republic of China

ARTICLE INFO

Article history:
Received 14 April 2011
Received in revised form 29 July 2011
Available online 2 October 2011

Keywords:
Complex networks
Centrality measures
Influential nodes
Spreading
SIR model

ABSTRACT

Identifying influential nodes that lead to faster and wider spreading in complex networks is of theoretical and practical significance. The degree centrality method is very simple but of little relevance. Global metrics such as *betweenness centrality* and *closeness centrality* can better identify influential nodes, but are incapable to be applied in large-scale networks due to the computational complexity. In order to design an effective ranking method, we proposed a semi-local centrality measure as a tradeoff between the low-relevant degree centrality and other time-consuming measures. We use the *Susceptible–Infected–Recovered* (SIR) model to evaluate the performance by using the spreading rate and the number of infected nodes. Simulations on four real networks show that our method can well identify influential nodes.

1. Introduction

The understanding of network structures, functions, and their relations has attracted much attention recently [1–5]. It is well-known that many mechanisms such as cascading, spreading, and synchronizing are highly affected by a tiny fraction of influential nodes [6–11]. How to find these influential nodes is of theoretical significance. Besides, identifying influential nodes has remarkable practical value: this is helpful in controlling rumor and disease spreading, and creating new marketing tools.

Degree centrality is a straightforward and efficient metric, however, it is less relevant since a node having a few high influential neighbors may have much higher influence than a node having a larger number of less influential neighbors. Although some well-known global metrics such as *betweenness centrality* and *closeness centrality* can give better results, due to the very high computational complexity, they are not easy to manage very large-scale online social webs. Recently, Lü et al. [12] have proposed a random-walk-based algorithm *LeaderRank* to identify leaders in social networks, which outperforms the well-known PageRank [13] in identifying the most influential nodes for opinion spreading and protecting from the spammers' attacks. LeaderRank [12], as well as PageRank [13], has good performance for directed networks, but does not work well for undirected networks (it will degenerate to degree centrality in undirected networks). In a word, the design of an effective ranking method to identify influential nodes is still an open issue.

In this paper, we proposed a semi-local centrality measure as a tradeoff between the low-relevant degree centrality and other time-consuming measures. To evaluate the algorithmic performance, we use the *Susceptible–Infected–Recovered* (SIR) model [14] to examine the spreading influence of the nodes ranked by different centrality measures. The simulations on four real networks show that our method can well identify influential nodes. Comparing with degree centrality, closeness centrality, and betweenness centrality methods, our method performs almost as good as the closeness centrality method

* Corresponding author.
E-mail address: linyuan.lue@unifr.ch (L. Lü).

0378-4371/$ – see front matter © 2011 Published by Elsevier B.V.
doi:10.1016/j.physa.2011.09.017

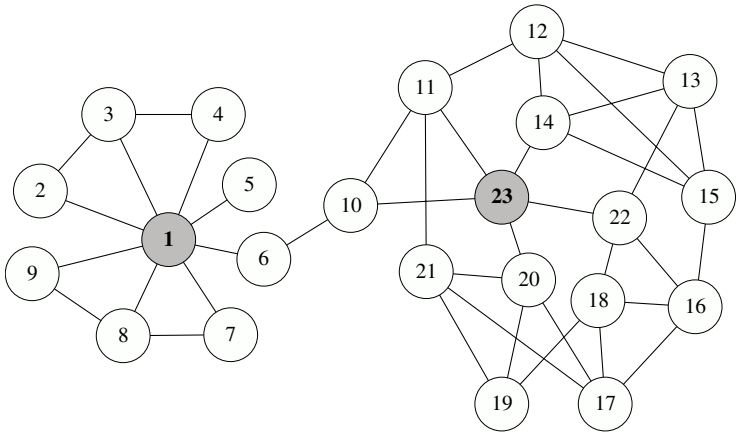


Fig. 1. An example network consisted of 23 nodes and 40 edges. Although node 23 has lower degree than node 1, its influence may be even higher.

while with much lower computational complexity, and much better than degree and betweenness centrality methods. Moreover, we investigate the relation between these centrality measures in terms of the influence of the top-ranked nodes.

Following parts are organized as follows. We introduce our new centrality measure and briefly review the definition of other centrality measures for comparison in Section 2. In Section 3, we use the SIR model to evaluate the performance in an example network. The data description is presented in Section 4.1, and the effectiveness of centrality measures and their correlations are respectively discussed in Sections 4.2 and 4.3. Conclusions are given in Section 5.

2. Centrality measures for node influences

Many centrality measures have been proposed to rank nodes in networks. A simple one is *degree centrality*, namely, a node with larger degree is likely to have higher influence (e.g., as an initially infected node, it is expected to spread more quickly and broadly) than a node with smaller degree. However, in some cases, this method fails to identify influential nodes since it considers only very limited information. For example, as shown in Fig. 1, although node 1 has the largest degree among all 23 nodes, the disease, if it origins at node 1, may not spread the fastest or the most broadly since all neighbors of node 1 have very low degree. In contrast, node 23 may be of higher influence although it has lower degree comparing with node 1.

Another group of methods considering the global information gives better ranking results, such as betweenness centrality and closeness centrality, two prominent geodesic-path-based ranking measures.

Betweenness is a centrality measure of a node in a network, usually defined as the fraction of shortest paths between node pairs that pass through the node of interest. Betweenness is, in some sense, a measure of the influence of a node over the information spread through the network or the expected load of a node in a transportation network [15,16]. For a network $G = (V, E)$ with $n = |V|$ nodes and $m = |E|$ edges, the betweenness centrality of node v, denoted by $C_B(v)$ is [17,18]:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{1}$$

where σ_{st} is the number of shortest paths between nodes s and t, and $\sigma_{st}(v)$ denotes the number of shortest paths between s and t which pass through node v.

Closeness of node v is defined as the reciprocal of the sum of geodesic distances to all other nodes of V [19]:

$$C_C(v) = \frac{1}{\sum_{t \in V \setminus v} d_G(v, t)}, \tag{2}$$

where $d_G(v, t)$ is the geodesic distance between v and t. Closeness can be considered as a measure of how long it will spread information from a given node to other reachable nodes in the network. Dangalchev [20] modified the definition to a general form, called *residual closeness*, which is more sensitive than the well-known measures of vulnerability [21], since it is able to reflect the effects of node removal even if this removal does not result in disconnected components. The *residual closeness* reads [20]:

$$C_R(v) = \sum_{t \in V \setminus v} 2^{-d_G(v,t)}. \tag{3}$$

Comparing with degree centrality, betweenness and closeness centrality measures can better quantify the influence of node, but they have higher computational complexity (Centralities based on PageRank or LeaderRank are even more relevant

Table 1
Simulations of effectiveness on the example network shown in Fig. 1. Initially, only one node is selected to be infected. $K(v)$ is the degree of node v. For each initial node, $F(t_c)$ is obtained by averaging over 100 implementations.

v	$K(v)$	N_v	Q_v	$C_L(v)$	$C_C(v)$	$C_B(v)$	$F(t_c)$
1	8	9	67	145	0.1368	242.00	7.34
2	2	8	17	92	0.0749	0.00	8.45
3	3	8	25	101	0.0772	1.00	7.93
4	2	8	17	92	0.0749	0.00	7.74
5	1	8	9	67	0.0727	0.00	8.24
6	2	11	18	104	0.1690	224.00	12.62
7	2	8	17	92	0.0749	0.00	8.65
8	3	8	25	101	0.0772	1.00	8.73
9	2	8	17	92	0.0749	0.00	7.95
10	3	9	37	111	0.1964	234.00	12.48
11	4	12	41	166	0.1795	89.73	12.71
12	4	9	38	157	0.1288	26.00	11.45
13	4	8	39	157	0.0953	5.67	12.01
14	4	9	40	166	0.1288	23.33	12.05
15	4	9	37	156	0.0982	10.0	12.87
16	4	11	39	158	0.1043	15.40	12.78
17	4	9	39	158	0.0982	10.13	12.96
18	4	9	40	148	0.0982	11.13	13.15
19	3	8	28	119	0.0925	3.07	12.99
20	4	10	40	158	0.1328	29.73	12.32
21	4	9	39	148	0.1288	31.33	12.96
22	4	12	42	170	0.1410	62.67	12.77
23	5	14	52	200	0.1964	163.8	13.63

but more time-consuming [12,13,22,23]). Calculating the shortest paths between all pairs of nodes in a network takes the complexity $O(n^3)$ with the Floyd's algorithm [24]. For a sparse graph, it will be more efficient with Johnson's algorithm [25], which takes $O(n^2 \log n + nm)$. For unweighted networks, calculating betweenness centrality takes $O(nm) = O(n^2\langle k\rangle)$ using Brandes' algorithm [26], where $\langle k\rangle$ is the average degree of the network. Since the online social networks usually contain millions of nodes or more, the calculation of betweenness and closeness centrality is very time-consuming or even not feasible.

Making the method more effective, we propose a local centrality measure as a tradeoff between low-relevant degree-centrality and other time-consuming measures. It considers both the nearest and the next nearest neighbors. The local centrality $C_L(v)$ of node v is defined as

$$Q(u) = \sum_{w\in\Gamma(u)} N(w), \tag{4}$$

$$C_L(v) = \sum_{u\in\Gamma(v)} Q(u), \tag{5}$$

where $\Gamma(u)$ is the set of the nearest neighbors of node u and $N(w)$ is the number of the nearest and the next nearest neighbors of node w. Take Fig. 1 as example, node 1 has eight nearest neighbors including nodes from 2 to 9 and one next nearest neighbor: the node 10, and thus $N(1) = 9$. The values of $N(w)$ for the nodes in Fig. 1 are presented in the third column of Table 1. According to Eq. (4), $Q(1) = N(2) + N(3) + N(4) + N(5) + N(6) + N(7) + N(8) + N(9) = 67$. Similarly, we can obtain the values of Q for the rest nodes which are shown in the fourth column of Table 1. Finally, according to Eq. (5), the local centrality of node 1 is equal to the sum of Q over all the nearest neighbors of node 1, namely $C_L(1) = Q(2) + Q(3) + Q(4) + Q(5) + Q(6) + Q(7) + Q(8) + Q(9) = 145$. The values of four centrality measures, namely the degree centrality, the local centrality, the closeness centrality, and the betweenness centrality of the nodes in Fig. 1 are respectively shown in the second, fifth, sixth, and seventh columns of Table 1.

Local centrality measure is likely to be more effective to identify influential nodes than degree centrality measure as it utilizes more information, while it has much lower computational complexity than the betweenness and closeness centralities. Since to calculate $N(w)$ requires traversing node $w's$ neighborhood within two steps, the computational complexity of local centrality is $O(n\langle k\rangle^2)$ which grows linearly with the size of a sparse network.

3. Evaluation with SIR model

To evaluate the performance of our ranking method, we use the SIR model to examine the spreading influence of top-ranked nodes [12,27,28]. In such a system, there are three compartments [14]: (i) Susceptible $S(t)$ represents the number of individuals susceptible to (not yet infected) the disease; (ii) Infected $I(t)$ denotes the number of individuals that have been infected and are able to spread the disease to susceptible individuals; (iii) Recovered $R(t)$ stands for individuals that have been recovered and will never be infected again. At each step, for each infected node, one randomly selected susceptible

Table 2
The basic topological features of the four real networks. n and m are the total numbers of nodes and links, respectively. $\langle k\rangle$ and $k_{\max}$ denote the average and the maximum degree. $\langle d\rangle$ is the average shortest distance. C and r are the clustering coefficient [35] and assortative coefficient [36], respectively. H is the degree heterogeneity, defined as $H = \frac{\langle k^2\rangle}{\langle k\rangle^2}$ [37].

Network	n	m	$\langle k\rangle$	$k_{\max}$	C	$\langle d\rangle$	r	H
Blogs	3982	6803	3.42	189	0.1409	6.227	−0.1330	4.038
Netscience	379	914	4.82	34	0.3706	6.061	−0.0817	1.663
Router	5022	6258	2.49	106	0.0058	6.393	−0.1384	5.503
Email	1133	5451	9.62	71	0.1101	3.716	0.0782	1.942

neighbor gets infected with probability λ (for simplicity, here we set $\lambda = 1$). Notice that this model is slightly different from the standard SIR model where all the neighbors of an infected node have the chance to be infected. The present mechanism is usually used to mimic the limited spreading capability of individuals [29,30]. Infected nodes recover with probability $1/\langle k\rangle$ at each step. Under this assumption, an infected node will in average contact $\langle k\rangle$ neighbors before he/she is recovered. The process stops when there is no infected node. To investigate the influence of a single node in the network, we set this node to be infected initially. The total number of infected and recovered nodes at time t, denoted by $F(t)$, can be considered as an indicator to evaluate the influence of the initially infected node at time t. Clearly, $F(t)$ increases with t, and finally gets stable, labeled by $F(t_c)$, where t_c corresponds to the time that there is no infected node in the network. Thus $F(t_c)$ evaluates the eventual influence of the initially infected node—higher $F(t_c)$ indicates a larger influence.

Take Fig. 1 for example, $F(t_c)$ for a single node is shown in Table 1. Node 23 that has the highest local centrality has the largest $F(t_c)$ (much larger than the spreading influence of node 1, which is the winner for degree centrality). We also list the results of betweenness and closeness centrality methods for comparison. With betweenness centrality, node 1 is ranked the first place, while with closeness centrality, node 23 is the top-1. Although node 10 has the same closeness centrality with node 23 and even higher betweenness centrality than that of node 23, its spreading influence is lower than node 23. These results demonstrate that the proposed local centrality can better capture the node's influence for spreading.

4. Experimental analysis

4.1. Data

Four real networks are used to evaluate the performance of centrality measures. (i) Blogs—the communication relationships between owners of blogs on the MSN (Windows Live) Spaces website [31]. (ii) Netscience—the network of co-authorships between scientists who are themselves publishing on the topic of networks. There are in total 1589 scientists in this collaboration network. We here consider the largest component with 379 scientists [32]. (iii) Router—the router-level topology of the Internet, collected by the *Rocketfuel Project* [33]. It has 5022 nodes and is well connected, while it is an extremely sparse network with an average degree only being 2.49. (iv) Email—the network of e-mail interchanges between members of the University Rovira i Virgili (Tarragona) [34]. The data of Blog and Email can be downloaded from http://www.cs.bris.ac.uk/~steve/networks/peacockpaper, and Netscience data can be downloaded from http://www.personal.umich.edu/~mejn/netdata. The basic topological properties of these four networks are shown in Table 2.

4.2. Effectiveness

We use the SIR model to compare the proposed local centrality method with degree, closeness, and betweenness centrality ones. In each implementation only one node is selected to be infected, and then the information (or disease) spreads in the network according to the SIR model described in Section 3. After n implementations (each node is selected to be the initially infected node once and only once), we investigate the relation between node's influence measured by $F(t)$ and its centrality value. Instead of considering the stable state of each node, we focus on the influence within a given time, since the spreading in early stage is more important in practice. Here we set $t = 10$ for further investigation. The results of four centralities on four networks are shown in Fig. 2. In Blogs, there is no clear correlation between $F(t)$ and the degree centrality, and the situation is even worse for betweenness centrality. For example, the $F(t)$ of some high-degree nodes are lower than that of some very-low-degree nodes. Comparatively speaking, local and closeness centralities perform better, as weakly positively correlated with $F(t)$. Especially, the node with higher local centrality is very likely to infect more nodes. In Router, none of these four measures can well capture the spreading influence. However, the local and closeness centralities are better than degree and betweenness centralities. In Netscience, betweenness centrality performs worst. There is strongly positive correlation between $F(t)$ and local centrality as well as closeness centrality. In Email, all these four measures perform good, and the local and closeness centralities are still better. In a word, by testing the correlation between spreading influence $F(t)$ and centralities, we show that the local and closeness centrality measures perform competitively good, and are much better than degree and betweenness centrality ones.

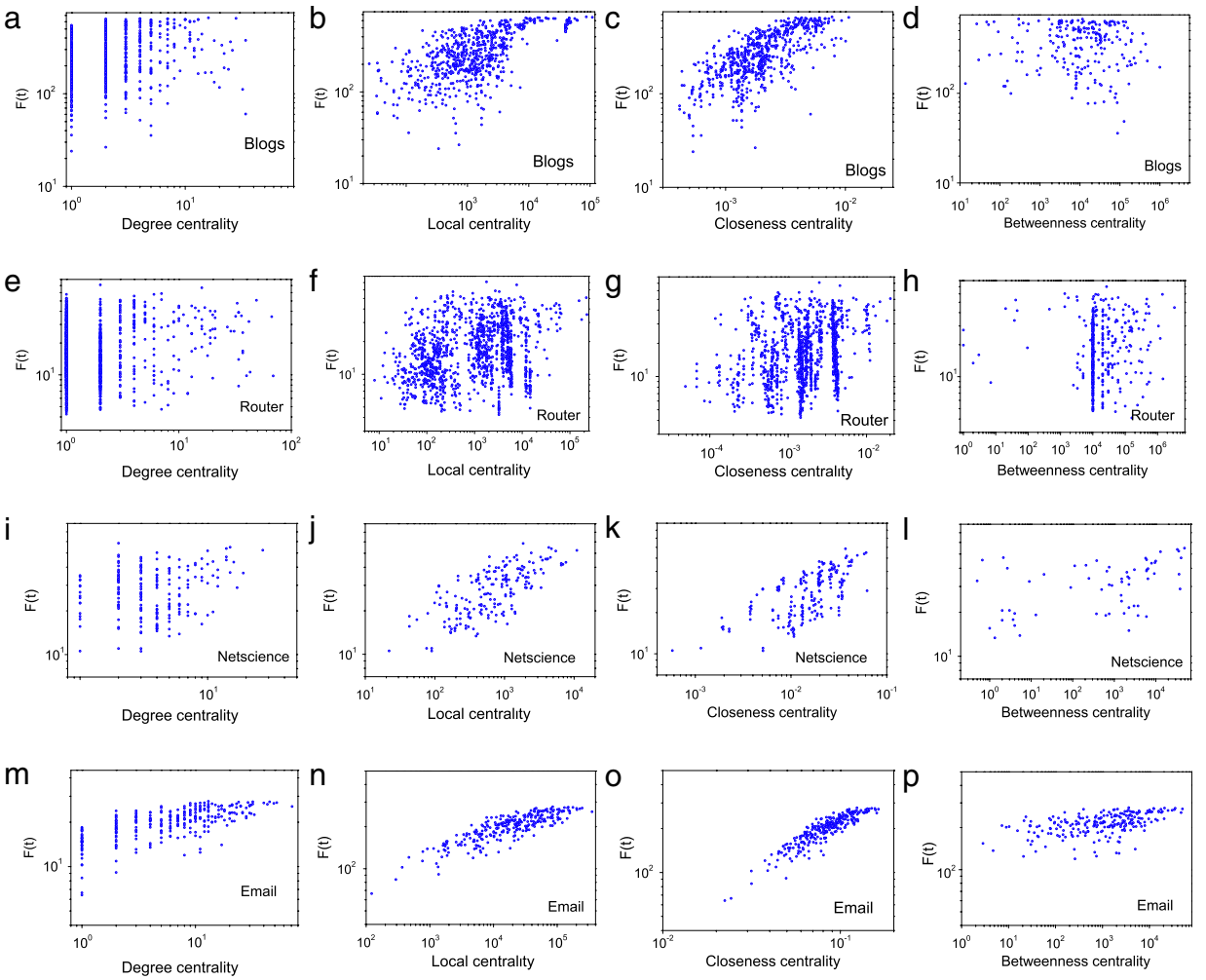


Fig. 2. The relation between node's influence measured by $F(t)$ ($t = 10$) and its centrality. Four rows respectively correspond to the results on four example networks, and four columns respectively correspond to four centrality measures. For each initial node, $F(t)$ is obtained by averaging over 100 independent runs.

In order to clarify the performance of each ranking method, Fig. 3 shows the average number of infected nodes (i.e., $\langle F(t)\rangle(t = 10)$) by the top-$L$ nodes as ranked by four centrality measures. The random ranking method is also presented for comparison. A good centrality measure should be downward sloping, namely the average number of infected nodes by the top-L nodes decreases with the increasing of L. Clearly, random ranking is the worst method. In Blogs and Netscience networks, the local centrality performs the best among all five methods, while in Router and Email networks, the closeness centrality performs the best, but the local centrality measure can still give comparatively good performance.

Furthermore, we compare the influence of the nodes that either appear in the top-10 list by local centrality or degree centrality (not appearing in both lists). Note that, without considering the effects of common nodes in both ranking lists, the differences of these two methods can be well distinguished. The simulations on the cumulative number of infected nodes, namely $F(t)$, as a function of time for four networks are shown in Fig. 4. The number of cumulative infected nodes increases with time and ultimately reach the steady value. For all these four networks, local centrality outperforms degree centrality for both spreading rate and the number of infected nodes $F(t_c)$. Fig. 5 gives two typical examples to explain why local centrality outperforms degree centrality in Router. Node α represented by a large solid circle in Fig. 5(a) has the largest local centrality value, while node β represented by a large solid circle in Fig. 5(b) has the largest degree. The two plots respectively show the local structure surrounding nodes α and β, which take into consideration only the nearest and the next nearest neighbors of α and β. Clearly, there are many connections among the neighbors of node α, but only a few connections among the neighbors of node β. Although node β directly connects many nodes, it cannot spread so far for most of its directed neighbors are less influential nodes (i.e., more than half of β's neighbors have degree 1, and in the remaining neighbors, more than half of them have degree 2). This is the reason why the spreading rate of node α is faster than that of node β, and the total number of infected nodes of node α is also larger than that of node β. Similarly, we also compare

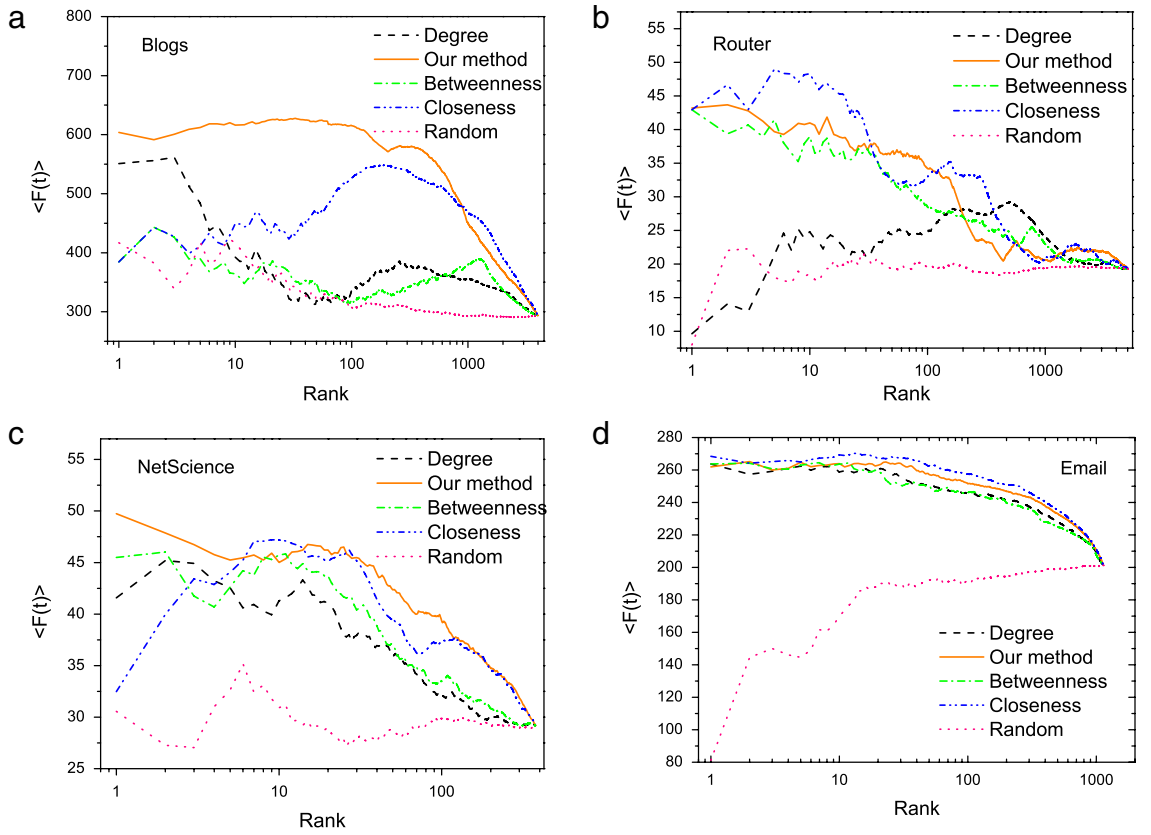


Fig. 3. The average number of $F(t)$ $(t = 10)$ of the top-L users as ranked by the five centrality measures, including degree centrality (dash line), betweenness centrality (dash–dot line), closeness centrality (dash–dot–dot line), random ranking method (dot line) and our method (solid line).

the local centrality with other two global centrality measures, closeness and betweenness via infecting the different top-10 nodes (the figures are omitted here). In Blogs, the local centrality can achieve almost the same number of eventually infected nodes $F(t_c)$ compared with betweenness centrality, while with a faster spreading rate (i.e., shorter convergence time). The situation is the same when comparing with closeness centrality. In Router and Netscience, the closeness and betweenness centralities outperform local centrality. In Email, the result for local and closeness (or betweenness) centralities is very similar to Fig. 4(d). That is to say all these measures have almost the same performance on Email, because these three centralities are all positively correlated with local centrality in this network (see next subsection for details). Furthermore, from the error bar of Fig. 4, we can also see that the results are not sensitive to the dynamic process on networks. Table 3 shows the top-10 nodes by local centrality and their ranks by other three centralities, as well as the number of total infected nodes $F(t_c)$. While in Table 4, we present the mean value of $F(t)$ $(t = 10)$ over the top-10 nodes on four centralities. One can observe that the local centrality and closeness centrality perform competitively good, and they are slightly better than the betweenness centrality. Of course, degree centrality is the worst.

It is noted that the performance of local centrality depends on the network structure. In our opinion, the local centrality is more suitable to be applied to heterogeneous networks where the ranking problem is worth following up. However, even in homogeneous tree-structure networks, our method is different from the degree centrality measure although they may generate very close ranking results. Take a four-layer full binary tree as an example. Although nodes in the second and the third layer have exactly the same degree, they own different local centrality values, 44 and 26 corresponding to the nodes of the second and the third layer, respectively. So, even in this tree-structure network, our method can successfully show that the nodes in the second layer are more influential than the ones in the third layer.

4.3. Relation between centrality measures

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a set of joint observations from two random variables X and Y respectively, such that all the values of (x_i) and (y_i) are unique. Any pair of observations (x_i, y_i) and (x_j, y_j) are said to be concordant if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. They are said to be discordant, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant. Then Kendall's Tau is

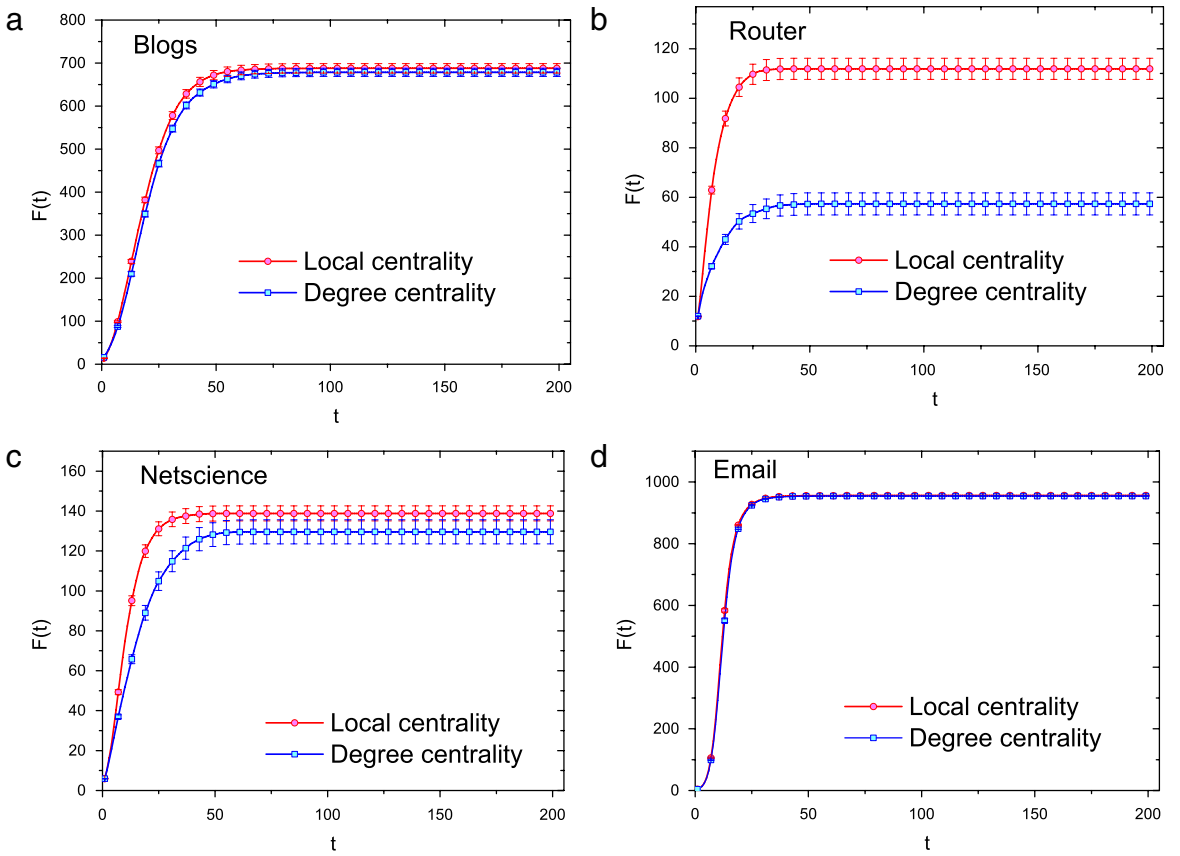
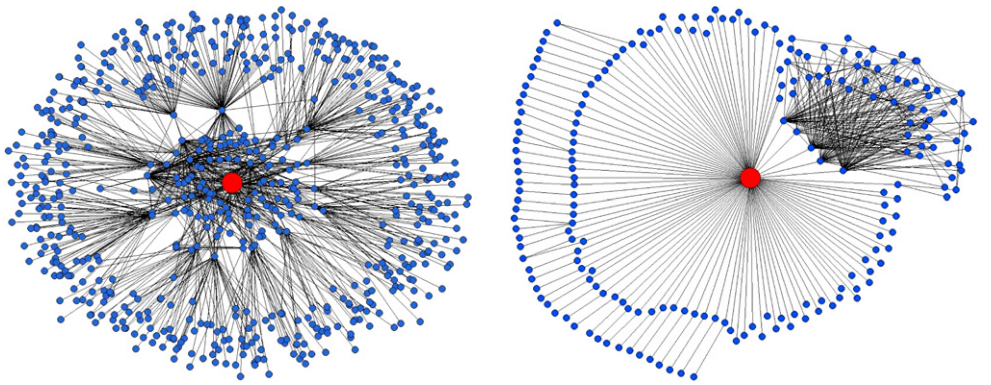


Fig. 4. The cumulative number of infected nodes as a function of time, with the initially infected nodes being those that either appear in the top-10 list by local centrality (circles) or degree centrality (squares), but not appearing in both list. Results are obtained by averaging over 100 implementations.



(a) Node α with the largest local centrality. (b) Node β with the largest degree centrality.

Fig. 5. Local structure surrounding the two representative nodes in Router. Node α represented by a large solid circle in (a) has the largest local centrality. Node β represented by a large solid circle in (b) has the largest degree centrality.

defined as [38]:

$$\tau = \frac{N_c - N_d}{\frac{1}{2}n(n-1)}, \tag{6}$$

where N_c and N_d are the number of concordant and discordant pairs, respectively.

Fig. 6 shows the relations between the local centrality and other three centralities on four networks, where Kendall's Tau is used to measure the correlation between local centrality and other three measures. Generally, local centrality has the strongest correlation with closeness centrality, and the weakest correlation with betweenness centrality. In Fig. 6, each

Table 3
The top-10 ranked nodes by local centrality (L) and their corresponding ranks by degree (D), closeness (C) and betweenness (B) centralities. $F(t_c)$ is obtained by averaging over 100 implementations.

Blogs					Router				
L	D	C	B	$F(t_c)$	L	D	C	B	$F(t_c)$
1	1	3	3	525.18	1	5	16	31	7.34
2	15	11	45	671.01	2	9	2	2	8.45
3	3	7	16	520.92	3	7	7	4	7.93
4	37	15	66	652.71	4	6	19	36	7.74
5	16	16	50	679.46	5	11	1	1	8.24
6	87	57	741	619.74	6	15	4	5	12.62
7	96	63	832	660.57	7	20	9	24	8.65
8	88	22	140	610.62	8	16	20	52	8.73
9	109	56	184	653.00	9	37	8	20	7.95
10	135	78	707	624.29	10	31	21	47	12.48
NetScience					Email				
L	D	C	B	$F(t_c)$	L	D	C	B	$F(t_c)$
1	1	19	12	47.08	1	1	3	2	255.70
2	2	7	6	52.38	2	3	4	10	268.18
3	4	77	50	44.16	3	2	1	1	271.70
4	8	81	21	43.30	4	4	40	22	244.38
5	29	85	101	41.92	5	5	2	3	265.24
6	44	86	135	47.56	6	19	11	61	283.24
7	45	87	136	41.40	7	7	19	15	245.54
8	46	88	137	46.78	8	6	5	8	268.64
9	47	89	138	42.58	9	9	21	16	264.62
10	30	22	15	51.84	10	12	33	36	270.24

Table 4
Mean value of $F(t)$ over top-10 nodes on four centralities.

Network	L	D	C	B
Blogs	621.75	373.08	419.28	361.75
Netscience	45.90	41.75	47.21	45.30
Router	40.76	23.81	48.18	37.95
Email	264.75	261.91	267.88	262.88

point indicates a node in the network, and its color represents the influence of this node in 10 steps, namely $F(10)$. In Email, the degree, closeness and betweenness centralities are all positively correlated with local centrality. Especially the closeness centrality is strongly positively correlated with local centrality. That is to say, the nodes with large closeness centralities are expected to have large local centralities. In addition, we can see that the nodes with higher closeness centralities and local centralities have higher influence (as indicated by the color). This may be the reason why the spreading rate and the number of infected nodes are nearly the same with four centralities in Email. Overall speaking, the correlation between closeness and local centrality is more strong than other two cases—local centrality vs. degree and local centrality vs. betweenness.

In Blogs, some small-degree nodes have much higher influence than large-degree nodes. And the highly influential nodes are likely to have high local centrality. We choose two typical examples. The local structure including the nearest and the next nearest neighbors of these two nodes labeled by α and β are respectively shown in Fig. 7(a) and (b). Node α has small degree centrality (i.e., 4) but large local centrality (i.e., 14,008), while node β has large degree centrality (i.e., 50) but small local centrality (i.e., 9388). Although node α has only four neighbors, these neighbors have many connections with other nodes. So, if node α is infected, it can affect more nodes through its neighbors. In contrast, since node β connects many 1-degree nodes, the disease or information cannot spread further. The spreading results of these two nodes are shown in Fig. 7(c). It can be seen that node α spreads faster than node β and reaches a much higher value of $F(t_c)$.

Another interesting phenomena in Blogs is that although two nodes have almost the same local centralities, the low-degree node has higher influence than the high-degree node. Two typical examples are shown in Fig. 8. The local structures shown in Fig. 8(a) and (b) include the neighbors within two steps of these two nodes labeled by α and β. The degree centrality of node α is 7 and its local centrality is 6434. In comparison, node β has much larger degree equal to 54 and its local centrality is 6490. The spreading results of these two nodes are shown in Fig. 8(c). It can be seen that node α spreads faster than node β and can infect larger number of nodes (i.e., larger $F(t_c)$). The reason is that although the degree of node α is small, it connects a node with large local centrality and its other neighbors also have many connections with other nodes, while node β connects many nodes whose local centrality are very low. In a word, compared with the local centrality, degree centrality is a much worse predictor for a node's spreading influence. And if a node itself is of high local centrality, or it is of small degree yet neighboring to a high-local-centrality node (in such a case, it has higher probability to infect this high-local-centrality node), it is very likely to have high spreading influence.

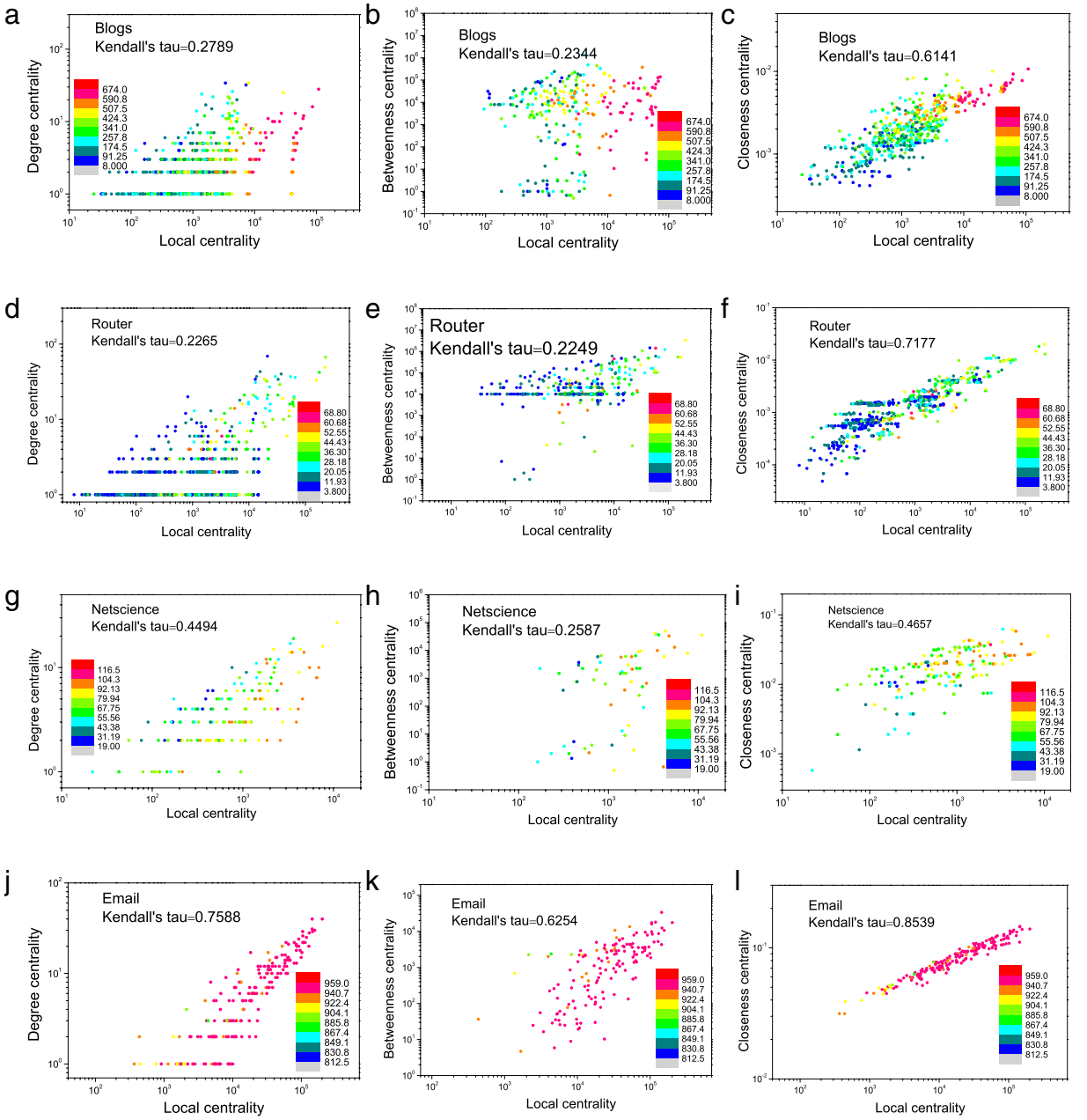


Fig. 6. The relations between local centrality and degree, betweenness and closeness centralities on four example networks. Each data point denotes a node, and its color represent the $F(t)$ value ($t = 10$) of this node. The values are obtained by averaged over 100 independent runs.

5. Conclusions

We focused on identifying influential nodes in undirected networks and proposed a local centrality measure as a tradeoff between the low-relevant degree centrality and other time-consuming measures. To evaluate the performance, we used the SIR model to estimate the spreading influence of the top-ranked nodes by different centrality measures. It is expected that with the influential nodes being initially infected the spreading rate and the number of infected nodes are higher than with the random nodes being initially infected. The experimental results on four real networks (the network of MSN blogs, the co-authorship network, the Internet at router level, and the email communication network) show that our method can well identify influential nodes. Comparing with other three well-known centrality measures, the newly proposed measure performs much better than degree and betweenness centrality ones, and almost as good as the closeness centrality measure while with much lower computational complexity.

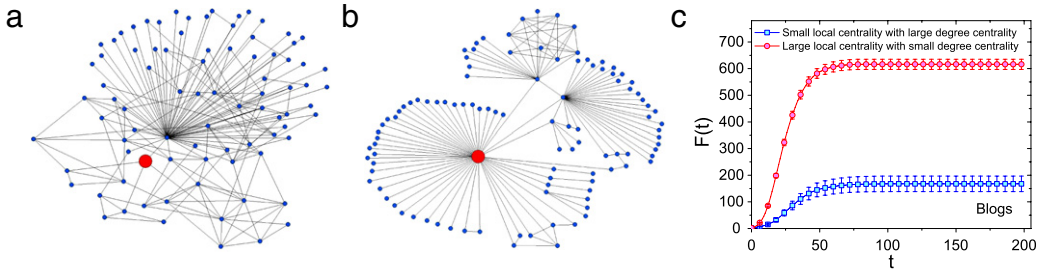


Fig. 7. (a) The local structure surrounding node α in Blogs. Node α has small degree (4) but large local centrality (14,008). (b) The local structure surrounding node β in Blogs. Node β has large degree (50) but small local centrality (9388). (c) The cumulative number of infected nodes as a function of time, with the initially infected nodes being nodes α (circles) and β (squares) respectively. Each point in (c) is obtained by averaging over 100 independent implementations.

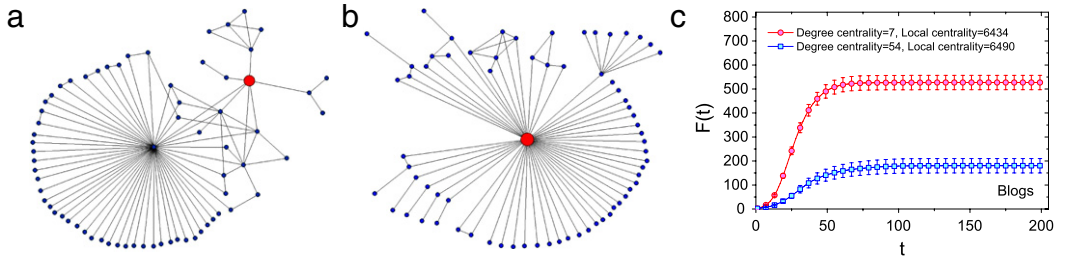


Fig. 8. (a) The local structure surrounding node α in Blogs. (b) The local structure surrounding node β in Blogs. Node α and node β have near the same local centrality while node α has only 7 neighbors and node β has 54 neighbors. (c) The cumulative number of infected nodes as a function of time, with the initially infected nodes being node α (circles) and β (squares) respectively. Each point in (c) is obtained by averaging over 100 independent implementations.

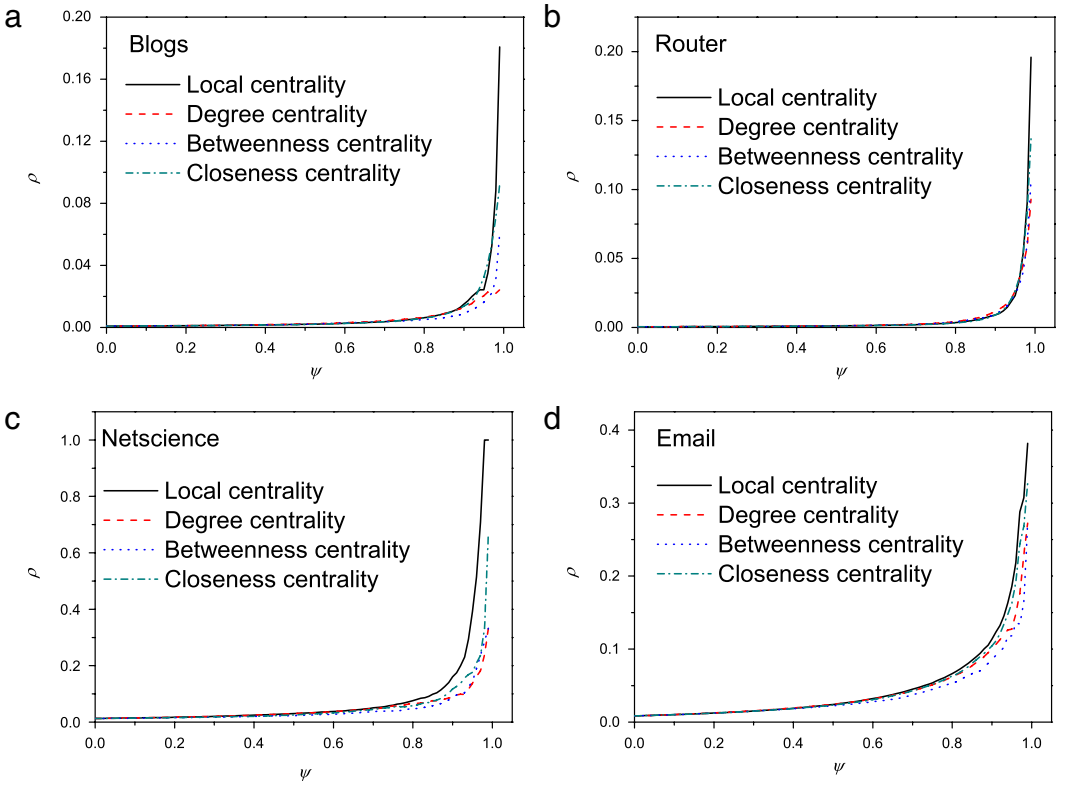


Fig. 9. Ranking-based rich-club phenomenon of the four methods. $\rho(\psi)$ denotes the density of connections among the $(1-\psi)$ top-ranked nodes. All the methods display rich-club phenomenon, where our method is the most remarkable one.

In addition to identify the influential nodes in spreading dynamics, the ranking algorithms can also be applied in revealing structural features, as well as the hidden relationships between structure and function of networks. In Fig. 9, we report the ranking-based rich-club phenomenon [39], where $\rho(\psi)$ is the density of connections among the $(1-\psi)$ top-ranked nodes. For example, $\rho(0.8)$ means the density of connections among the top-20% nodes. The monotonic rise of $\rho(\psi)$ indicates that all the four methods display rich-club phenomenon, where our method is the most remarkable one. That is to say, the top-ranked nodes by our method are more closely connected than by others. Since the local clustering are recently known to be beneficial to online information spreading [40,41], our method may also perform the best in digging out the most influential spreaders in online society.

Inspired by the local centrality measure, an interactive formulation can be obtained as $\vec{C}_{t+1} = A \cdot \vec{C}_t$, where A is the adjacency matrix of the network and t is the interactive step. Therefore, the local centrality measure is equivalent to $\vec{C}_2$ with initially setting $\vec{C}_0(u) = N(u)$. Of course, one can tune the step or change the initial condition to obtain a better ranking results, similar to the method applied in Ref. [42]. Although the local centrality is proposed aiming at identifying the influencers in undirected network, it can be applied to directed network as well with a modified definition of $N(w)$. Of course, for directed network, $N(w)$ should be the number of the nearest and next nearest upstream nodes of node w.

The sizes of many typical social and technical networks are increasing on and on, and thus the design of efficient and adaptive ranking methods will be a long-term challenge. We believe that this paper may shed some light on this direction.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 90924011 and 60903073, and the International Scientific Cooperation and Communication Project of Sichuan Province in China under Grant No. 2010HH0002. L.L. and Y.C.Z. acknowledge the Swiss National Science Foundation under Grant No. 200020-132253.

References

[1] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (2002) 47.
[2] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2003) 167.
[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Phys. Rep. 424 (2006) 175.
[4] J. Zhang, C. Zhou, X. Xu, M. Small, Mapping from structure to dynamics: a unified view of dynamical processes on networks, Phys. Rev. E 82 (2010) 026116.
[5] A. Zeng, L. Lü, Coarse graining for synchronization in directed networks, Phys. Rev. E 83 (2011) 056123.
[6] A.E. Motter, Y.-C. Lai, Cascade-based attacks on complex networks, Phys. Rev. E 66 (2002) 065102.
[7] T. Zhou, B.-H. Wang, Catastrophes in scale-free networks, Chin. Phys. Lett. 22 (2005) 1072.
[8] R. Pastor-Satorras, A. Vespignani, Immunization of complex networks, Phys. Rev. E 65 (2002) 036104.
[9] M. Zhao, T. Zhou, B.-H. Wang, W.-X. Wang, Enhanced synchronizability by structural perturbations, Phys. Rev. E 72 (2005) 057102.
[10] L. Zemanová, C. Zhou, J. Kurths, Structural and functional clusters of complex brain networks, Physica D 224 (2006) 202.
[11] G. Zamora-López, C. Zhou, J. Kurths, Cortical hubs form a module for multisensory integration on top of the hierarchy of cortical networks, Front Neuroinform. 4 (2010) 1.
[12] L. Lü, Y.-C. Zhang, C.H. Yeung, T. Zhou, Leaders in social networks, the delicious case, PLoS ONE 6 (2011) e21202.
[13] S. Brin, L. Page, The anatomy of a largescale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1998) 107.
[14] R.M. Anderson, R.M. May, B. Anderson, Infectious Diseases of Humans: Dynamics and Control, Oxford University Press, USA, 1992.
[15] R. Guimerà, A. Días-Guilera, F. Vega-Redondo, A. Cabrales, A. Arenas, Optimal network topologies for local search with congestion, Phys. Rev. Lett. 89 (2002) 248701.
[16] G. Yan, T. Zhou, B. Hu, Z.-Q. Fu, B.-H. Wang, Efficient routing on complex networks, Phys. Rev. E 73 (2006) 046108.
[17] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1977) 35.
[18] L.C. Freeman, Centrality in social networks conceptual clarification, Social Netw. 1 (1979) 215.
[19] G. Sabidussi, The centrality index of a graph, Psychometrika 31 (1966) 581.
[20] C. Dangalchev, Residual closeness in networks, Physica A 365 (2006) 556.
[21] P. Holme, B.J. Kim, C.N. Yoon, S.K. Han, Attack vulnerability of complex networks, Phys. Rev. E 65 (2002) 056109.
[22] F. Radicchi, S. Fortunato, B. Markines, A. Vespignani, Diffusion of scientific credits and the ranking of scientists, Phys. Rev. E 80 (2009) 056103.
[23] S.H. Lee, P.J. Kim, Y.Y. Ahn, H. Jeong, Googling social interactions: web search engine based social network construction, PLoS ONE 5 (2010) e11233.
[24] R.W. Floyd, Algorithm 97: shortest path, Commun. ACM 5 (1962) 345.
[25] D.B. Johnson, Efficient algorithms for shortest paths in sparse networks, J. ACM 24 (1977) 1.
[26] U. Brandes, A faster algorithm for betweenness centrality, J. Math. Sociol. 25 (2001) 163.
[27] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, Nat. Phys. 6 (2010) 888.
[28] A. Garas, P. Argyrakis, C. Rozenblat, M. Tomassini, S. Havlin, Worldwide spreading of economic crisis, New J. Phys. 12 (2010) 113043.
[29] T. Zhou, J.-G. Liu, W.-J. Bai, G. Chen, B.-H. Wang, Behaviors of susceptible-infected epidemics on scale-free networks with identical infectivity, Phys. Rev. E 74 (2006) 056109.
[30] R. Yang, B.H. Wang, J. Ren, W.J. Bai, Z.W. Shi, W.X. Wang, T. Zhou, Epidemic spreading on heterogeneous networks with identical infectivity, Phys. Lett. A 364 (2007) 189.
[31] N. Xie, Social network analysis of blogs, M.Sc. Dissertation, University of Bristol, 2006.
[32] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[33] N. Spring, R. Mahajan, D. Wetherall, T. Anderson, Measuring ISP topologies with rocketfuel, IEEE/ACM Trans. Netw. 12 (2004) 2.
[34] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, A. Arenas, Self-similar community structure in a network of human interactions, Phys. Rev. E 68 (2003) 065103.
[35] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (1998) 440.
[36] M.E.J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (2002) 208701.
[37] H.B. Hu, X.F. Wang, Unified index to quantifying heterogeneity of complex networks, Physica A 387 (2008) 3769.
[38] M. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81.
[39] S. Zhou, R.J. Mondragon, The rich-club phenomenon in the internet topology, IEEE Commun. Lett. 8 (2004) 180.
[40] D. Centola, The spread of behavior in an online social network experiment, Science 329 (2010) 1194.
[41] L. Lü, D.-B. Chen, T. Zhou, Small world yields the most effective information spreading, 2011. arXiv:1107.0429.
[42] W.-P. Liu, L. Lü, Link prediction based on local random walk, EPL 89 (2010) 58007.