

# Accurate ranking of influential spreaders in networks based on dynamically asymmetric link weights

Ying Liu,<sup>1,2,\*</sup> Ming Tang,<sup>1,3,†</sup> Younghae Do,<sup>4</sup> and Pak Ming Hui<sup>5</sup>

<sup>1</sup>*Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China*

<sup>2</sup>*School of Computer Science, Southwest Petroleum University, Chengdu 610500, China*

<sup>3</sup>*School of Information Science Technology, East China Normal University, Shanghai 200241, China*

<sup>4</sup>*Department of Mathematics, Kyungpook National University, Daegu 702-701, South Korea*

<sup>5</sup>*Department of Physics, Chinese University of Hong Kong, Shatin, Hong Kong SAR, China*

(Received 26 April 2017; revised manuscript received 3 August 2017; published 31 August 2017)

We propose an efficient and accurate measure for ranking spreaders and identifying the influential ones in spreading processes in networks. While the edges determine the connections among the nodes, their specific role in spreading should be considered explicitly. An edge connecting nodes  $i$  and  $j$  may differ in its importance for spreading from  $i$  to  $j$  and from  $j$  to  $i$ . The key issue is whether node  $j$ , after infected by  $i$  through the edge, would reach out to other nodes that  $i$  itself could not reach directly. It becomes necessary to invoke two unequal weights  $w_{ij}$  and  $w_{ji}$  characterizing the importance of an edge according to the neighborhoods of nodes  $i$  and  $j$ . The total asymmetric directional weights originating from a node leads to a novel measure  $s_i$ , which quantifies the impact of the node in spreading processes. An  $s$ -shell decomposition scheme further assigns an  $s$ -shell index or weighted coreness to the nodes. The effectiveness and accuracy of rankings based on  $s_i$  and the weighted coreness are demonstrated by applying them to nine real-world networks. Results show that they generally outperform rankings based on the nodes' degree and  $k$ -shell index while maintaining a low computational complexity. Our work represents a crucial step towards understanding and controlling the spread of diseases, rumors, information, trends, and innovations in networks.

DOI: [10.1103/PhysRevE.96.022323](https://doi.org/10.1103/PhysRevE.96.022323)

## I. INTRODUCTION

The structural properties of complex networks and the intricate interplay between the structure and spreading dynamics lead to highly diversified spreading capabilities among individual nodes. From the perspective of the severity of a spreading process, the most influential spreaders are those resulting in a much larger final infected proportion of the whole system when the spread of a disease or a piece of information originates from them than from other nodes. Centrality measures such as the degree [1], betweenness [2], closeness [3], eigenvector centrality [4], and  $k$ -shell coreness [5] have been used to identify the most influential spreaders. The degree is the simplest measure. In social networks, for example, an individual with a large degree has more direct contact with other people and is thus likely to be more influential than one with a small degree for transmitting disease or information. Subsequent research indicates that the core nodes as identified by the  $k$ -shell decomposition are the most influential spreaders [6]. Algorithms based on other centrality measures have been proposed to improve the accuracy of identifying influential spreaders [7–10]. They include the neighborhood coreness [11], improved eigenvector centrality [12,13],  $H$  index [14,15], and nonbacktracking centrality [16].

Methods other than centrality-based algorithms have also been proposed for predicting how influential a node can be in a spread. For example, by counting the number of possible infection paths of various lengths, the final infection range can be estimated for a spread originated from any node [17].

The degree distribution of clusters of infected nodes after certain transmission events leads to a node property called the expected force, which can be applied to predict the spreading influence of all nodes under different epidemiological models [18]. The dynamic-sensitive centrality is able to locate influential nodes from both topological features and the dynamical parameters, such as the infection and recovery rates in a susceptible-infected-recovered (SIR) spreading model [19]. In the  $k$ -truss decomposition, which is a triangle-based extension of  $k$ -shell decomposition, the maximal  $k$ -truss subgraph contains the most influential spreaders [20].

In most studies on identifying influential spreaders so far, the networks are taken to be unweighted and undirected. Each edge is treated as being equivalent in its function, as in the centrality and ranking methods [10]. However, edges in a network could be quite different [21]. In weighted networks, the weight of an edge reflects the strength of the interaction between the connected nodes, as in situations concerning the number of communications, size of trade, intimacy of friendship, frequency of cooperation, etc. [22–24]. In addition, edges may not be equally important in keeping the network robust [25]. An example is the small influence on network robustness in food web networks when redundant links are removed [26]. In terms of network functionality, differences among edges are also observed. For example, removing redundant links has no effect on network synchronization [27], but closing specific routes in air transportation networks can minimize the spreading of a disease [28]. To quantify the weight of an edge, a class of measures relying on its importance in the network structure has been defined [29]. For example, the edge betweenness counts the number of shortest paths between any two nodes that go through the edge and it can be regarded as the weight of an edge [30]. Immunizing edges

\*shinningliu@163.com

†tangminghan007@gmail.com

of high betweenness was found to be effective in suppressing epidemics [31], but deleting such edges in scale-free networks would enhance the transmission efficiency dramatically [32]. In a global air transportation network, the strength of an edge that reflects the volume of passengers traveling between two airports was found to correlate positively with the product of the degree of the connected nodes. Thus, a measure  $w_{ij} = (k_i k_j)^\theta$  [33] was introduced as the weight of an edge. This measure has been adopted in many works for distinguishing the importance among edges in unweighted networks [34–36].

In the present work we propose a measure to quantify the importance of an edge when spreading processes are concerned. As spreading is necessarily directional, e.g., only an infected node would spread a disease to a neighboring susceptible node but not the other way round, our measure stresses the importance of an edge in the spreading dynamics in the vicinity of the two nodes connected by the edge and it has the general property of  $w_{ij} \neq w_{ji}$ . The sum of asymmetric weights of links originated from a node defines a measure of the strength  $s_i$  of a node  $i$ , which is shown to be an efficient quantity for identifying influential spreaders with a low computational complexity. Based on the node strength, an  $s$ -shell decomposition scheme is proposed for assigning an  $s$ -shell index to every node, which provides a more accurate ranking of the nodes in their influence in spreading processes.

The paper is organized as follows. In Sec. II the degree centrality,  $k$ -shell index, the spreading model used in the study, and the methods of evaluating the performance of measures for identifying influential spreaders are introduced for completeness. In Sec. III we propose a measure that focuses on the importance of an edge in the dynamics of a spreading process. The measure is then applied to define a node strength for every node. An  $s$ -shell decomposition method that emphasizes the importance of a node in the spreading dynamics is proposed. In Sec. IV we apply the node strength and  $s$ -shell index to rank and identify influential spreaders in nine real-world networks and demonstrate their effectiveness. A summary is given in Sec. V.

## II. CENTRALITIES, SPREADING MODEL, AND EVALUATION METHODS

We review briefly the degree centrality and the  $k$ -shell index for completeness. They are efficient measures for identifying influential spreaders [37–39]. We will compare the performance of our presently defined node strength and  $s$ -shell index with these methods. The SIR model is adopted to simulate the spreading dynamics on networks. To quantify the performance of our measures in predicting the influence of the nodes and identifying influential spreaders, Kendall's  $\tau$  correlation and the imprecision function are introduced.

### A. Degree and $k$ -shell centrality

In a graph  $G = (V, E)$ , where  $V$  is the set of nodes and  $E$  is the set of edges, the degree  $k_i$  of a node  $i$  is the number of links it carries. It is given by  $k_i = \sum_j a_{ij}$ , where  $a_{ij}$  is an element of the adjacent matrix, with  $a_{ij} = 1$  if there is a link between nodes  $i$  and  $j$  and  $a_{ij} = 0$  otherwise. The  $k$ -shell decomposition method decomposes the network into

hierarchical shells in a progressive process. Initially, nodes with degree  $k = 1$  are removed from the network together with their links. After the process, nodes with only one link left may appear. These nodes and their links are then removed and the process is repeated until there are no nodes left in the network with only one link. The removed nodes and links form the 1-shell, and these nodes are assigned an index  $k_S = 1$ . Next, nodes with degree  $k \leq 2$  are removed in a similar way and the set of removed nodes are assigned an index  $k_S = 2$ . This pruning process is continued until all nodes are removed and assigned a  $k_S$  index. This index is called the  $k$ -shell index or coreness of a node. It represents the core position of a node in the network. Nodes with a large  $k_S$  are considered to be at the core of the network, while nodes with a small  $k_S$  form the peripheral part of the network.

Nodes with a large degree and large coreness are considered the most influential spreaders in networks. These measures have a low computational complexity of  $O(E)$  and  $O(N + E)$ , respectively, where  $N$  and  $E$  are the number of nodes and edges in the network, respectively. By using the bin-sort structure, the complexity of the  $k$ -shell decomposition can even reduce to  $O(E)$  [40].

### B. The SIR model

The SIR model is chosen to simulate spreading on complex networks. In the model, the nodes have three possible states: susceptible, infected, and recovered. At each time step, the infected nodes infect their susceptible neighbors with a probability  $\lambda$  and then recover with a probability  $\beta$ . To quantify the influence of each node on spreading, we let one node, say, node  $i$ , be infected and all the other nodes be susceptible initially. The SIR dynamics proceeds from the seed-infected node to other nodes until there is no infected node in the network. The recovered nodes at the end are those once infected and the fraction of recovered nodes gives the final infected range of the initial seed. For an initially infected node  $i$ , the spreading dynamics is repeated for 100 times. The average infected range  $M_i$  of node  $i$  is recorded and taken to reflect the influence or the spreading efficiency of the node  $i$ . This quantity can be obtained for any node  $i$  in the network and used as a measure to rank the nodes on their importance in the spreading dynamics. This dynamics-based list is taken to be the exact ranking that gauges the accuracy of other topology-based measures.

While the final infected ranges for the nodes vary with the parameters  $\lambda$  and  $\beta$  in the SIR model, the relative ranking of spreading efficiency of the nodes remains unchanged in a wide range of infection probabilities [38]. Thus, we take the recovered probability to be  $\beta = 1$  for simplicity. The infection probability  $\lambda$  should be chosen more carefully. On the one hand, it should be above the epidemic threshold to ensure that the disease can spread to a large part of the network [41]. On the other hand, too large an infection probability gives spreading efficiencies of the nodes that are too close to each other to clearly distinguish their relative importance. In the results that follow, we choose an infection probability  $\lambda$  that is above the epidemic threshold and makes the final infected range amount to 1%–20% of the system for most nodes as the spreading origins [6].

### C. Kendall's $\tau$ correlation and imprecision function

Two figures of merit are used to quantify the performance of different topology-based measures for predicting the spreading efficiency of the nodes. Kendall's  $\tau$  correlation coefficient measures the ranking consistency of two lists that rank the same set of objects. By referring to the number of concordant ranking pairs and the number of discordant ranking pairs in two ranking lists of  $N$  objects, the correlation coefficient is evaluated as

$$\tau = \frac{\sum_{i < j} \text{sgn}[(x_i - x_j)(y_i - y_j)]}{\frac{1}{2}N(N-1)}, \quad (1)$$

where  $\text{sgn}(x)$  is the sign function, which returns 1 if  $x > 0$ ,  $-1$  if  $x < 0$ , and 0 if  $x = 0$ , and the summation is over all distinguished pairs  $i$  and  $j$ . Here  $x_i$  is the rank of node  $i$  in ranking list 1, while  $y_i$  is the rank of node  $i$  in ranking list 2. In the present context, list 1 is a topology-based ranking and list 2 is the SIR dynamics-based ranking. If  $(x_i - x_j)$  has the same sign as  $(y_i - y_j)$ , the two lists give the same relative ranking of node  $i$  and node  $j$ . Therefore, a large  $\tau$  implies a more concordant relation between two methods of ranking the nodes.

For spreading processes, it is also important to quantify the accuracy in pinpointing the most influential spreaders. For a topology-based measure  $\theta$ , e.g., some kind of node centrality, let  $M_\theta(p)$  be the average spreading efficiency of the  $pN$  nodes carrying the highest measure  $\theta$ . Similarly, let  $M_{\text{eff}}(p)$  be the average spreading efficiency of  $pN$  nodes carrying the highest actual spreading efficiency according to the SIR dynamics. The imprecision function [6]

$$\varepsilon_\theta(p) = 1 - \frac{M_\theta(p)}{M_{\text{eff}}(p)} \quad (2)$$

quantifies how close to the actual spreading is the average spreading of the  $pN$  nodes based on centrality measure. A smaller  $\varepsilon_\theta$  represents a higher accuracy of  $\theta$  in identifying the most influential spreaders.

### III. DYNAMICAL IMPORTANCE OF EDGES AND WEIGHTED NODE CENTRALITY

The dynamical importance of an edge is analyzed by focusing on the spreading dynamics and the edge's local structure. This leads to the necessity of assigning bidirectional and asymmetric weights to an edge. A node strength  $s$  can then be defined to quantify the impact of a node on spreading. An  $s$ -shell decomposition method is proposed to be a reliable way of ranking the nodes for spreading processes.

#### A. Dynamical importance of edges

Figure 1 shows part of a network. When a disease originates from node  $i$  and spreads along the edge  $e_{ij}$ , node  $j$  will be infected first. Once node  $j$  is infected, it could spread to other parts of the network through node  $j$ 's outgoing edges, which are edges that connect node  $j$  to nodes that are not in  $i$ 's neighborhood. The number of outgoing edges from  $j$  is denoted by  $k_j^{\text{out}}$  and it is 3 in the example of Fig. 1. Note that  $k_j^{\text{out}}$  should depend on the node  $i$ , as  $j$  must be a neighboring node of  $i$ . In contrast, the edge  $e_{ik}$  has zero outgoing edges after

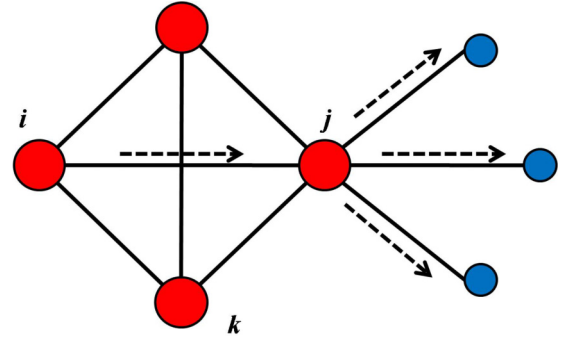


FIG. 1. Local structure of a network emphasizing the role of the link  $e_{ij}$  in spreading a disease from node  $i$  to node  $j$  and then to reach out to nodes that node  $i$  itself cannot reach. The same link  $e_{ji}$ , however, plays a different role as it does not help spread the disease to nodes beyond the reach of node  $j$  after it infects node  $i$ . The asymmetry requires the assignment of directional weights with  $w_{ij} \neq w_{ji}$ .

it is infected by node  $i$ . Therefore, the edge  $e_{ij}$  is expected to be more important in that it is more likely to lead to a larger infected area than confining the infection to node  $i$ 's neighborhood as  $e_{ik}$  does [18,39]. We are therefore motivated to introduce a measure to distinguish the different importance of edges in a spreading process, even though the links may be unweighted in the construction of the network.

For our purpose, we define a weight  $w_{ij}$  for an edge  $e_{ij}$  by

$$w_{ij} = 1 + (k_i k_j^{\text{out}})^a \quad (3)$$

to represent its importance in a spreading process from node  $i$  to node  $j$ . The first term stands for the basic effect of infecting the direct neighbor  $j$  of node  $i$  through the edge. We test the result of neglecting this element and find that the ranking accuracy is not as good as keeping it, implying that the basic effect of an edge on the direct neighbor should not be neglected. The factor  $k_j^{\text{out}}$  is included to reflect the potential impact of node  $i$  through infecting its neighbor  $j$ . The product  $k_i k_j^{\text{out}}$  include the degree of node  $i$  in the consideration. The idea can be illustrated by considering a leaf node (node of degree 1) connected to a hub (a node of large degree), the number of outgoing links is very large for this leaf node. However, its impact is not necessarily high because only when it and its neighboring hub are infected could the infection spread to the other part of the network. The parameter  $a$  serves to tune the contribution of  $k_i k_j^{\text{out}}$  to the importance of edge  $e_{ij}$ .

The presence of  $(k_i k_j^{\text{out}})^a$  emphasizes the *asymmetric* importance of an edge. The weight  $w_{ij}$  is different from  $w_{ji}$  for the same link connecting node  $i$  and node  $j$ . From Eq. (3),  $w_{ji} = 1 + (k_j k_i^{\text{out}})^a$ , which measures the importance of the edge  $e_{ji}$  when a spread goes from node  $j$  to node  $i$  along the link  $e_{ji}$  and then moves on to other parts of the network. Note that  $w_{ij} \neq w_{ji}$  generally as they are defined by considering the neighborhoods of the neighbors of node  $i$  and node  $j$ , respectively. Given a network,  $w_{ij}$  and  $w_{ji}$  can be evaluated entirely based on the network topology and they label every edge to better reflect the bidirectional and yet asymmetric contributions of the edge in spreading processes. There are many other ways to define the edge weight from the

network topology [30,33]. For example, the edge multiplicity is defined as the number of triangles an edge participates in Ref. [42]. This method defines the weight symmetrically in both directions and focuses on the influence of clustering, i.e., counting the number of nodes that are both first and second neighbors. This idea is partly in contrast to our method of weakening the effect of nodes that are both first and second neighbors. While the edge multiplicity focuses on the effect of spreading in the first neighborhood, our way of defining the link impact considers further steps of spreading out of the first neighborhood, thus resulting in asymmetrically defined weights.

### B. Node strength and $s$ -shell decomposition

It will be advantageous to introduce a node-level quantity analogous to the degree to quantify the importance of a node in spreading dynamics. This will put the computational complexity at the same level as those based on the degree and  $k$ -shell decomposition. Motivated by the idea of a weighted degree [22] that the strength of a node in a weighted graph is the sum of the weights of its edges, we define the strength  $s_i$  of a node  $i$  by

$$s_i = \sum_{j \in \Gamma_i} w_{ij}, \quad (4)$$

where the summation is over the nodes  $j$  belonging to the neighborhood  $\Gamma_i$  of node  $i$ . Invoking  $w_{ij}$  in the definition of  $s_i$  makes it a better measure in quantifying a node's importance in spreading dynamics.

We propose an  $s$ -shell decomposition method as an extension of the  $k$ -shell decomposition. The algorithm is as follows. With the strengths  $s_i$  evaluated for all nodes, the algorithm starts with removing the nodes with the smallest strength  $s_m$  and the links associated with the nodes. Let node  $i$  be removed. The strength of its neighboring node  $j$  is then updated to  $s_j - w_{ji}$  as the edge  $e_{ij}$  is removed. Continue to remove nodes with strength less than or equal to  $s_m$  until no such nodes remains. The deleted nodes are assigned an  $s$ -shell index of  $s_s = 1$ , where the variable emphasizes that the decomposition is based on the nodes' strength and the subscript represents a shell. Next find the smallest strength in the remaining network and remove nodes in the same way as before; the nodes so removed are assigned the index  $s_s = 2$ . This pruning process is continued until all nodes are removed and assigned an  $s_s$  index. This  $s$ -shell index of a node reflects the order (hierarchy) of a node being removed in the pruning process and can be regarded the weighted coreness of the node, emphasizing its importance in the spreading dynamics.

## IV. PERFORMANCE IN IDENTIFYING INFLUENTIAL SPREADERS IN REAL-WORLD NETWORKS

To examine the effectiveness of using the node strength and weighted coreness in identifying influential spreaders, we apply the measures to nine real-world networks as listed in Table I. The networks studied are CA-Hep (giant connected component of collaboration network of arXivs in high-energy physics theory) [43], Astro (collaboration network of astrophysics scientists) [44], Emailcontact (email contacts at

TABLE I. Properties of the real-world networks studied in this work. Structural properties of the number of nodes  $N$ , number of edges  $E$ , average degree  $\langle k \rangle$ , degree assortativity  $r$ , clustering coefficient  $C$ , epidemic threshold  $\lambda_c$ , infection probability  $\lambda$  used in the SIR dynamics, and the optimal value of  $a$  as given by Kendall's  $\tau$  correlation coefficient  $a_{\text{opt}}$ .

Network	$N$	$E$	$\langle k \rangle$	$r$	$C$	$\lambda_c$	$\lambda$	$a_{\text{opt}}$
CA-Hep	8638	24806	5.7	0.239	0.482	0.08	0.12	1.0
Astro	14845	119652	16.1	0.228	0.670	0.02	0.05	0.9
Emailcontact	12625	20362	3.2	-0.387	0.109	0.01	0.10	0.3
PGP	10680	24340	4.6	0.240	0.266	0.06	0.19	1.0
Blog	3982	6803	3.4	-0.133	0.284	0.08	0.27	0.9
AS	22963	48436	4.2	-0.198	0.230	0.004	0.13	0.2
Router	5022	6258	2.5	-0.138	0.012	0.08	0.27	0.7
Hamster	2000	16097	16.1	0.023	0.540	0.02	0.04	0.8
Netsci	379	914	4.8	-0.082	0.741	0.14	0.30	0.8

Computer Science Department of University College London) [6], PGP (an encrypted communication network) [45], Blog [the communication relationships between owners of blogs on the MSN (Windows Live) Spaces website] [46], AS (Internet at the autonomous system level) [47], Router (the router level topology of the Internet, collected by the Rocketfuel Project) [48], Hamster (friendships and family links between users of the website hamsterster.com) [49], and (9) Netsci (collaboration network of network scientists) [50]. Our measures are found to outperform predictions based on the degree centrality and  $k$ -shell decomposition, as we now show.

### A. Performance of node strength

From the structure of each network, every node carries a degree  $k_i$  and a node strength  $s_i$ . Using the SIR dynamics, the spreading efficiency  $M_i$  of each node can be obtained by simulations. Figure 2 compares the correlations between the spreading efficiency with the node strength and with the degree in nine real-world networks. Here we set  $a = 0.5$  in Eq. (3) in determining  $w_{ij}$  for the edges. The sensitivity to the parameter  $a$  will be discussed later. The strength and the degree are both positively correlated with the spreading efficiency. The merit of using the strength over the degree as a measure is that its value covers a much wider range and it can distinguish the spreading efficiency more specifically. This advantage is built into the definition of the node strength as it captures the key elements in the spreading dynamics.

The node strength provides a ranking of the nodes. This list can be compared with the list based on the actual spreading efficiency by calculating Kendall's  $\tau$  correlation coefficient. We calculate the ranking correlation of nodes' spreading efficiency and their strength for different values of  $a$  and obtained  $\tau(a)$ , as shown in Fig. 3 (squares). For  $a = 0$  [see Eq. (4)],  $s_i$  reduces to the degree  $k_i$  and thus  $\tau(a = 0)$  measures the correlation between the rankings based on the degree and the spreading efficiency. Note that  $\tau$  is significantly enhanced for  $a > 0$ , implying that the node strength, which includes the bidirectional and asymmetric weights of the edges, ranks the nodes more accurately. Results in Fig. 3 further show that there exists an optimal value of  $a$  for each network at which  $\tau$  is

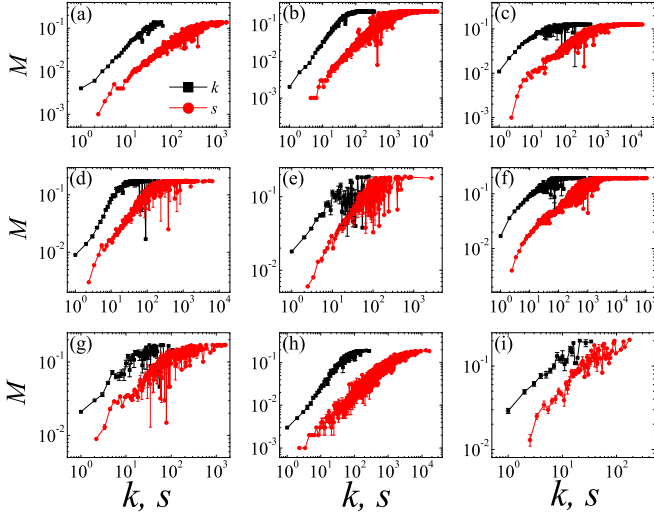


FIG. 2. Mean spreading efficiency  $M$  of nodes as classified by their degree  $k$  (black squares) or their node strength  $s$  (red circles) in nine real-world networks. As the node strengths are real numbers instead of integers, data are grouped in intervals of size unity. The corresponding average spreading efficiency and node strength in each interval are displayed, starting from the minimal strength. The real-world networks are (a) CA-Hep, (b) Astro, (c) Emailcontact, (d) PGP, (e) Blog, (f) AS, (g) Router, (h) Hamster, and (i) Netsci.

a maximum. The optimal value of each network is given in Table I, together with the other network properties. Figure 3 also shows the  $\tau(a)$  obtained by ranking the nodes according to the  $s$ -shell index. The results will be discussed later.

Figure 4 shows the imprecision function of the ranking based on the node strength, together with the results based on the degree and  $k$ -shell index for comparison. Recall that a lower imprecision implies a higher accuracy in identifying

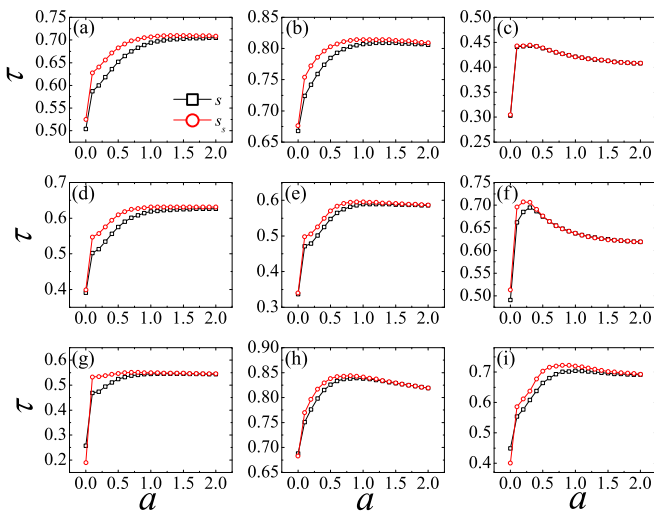


FIG. 3. Kendall's  $\tau$  correlation coefficients evaluated between the actual spreading efficiency of the nodes and the ranking based on node strength (black square) and based on the  $s$ -shell index (red circle) for different values of the parameter  $a$  in Eq. (3). The real-world networks are (a) CA-Hep, (b) Astro, (c) Emailcontact, (d) PGP, (e) Blog, (f) AS, (g) Router, (h) Hamster, and (i) Netsci.

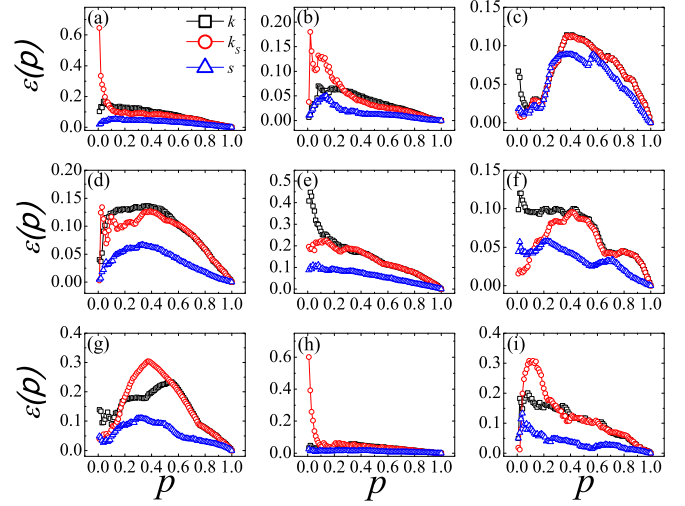


FIG. 4. Imprecision of rankings based on the degree ( $k$ , black squares),  $k$ -shell index ( $k_s$ , red circles), and node strength ( $s$ , blue triangles) evaluated at the optimal value of  $a$  as a function of  $p$  for nine real-world networks. The node strength provides a better measure for identifying influential spreaders. The real-world networks are (a) CA-Hep, (b) Astro, (c) Emailcontact, (d) PGP, (e) Blog, (f) AS, (g) Router, (h) Hamster, and (i) Netsci.

the influential spreaders. The node strength (triangles) give an imprecision that is less than 0.1 for all  $p$  in nearly all cases. Only in the network Netsci is the imprecision slightly larger than 0.1 for a few values of  $p$ . The results show that the node strength outperforms the degree (squares) in almost all networks. Only in the network Hamster do the imprecisions based on node strength and on the degree become comparable, but they are both small. The node strength is therefore a better index for pinpointing the influential spreaders than the degree. More noticeable is that the node strength performs even better than the  $k$ -shell index in most cases, except at some small values of  $p$  in the AS and Netsci networks. The  $k$ -shell index is regarded as an efficient measure for identifying influential spreaders and it is widely used in ranking algorithms. However, the assignment of a  $k$ -shell index requires a complete network structure rather than indices relying solely on the local structure such as the degree or the node strength. The node strength introduced here provides not only a more accurate measure, but also a computationally efficient method in handling large-scale networks.

## B. Performance of weighted coreness

The  $k$ -shell index works better than the degree in identifying influential spreaders [39]. Here we investigate how the  $s$ -shell index  $s_s$  or weighted coreness works in comparison to the other measures. The results of Kendall's  $\tau$  correlation of  $s_s$  ranking in Fig. 3 suggest that it is a better measure than using the node strength in eight systems out of nine. In the Emailcontact network,  $s_s$  and  $s$  rankings work equally well. In fact, the  $s_s$  and  $s$  rankings approach the same value of  $\tau$  as  $a$  increases. Given that the optimal values of  $a$  in the networks are less than or equal to 1, the weighted coreness gives a better ranking. Note that the  $a = 0$  case gives the value of  $\tau$  corresponding to

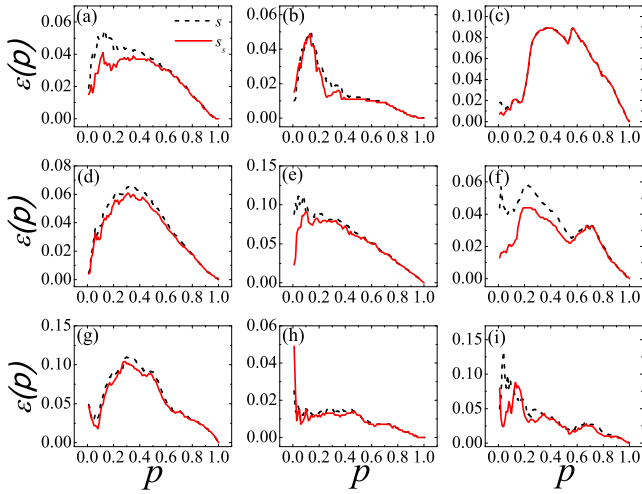


FIG. 5. Imprecision of rankings based on the node strength ( $s$ , black dash) and weighted coreness ( $s_s$ , red solid) obtained by  $s$ -shell decomposition as a function of  $p$ . The weighted coreness  $s_s$  provides a further improvement over  $s$  in identifying influential spreaders. The real-world networks are (a) CA-Hep, (b) Astro, (c) Emailcontact, (d) PGP, (e) Blog, (f) AS, (g) Router, (h) Hamster, and (i) Netsci.

the  $k$ -shell index  $k_s$ . Using  $s_s$  to rank the nodes always gives a higher  $\tau$  than the  $a = 0$  value, implying that the  $s$ -shell index is also a better measure than the  $k$ -shell index.

The imprecision functions of rankings using  $s_s$  and  $s$  are compared in Fig. 5. Their performances are comparable and they both work better than measures based on the degree alone (see Fig. 4). Looking closer, the lower imprecision of  $s_s$  ranking in six (CA-Hep, PGP, Blog, AS, Router, and Netsci) out of nine cases suggests that the  $s$ -shell decomposition method is more accurate in identifying influential spreaders in real-world networks. Even in the networks of Astro, Emailcontact, and Hamster in which  $s_s$  and  $s$  work almost equally well, the imprecision of  $s_s$  is slightly lower than or equal to that of  $s$ . Only in the Hamster network does  $s$  work slightly better than  $s_s$  at  $p = 0.01$ ; even so the imprecision functions are small (under 0.05) on the absolute scale.

We analyze the reason why the proposed node strength and  $s$ -shell index outperform the degree and  $k$ -shell index. First, the method proposed herein can identify weak nodes, which have a low degree but are influential spreaders [51]. For example, a node  $i$  of degree 2 has two neighbors with large outgoing links. The edges connecting  $i$  and its neighbors will be assigned a large link weight, thus resulting in a large strength, as well as a large  $s$ -shell index. If evaluated by degree and the  $k$ -shell index, node  $i$  will be ranked low as a less important node. Second, the methods proposed herein can identify nodes that are in fact not influential spreaders but are ranked high by degree and  $k$ -shell index, which form a local clique structure [39]. For example, in the real-world network CA-Hep, there exists a cliquelike local structure composed of 32 mutually connected nodes with very few outgoing links. In our methods, the edges of these nodes will be assigned low weights, due to the very few or even no outgoing links, thus having a low strength and  $s$ -shell index. If

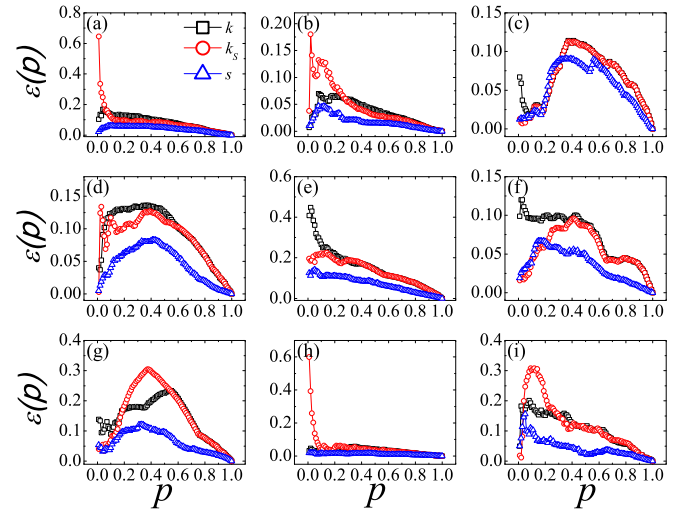


FIG. 6. Imprecision of rankings based on the degree ( $k$ , black squares),  $k$ -shell index ( $k_s$ , red circles), and node strength ( $s$ , blue triangles) evaluated by  $a = 0.5$  for nine real-world networks as a function of  $p$ . The real-world networks are (a) CA-Hep, (b) Astro, (c) Emailcontact, (d) PGP, (e) Blog, (f) AS, (g) Router, (h) Hamster, and (i) Netsci.

evaluated by the  $k$ -shell index, these nodes have the maximal  $k$ -shell index of 31 in the network.

Though we use the SIR spreading dynamics to obtain the spreading influence of each node, which can be interpreted as a percolation process, previous researches have pointed out the correlation between the existence of a large component and the ratio of the average number of second and first neighbors [52,53]. In Ref. [53] the authors indicated that clustering strongly affects the percolation properties and size of a giant component. By avoiding the edges that are responsible for clustering, the percolation properties of clustered random networks can be computed correctly. This idea is very similar to our way of defining the dynamical importance of edges, which synthesizes the first and second neighbors but neglects the nodes that are both first and second neighbors. This guarantees that the contribution of triangles is not overcounted. How to use the percolation theory to distinguish the influence of each node requires further exploration [54,55]. For example, researchers used the optimal percolation and message-passing method to predict the spreading influence of nodes and the proposed index counts the degree of nodes that is at a certain distance from the central node [51].

### C. Robustness of proposed weighted centrality

So far, we have used the optimal value of  $a$  to evaluate  $w_{ij}$  and  $s_i$  and compared our results with other measures. However, the optimal value is not often known precisely in real applications. It will be useful to examine the performance of the node strength  $s_i$  for some arbitrarily chosen value of  $a$ . Let us set  $a = 1/2$  so that the term  $(k_i k_j^{\text{out}})^a$  in  $w_{ij}$  represents a geometric mean. The comparison in Fig. 6 of the imprecision function shows that  $s_i$  ranking gives a lower imprecision than the degree and  $k$ -shell index. An interesting point is that the imprecision of node strength evaluated at  $a = 1/2$  is even

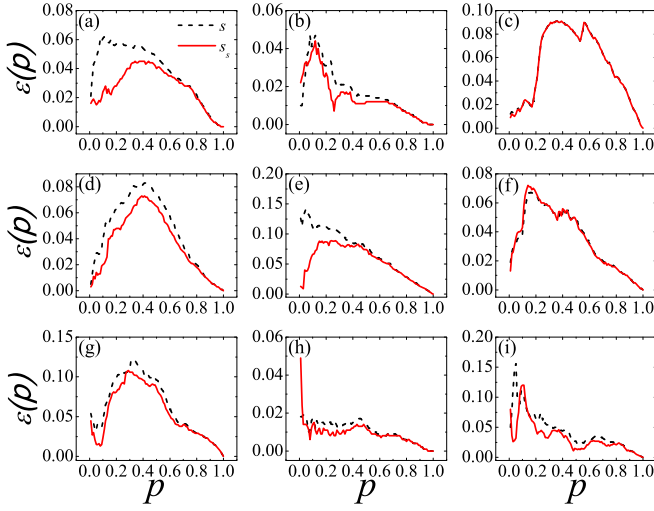


FIG. 7. Imprecision of rankings based on the node strength ( $s$ , black dashed line) and weighted coreness ( $s_s$ , red solid line) obtained by  $s$ -shell decomposition for  $a = 1/2$  as a function of  $p$ . The real-world networks are (a) CA-Hep, (b) Astro, (c) Emailcontact, (d) PGP, (e) Blog, (f) AS, (g) Router, (h) Hamster, and (i) Netsci.

lower than that evaluated at the optimal value of  $a$  in the AS network for  $p < 0.1$ . The result indicates that although the best performing overall ranking correlation coefficient occurs at some optimal  $a$ , the same value does not necessarily give the best identification of the most influential spreaders.

Figure 7 compares the effectiveness of the node strength and the  $s$ -shell decomposition for  $a = 1/2$ . Again, the  $s$ -shell index works better in most cases. In fact, the results resemble those in Fig. 5 when the optimal  $a$  is used. These results further support the assertion that the node strength and the corresponding  $s$ -shell index are better measures for spreading processes than methods based on the degree. Between them, the  $s$ -shell index performs slightly better, but evaluating the index requires more computing effort than the node strength alone.

## V. CONCLUSION

The roles of nodes and edges in deciding the structural properties of a network should be carefully distinguished from their roles in determining the extent of spreading processes. Although an edge between nodes  $i$  and  $j$  certainly helps spread a disease, its role may be different when the infection goes from  $i$  to  $j$  than in the other direction. This is because what matters is whether the node  $j$ , after infected by  $i$ , would reach out to other nodes that node  $i$  itself could not reach. If so, the link carries a greater importance for infection from  $i$  to  $j$ , which is quantified by a higher weight  $w_{ij}$  for the link. It is therefore necessary to invoke asymmetric and bidirectional weights with  $w_{ij} \neq w_{ji}$  for a link so as to capture the dynamics in spreading processes. Here we introduced a form of  $w_{ij}$  [see Eq. (3)] and showed that it facilitates accurate ranking in the node's importance. Pictorially, the network is better described by the nodes connected by links with asymmetric weights in different directions when spreading dynamics is concerned.

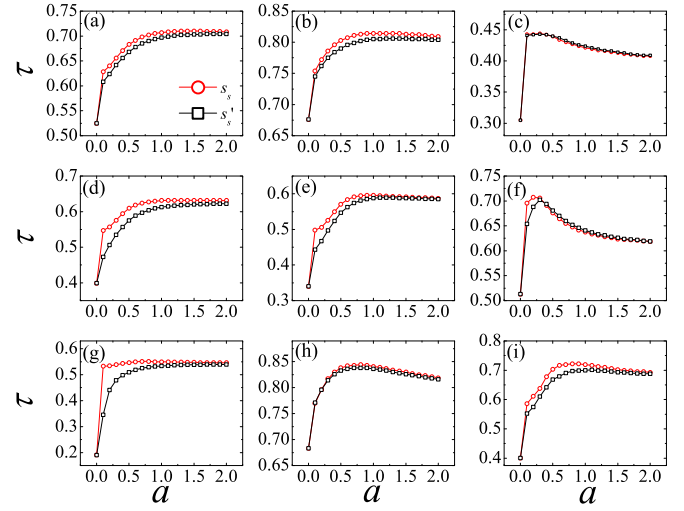


FIG. 8. Kendall's  $\tau$  correlation coefficients evaluated between the actual spreading efficiency of the nodes and the rankings based on weighted coreness  $s_s$  (red circles) as obtained by assigning asymmetric weighting  $w_{ij}$  of Eq. (3) to the links and based on weighted coreness  $s'_s$  (black square) by assigning symmetric weighting  $w'_{ij}$  of Eq. (5) to the links. The necessity of invoking asymmetric weights is demonstrated by the higher accuracy in the ranking based on  $s_s$  rather than on  $s'_s$ . The real-world networks are (a) CA-Hep, (b) Astro, (c) Emailcontact, (d) PGP, (e) Blog, (f) AS, (g) Router, (h) Hamster, and (i) Netsci.

To establish the effectiveness of our method, the weights of the links were used to construct a node strength  $s$  that predicts the importance of a node in spreading processes. An  $s$ -shell decomposition scheme based on the node strength was then introduced. The  $s$ -shell index  $s_s$  of the nodes provides another way of ranking them. Applying  $s$  and  $s_s$  rankings to nine real-world networks, it was found that our measures generally outperform the standard rankings based on the degree of the nodes and the  $k$ -shell decomposition method. Superiority is shown in both the overall performance of the ranking as indicated by Kendall's  $\tau$  correction coefficient and in identifying the influential spreaders as indicated by the imprecision.

The success of our measure relies on the asymmetry in the weights contained in  $w_{ij}$  and  $w_{ji}$ . To stress the point, we constructed a related network with weighted links, but the weights are symmetric by assigning a weight  $w'_{ij}$  to a link according to

$$w'_{ij} = \frac{1}{2}(w_{ij} + w_{ji}), \quad (5)$$

with  $w_{ij}$  given by Eq. (3). The weights  $w'_{ij}$  can then be used to assign a strength  $s'$  to the nodes and a corresponding  $s$ -shell decomposition based on  $s'$  can be carried out to assign an index  $s'_s$  to each node. Figure 8 compares Kendall's  $\tau$  correlation of rankings based on  $s_s$  and  $s'_s$  with the actual SIR spreading efficiency for different values of the parameter  $a$ . In all cases, the measure with asymmetric weights  $s_s$  works better than that without the asymmetry. In the Emailcontact and Hamster networks, the two measures are equally accurate. The results confirm that it is important to include the different roles of

a link in spreading a disease between nodes  $i$  and  $j$  in two different directions into the construction of a reliable measure.

In summary, we proposed a node strength as an alternative centrality measure for efficient and accurate identification of influential spreaders. The idea of examining the functionality of a link in spreading in either direction is a general one and thus could be further developed for ranking a set of objects. As the current method only considers one step, it may be extended to further steps. Consider if node  $j$  in Fig. 1 has three outgoing nodes that are leaf nodes, while node  $k$  has only one outgoing node which further connects to many nodes. Whether taking further steps into consideration is a better choice for defining the dynamical link weights needs to be verified in future research. Previous results indicated that taking the two-step neighborhood into consideration balanced the calculation cost and ranking accuracy the best in a ranking method [56]. Although we work on defining the link weight from the network structure, it is worth considering how to integrate the artificial link weight with real weight on weighted networks. We think that the formation of a network structure depends on the function of the network elements of nodes and links.

Thus, the artificial link weight defined by the structure and the real weight on weighted networks that reflects the function may be highly correlated. How to merge and balance these two features in weighted networks is a valuable question to be explored in future work. In addition, we used the SIR model as the spreading dynamics. However, the idea of invoking asymmetric weights  $w_{ij} \neq w_{ji}$  for a link remains valid for other processes such as rumor spreading and information diffusion, although the exact form of the weights may depend on the details of the process under consideration.

#### ACKNOWLEDGMENTS

This work was jointly funded by the National Natural Science Foundation of China (Grants No. 11575041 and No. 61672238), the Scientific Research Starting Program of Southwest Petroleum University (Grant No. 2014QHZ024), the Data Intelligence Academic Innovation Team of Southwest Petroleum University (Grant No. 2015CXTD06), and the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2015J153).

- 
- [1] L. C. Freeman, *Soc. Networks* **1**, 215 (1978).
  - [2] L. C. Freeman, *Sociometry* **40**, 35 (1977).
  - [3] G. Sabidussi, *Psychometrika* **31**, 581 (1966).
  - [4] P. Bonacich and P. Floyd, *Soc. Networks* **23**, 191 (2001).
  - [5] B. Bollobás, in *Graph Theory and Combinatorics: Proceedings of the Cambridge Combinatorial Conference in Honour of Paul Erdős*, edited by B. Bollobás (Academic, New York, 1984).
  - [6] M. Kitsak, L. K. Gallos, F. Havlin, L. Liljeros, H. E. Mucnik, H. E. Stanley, and H. A. Makse, *Nat. Phys.* **6**, 888 (2010).
  - [7] S. Pei and H. A. Makse, *J. Stat. Mech.* (2013) P12002.
  - [8] D.-B. Chen, H. Gao, L. Lü, and T. Zhou, *PLoS One* **8**, e77455 (2013).
  - [9] Z.-M. Ren, A. Zeng, D.-B. Chen, H. Liao, and J.-G. Liu, *Europhys. Lett.* **106**, 48005 (2014).
  - [10] L. Lü, D.-B. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, *Phys. Rep.* **650**, 1 (2016).
  - [11] J. Bae and S. Kim, *Physica A* **395**, 549 (2014).
  - [12] T. Martin, X. Zhang, and M. E. J. Newman, *Phys. Rev. E* **90**, 052808 (2014).
  - [13] A. J. Alvarez-Socorro, G. C. Herrera-Almarza, and L. A. González-Díaz, *Sci. Rep.* **5**, 17095 (2015).
  - [14] J. E. Hirsch, *Proc. Natl. Acad. Sci. USA* **102**, 16569 (2005).
  - [15] L. Lü, T. Zhou, Q. M. Zhang, and H. E. Stanley, *Nat. Commun.* **7**, 10168 (2016).
  - [16] F. Radicchi and C. Castellano, *Phys. Rev. E* **93**, 062314 (2016).
  - [17] F. Bauer and J. T. Lizier, *Europhys. Lett.* **99**, 68007 (2012).
  - [18] G. Lawyer, *Sci. Rep.* **5**, 8665 (2015).
  - [19] J.-G. Liu, J.-H. Lin, Q. Guo, and T. Zhou, *Sci. Rep.* **6**, 21380 (2016).
  - [20] F. D. Malliaros, M.-E. G. Rossi, and M. Vazirgiannis, *Sci. Rep.* **6**, 19307 (2016).
  - [21] D. Grady, C. Thiemann, and D. Brockmann, *Nat. Commun.* **3**, 864 (2012).
  - [22] T. Opsahl, F. Agneessens, and J. Skvoretz, *Soc. Networks* **32**, 245 (2010).
  - [23] A. Garas, F. Schweitzer, and S. Havlin, *New J. Phys.* **14**, 083030 (2012).
  - [24] M. Eidsaa and E. Almaas, *Phys. Rev. E* **88**, 062819 (2013).
  - [25] Z.-X. Wu and P. Holme, *Phys. Rev. E* **84**, 026106 (2011).
  - [26] S. Allesina, A. Bodini, and M. Pascual, *Philos. Trans. R. Soc. B* **364**, 1701 (2009).
  - [27] C.-J. Zhang and A. Zeng, *Physica A* **402**, 180 (2014).
  - [28] N. N. Chung, L. Y. Chew, J. Zhou, and C. H. Lai, *Europhys. Lett.* **98**, 58004 (2012).
  - [29] D. Brockmann and D. Helbing, *Science* **342**, 1337 (2013).
  - [30] U. A. Brandes, *J. Math. Sociology* **25**, 163 (2001).
  - [31] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, *Phys. Rev. E* **65**, 056109 (2002).
  - [32] G.-Q. Zhang, D. Wang, and G.-J. Li, *Phys. Rev. E* **76**, 017101 (2007).
  - [33] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, *Proc. Natl. Acad. Sci. USA* **101**, 3747 (2004).
  - [34] W.-X. Wang, B.-H. Wang, B. Hu, G. Yan, and Q. Ou, *Phys. Rev. Lett.* **94**, 188702 (2005).
  - [35] M. Tang and T. Zhou, *Phys. Rev. E* **84**, 026116 (2011).
  - [36] H. F. Zhang, J. R. Xie, M. Tang, and Y. C. Lai, *Chaos* **24**, 043106 (2014).
  - [37] C. Castellano and R. Pastor-Satorras, *Sci. Rep.* **2**, 371 (2012).
  - [38] Y. Liu, M. Tang, T. Zhou, and Y. Do, *Sci. Rep.* **5**, 9602 (2015).
  - [39] Y. Liu, M. Tang, T. Zhou, and Y. Do, *Sci. Rep.* **5**, 13172 (2015).
  - [40] V. Batagelj and M. Zaveršnik, *Adv. Data Anal. Class.* **5**, 129 (2011).
  - [41] W. Wang, M. Tang, H. E. Stanley, and L. A. Braunstein, *Rep. Prog. Phys.* **80**, 036603 (2017).
  - [42] M. A. Serrano and M. Boguñá, *Phys. Rev. E* **74**, 056114 (2006).
  - [43] J. Leskovec, J. Kleinberg, and C. Faloutsos, *ACM Trans. Knowl. Discov. Data* **1**, 1 (2007).
  - [44] M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **98**, 404 (2001).
  - [45] M. Boguñá, R. Pastor-Satorras, A. Diaz-Guilera, and A. Arenas, *Phys. Rev. E* **70**, 056122 (2004).



- [46] N. Xie, Social network analysis of blogs, M.Sc. thesis, University of Bristol, 2006.
- [47] M. E. J. Newman, Network data, available at <http://www-personal.umich.edu/%7Emejn/netdata> (accessed 12/12/2012).
- [48] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson, *IEEE/ACM Trans. Networking* **12**, 2 (2004).
- [49] J. Kunegis, Hamsterster full network dataset, KONECT, available at <http://konect.uni-koblenz.de/networks/petster-hamster> (accessed 01/03/2014).
- [50] M. E. J. Newman, *Phys. Rev. E* **74**, 036104 (2006).
- [51] F. Morone and H. A. Makse, *Nature (London)* **524**, 65 (2015).
- [52] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Phys. Rev. E* **64**, 026118 (2001).
- [53] M. A. Serrano and M. Boguñá, *Phys. Rev. Lett.* **97**, 088701 (2006).
- [54] S. N. Dorogovtsev and A. V. Goltsev, *Rev. Mod. Phys.* **80**, 1275 (2008).
- [55] F. Radicchi and C. Castellano, *Phys. Rev. E* **93**, 030302 (2016).
- [56] Y. Liu, M. Tang, T. Zhou, and Y. Do, *Physica A* **452**, 289 (2016).