Physica A 492 (2018) 2237–2248

Contents lists available at ScienceDirect

Physica A

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

Identifying multiple influential spreaders based on generalized closeness centrality

Huan-Li Liu [a], Chuang Ma [a,*], Bing-Bing Xiang [a], Ming Tang [b,c], Hai-Feng Zhang [a]

[a] School of Mathematical Science, Anhui University, Hefei 230601, China
[b] School of Information Science Technology, East China Normal University, Shanghai 200241, China
[c] Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 610054, China

HIGHLIGHTS

- An algorithm is proposed to identify multiple influential spreaders in complex networks.
- A generalized closeness index (GCC) is given to maximize the distance among spreaders.
- Finding the multiple nodes with the highest GCC can be approximately solved by K-means method.
- The performance is validated by different spreading processes on different networks.

ARTICLE INFO

Article history:
Received 13 April 2017
Received in revised form 21 September 2017
Available online 21 November 2017

Keywords:
Complex networks
Multiple influential spreaders
Generalized closeness centrality
K-means method

ABSTRACT

To maximize the spreading influence of multiple spreaders in complex networks, one important fact cannot be ignored: the multiple spreaders should be dispersively distributed in networks, which can effectively reduce the redundance of information spreading. For this purpose, we define a generalized closeness centrality (GCC) index by generalizing the closeness centrality index to a set of nodes. The problem converts to how to identify multiple spreaders such that an objective function has the minimal value. By comparing with the K-means clustering algorithm, we find that the optimization problem is very similar to the problem of minimizing the objective function in the K-means method. Therefore, how to find multiple nodes with the highest GCC value can be approximately solved by the K-means method. Two typical transmission dynamics—epidemic spreading process and rumor spreading process are implemented in real networks to verify the good performance of our proposed method.

How to identify the influential spreaders in social networks is of theoretical and practical significance [1–8], since it is crucial for developing efficient strategies to hinder epidemic spreading, accelerate information diffusion, promote new products, and so on. So far many effective methods have been proposed to identify the influential spreaders in social networks, such as degree centrality (DC) index [9], betweenness centrality (BC) index [10], closeness centrality (CC) index [11] and k-shell (KS) decomposition method [12]. However, most of these methods may be particularly suitable for the case in which single spreader is considered.

Many times, the spreading of rumors, ideas, opinions, or the promotion of advertisements may be initiated from different nodes simultaneously [13]. Unlike the single spreader case, the distance between multiple spreaders is a crucial factor in determining whether the spreading process can be widely diffused [14–16]. Recently, how to identify multiple influential spreaders in social networks has attracted some attention. For instance, Flaviano et al. provided a framework to find a set of

* Corresponding author.
E-mail address: chuang_m@126.com (C. Ma).

https://doi.org/10.1016/j.physa.2017.11.138
0378-4371/

optimal influencers in networks by mapping the problem onto optimal percolation [17]. In the method, the node with the highest value of *CI* is removed, the value of *CI* for each node in the residual network is recalculated. Chen et al. proposed a degree discount heuristics to maximize the spreading influence, where the degree of a considered node is discounted if it has edges linking to the prior selected spreaders. Also, the discount degree of each node in residual network should be recalculated and re-ranked after one spreader is selected. Hu et al. designed a method to identify multiple influential spreaders by finding the community hubs [16]. Evidently, when the number of spreaders is far more than the number of communities, which may fail in ensuring the large distance among the multiple spreaders. We also proposed a method by introducing graph coloring method into the influential spreader identification, and found that the method can enhance the spreading influence [18]. Sometimes, the method cannot guarantee the distance among the multiple spreaders is far enough, for this purpose, the method is improved in Ref. [19]. However, current advances in how to identify multiple influential spreaders are still in the initial phase.

For *a given node*, one can say that the node is more "important" if the node is in the *center* of networks, i.e., the average distance from the node to all other nodes is very short, leading to large value of CC index [11]. Many experimental results have indicated that CC index is a good index to measure the influence of nodes (Of course, no one centrality index outperforms the other centrality indices in all networks). Now for a set of nodes, we can intuitively assume that the set of nodes is more "important" if the distance from the set to all other nodes is shorter. So we can define a generalized closeness centrality (GCC) index regarding to *a set of nodes*, which assumes that the set of nodes is more important if the distance from all other nodes to the set is shorter. Therefore, the problem converts to how to find a set of nodes such that the distance from the set to other nodes is minimal (i.e., an optimization problem to minimize the distance). Theoretically, it is hard to obtain the exact solution. Luckily, we find that our defined objective function is similar to the objective function in the K-means method. So the optimization problem in our method can be approximated by the K-means method. In the K-means method, the center in each cluster can approximately ensure the value of objective function is minimal.

1. Generalized closeness centrality index

We consider an undirected and un-weighed network $G(V, E)$, where V is the set of nodes and E is the set of links. Define the "distance" between two nodes v_i and v_j as $\delta(v_i, v_j)$, then the distance of node v_i to all nodes is calculated as:

$$\delta_i = \sum_{j=1}^{N} \delta(v_i, v_j). \tag{1}$$

Here we do not use $d(v_i, v_j)$ to denote distance between two nodes v_i and v_j because we want to address that the "distance" defined here has a wider meaning, which can be the shortest path length between two nodes or the dissimilarities between two nodes. For instance, if we use the spectral of the Laplacian matrix to characterize the attributes of nodes, then the distance can be defined as Eq. (11), which is the dissimilarity rather than the shortest path length.

Consequently, smaller value of δ_i means that node v_i is much "closer" (or more similar) to other nodes. In particular, if $\delta(v_i, v_j)$ is defined as the shortest path length between v_i and v_j, denoted by $d(v_i, v_j)$, then $1/\delta_i$ is the CC index of node i.

To find *one* most important node in networks, one can find a node with the minimal distance, i.e.,

$$\arg\min_{v_i} \sum_{v_j \in V} \delta(v_i, v_j), \quad (v_i \in V). \tag{2}$$

In a similar way, if we want to find a set containing n_0 influential spreaders, denoted as $\tilde{V}$, the distance from $\tilde{V}$ to network is minimal. In other words, the generalized closeness centrality (termed, GCC index), given as:

$$GCC(\tilde{V}) = \frac{1}{\sum_{v_j \in V} \delta(\tilde{V}, v_j)}, \quad (\tilde{V} \subseteq V) \tag{3}$$

is the maximal.

Our main purpose is to find the set with the highest GCC value, namely, with the goal to minimize the following objective function:

$$D(\tilde{V}) = \sum_{v_j \in V} \delta(\tilde{V}, v_j), \quad (\tilde{V} \subseteq V). \tag{4}$$

Here the distance from set $\tilde{V}$ to a node v_j is defined as:

$$\delta(\tilde{V}, v_j) = \min_{\tilde{v}_l \in \tilde{V}} \delta(\tilde{v}_l, v_j), \quad \tilde{v}_l \in \tilde{V}. \tag{5}$$

As a result, Eq. (4) can be rewritten as:

$$D(\tilde{V}) = \sum_{v_j \in V} \min_{v_l \in \tilde{V}} \delta(\tilde{v}_l, v_j), \quad (\tilde{V} \subseteq V). \tag{6}$$

Our goal is to minimize the objective function in Eq. (6), namely,

$$\arg\min_{\tilde{V}} \sum_{v_j \in V} \min_{v_l \in \tilde{V}} \delta(\tilde{v}_l, v_j), \quad (\tilde{V} \subseteq V). \tag{7}$$

In principle, we should calculate the distance from $\tilde{V}$ to the residual network rather than the whole network. However, the distance from the nodes belonging to $\tilde{V}$ to $\tilde{V}$ is zero. Thus, it is unnecessary to stress the residual network in particular.

Let $\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_{n_0}$ be the n_0 nodes in $\tilde{V}$, and S_l be the subset of V whose nodes have the minimal distance to node $\tilde{v}_l \in \tilde{V}$ (that is to say, for a randomly selected node v_j, if it has the shortest distance to node $\tilde{v}_l$ among all n_0 nodes, then $v_j \in S_l$. In this way, each node can be classified into a subset S_l centering on node $\tilde{v}_l$), then the optimization problem in Eq. (7) can be further rewritten as:

$$\arg\min_{\tilde{V}} \sum_{\tilde{v}_l \in \tilde{V}} \sum_{v_j \in S_l} \delta(v_j, \tilde{v}_l), \quad (\tilde{V} \subseteq V). \tag{8}$$

2. K-means method

Though the optimization problem of the objective function has been presented in Eq. (8), it is very hard to obtain the exact solution. Fortunately, Eq. (8) is very similar to the optimization problem in the K-means clustering algorithm, which is defined as [20]:

$$\arg\min_{S} \sum_{l=1}^{n_0} \sum_{x_j \in S_l} \|X_j - \mu_l\|^2, \tag{9}$$

where X_j is the attribute/vector of the jth object. The set $S = \{S_1, S_2, \ldots, S_{n_0}\}$ divides the total objects into n_0 clusters, and μ_l is the center of set S_l (i.e., the distance from μ_l to all nodes in S_l is the smallest).

Denote $\|X_j - \mu_l\|^2$ as $\delta(X_j, \mu_l)$, Eq. (9) becomes:

$$\arg\min_{S} \sum_{l=1}^{n_0} \sum_{X_j \in S_l} \delta(X_j, \mu_l). \tag{10}$$

By comparing Eqs. (8) and (10), one can see that if X_j and μ_l correspond to v_j and $\tilde{v}_l$, respectively. Then the two equations are almost the same. The only difference is that the center μ_l may not locate in networks, namely, the centers are not the nodes of networks. In the next context, we will address how to solve such a problem when implementing the K-means method in networks.

In order to use the K-means method to approximate the optimization problem in Eq. (8), we need to define X_i to characterize the attribute of node v_i in the network. Spectral clustering theory has proven that, by using some mathematical transformations, the optimization problems of graph partition and clustering can be approximately characterized by the eigenvectors of the Laplacian matrix [21,22]. As a typical algorithm in clustering, the attributes used in the K-means method can be constructed based on the spectral of the Laplacian matrix. For example, Benson et al. developed a generalized framework for clustering networks by mapping the higher-order motifs into the weighted networks, then the K-means method used for clustering networks was implemented based on the spectral of the Laplacian matrix [23]. For this purpose, we let $\mathbf{A}$ be the adjacency matrix of a given network with N nodes, and define a diagonal matrix $\mathbf{D}$ whose diagonal element is the degree of each node, then the Laplacian matrix is given as $\mathbf{L} = \mathbf{D} - \mathbf{A}$. For the Laplacian matrix $\mathbf{L}$, we can calculate the first hth eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_h$ and their corresponding eigenvectors $Y_1, Y_2, \ldots, Y_h$, where $Y_i = \{y_{i1}, y_{i2}, \ldots, y_{iN}\}^T$. For simplicity, we mainly set $h = n_0$ in this paper, that is to say, the number of eigenvectors is the same to the number of chosen spreaders. We also investigate the sensitivity of h in Fig. 7, and we find that the results are generally robust to h once the value of h is not very small. By letting $X_i = \{y_{1i}, y_{2i}, \ldots, y_{hi}\}^T$ be the attribute of node v_i, thus, the distance between two nodes can be defined as:

$$\delta(v_i, v_j) = \|X_i - X_j\|^2. \tag{11}$$

Here the "distance" between two nodes is defined as their attribute's dissimilarity rather than the shortest path length.

As a result, Eq. (8) can be rewritten as

$$\arg\min_{\tilde{V}} \sum_{\tilde{v}_l \in \tilde{V}} \sum_{v_j \in S_l} \|X_j - X_l\|^2, \quad (\tilde{V} \subseteq V). \tag{12}$$

As suggested in Eq. (10), the above optimal solution can be approximately solved by the K-means method, and n_0 centers $\{\mu_1, \mu_2, \ldots, \mu_{n_0}\}$ can be easily obtained when implementing the K-means method on $\{X_1, X_2, \ldots, X_N\}$.

Inspired by the above comparison, we can use the K-means method to approximately solve the optimal solution in Eq. (8). Nevertheless, one should notice that these centers obtained by the K-means method may not be the nodes of the network.

To simplify the algorithm and to avoid the degree of chosen center too small, we first choose a node who is closest to its center, and then compare its degree with its neighbors who are also in the same cluster. Then choose the node with the highest degree as the center. If more than one node is found, one of them is randomly selected.

Therefore, the steps of obtaining n_0 influential nodes in social networks can be summarized as:

Step 1: Input the adjacency matrix **A** and determine the Laplacian matrix **L**, and then calculate the corresponding eigenvectors of the first hth eigenvalues,

Step 2: Define the node's attribute as $X_i = \{y_{1i}, y_{2i}, \ldots, y_{hi}\}^T, \ \ i = 1, \ldots, N$;

Step 3: Obtain the n_0 number of centers $\mu_1, \mu_2, \ldots, \mu_h$ by implementing the K-means method on $\{X_1, X_2, \ldots, X_N\}$. There are two iterative steps in the K-means method [20]: **Assignment step**: each observation is assigned to its closest cluster center; **Update step**: Calculate the new means to be the centroids of the observations in the new clusters. The algorithm converges when the assignments no longer change;

Step 4: For each center μ_i, find a node who is closest to the center μ_i, and compare its degree with its neighbors who are also in the same cluster. Then the node $\tilde{v}_i$ with the highest degree is viewed as the center. Namely, set $\{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_{n_0}\}$ is viewed as the n_0 multiple influential spreaders.

An illustration of our algorithm on a toy model is presented in Fig. 1.

3. Preliminaries

3.1. Centrality indices

Several widely used centrality indices are summarized as follows to compare with our method. The degree centrality (DC) of a node i is defined as the number of nearest neighbors, namely

$$DC(i) = \sum_{j=1}^{N} a_{ij}, \tag{13}$$

where a_{ij} is the element of adjacency matrix **A**.

The betweenness centrality (BC) of a node i is defined as the fraction of all shortest paths travel through the node, which is denoted as

$$BC(i) = \sum_{s \neq i \neq l} n_{sl}^i / n_{sl} \tag{14}$$

with n_{sl} and n_{sl}^i be the number of shortest paths between nodes s and l, and the number of shortest paths between s and l passing through node i.

The k-shell (KS) decomposition method is implemented by the following steps: Firstly, all the nodes with degree $k = 1$ are removed until all nodes' degrees are larger than one. All of these removed nodes are 1-shell. Then *recursively* remove the nodes with $k = 2$, and keep deleting the existing nodes until all nodes' degrees are larger than two, and include them into 2-shell. This procedure continues until all nodes have been assigned to a k-shell [12,24].

Recently, a quantity called "Collective Influence" (CI) was proposed by Morone and Makse [17], which can be used as an index to select multiple influential spreaders. Define $\partial Ball(i, s)$ be the frontier of a ball whose center is node v_i and the length of radius is s (defined as the shortest path). Then the CI index of node v_i at level s is defined as:

$$CI_s(i) = (k_i - 1) \sum_{j \in \partial Ball(i,s)} (k_j - 1), \tag{15}$$

where k_i is the degree of node v_i and s is a predefined nonnegative integer. In this paper we set $s = 3$. Initially, the node with the highest value of CI_s is selected as spreader. Recalculate the values of CI_s for the remaining nodes and select the new top CI_s node. Repeat the process until n_0 spreaders have been selected.

The degree discount heuristics (DDH) was proposed by Chen et al., the general idea is as follows [25]: let v_i be a neighbor of node v_j. If v_j has been selected as a spreader, when considering v_i as a new spreader based on its degree, we should not count the edge e_{ij} toward node v_j. Thus we should discount v_i's degree by one due to the presence of v_j in the spreader set, and we do the same discount on v_i's degree for every neighbor of v_i that is already in the spreader set.

By generalizing the graph coloring in graph theory into complex networks, Zhao et al. used the well-known Welsh–Powell algorithm to divide each network into several independent sets, where any two nodes in one independent set has no direct edge [18]. Then choose the top-n_0 nodes in the largest independent set based on a given centrality index as multiple spreaders. In this paper, we compare the coloring method based on the DC index (labeled as DCC), the BC index (labeled as BCC) and the KS index (labeled as KSC).

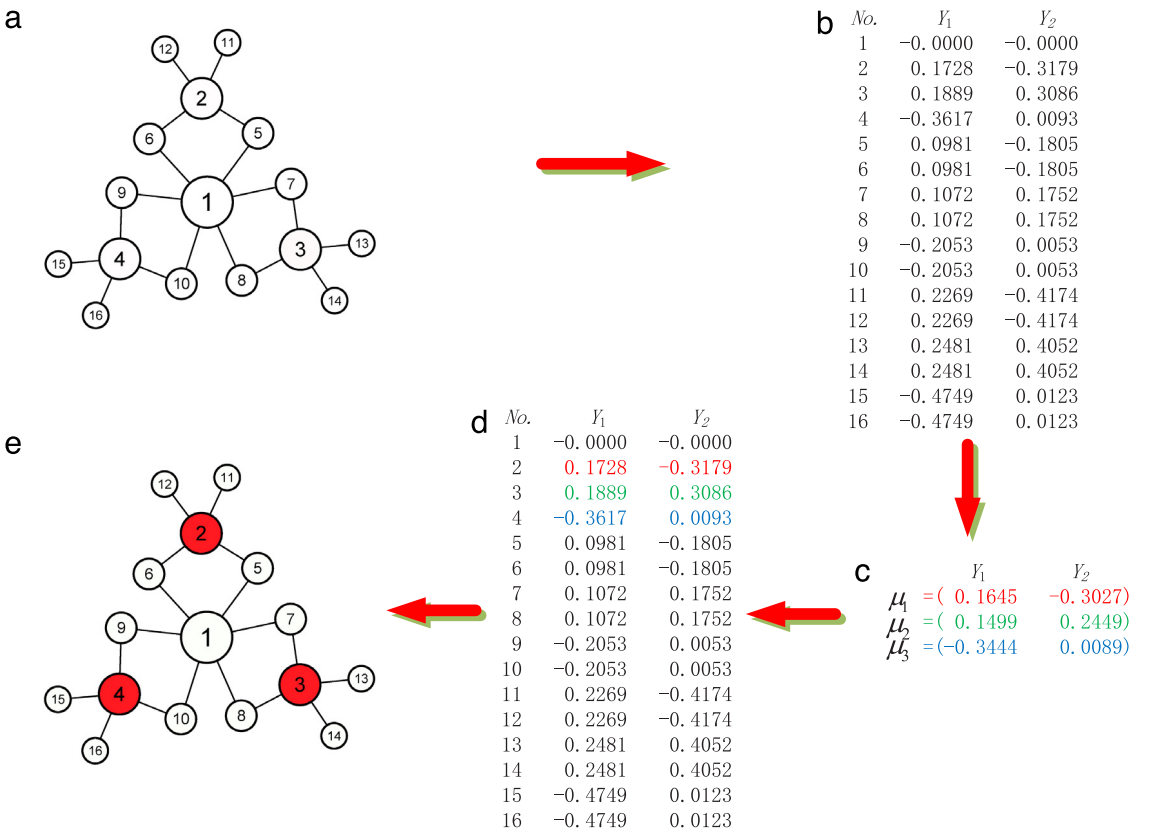

Fig. 1. (Color online) An illustration of our algorithm, where three influential spreaders should be detected. (a) a toy model with 16 nodes. (b) the eigenvectors of the first two non-zero eigenvalues of the Laplacian matrix L are calculated. So the attribute of a node v_i is composed of two elements. (c), the three centers obtained from the K-means method. (d) three nodes are chosen according to **step 4**, i.e., node 2, 3 and 4. (e) the three spreaders are marked as red nodes.

3.2. Spreading models

Previous literatures have proven that the influence of nodes is sensitively dependent on the spreading dynamics [26–28]. For instance, Borge-Holthoefer et al. have proven that since the central nodes act as rumor firewalls which may hinder rather enhance the spreading of rumor. Therefore, the nodes with high k-shell index do not imply high spreading influence [4]. In view of this, we use two typical spreading processes—SIR epidemic spreading process and rumor spreading process to systematically validate the effectiveness of our method in this paper.

For the classical SIR spreading process, where an infected node (I) can infect its susceptible neighbors (S) with transmission rate β, and then becomes recovered (R) with recovery rate μ. Therefore, each infected node can contact *all* of its neighbors at per time step (we call all-contact SIR spreading process) [29]. In reality, at per time step, one person can contact one neighbor at most, taking the sex activity and the telephone marketing activity as examples. As a result, we consider a modified SIR spreading process where infected nodes only contact *one* neighbor at per time step (we call single-contact SIR spreading process), which is similar to the contact process [30]. In this paper, we set $\mu = 1$ for the all-contact SIR spreading process and $\mu = 0.1$ for the single-contact SIR spreading process.

For the rumor spreading process, at each time step, on the one hand, a spreader can infect an ignorant with a rate β, on the other hand, the spreader becomes a stifler with a rate μ once it contacts a spreader or stifler neighbor. The most remarkable difference between the two spreading processes is: for the SIR spreading process, where infected nodes become recovered nodes by themselves. However, for the rumor spreading process, where a spreader becomes a stifler only when they contact at least a spreader or stifler neighbor. Because of this main difference, the traditional methods may fail in identifying the influential spreaders. We also set $\mu = 1.0$ for the rumor spreading process [31,32].

3.3. Data description

To evaluate the performance of our method, we implement several indices in four empirical networks, including the Email (e-mail network of University at Rovira i Virgili, URV) [33], the PB-A network of the US political blogs [34], the Power—An

Table 1
The basic topological features and the epidemic threshold of four empirical networks. N and M are the total numbers of nodes and links, respectively. C and r are the clustering coefficient and the assortative coefficient, respectively. β_c is the epidemic threshold, defined as $\beta_c = \frac{\langle k\rangle}{\langle k^2\rangle}$.

Network	N	M	C	r	β_c
Email	1133	5451	0.220	0.078	0.053
PB	1222	16724	0.36	−0.221	0.0123
Power	4941	6594	0.107	0.003	0.0258
Router	5022	6258	0.033	−0.138	0.072

electrical power grid of the western US [35], the Router (the router-level topology of the Internet) [36]. For simplicity, these networks are treated as undirected and unweighed networks in this work. The detailed information about these empirical networks are presented in Table 1.

4. Main results

A relative ratio Δ is defined to compare the effectiveness of different methods, which is denoted as [18]

$$\Delta = \frac{R_i - R_{DC}}{R_{DC}}, \tag{16}$$

where R_i and R_{DC} are the final number of recovered nodes (or stiflers for rumor spreading process) for a used method and the DC method, respectively. $\Delta > 0$ means that the performance of the used method is better than the DC method.

We first compare our GCC index with the other indices for the single-contact SIR spreading process in Fig. 2 under different situations. As shown in Fig. 2, the GCC index has the best performance in facilitating the spreading influence regardless of the different network structures, the different values of n_0 and the different transmission rates β (the epidemic threshold $\langle k\rangle/\langle k^2\rangle$ for different networks are also summarized in Table 1). However, for the other indices, no one has the dominant advantage under different situations. Since the spreading influence of single-contact SIR spreading process is mainly determined by the average distance among multiple spreaders, but the effect of node degree is not significant. From Fig. 5, one can observe that the GCC index can lead to the largest average distance among the multiple spreaders, inducing the best performance of GCC index.

Next we study the performances of different indices for the all-contact SIR spreading process. Fig. 3 indicates the performance of the GCC index is still generally better than the other indices except for the DCC index or DDH index. For the Power network, the performances of the DCC index and DDH are always better than the other indices. The reasons can be mainly summarized as follows: Unlike the single-contact SIR spreading process, the spreading influence of the all-contact SIR spreading process is also sensitively dependent on the degree of spreaders. For the Power and Router networks, when the multiple spreaders are chosen based on the DCC index or DDH index, their average distance is rather large (see Fig. 5(c) and (d)) and their degree is large too (see Fig. 6(c) and (d)), namely, the DCC index and DDH can ensure two important conditions for all-contact SIR case: the chosen nodes are rather "important" and their distance are rather dispersive, leading to the better performance of DCC index and DDH index. For our proposed GCC index, though the chosen multiple spreaders have the largest average distance $\bar{d}$ (see Fig. 5), the average degree of the multiple spreaders is not large enough (see Fig. 6). Combining above explanations, we can understand why the performance of GCC index in the all-contact SIR spreading process is not always better than the DCC index or DDH index.

The comparison of different indices regarding to rumor spreading process is presented in Fig. 4: like the all-contact SIR case, the performance of GCC index is the best when different indices are compared in the Email network and the PB network (the top two rows in Fig. 4). For the Power network (the third row in Fig. 4), the performance of the DDH index is the best, and the performance of GCC index is closer to that of the DDH index with the increasing of the value of n_0. As shown in Fig. 5(c), the average distance $\bar{d}$ based on the DDH index is very close to that of the GCC. Moreover, Fig. 6(c) denotes that the average degree $\langle k\rangle_m$ based on the DDH index is the largest. So the performance of the DDH is the best. For the Router network, the performance of the CI index is the best, and the GCC index is the second best. The reason can be similarly explained as the case of the Power network. However, we can address that our GCC index is a good index from a overall perspective. Moreover, we should address that the value of the DDH or CI should be recalculated after one node is chosen, that is to say, the multiple spreaders are chosen *one by one*. They are remarkably different from our GCC index, in which all multiple spreaders are *simultaneously* chosen.

To explain the observed phenomena in Figs. 2–4, the average distance $\bar{d}$ and the average degree $\langle k\rangle_m$ as functions of n_0 are plotted in Figs. 5 and 6, respectively. Fig. 5 demonstrates that the GCC index can make sure that the value of $\bar{d}$ is the largest. Fig. 6 indicates that the average degree of the GCC index is smaller than the other indices. Combining the remarkable performances of the GCC index, one can find that the distance among multiple spreaders is a particularly important factor in facilitating spreading process, which is significantly different from the single spreader case. The results also suggest that we need not always focus on the "important persons/webs" to accelerate information diffusion or promote new products when multiple targets are considered, proper selection of multiple spreaders can effectively reduce the cost of promotion.

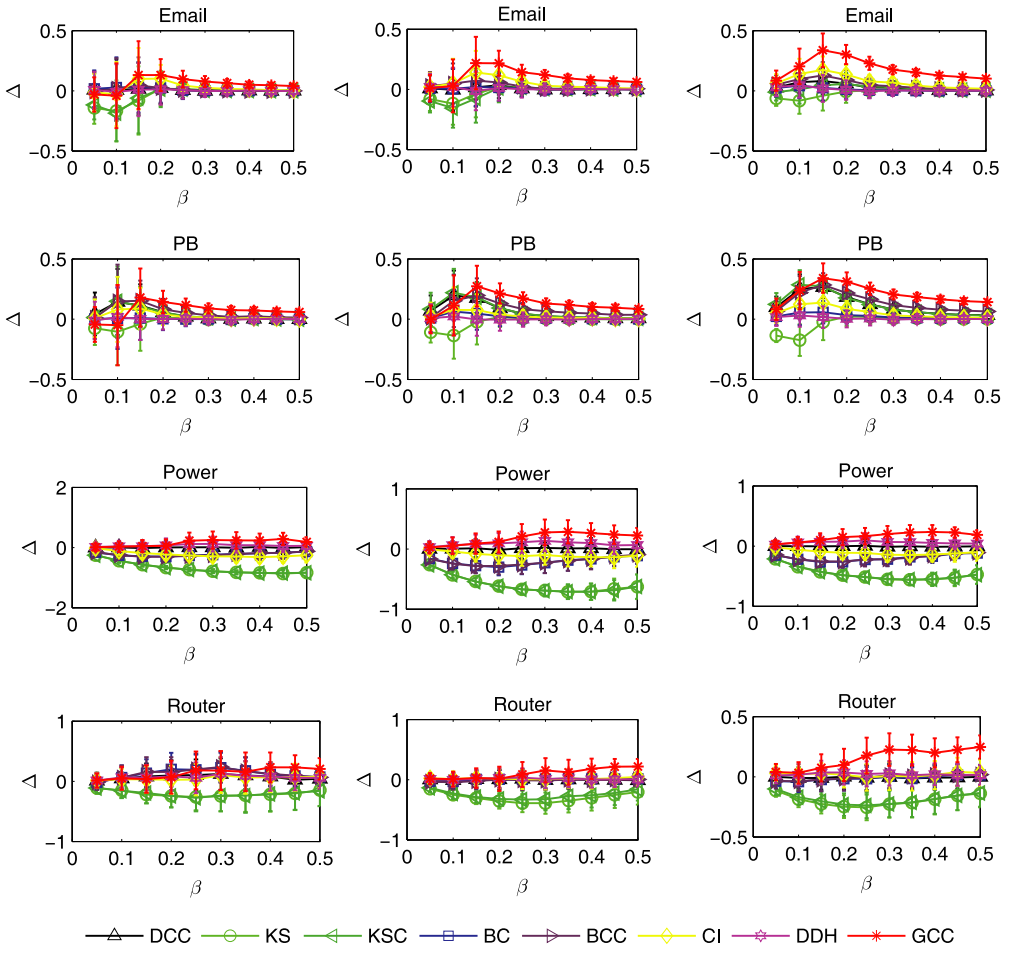

Fig. 2. (Color online) For single-contact SIR spreading process, the relative ratios Δ for different indices as functions of transmission rate β are compared in four real networks. Left panels: the number of spreaders $n_0 = 30$; Middle panels: the number of spreaders $n_0 = 50$; Right panels: the number of spreaders $n_0 = 90$. Here recovery rate $\mu = 0.1$. The error bars are given by the standard deviation. The results are averaged over 500 independent runs.

Furthermore, we systematically investigate the effect of n_0 on the performances of different indices in Fig. 7, one can see that the GCC index has no any advantage when the value of n_0 is small. In other words, our method is not superior to other indices when the number of spreaders is too few. Similar to the single spreader case, the influence of spreader itself (i.e., such as degree) plays significant role in influence maximization when spreaders is few. However, with the increasing of the value of n_0, one can observe that the advantage of the GCC index becomes more obvious and then better than the other indices generally.

Our main results are obtained by assuming that the number of chosen eigenvectors h equals to the number of multiple spreaders n_0, so it is necessary to check the impact of h on the effectiveness of our method. As shown in Fig. 8, the results in the Email and PB networks illustrate that increasing the value of h can slightly increase the value of R_{GCC} (the final spreading fraction when the multiple spreading are chosen based on our GCC method). However, for the Power and Router networks, the monotonicity is not evident regardless of single-contact SIR spreading process, all-contact SIR spreading process or rumor spreading process. Therefore, it is a reasonable to choose $h = n_0$ in our paper.

5. Conclusions

In this paper, inspired the CC index for a node, we have defined a GCC index for a set of nodes to identify multiple influential spreaders. Our analysis suggested that how to obtain a set of nodes with highest GCC value corresponds to how to obtain an optimal solution for an objective function. Since this problem is similar to finding the optimization problem in the K-means method, the influential nodes can be approximately detected by the K-means method. By implementing SIR spreading process (single-contact and all-contact) and rumor spreading process in real networks, we found that the GCC

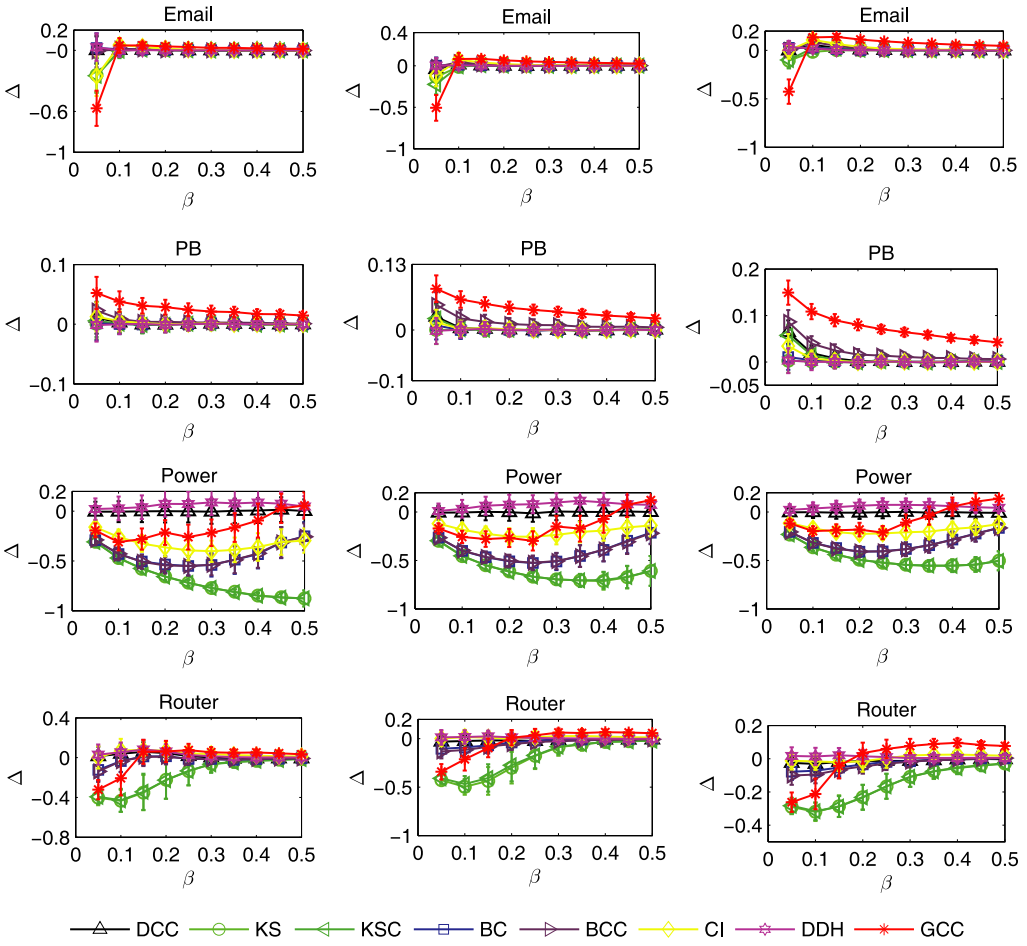

Fig. 3. (Color online) For all-contact SIR spreading process, the relative ratios Δ for different indices as functions of transmission rate β are compared in four real networks. Left panels: the number of spreaders $n_0 = 30$; Middle panels: the number of spreaders $n_0 = 50$; Right panels: the number of spreaders $n_0 = 90$. Here recovery rate $\mu = 1.0$. The error bars are given by the standard deviation. The results are averaged over 500 independent runs.

index has a good performance in identifying multiple influential spreaders. The reason is that the average distance among selected spreaders plays significant roles in facilitating the spreading influence. Our method makes sure that the chosen multiple spreaders are dispersively distributed in networks. Moreover, the performance of the GCC index in the single-contact SIR spreading process is extraordinarily remarkable, since which is not sensitively dependent on the importance of node itself. One should note that each index is proposed from a certain perspective, so there is no one index always better than all the other indices. For example, the node with the largest CC index or CI index cannot always yield the maximization influence under different situations. Likewise, our proposed GCC index ensures its performance is better than the other indices in many cases, so we can claim that the GCC index is a good method in identifying multiple spreaders.

There are some improvements should be considered, such as, how to further improve the method such that the average degree of spreaders is also large, how to generalize the method to weighted or directed networks, and so forth. Meanwhile, our GCC index is proposed from the structural perspective, as we have mentioned, the problem of influence maximization is also sensitively dependent on the spreading dynamics. Therefore, a more rigorous index regarding of the multiple spreaders case should be proposed from structural perspective as well as spreading dynamics perspective. It is a key problem in this field in further works.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61473001, 11575041), and partially supported by Anhui Technology Foundation for Overseas Chinese Talents (Grant No. 05015139).

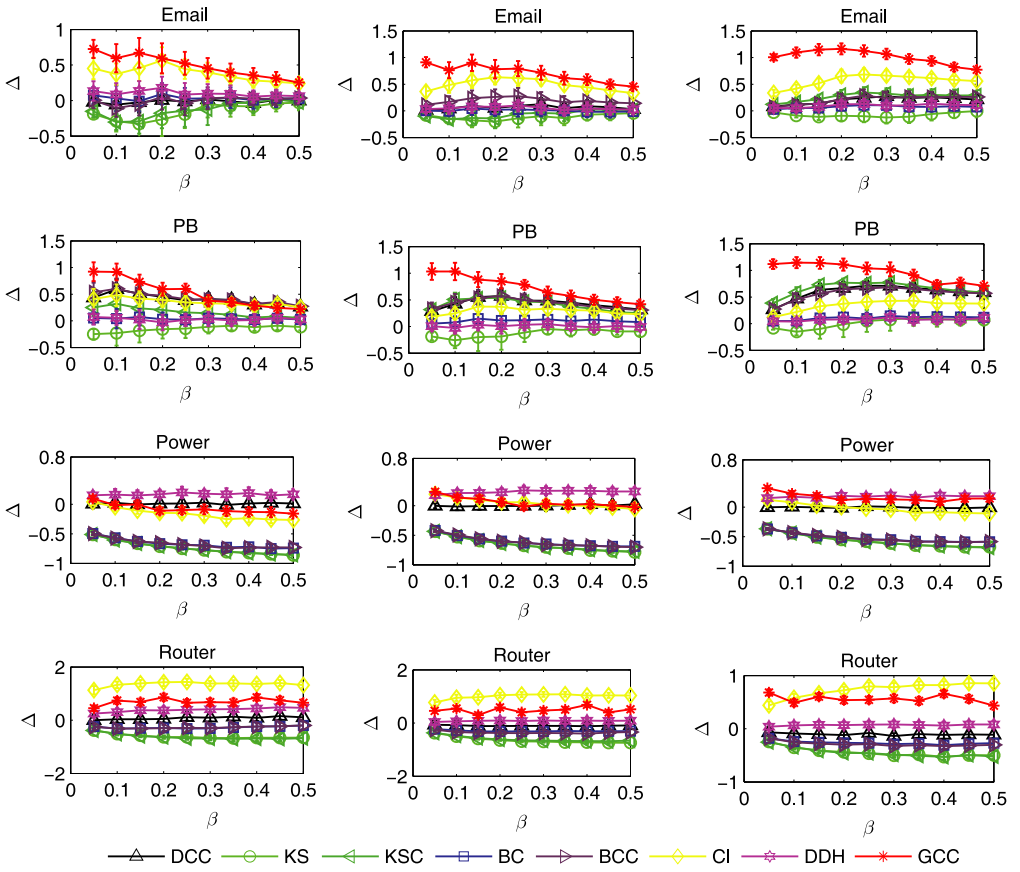

Fig. 4. (Color online) For rumor spreading process, the relative ratios Δ for different indices as functions of transmission rate β are compared in four real networks. Left panels: the number of spreaders $n_0 = 30$; Middle panels: the number of spreaders $n_0 = 50$; Right panels: the number of spreaders $n_0 = 90$. Here recovery rate $\mu = 1.0$. The error bars are given by the standard deviation. The results are averaged over 500 independent runs.

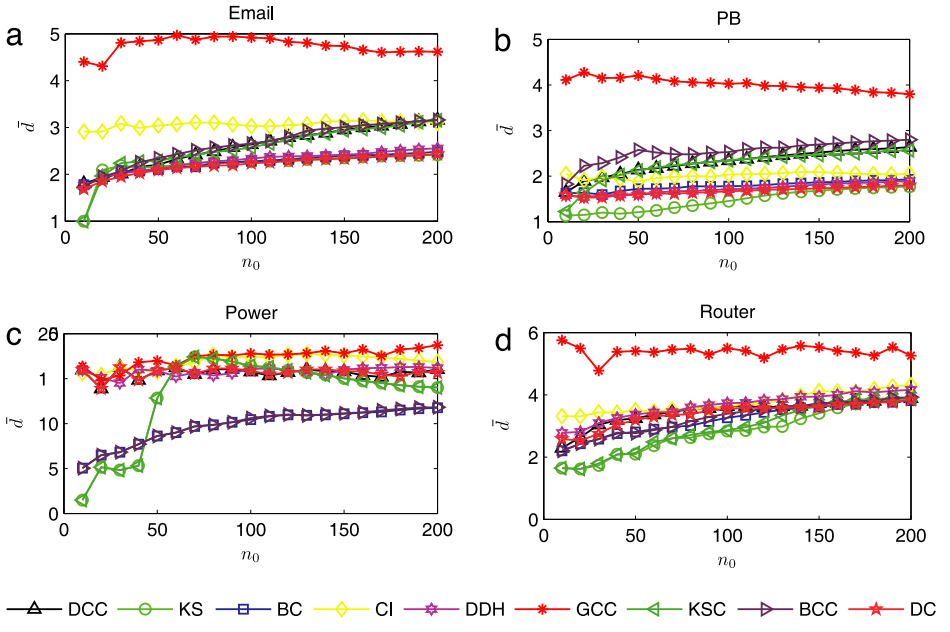

Fig. 5. (Color online) For different methods, the average distance $\bar{d}$ among the chosen multiple spreaders as function of n_0 is compared in four real networks. The results are averaged over 500 independent runs.

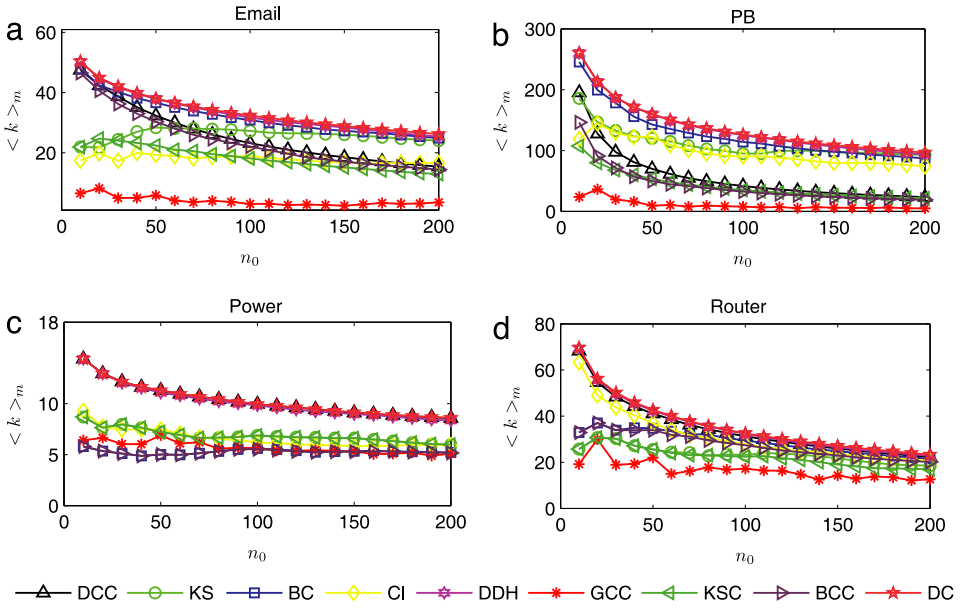

Fig. 6. (Color online) For different methods, the average degree $\langle k \rangle_m$ of the chosen multiple spreaders as function of n_0 is compared in four real networks. The results are averaged over 500 independent runs.

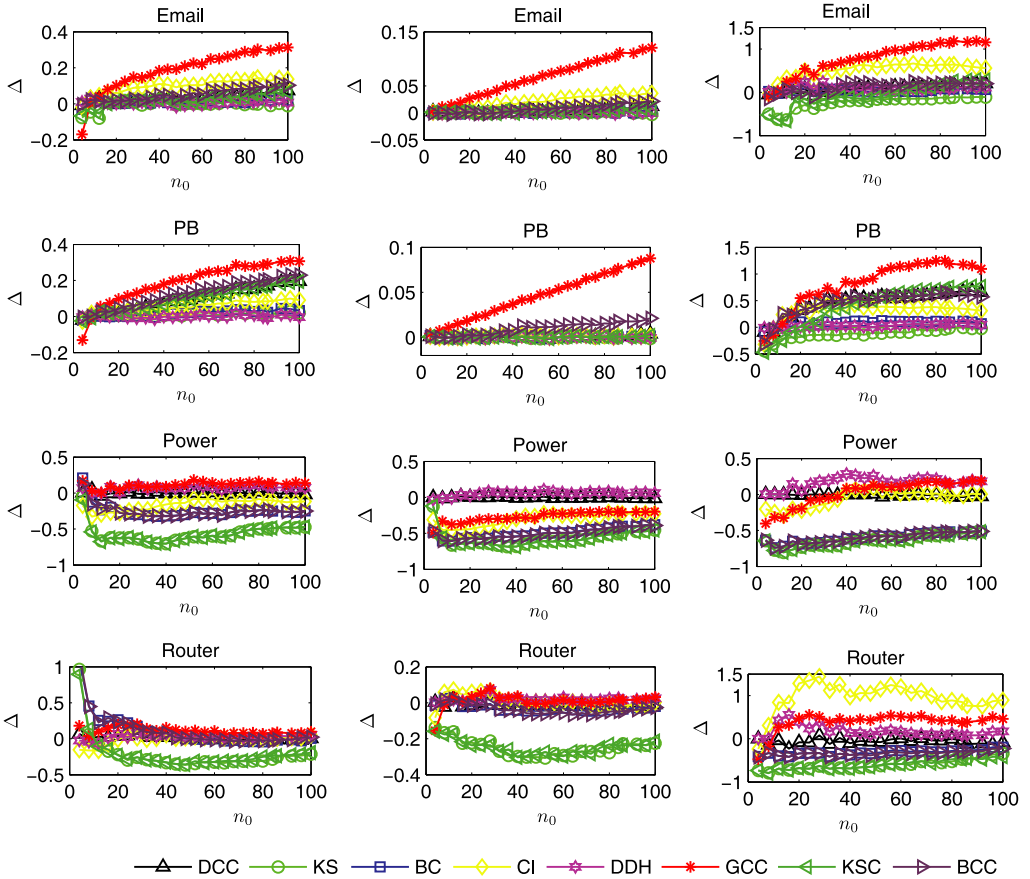

Fig. 7. (Color online) The effect of the number of spreaders on the performances of different indices. Left panels: single-contact SIR spreading process; Middle panels: all-contact SIR spreading process; Right panels: rumor spreading process. Here $\beta = 0.2$ for the three spreading models. The results are averaged over 500 independent runs.

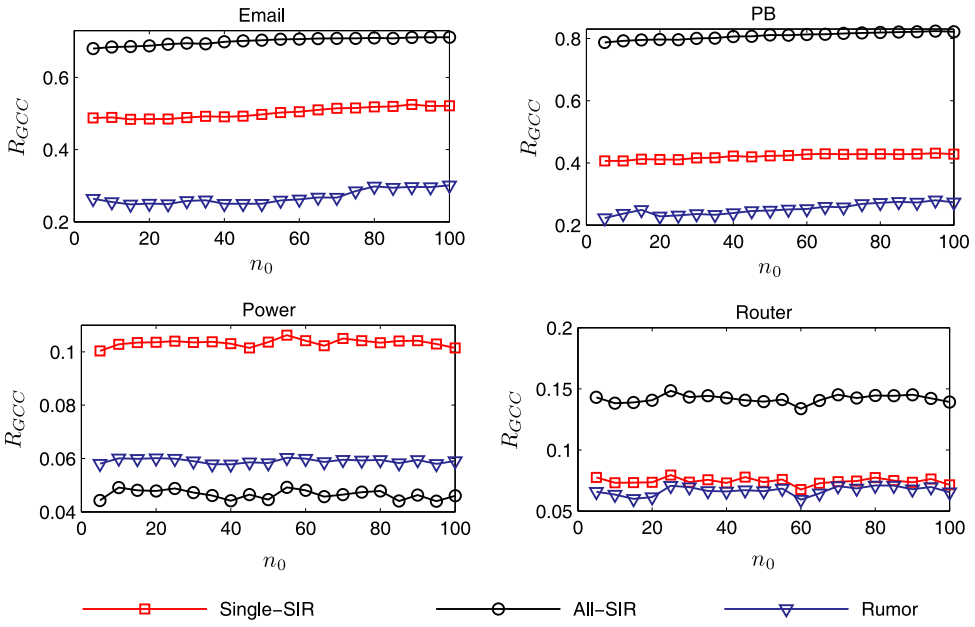

Fig. 8. (Color online) For single-contact SIR spreading process, all-contact SIR spreading process and rumor spreading process, the final spreading fraction R_{GCC} as function of the value of h. Here $\beta = 0.2$ for the three spreading models and $n_0 = 100$.

References

[1] P. Basaras, D. Katsaros, L. Tassiulas, Detecting influential spreaders in complex, dynamic networks, Computer 46 (4) (2013) 0024–29.
[2] J. Zhang, X.-K. Xu, P. Li, K. Zhang, M. Small, Node importance for dynamical process on networks: A multiscale characterization, Chaos 21 (1) (2011) 016107.
[3] L. Lü, Y.-C. Zhang, C.H. Yeung, T. Zhou, Leaders in social networks, the delicious case, PloS One 6 (6) (2011) e21202.
[4] J. Borge-Holthoefer, Y. Moreno, Absence of influential spreaders in rumor dynamics, Phys. Rev. E 85 (2) (2012) 026116.
[5] J.-G. Liu, Z.-M. Ren, Q. Guo, Ranking the spreading influence in complex networks, Physica A 392 (18) (2013) 4154–4159.
[6] K. Klemm, M.Á. Serrano, V.M. Eguíluz, M. San Miguel, A measure of individual role in collective dynamics, Sci. Rep. 2 (2012) 292.
[7] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, T. Zhou, Vital nodes identification in complex networks, Phys. Rep. 650 (2016) 1–63.
[8] Z.-M. Ren, A. Zeng, D.-B. Chen, H. Liao, J.-G. Liu, Iterative resource allocation for ranking spreaders in complex networks, Europhys. Lett. 106 (4) (2014) 48005.
[9] P. Bonacich, Factoring and weighting approaches to status scores and clique identification, J. Math. Sociol. 2 (1) (1972) 113–120.
[10] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry (1977) 35–41.
[11] G. Sabidussi, The centrality index of a graph, Psychometrika 31 (4) (1966) 581–603.
[12] 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 (11) (2010) 888–893.
[13] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, 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.
[14] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496.
[15] Z.-L. Hu, J.-G. Liu, G.-Y. Yang, Z.-M. Ren, Effects of the distance among multiple spreaders on the spreading, Europhys. Lett. 106 (1) (2014) 18002.
[16] Z.-L. Hu, Z.-M. Ren, G.-Y. Yang, J.-G. Liu, Effects of multiple spreaders in community networks, Internat. J. Modern Phys. C 25 (05) (2014) 1440013.
[17] F. Morone, H.A. Makse, Influence maximization in complex networks through optimal percolation, Nature 524 (2015) 65–68.
[18] X.-Y. Zhao, B. Huang, M. Tang, H.-F. Zhang, D.-B. Chen, Identifying effective multiple spreaders by coloring complex networks, Europhys. Lett. 108 (6) (2015) 68005.
[19] L. Guo, J.-H. Lin, Q. Guo, J.-G. Liu, Identifying multiple influential spreaders in term of the distance-based coloring, Phys. Lett. A 380 (2016) 837–842.
[20] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, Oakland, CA, USA, 1967, pp. 281–297.
[21] A.Y. Ng, M.I. Jordan, Y. Weiss, et al., On spectral clustering: Analysis and an algorithm, Adv. Neural Inf. Process. Syst. 2 (2002) 849–856.
[22] M. Newman, Networks: An Introduction, Oxford university press, 2010.
[23] A.R. Benson, D.F. Gleich, J. Leskovec, Higher-order organization of complex networks, Science 353 (6295) (2016) 163–166.
[24] Y. Liu, M. Tang, T. Zhou, Y. Do, Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition, Sci. Rep. 5 (2015) 9602.
[25] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 199–208.
[26] Y. Liu, M. Tang, T. Zhou, Y. Do, Improving the accuracy of the k-shell method by removing redundant links: from a perspective of spreading dynamics, Sci. Rep. 5 (2015) 13172.
[27] J.-G. Liu, J.-H. Lin, Q. Guo, T. Zhou, Locating influential nodes via dynamics-sensitive centrality, Sci. Rep. 6 (2016) 21380.
[28] J. Wang, X. Hou, K. Li, Y. Ding, A novel weight neighborhood centrality algorithm for identifying influential spreaders in complex networks, Physica A 475 (2017) 88–105.
[29] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Modern Phys. 87 (3) (2015) 925.
[30] R. Yang, T. Zhou, Y.-B. Xie, Y.-C. Lai, B.-H. Wang, Optimal contact process on complex networks, Phys. Rev. E 78 (6) (2008) 066109.

[31] Y. Moreno, M. Nekovee, A.F. Pacheco, Dynamics of rumor spreading in complex networks, Phys. Rev. E 69 (6) (2004) 066130.
[32] Z. Liu, Y.-C. Lai, N. Ye, Propagation and immunization of infection on general networks with both homogeneous and heterogeneous components, Phys. Rev. E 67 (3) (2003) 031911.
[33] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, A. Arenas, Self-similar community structure in a network of human interactions, Phys. Rev. E 68 (6) (2003) 065103.
[34] S.D. Reese, L. Rutigliano, K. Hyun, J. Jeong, Mapping the blogosphere professional and citizen-based media in the global news arena, Journalism 8 (3) (2007) 235–261.
[35] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440–442.
[36] N. Spring, R. Mahajan, D. Wetherall, T. Anderson, Measuring isp topologies with rocketfuel, IEEE/ACM Trans. Netw. 12 (1) (2004) 2–16.