Identifying influential spreaders in complex networks based on gravity formula

Ling-ling Ma[a] Chuang Ma[a] Hai-Feng Zhang[a,b,c 1]

[a] School of Mathematical Science, Anhui University, Hefei 230601, P. R. China

[b] Research centre of information supply & assurance, Anhui University, Hefei 230601, P. R. China

[c] Department of Communication Engineering, North University of China, Taiyuan, Shan'xi 030051, P. R. China

Abstract

How to identify the influential spreaders in social networks is crucial for accelerating/hindering information diffusion, increasing product exposure, controlling diseases and rumors, and so on. In this paper, by viewing the k-shell value of each node as its mass and the shortest path distance between any two nodes as their distance, then inspired by the idea of the gravity formula, we propose a gravity centrality index to identify the influential spreaders in complex networks. The comparison between the gravity centrality index with some well-known centralities, such as degree centrality, betweenness centrality, closeness centrality, and k-shell centrality, and so forth, indicates that our method can effectively identify the influential spreaders in real networks as well as artificial networks. We also use the classical Susceptible-Infected-Recovered (SIR) epidemic model to verify the good performance of our method.

Key words: Complex network, Influential spreader, Gravity formula.

1 Introduction

To effectively identify influential spreaders in social networks is of theoretical and practical significance [1,2,3,4,5,6,7,8,9,10,11], since it is crucial for developing efficient strategies to control epidemic spreading, accelerate information diffusion, promote new products, and so on. In view of this, many centrality indices have been proposed to address this problem, including degree centrality [12], betweenness centrality [13], neighborhood centrality [14] and closeness

[1] Corresponding author: haifengzhang1978@gmail.com

centrality [15], etc. In particular, Kitsak *et al.* proposed a k-shell decomposition to identify the most influential spreaders based on the assumption that nodes in the same shell have similar influence and nodes in higher shells are likely to infect more nodes, which is found to be better than the degree centrality index in many real networks [1]. However, recent researches have demonstrated that the nodes within the same k-core often have distinct influences, and this method may fail in some networks without core-like structure, e.g., Barasási-Albert network [16]. Thus, after this, some methods are proposed to further improve the performance of the k-shell method. For example, Zeng *et al.* proposed a mixed degree decomposition method by incorporating the residual degree and the exhausted degree [17]; Chen *et al.* devised a semi-local index by considering the next nearest neighborhood [18]; Lin *et al.* presented an improved ranking method by taking into account the shortest path distance between a target node and the node set with the highest k-core value [19]; Recently, Bae *et al.* defined a novel measure–coreness centrality index, which is given by summing all neighbors' k-shell values [20].

In general, a node's influence is not only dependent on its nearest neighbors but also on the nodes who are not the nearest neighbors, meanwhile, their interaction influence commonly decreases with their shortest path distance. If the k-shell value of each node is viewed as its mass, and the shortest path distance between two nodes is defined as their distance, then we can use the idea of gravity formula to measure the influence of nodes. Inspired by these factors, in the work, we propose a new centrality index to measure the influence of nodes, which is called gravity centrality index. We apply the susceptible-infectious-recovered (SIR) spreading dynamics to evaluate the effectiveness of our proposed method, the experimental results indicate that gravity centrality index can better evaluate the influence of nodes than the ones generated by degree centrality, betweenness centrality, k-shell centrality, closeness centrality, and so on.

The layout of the paper is as follows: In Sec. 2, we first briefly review several typical centrality indices which are used to compared in this work, and present the description of our method. Then the experimental results are presented in Sec. 3. Finally, Conclusions and discussions are summarized in Sec. 4.

2 Model

A undirect network is represented by $G = (N, M)$ with N nodes and M edges, and its structure can be described by an adjacent matrix $A = (a_{ij})_{N\times N}$ where $a_{ij} = 1$ if node i is connected to node j, and $a_{ij} = 0$ otherwise.

Here we briefly review the definitions of several centrality indices that will be discussed in this work.

The degree centrality (DC) of a node is defined as the number of nearest neighbors. The betweenness centrality(BC) of a node is defined as the fraction of all shortest paths travel through the node. The closeness centrality (CC) of a node is defined as the reciprocal of the sum of the lengths of the geodesic distance to every other node. The k-shell decomposition method(ks) is implemented by the following steps: Firstly, remove all nodes with degree one, and keep deleting the existing nodes until all nodes' degrees are larger than one. All of these removed nodes are assigned 1-shell. Then recursively remove the nodes with degree two and include them to 2-shell. This procedure continues as until all nodes in the networks have been assigned to one of the shells [17].

To improve the exactness of k-shell method, the mixed degree decomposition (MDD) method was proposed by Zeng *et al.* [17]. The mixed degree $k_m(i)$ for a node i is defined by considering the residual degree $k_r(i)$ and the exhausted degree $k_e(i)$ simultaneously, which is written as:

$$k_m(i) = k_r(i) + \lambda * k_e(i). \tag{1}$$

At each step of the MDD procedure, the nodes are removed according to the mixed degree, and the mixed degrees of remaining nodes are also updated. where λ is a tunable parameter between 0 and 1. As in Ref. [17], we take $\lambda = 0.7$ in this work.

Recently, Baus *et al.* designed a ranking method–neighborhood coreness C_{nc} by considering the degree and the coreness of a node simultaneously, the $C_{nc}(i)$ for a node i is defined as [20]

$$C_{cn}(i) = \sum_{j \in \Lambda_i} ks(j), \tag{2}$$

where Λ_i is the neighbor node set of node i. They further developed an extended neighborhood coreness C_{nc+}, which is described as:

$$C_{nc+}(i) = \sum_{j \in \Lambda_i} C_{cn}(j). \tag{3}$$

It is fact that, on the one hand, the influence of a node is increased if its neighbors (here the neighbors of a node do not just includes its nearest neighbors, which may also include next nearest neighbors, next-next nearest neighbors, etc.) have higher value of ks; on the other hand, the interaction effect between two nodes decreases with their distance. Enlighten by the idea of gravity formula, we can view the k-shell value of node i as its mass, and the shortest path distance between two nodes in network is viewed as their distance. In this way, the influence of node i is measured by (labeled G):

$$G(i) = \sum_{j \in \psi_i} \frac{ks(i)ks(j)}{d_{ij}^2}, \tag{4}$$

where d_{ij} is the shortest path distance between node i and node j. ψ_i is the neighborhood set whose distance to node i is less than or equal to a given value r. Since real networks are often very large, in the paper, we let $r = 3$, i.e, only nearest neighbors, next nearest neighbors and the next-next nearest neighbors are considered.

An extended gravity index is further developed based on Eq. (4), which is defined as (labeled G_+):

$$G_+(i) = \sum_{j \in \Lambda_i} G(j), \tag{5}$$

also Λ_i is the nearest neighborhood of node i.

Since the k-shell decomposition method can be efficiently implemented with the linear time complexity of $O(m)$, where m is the number of edges in the network [20], and our method itself is a semi-local index. As a result, our method based on G or G_+ index is more efficient than the ranking algorithms based on betweenness centrality as well as closeness centrality.

3 Experimental results

In this section, we compare the effectiveness of other indices with our proposed G or G_+ index from different aspects and on different networks, including real networks as well as artificial networks.

We employ the standard susceptible-infected-removed (SIR) model [21] to estimate the spreading influence of the nodes (labeled by R). In detail, to check the spreading influence of one given node, we set this node as an infected node and the other nodes are susceptible nodes. At each time step, each infected node can infect its susceptible neighbors with infection probability β, and then it recovered from the diseases with probability μ. In this paper, we set $\mu = 1.0$. This process repeats until there has no any infected nodes. At last, the number of recovered nodes is used to reflect the *real* influence of the node. To guarantee the reliability of the results, all of them are averaged over 1000 independent realizations.

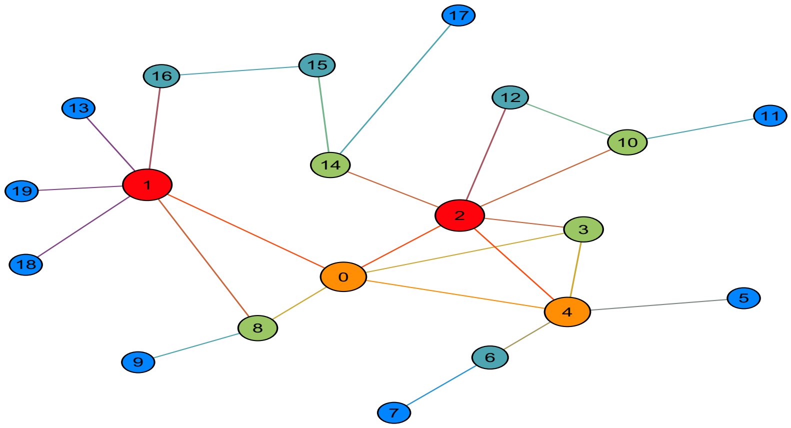

Fig. 1. (Color online) An example network consisted of 20 nodes and 25 edges. Nodes with larger degrees have larger size.

First, a small example network consisting $N = 20$ nodes and $M = 25$ edges is given in Fig. 1 to intuitively compare these indices, the ranking lists from different indices are presented in table 1

Table 1
The ranking lists determined by different indices. Degree centrality: DC; mixed Degree decomposition: MDD; gravity centrality: G: extended gravity centrality: G_+; extended neighborhood coreness defined in Eq. (3): C_{nc+}; k-shell decomposition:ks; betweeness centrality: BC; closeness centrality: CC; the node spreading influence evaluated by SIR model: R, by taking $\beta = 0.25$.

Rank	DC	MDD	G	G_+	C_{nc+}	ks	BC	CC	R
1	1, 2	1, 2	2	2	0	0, 2, 3, 4	2	0	2
2	0,4	0, 4	0	0	2	1, 8, 10, 12, 14-16	0	2	0
3	3, 8, 10, 14	3	4	4	4	others	1	4	4
4	6, 12, 15, 16	8	3	3	3	—	4	1	1
5	others	10,14	1	1	1	—	14	3	3
6	—	12, 15, 16	8	8	8	—	6, 8, 10	8	8
7	—	6	14	—	10	—	16	14	14
8	—	others	10	—	12, 14	—	15	10	10
9	—	—	—	—	—	—	others	—	—

From table 1, one can see that, the *ks* centrality cannot well distinguish the influence of nodes, even in the same shell, the nodes' influence may be totally

Table 2
Basic Structural properties. N and M are the number of nodes and edges, respectively. β_{th} is the epidemic threshold. H is degree heterogeneity, given by $\langle k^2\rangle/\langle k\rangle^2$. r is degree assortativity. C is clustering coefficient. L is average shortest path length.

Network	N	M	β_{th}	H	r	C	L
Facebook	324	2218	0.047	1.567	0.247	0.465	3.054
Netsci	379	914	0.125	1.663	-0.082	0.741	6.042
Email	1133	5451	0.053	1.942	0.078	0.220	3.606
TAP	1373	6833	0.061	1.644	0.579	0.529	5.224
Y2H	1458	1948	0.140	2.667	-0.209	0.071	6.812
Blogs	3982	6803	0.072	4.038	-0.133	0.284	6.252
Router	5022	6258	0.072	5.503	-0.138	0.012	6.449
HEP	5835	13815	0.110	1.926	0.185	0.506	7.026

different. Moreover, the result indicates that, our proposed G and G_+ index can effectively identify the influence of nodes, i.e., the ranking list determined from G or G_+ index is in good agreement with the ranking list obtained from SIR model in the last row.

To validate the effectiveness of the G or G_+ index, we apply it to 8 real networks, including Facebook (Slavo Zitnik's friendship network in Facebook) [22], Netsci (collaboration network of network scientists), Email (communication), TAP(yeast protein-protein binding network generated by tandem affinity purification experiments),Y2H (yeast protein-protein binding network generated using yeast two hybridization), Blogs (the communication relationships between owners of blogs), Router (the router-level topology of the Internet), HEP (collaboration network of high-energy physicists) [17]. For simplicity, these networks are treated as non-directed and non-weighted networks in this work. The detailed information about these 8 real networks are presented in table 2.

A good index should be able to distinguish the nodes' differences, that is to say, few nodes have the same ranking value. For example, as illustrated in table 1, G or G_+ index is good at distinguishing the nodes' differences, which is much better than the ks index. So to quantitatively measure the monotonicity of a ranking list X, a monotonicity index $M(X)$ is given as [20]:

$$M(X) = [1 - \frac{\sum_{r\in V} N_r(N_r-1)}{N(N-1)}]^2, \tag{6}$$

where N is the size of network, and N_r is the number of nodes with the same index value r. If $M(X) = 1$, which means that the ranking method is perfectly monotonic and each node is categorized a different index value; otherwise, all nodes are in the same rank as $M(X) = 0$. The monotonicity M for different ranking methods is summarized in Table 3. Generally, the results suggest that G or G_+ index can give higher value of M, moreover, $M(G)$ and $M(G_+)$ are very near 1 in some networks. Therefore, gravity method can better distinguish the node's influence than other indices.

Table 3
M (.) is the monotonicity of the corresponding measures.

Network	M(DC)	M(MDD)	M(G)	M(Cnc+)	M(ks)	M(BC)	M(CC)
Facebook	0.9315	0.9729	0.9999	0.9995	0.8445	0.9855	0.9953
Netsci	0.7642	0.8215	0.9949	0.9893	0.6421	0.3387	0.9928
Email	0.8874	0.9229	0.9999	0.9991	0.8088	0.9400	0.9988
TAP	0.8991	0.9599	0.9994	0.9981	0.8380	0.9238	0.9988
Y2H	0.4884	0.5304	0.9966	0.9633	0.2972	0.5063	0.9957
Blogs	0.5654	0.5906	0.9976	0.9868	0.4670	0.4004	0.9973
Router	0.2886	0.3009	0.9967	0.9657	0.0691	0.2983	0.9961
HEP	0.7654	0.8314	0.9998	0.9917	0.6303	0.5651	0.9998

The Kendall's tau rank correlation coefficient τ is used to measure the correlation one topology-based ranking list and the real spreading capability R. Let (x_i, y_i) and (x_j, y_j) be a randomly selected pair of joint observations from ranking lists X and Y, respectively. If one has $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$, the observations (x_i, y_i) and (x_j, y_j) are said to be concordant. If $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$, they are said to be discordant. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant [23]. τ is defined as

$$\tau = \frac{N_1 - N_2}{0.5N(N-1)}, \tag{7}$$

where N_1 and N_2 are the number of concordant pairs and discordant pairs, respectively.

When we employ SIR model to check the spreading influence of nodes, the infection probability β should not be too small or too large, if β is too small, the epidemic cannot successfully spread, so the spreading capability of each node cannot be measured. On the contrary, if β is too large, the epidemic can easily outbreak on almost whole network, as a result, the spreading capability of each node cannot be distinguished too. Thus, in this work, we first obtain the epidemic threshold β_{th} for each network, previous work has proven that

$\beta_{th} \sim \langle k \rangle / \langle k^2 \rangle$ with $\langle k \rangle$ and $\langle k^2 \rangle$ be the average degree and the second order average degree [21], respectively. The value of β_{th} for different networks is given in table 2 too. Then, we choose the value of β to be slightly larger than the threshold β_{th} when computing τ for different indices. The results in table 4 manifests that our method outperforms the other methods in most cases.

Table 4
$\tau(.)$ is correlation of corresponding methods for given β.

Network	β	τ_{DC}	τ_{MDD}	τ_G	τ_{G_+}	$\tau_{C_{nc+}}$	τ_{ks}	τ_{BC}	τ_{CC}
Facebook	0.050	0.771	0.798	0.859	0.902	0.904	0.732	0.3686	0.727
Netsci	0.130	0.597	0.617	0.823	0.848	0.839	0.520	0.311	0.336
Email	0.070	0.767	0.786	0.882	0.926	0.924	0.778	0.621	0.816
TAP	0.065	0.724	0.744	0.866	0.894	0.867	0.690	0.271	0.525
Y2H	0.160	0.442	0.460	0.819	0.826	0.818	0.406	0.410	0.698
Blogs	0.075	0.524	0.531	0.821	0.751	0.782	0.481	0.389	0.570
Router	0.075	0.326	0.323	0.774	0.781	0.765	0.185	0.316	0.629
HEP	0.110	0.485	0.504	0.778	0.850	0.728	0.483	0.344	0.773

To further estimate how the infection probability β affects the effectiveness of different methods, the correlation value τ as a function of β for different methods is shown in Fig. 2. As described in Fig. 2, in most cases, G or G_+ index provides better performance than the other index when $\beta > \beta_{th}$ (the values of β_{th} for different networks are illustrated by the dot lines in Fig. 2). However, Fig. 2 clearly indicates that though the high time complexity of the betweenness centrality, it is not a good index to measure the influence of nodes in these networks. Meanwhile, the effect of MDD method on identifying the node's influence is almost the same as the degree centrality.

Besides the real networks, we also compare the performance of our method with other methods on the two typical artificial networks–Barabás-Albert (BA) networks [16] and the Watts-Strogatz (WS) small-world networks [24] with $N = 1000$. Starting from a connected network with m_0 nodes to construct a BA network, at each step, a new node is added to the network and is connected to m existing nodes according to the preferential attachment mechanism, where $m \leq m_0$ [16]. We set the number of nodes $m_0 = 10$ in this paper. The WS small-world model considers a ring nearest neighbor coupled network with N nodes. Each node symmetrically connects to its $2K$ nearest neighbors. Starting from it, a fraction p of edges in the network are rewired, by visiting all K clock-wise edges of each node and reconnecting them, with probability p, to a randomly chosen node [24]. During the rewiring process, self-connection and reconnection are forbidden.

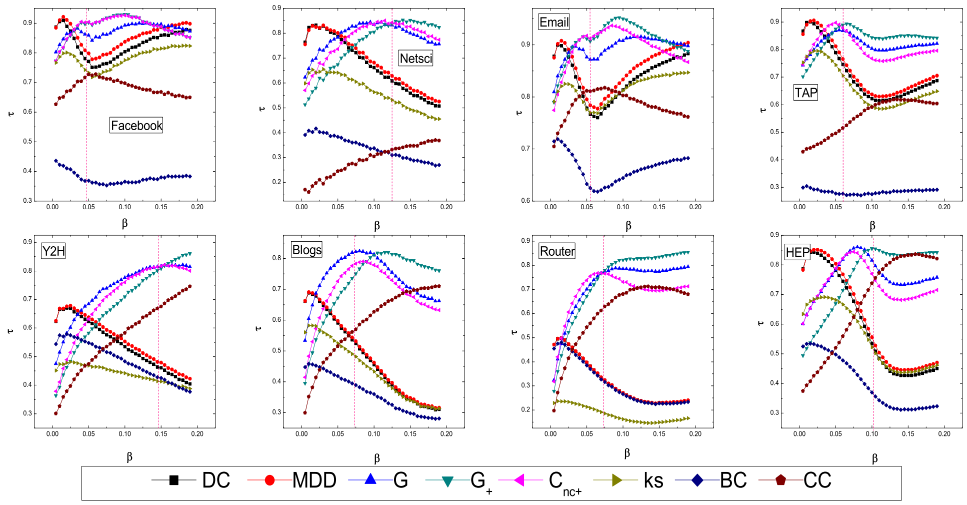

Fig. 2. (Color online) The value of τ obtained by comparing the ranking list generated by the SIR model and the ranking lists generated by the topology-based method on Facebook, Netsci, Email, TAP, Y2H, Blogs, Router and HEP. The dot lines correspond to the epidemic threshold.

For BA networks (see Fig. 3 (a) and (b)), one can see that the performances of G, G_+ and C_{nc+} indices are almost the same. The reason is that the three indices are all the improved methods of k-shell decomposition, however, all nodes in BA network are almost classified into the same shell when using the the k-shell decomposition (so we do not calculate the case of ks in Fig. 3). Moreover, the results show that the three indices are better than CC index and are much better than DC, BC and MDD indices. For WS networks (see Fig. 3 (c) and (d)), whose degree distribution shows Poisson distribution, i.e, their degrees are not so different. In this case, it is difficult for traditional DC index to distinguish the influence of nodes. However, as shown in Fig. 3(c) and (d), as $\beta > \beta_{th}$, the performances of G and G_+ methods are still better than the other method. In particular, for WS network, one can observe that the performances of G and G_+ indices are much better than the C_{nc+} index when $\beta > \beta_{th}$. The results in Fig. 3 suggest that our method can not only identify the influential nodes on real networks but also on artificial networks.

4 Conclusions and discussions

In summary, in this paper, we have proposed a gravity method to identify the influential spreaders in complex networks. In the model, each node's k-shell value is considered as its mass and the shortest path distance between two nodes is viewed as their distance. The idea of the gravity method comes from the well-known gravity formula, which is very dramatic and impressive.

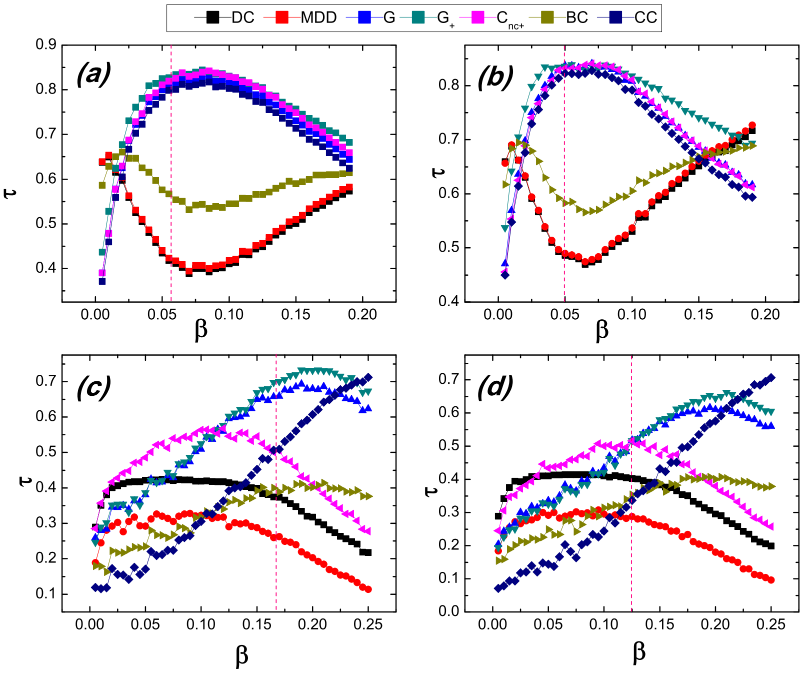

Fig. 3. (Color online) (a) BA: $m = 3$; (b) BA: $m = 4$;(c) WS: $K = 3$, $p = 0.05$; (d) WS: $K = 4$, $p = 0.05$. The pink dot lines correspond to the epidemic threshold.

What's more, the gravity model can reflect the facts that, on the one hand, the interaction influence between two nodes is proportional to their corresponding k-shell values; on the other hand, the influences of the neighbors decreases with their distance. Meanwhile, to lower the time complexity, we just considered the nodes interaction influences within three steps, i.e, their shortest path distance is less than or equal to 3. We employ our method on some real networks and artificial networks, by calculating the monotonicity index M, we found that our method can better distinguish the difference of node influence than other methods. Also, by computing Kendalls tau rank correlation coefficient τ, we have shown that, in most cases, our method has a better performance in evaluating the node's influence than other methods. Therefore, our method provides an effective way to identify the influential spreaders in social networks.

Some extensions may be made based on this method. For example, by defining the combination of node's degree and node's strength as the weighted degree of a node in weighted networks, Garas et al. have proposed a new k-shell decomposition method for weighted networks [25]. Therefore, once the new k-shell value for each node in weighted network is assigned, our method can be simply generalized to weighted networks [26]. Also, if we view the closeness centrality, degree centrality, betweenness centrality, and so forth as the mass of a node, then the gravity method may be further generalized.

Acknowledgments

This work is funded by the National Natural Science Foundation of China (Grant Nos. 61473001, 11331009), and partially supported by Graduate Student Academic Innovation Foundation of Anhui University (Grant Nos. yfc100015).

References

[1] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, H. A. Makse, Identification of influential spreaders in complex networks, Nature Physics 6 (11) (2010) 888–893.

[2] P. Basaras, D. Katsaros, L. Tassiulas, Detecting influential spreaders in complex, dynamic networks, Computer 46 (4) (2013) 0024–29.

[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, A. Rivero, Y. Moreno, Locating privileged spreaders on an online social network, Physical Review E 85 (6) (2012) 066123.

[5] D. Wei, X. Deng, X. Zhang, Y. Deng, S. Mahadevan, Identifying influential nodes in weighted networks based on evidence theory, Physica A 392 (10) (2013) 2564–2575.

[6] J.-G. Liu, Z.-M. Ren, Q. Guo, Ranking the spreading influence in complex networks, Physica A 392 (18) (2013) 4154–4159.

[7] D.-B. Chen, R. Xiao, A. Zeng, Y.-C. Zhang, Path diversity improves the identification of influential spreaders, EPL (Europhysics Letters) 104 (6) (2013) 68006.

[8] Z.-M. Ren, A. Zeng, D.-B. Chen, H. Liao, J.-G. Liu, Iterative resource allocation for ranking spreaders in complex networks, EPL (Europhysics Letters) 106 (4) (2014) 48005.

[9] 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.

[10] Y. Liu, M. Tang, T. Zhou, Y. Do, Core-like groups result in invalidation of identifying super-spreader by k-shell decomposition, Scientific Reports 5 (2015) 9602.

[11] X.-Y. Zhao, B. Huang, M. Tang, H.-F. Zhang, D.-B. Chen, Identifying effective multiple spreaders by coloring complex networks, EPL (Europhysics Letters) 108 (6) (2014) 68005.

[12] P. Bonacich, Factoring and weighting approaches to status scores and clique identification, Journal of Mathematical Sociology 2 (1) (1972) 113–120.

[13] L. C. Freeman, A set of measures of centrality based on betweenness, Sociometry (1977) 35–41.

[14] S. Maslov, K. Sneppen, Specificity and stability in topology of protein networks, Science 296 (5569) (2002) 910–913.

[15] G. Sabidussi, The centrality index of a graph, Psychometrika 31 (4) (1966) 581–603.

[16] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.

[17] A. Zeng, C.-J. Zhang, Ranking spreaders by decomposing complex networks, Physics Letters A 377 (14) (2013) 1031–1035.

[18] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Physica A 391 (4) (2012) 1777–1787.

[19] J.-H. Lin, Q. Guo, W.-Z. Dong, L.-Y. Tang, J.-G. Liu, Identifying the node spreading influence with largest k-core values, Physics Letters A 378 (45) (2014) 3279–3284.

[20] J. Bae, S. Kim, Identifying and ranking influential spreaders in complex networks by neighborhood coreness, Physica A 395 (2014) 549–559.

[21] Y. Moreno, R. Pastor-Satorras, A. Vespignani, Epidemic outbreaks in complex heterogeneous networks, The European Physical Journal B 26 (4) (2002) 521–529.

[22] N. Blagus, L. Šubelj, M. Bajec, Self-similar scaling of density in complex real-world networks, Physica A 391 (8) (2012) 2794–2802.

[23] W. R. Knight, A computer method for calculating kendall's tau with ungrouped data, Journal of the American Statistical Association 61 (314) (1966) 436–439.

[24] D. J. Watts, S. H. Strogatz, Collective dynamics of small-worldnetworks, Nature 393 (6684) (1998) 440–442.

[25] A. Garas, F. Schweitzer, S. Havlin, A k-shell decomposition method for weighted networks, New Journal of Physics 14 (8) (2012) 083030.

[26] Q. Li, T. Zhou, L. Lü, D. Chen, Identifying influential spreaders by weighted leaderrank, Physica A 404 (2014) 47–55.