Physica A 469 (2017) 654–664

Contents lists available at ScienceDirect

Physica A

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

Identifying important nodes by adaptive LeaderRank

Shuang Xu[a,b], Pei Wang[b,c,*]

[a] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China
[b] School of Mathematics and Statistics, Henan University, Kaifeng 475004, China
[c] Laboratory of Data Analysis Technology, Henan University, Kaifeng 475004, China

HIGHLIGHTS

- A new node ranking algorithm named adaptive LeaderRank is proposed.
- The new algorithm is with competitive prediction accuracy and resolution.
- The new algorithm can well adapt to network perturbations or noisy data.

ARTICLE INFO

Article history:
Received 3 July 2016
Received in revised form 26 August 2016
Available online 25 November 2016

Keywords:
Complex network
Important node
LeaderRank
H-index
Topological perturbation
Dynamical process

ABSTRACT

Spreading process is a common phenomenon in complex networks. Identifying important nodes in complex networks is of great significance in real-world applications. Based on the spreading process on networks, a lot of measures have been proposed to evaluate the importance of nodes. However, most of the existing measures are appropriate to static networks, which are fragile to topological perturbations. Many real-world complex networks are dynamic rather than static, meaning that the nodes and edges of such networks may change with time, which challenge numerous existing centrality measures. Based on a new weighted mechanism and the newly proposed H-index and LeaderRank (LR), this paper introduces a variant of the LR measure, called adaptive LeaderRank (ALR), which is a new member of the LR-family. Simulations on six real-world networks reveal that the new measure can well balance between prediction accuracy and robustness. More interestingly, the new measure can better adapt to the adjustment or local perturbations of network topologies, as compared with the existing measures. By discussing the detailed properties of the measures from the LR-family, we illustrate that the ALR has its competitive advantages over the other measures. The proposed algorithm enriches the measures to understand complex networks, and may have potential applications in social networks and biological systems.

1. Introduction

Network theory offers us a platform and tool to investigate the complex interactions among individuals in detail [1–8]. Spreading process, such as viral transmission, rumour spreading and knowledge diffusion, is a common phenomenon, whose investigation in complex network [9–14] has been brought to widely attention. Recently, research communities pay much emphasis on identifying and ranking influential spreaders, which help us to better understand the spreading process and

[*] Corresponding author.
E-mail addresses: wp0307@126.com, wangpei@henu.edu.cn (P. Wang).

http://dx.doi.org/10.1016/j.physa.2016.11.034
0378-4371/© 2016 Elsevier B.V. All rights reserved.

develop strategies to prevent or maximize spreading. Therefore, great efforts have been made to develop novel centrality measures or indicators [15–26].

Degree is a simple centrality, but of less accuracy in some circumstances [18]. Closeness [15] and betweenness [16] are based on the shortest path, which can often improve the degree centrality, but suffer from computational complexity. Semi-local centrality, proposed by Chen et al. [17], aims at making a trade-off between relevance and computational complexity. Kitsak et al. [18] argued that, in contrast to common beliefs, the node position (coreness) is a pretty indicator of epidemic spreading. That is, nodes located within the core of the network as identified by the k-core decomposition [27] are rich spreaders, and those located at the periphery of the network are poor ones. But it is shown that the low resolution of coreness limited its application [19,20]. Recently, Lü et al. [21] used H-index to measure the node importance. It is suggested that the H-index is a good trade-off that in many cases it can better distinguish node influence than degree and coreness [21]. What is more, Lü et al. construct an operator iteration process, showing that degree, H-index and coreness are the initial, intermediate and steady states of the sequences, respectively. In the following, the degree, H-index, coreness are denoted as measures of the H-family. For directed networks, diffusion-based algorithms have been extensively investigated, such as the well-known PageRank [22].

In the year 2011, a variant of the PageRank, called LR [23], was proposed, where the authors applied the standard random walk process after adding a new node who connected all other nodes with a bidirectional edge to the original network. It is reported that the LR outperforms the PageRank for its faster convergence, exacter identification results and higher robustness [23]. Weighted-LeaderRank (WLR) [24] is a proper extension of the LR, and it uses a biased random walk. Numerical experiments show that the WLR can considerably improve the original LR. The LR and the WLR consist of the measures for the LR-family.

The mentioned centrality measures have their virtues and faults. Most of the measures depend on the structures of the network and assume that the networks are static. Considering the fact that many real-world complex networks are dynamic rather than static, the time-invariant and robust centralities appropriate for dynamic networks are in demand. Motivated by the above questions, we put forward a new algorithm called adaptive LeaderRank (ALR), which is also based on biased random walk, but whose stochastic matrix is totally different from the WLR's. Simulations show that the ALR can adapt to dynamic networks and own both competitively prediction accuracy and high resolution. We further compare properties of the LR-family including the LR, the WLR and the ALR. The rests of the paper are organized as follows. We overview the LR and the WLR, and then introduce our new algorithm in Section 2. We describe data and some methods in Section 3. The main results are given in Section 4. Discussions and conclusions will be in the last Section 5.

2. New algorithm

2.1. LR and its generalization

Motivated by identifying influential leaders in social networks [23], Lü et al. proposed the LR algorithm. For a directed network with N nodes and M edges, to make the network strongly connected, add a new node, called ground node, and link all others by bidirectional edges. The new network is strongly connected, which is with $N+1$ nodes and $M+2N$ edges. Matrix $A=(a_{ij})$ captures the network's wiring diagram. $a_{ij}=1$ if node i points to node j and it means that user j is a fan of i in social networks. The LR assigns a score to every node, where score implies the importance. At first, scores are given by $s_g(0)=0$ for ground node and $s_i(0)=1$ for other nodes. Then, at the subsequent steps, scores are updated by

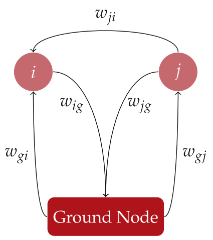
$$s_i(t+1)=\sum_{j=1}^{N+1}\frac{a_{ji}}{k_j^{out}}s_j(t) \tag{1}$$

where k_j^{out} is the out-degree of node j. It is proved that the iteration process will be converged quickly and it is reported that the LR outperforms the PageRank in terms of ranking effectiveness, as well as robustness against noisy data [23].

Following the LR, Li et al. developed the WLR by making nodes with more fans getting more scores from the ground node [24], since they argued that the number of fans (in-degree) is an important local indicator for a user's influence in spreading behaviours. In detail, nodes' importance scores are updated by

$$s_i(t+1)=\sum_{j=1}^{N}\frac{a_{ji}}{k_j^{out}}s_j(t)+\frac{w_{gi}}{\sum\limits_{k=1}^{N+1}w_{gk}}s_g(t), \tag{2}$$

where $w_{gi}=(k_i^{in})^\alpha$. Without loss of generality, we set $\alpha=1$ in the following investigation. It is shown that the WLR is more accurate and robust than the LR [24].



LR	WLR	ALR
$w_{ji} = a_{ji}$	$w_{ji} = a_{ji}$	$w_{ji} = a_{ji}h_i$
$w_{gi} = a_{gi}$	$w_{gi} = (k_i^{out})^\alpha$	$w_{gi} = a_{gi}h_i$
$w_{gj} = a_{gj}$	$w_{gj} = (k_j^{out})^\alpha$	$w_{gj} = a_{gj}h_j$
$w_{ig} = a_{ig}$	$w_{ig} = a_{ig}$	$w_{ig} = a_{ig}h_g$
$w_{jg} = a_{jg}$	$w_{jg} = a_{jg}$	$w_{jg} = a_{jg}h_g$

Fig. 1. Illustration of measures in the LR-family. Only two ordinary nodes and the ground node are shown as a toy example. It is noted that a_{ji} represents the topological connection and w_{ji} denotes the weight of the transition probability rather than the topological weight.

2.2. *The new algorithm: adaptive LeaderRank*

Hereinafter, in order to further improve the LR and the WLR, we introduce another weighted mechanism. We also apply a biased random walk process, where the transition probability is proportional to the H-index. We first introduce an operator $\mathcal{H}$ [21], which acts on a finite number of real numbers $(x_1, x_2, \ldots, x_n)$ and returns an integer $y = \mathcal{H}(x_1, x_2, \ldots, x_n) > 0$, where y is the maximum integer such that there exist at least y elements in $(x_1, x_2, \ldots, x_n)$, each of which is no less than y. As to nodes in a complex network, the H-index of node i denotes $h_i = \mathcal{H}(k_{j_1}, k_{j_2}, \ldots, k_{j_{k_i}})$, where k_i is the degree of node i. It is suggested that the H-index is a pretty indicator for identifying influential nodes in undirected networks and is not sensitive to changes of degree [21]. Therefore, we apply it to the weighted mechanism, expecting to improve the prediction accuracy and robustness of the existing measures.

For a network with adjacent matrix $A = (a_{ij})$, our algorithm is designed as follows. Firstly, obtaining the H-index for all nodes of the network. Then, adding a ground node to the network, which connects with all ordinary nodes via bidirectional edges, and setting its H-index as 1, namely, $h_g = 1$. After that, applying the biased random walk dynamic to each node according to the following updating rule:

$$s_i(t+1) = \sum_{j=1}^{N+1} \frac{a_{ji}h_i}{\sum\limits_{k=1}^{N+1} a_{jk}h_k} s_j(t). \tag{3}$$

Finally, the steady state of $s_i(t)$ can evaluate the relative influence or importance of node i. In summary, the algorithm works as follows:

Step 1 : Compute the H-index of nodes in the network;
Step 2 : Add a ground node to the network. Ground node connects with all ordinary nodes via bidirectional edges. And we define H-index for ground node $h_g = 1$;
Step 3 : Set initial scores. We set it as 1 for ordinary nodes and 0 for ground node;
Step 4 : Update scores. Successively updating the scores for all the nodes according to Eq. (3) until the scores reach its steady state.

The LR corresponds to the standard random walk, where the probability that random walker moves to each of its neighbours are identical. The WLR corresponds to a biased random walk. In this process, the probability that random walker moves from ordinary nodes to each of its neighbours are identical; while, it is not the case if random walker is from ground node, that is, it tends to move to nodes with higher out-degree. Therefore, there is a weighted mechanism. The ALR corresponds to another biased random walk, where random walker always prefer nodes with higher H-index, regardless of whether the random walker locates at the ground node or ordinary ones. The update rules Eqs. (1)–(3) can be written into the following general form.

$$s_i(t+1) = \sum_{j=1}^{N+1} \frac{w_{ji}}{\sum\limits_{k=1}^{N+1} w_{jk}} s_j(t), \tag{4}$$

where w_{ji} is the weight. Larger weight w_{ji} means higher probability that random walker moves from j to i. The differences of the LR, WLR and the ALR mechanisms are shown in Fig. 1. $w_{ji} = a_{ji}$ for all nodes in the LR. In the WLR, $w_{ji} = a_{ji}$ if j is an ordinary node and $w_{ji} = (k_i)^\alpha$ if j is the ground node. In the ALR, $w_{ji} = a_{ji}h_i$ for all nodes and $h_g = 1$ for ground node.

In the following, we will show that the new algorithm is very effective in ranking influential nodes, it is with good prediction accuracy and high resolution. Moreover, one of the most important advantages of the ALR is that it can highly adapt to topological perturbations. Simulations reveal that the ALR can better adapt to the change of degree distributions than the WLR and the LR. The ALR can also better adapt to zombie fans than the WLR and the LR in social networks, it can better distinguish fake influential spreaders from the truth ones. Thus, we call the new algorithm as adaptive LeaderRank (ALR), which is a new member of the LR-family.

Table 1
Topological features of six real-world networks. N and M are numbers of nodes and links. $\langle k \rangle$ denote the average degree. C and r represent the clustering [28] and assortative coefficients [29], respectively. β_c is approximated epidemic threshold of the SIR model [30–32], obtained by the heterogeneous mean-field theory. β denotes the infected probability.

Type	Networks	N	M	$\langle k \rangle$	C	r	β_c	β	Ref.
Communication	Email	1133	5 451	9.62	0.254	0.078	0.0565	0.0565	[33,34]
	UCsocial	1899	20 296	14.57	0.099	−0.188	0.0183	0.0183	[33,35]
HumanSocial	Jazz	198	2 742	27.70	0.633	0.020	0.0266	0.0532	[33,36]
	JM	2539	12 969	8.24	0.152	0.251	0.1053	0.1053	[33,37]
Coauthorship	NetSci	379	914	4.82	0.798	−0.082	0.1424	0.1424	[38]
Infrastructure	OpenFlights	2939	30 501	10.67	0.569	0.051	0.0183	0.0183	[33,39]

3. Materials and methods

3.1. Data

To validate the efficiency of the ALR, we perform numerical simulations on six real-world networks. The Email network is the email communication network at the University Rovira i Virgili in Tarragona in the south of Catalonia in Spain. Nodes are users and each edge represents that at least one email has been sent. The UCsocial network represents the message network between the users of an online community of students from the University of California, Irvine, where nodes represent users and directed edges represent message flows. The Jazz network is the collaboration network between Jazz musicians. Each node corresponds to a Jazz musician and an edge denotes that two musicians have played together in a band. The JM was created from a survey that took place in 1994/1995. Each student was asked to list his 5 best female and his 5 male friends. A node represents a student and an edge between two students shows that the source node (student) chose the sink node (student) as a friend. The NeiSci is a co-authorship network of scientists working on network theory and experiment. The OpenFlights contains flights between airports of the world. A directed edge represents a flight from one airport to another.

The statistical information of the considered networks are shown in Table 1. Except the E-mail, Jazz and NetSci, the other networks are all directed. The sizes of the considered networks range from hundreds to thousands of nodes. The average degrees for the six networks range from 4.82 to 27.70. The NeiSci, Jazz and OpenFlights networks are all with very high clustering coefficients. However, the clustering coefficient for the UCsocial network is comparatively very low. Moreover, except the UCsocial network and the NetSci network, the rest networks are all assortative, namely, highly connected nodes tend to be connected with each other. The NetSci and the JM networks are with quite high epidemic threshold, which indicates it will be difficult to spread information on the two networks, as compared with the other networks.

3.2. The susceptible–infected–recovered (SIR) model

The SIR model is widely researched in epidemics and information spreading. In the SIR model, there are three states for all nodes. An infected node will recover with probability λ and its neighbours will be infected with probability β. In the following simulations, we set $\lambda = 0.1$ and β as shown in Table 1. To speed the spreading dynamics in the Jazz network, we set a relative larger β value in Table 1. $\beta_c = \langle k \rangle / (\langle k^2 \rangle - \langle k \rangle)$ is a approximation of epidemic threshold via degree-based mean-field approach [13]. Epidemic strength is defined as β/λ, if epidemic strength was higher than the epidemic threshold β_c, then the information or disease can be spreading, while the infected numbers will be exponential decreased if $\beta/\lambda < \beta_c$ [30–32]. The chosen β and γ guarantee that $\beta/\lambda > \beta_c$, and information can be spread on the networks.

3.3. Correlation coefficient

Kendall τ_b correlation coefficient [40] is a popular rank correlation statistical measure. Considering n samples of two variables $x = (x_1, x_2, \dots, x_n)^T$ and $y = (y_1, y_2, \dots, y_n)^T$, paired samples (x_i, y_i) and (x_j, y_j) are concordant if $(x_i - x_j)(y_i - y_j) > 0$, discordant if $(x_i - x_j)(y_i - y_j) < 0$, or they are neither concordant nor discordant if $(x_i - x_j)(y_i - y_j) = 0$. In fact, if $(x_i - x_j)(y_i - y_j) = 0$, one can deduce that $x_i = x_j$ or $y_i = y_j$, and we call $x_i = x_j$ and $y_i = y_j$ as ties of x and y, respectively. There are totally $n(n-1)/2$ pairs of samples. There are three kinds of definition of Kendall correlation coefficients, where τ_b is adopted in this paper for its adjustments of ties. Based on the number of concordant and discordant pairs, the Kendall τ_b correlation coefficient is defined as

$$\tau_b = \frac{2(N_c - N_d)}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}, \tag{5}$$

where N_c and N_d are the number of concordant and discordant pairs, respectively. $n_0 = n(n-1)/2$, $n_1 = \sum_i t_i(t_i - 1)/2$, $n_2 = \sum_j u_j(u_j - 1)/2$, where t_i or u_j is the number of tied values in i'th or j'th group of ties for x or y.

3.4. Two-sample Kolmogorov–Smirnov test

The Kolmogorov–Smirnov test [41] is used to test whether two underlying one-dimensional probability distributions differ from each other. This method can be applied for data with any types of distributions. The null hypothesis is that two samples are drawn from the same distribution. The Kolmogorov–Smirnov statistic is defined as

$$D_{12} = \sup_x |F_1(x) - F_2(x)|, \tag{6}$$

where $F_1(x)$ and $F_2(x)$ are the empirical distribution functions of the two samples, respectively. sup is the supremum function. The null hypothesis will be rejected if P-value is lower than a given significant level (which is often taken as 0.05 or 0.01).

4. Performance of the new algorithm

4.1. Properties of the LR-family

As an index to well predict node influence, it is better to satisfy the following four rules:

(1) Sensitivity: rich-spreaders should not be assigned with low index values or scores, namely, the index should avoid type II error;

(2) Specificity: poor-spreaders should not be assigned with high scores, namely, the index should avoid type I error;

(3) Discrimination: the obtained score vector for all nodes is supposed to be with high resolution, namely, the index should avoid the appearance of too much ties;

(4) Robustness: top-ranked nodes should not be sensitive to data noise.

The first two rules guarantee the index to be with high identification accuracy, that is, high score nodes will be rich-spreader and rich-spreaders will not be with low scores. The third rule guarantees scores are unique in order to better distinguish among different individuals. The last rule facilitates the index to be appropriate to real-world networks with false positive data.

Hereinafter, we analyse and compare the performances of the new index with the other indexes of the LR-family.

4.1.1. Accuracy

Based on the SIR model, we consider each node as the single initial infectious seed, and the number of recovered individuals at stable states are used to evaluate node importance and spread range, denoted as $R_i (i = 1, 2, \ldots, N)$. We first compare the ALR with the WLR on specificity. If there are n different nodes between two top-L lists identified by the WLR and the ALR, we compare their average spread range. For example, if the top-5 list for the WLR is {1,2,3,4,5}, and {1,2,6,7,8} for the ALR, then there are 3 different nodes. We will compare $\overline{R}_{WLR}(5) = (R_3 + R_4 + R_5)/3$ and $\overline{R}_{ALR}(5) = (R_6 + R_7 + R_8)/3$. In the following, we set the cases for the WLR as benchmarks, and define the relative spread range as

$$\eta(L) = \frac{\overline{R}_{ALR}(L) - \overline{R}_{WLR}(L)}{\overline{R}_{WLR}(L)}. \tag{7}$$

If $\eta > 0$, then the ALR will be with higher specificity than the WLR, and therefore it will be better than the WLR.

Table 2 shows the relative spread range for the six networks, results for each network are averaged over 500 independent simulation runs. Our results indicate the performance of the ALR and the WLR are similar, that is, nodes with high ALR scores indeed have similar spread range with the WLR. The specificity of the ALR is not inferior to the WLR, it is even better than the WLR under some cases, such as the JM network with $L = 5$, the UCsocial with $L = 5$, 10 and 20, the NetSci with $L = 5$ and 10. Fig. 2 shows the evolutions of spread ranges for the top-ranked nodes of the six networks. From Fig. 2, we find that the performance of the ALR and the WLR are similar. For the JM network, the average infected size of the ALR is relatively higher than that under the WLR, while the NetSci is on the contrary, and there are no much differences for the other networks.

Now we further investigate the correlations of the new algorithm with existing ones by using the Kendall τ_b. For the ranks from the ALR and the other measures, we compute the pairwise Kendall τ_b coefficients, as shown in Table 3. Table 3 reveals the Kendall τ_b correlation coefficients among the ALR and the LR, the degree, the H-index, the coreness are lower than 0.8 for some networks, which indicates the new algorithm is different from the existing ones. The Kendall τ_b values between the ALR and the WLR are all higher than 0.84 for the six networks, which further reveals that there are certain similarities between the ALR and the WLR. However, as a new index, it should has its advantages over the existing measures. Therefore, we will further clarify the new features of the proposed algorithms in the following subsections.

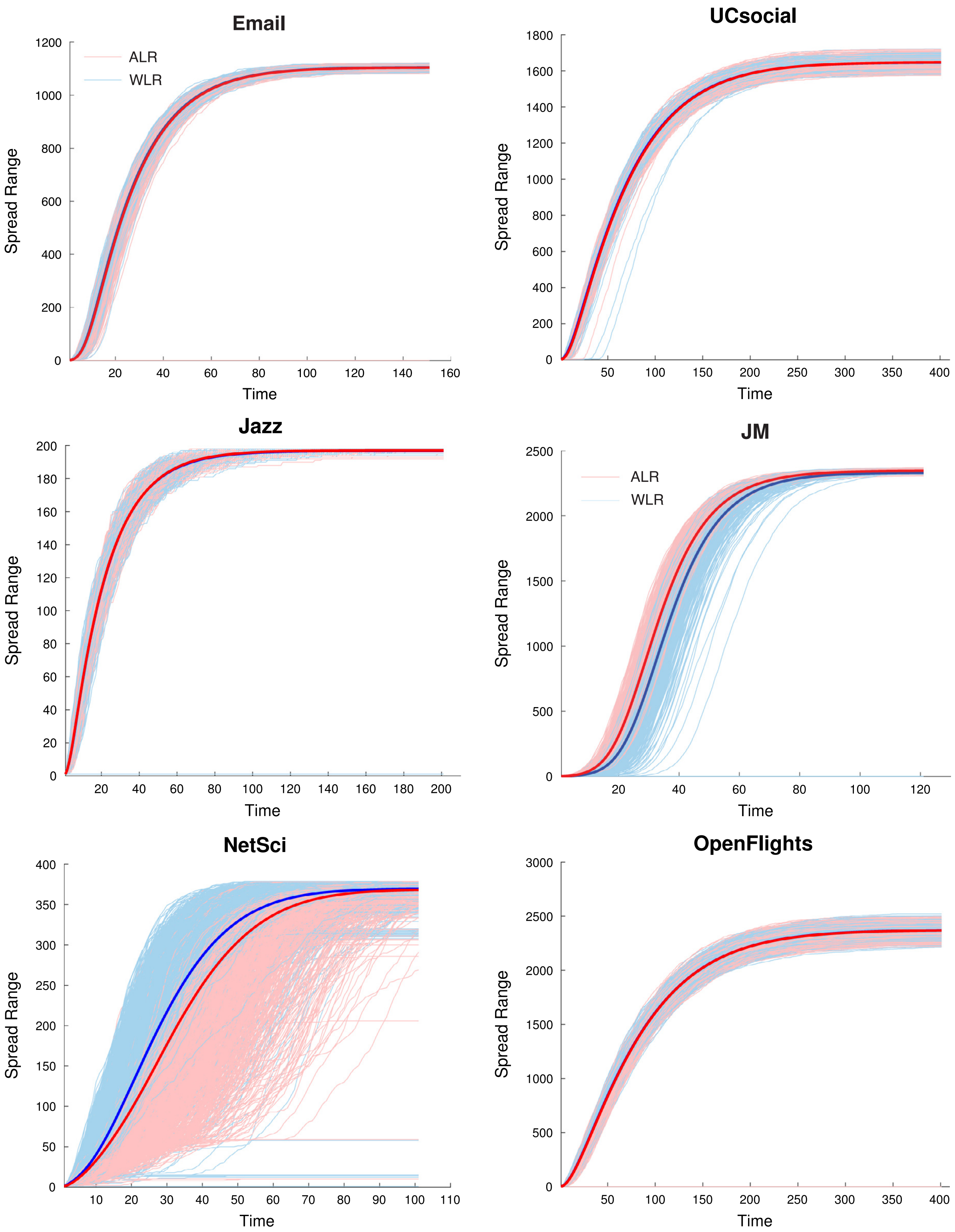


Fig. 2. Evolutions of spread ranges for the top-ranked nodes in the six networks. The JM is averaged over the top-5 ranked nodes, while the other networks are all averaged over the top-10 ranked nodes. The bold lines in each figure correspond to the average results for the ALR and the WLR, respectively, which are averaged over 500 independent simulation runs.

Table 2
Values of $\eta(L)$. Data is averaged over 500 independent simulation runs. "NA" denotes null value, where the identified top-ranked nodes from the ALR and WLR are the same.

Networks	$L = 5$	$L = 10$	$L = 15$	$L = 20$
Email	NA	0.00	−0.01	−0.01
UCsocial	0.03	0.01	NA	0.05
Jazz	0.00	0.00	0.00	0.00
JM	0.02	NA	−0.08	−0.01
NetSci	0.03	0.01	−0.01	−0.06
OpenFlights	0.00	0.00	0.00	0.00

Table 3
Kendall τ_b correlation coefficients between ranks from the ALR and the other measures.

Networks	LR	WLR	Degree	H-index	Coreness
Email	0.8423	0.8970	0.8755	0.9124	0.8815
UCsocial	0.8346	0.8682	0.8653	0.8830	0.8606
Jazz	0.8743	0.9027	0.8869	0.9371	0.7907
JM	0.8083	0.8425	0.6204	0.6576	0.6669
NetSci	0.7757	0.8848	0.8233	0.8904	0.8488
OpenFlights	0.7180	0.8588	0.7472	0.7874	0.8096

4.1.2. Resolution

To reveal the resolution of the new index, via the entropy theory [42], we define the entropy E for the index score, which is defined as:

$$E = -\sum_{s\in S} p_s \log(p_s). \tag{8}$$

Here

$$p_s = \frac{n_s}{\sum\limits_{j\in S} n_j}.$$

S is the score vector and n_s is the number of ties with score s. If all elements of the score vector are identical, the entropy is zero; if all elements are unique and denote the number of unique score by n, the entropy is $\log(n)$. Hence, score vector with larger entropy has higher discrimination. It is noted that the spread range of nodes with identical score differ very much, which will reduce the efficiency of the algorithm.

To further investigate the resolution property, we define resolution loss as follows,

$$l = \sum_{s\in S}\sum_{j=1}^{n_s}\left(R(j) - \bar{R}_s\right)^2, \tag{9}$$

where $R(j)$ is the spread range of node j and $\bar{R}_s$ is the average spread range over nodes with score s. Actually, resolution loss is the total sum of residual squares. To facilitate the comparison among different networks, we divide the spread range by its maximum, that is, $R^* = R/\max(R)$.

Tables 4 and 5 show the entropy and the resolution loss of the LR-family and the H-family, respectively. It shows that the entropy of the LR-family are similar, but the resolution losses have much differences. The ALR tends to be with lower resolution losses, namely, nodes with identical ALR scores have very similar spread ranges. While as to the H-family, the entropy decreases from degree to H-index, to coreness. Because a distinct value of H-index or coreness tends to correspond to more nodes than degree. But we do not observe the same phenomenon on the resolution loss, which is one possible reason of why the H-index or coreness is more accurate than degree in some cases. In a word, the LR-family outperform the H-family on resolution, and the resolution loss of the ALR is always the lowest.

4.1.3. Robustness

In social network, the number of fans stands for a user's status and prestige. It is a common phenomenon on microblog, one of the most popular Chinese social platforms, that part of users purchase *zombie fans* out of vanity. Zombie fans on microblogs are defined as invalid accounts signed up by network companies for the purpose of increasing the number of fans for certain users, especially celebrity ones, and getting them more attention. The mechanism is similar with Sybil attack in computer security.

In fact, *zombie fans* do not improve user's spread ability, prestige and social status in essence. A good ranking algorithm should discriminate the fake rich-spreaders from the truth ones. In other word, a good index should be robust to data noise.

Table 4
Entropy of measures from the LR-family and the H-family.

Networks	LR	WLR	ALR	Degree	H-index	Coreness
Email	6.9968	6.9968	6.9968	3.178	2.7112	2.3197
UCsocial	7.2639	7.2617	7.3006	3.3725	2.9854	2.2877
Jazz	5.2393	5.2393	5.2603	3.9526	3.4164	2.5305
JM	7.6368	7.6272	7.7362	2.8111	2.3598	1.3205
NetSci	5.4631	5.4608	5.4608	2.3205	1.8573	1.7778
OpenFlights	7.5527	7.5527	7.5812	2.8663	2.5146	2.4086

Table 5
Resolution loss of measures from the LR-family and the H-family.

Networks	LR	WLR	ALR	Degree	H-index	Coreness
Email	0.0099	0.0099	0.0099	1.9106	6.1385	2.1742
UCsocial	2.0287	2.0287	0.3113	21.5986	33.8574	14.8873
Jazz	0.0581	0.0581	0.0570	0.2494	0.9665	0.1680
JM	8.1320	9.2964	1.7725	104.0008	122.7655	118.0551
NetSci	0.0470	0.0484	0.0450	6.2075	6.5653	7.0652
OpenFlights	0.0368	0.0368	0.0328	1.3901	1.2240	0.7103

To test the robustness of the new index, we create v *zombie fans* for nodes of the top-L list. For node i, r_i and $\tilde{r}_i$ denote the original rank and the new rank after adding zombie fans, respectively; the smaller rank change $|r_i - \tilde{r}_i|$ is, the more robust the algorithm will be. Considering the cases with $v = 20, 50, 100$ and $L = 50$, we compare the tolerance of the LR-family to Sybil attack. Fig. 3 shows the line chart of original rank versus the new rank, lines coinciding with the diagonal ones indicate that the corresponding algorithm will perfectly preserve its rank. Fig. 3 shows that the lines for the ALR index almost keep unchanged, and even coincide with the diagonal lines (e.g. $v = 20$ in OpenFlights), followed by the WLR, while the LR is worse than the ALR and the WLR. It is noticed that degree and in-degree will change very much if a node have many zombie fans, but H-index changes little. Therefore the H-index based algorithm (ALR) shows superiority on deleting the influence of zombie fans.

4.2. Relationship between the LR-family and degree distribution

Degree distribution is a basic characteristic of complex network. Many properties of complex networks have more or less relations with its degree distribution. Whether the indexes of the LR-family depend on its degree distribution? Considering a node with the LR-family score s_i, we apply randomization methods to the original network, resulting in i with new score $\tilde{s}_i$. To measure the difference, we define

$$d_s = \frac{1}{N}\sum_{i=1}^{N} |s_i - \tilde{s}_i|, \tag{10}$$

where N is network size. d_s represents the average increment of score. We consider two approaches of randomization on a complex network. The first one is the full randomization method (FRM), where edges are rewired at random, resulting in an ER random network. The second is a degree-preserving randomization method (DRM), where the in- and out- degree distributions keep unchanged. If $d_s(DRM)$ is less than $d_s(FRM)$, the algorithm will be more correlated with its degree distribution. Thus, we define the relative score increment as:

$$D = \frac{d_s(DRM)}{d_s(FRM)}. \tag{11}$$

D can be viewed as an independence coefficient between the ranking algorithm and its degree distribution. $D = 0$ means perfect dependence, and $D = 1$ means perfect independence. Results for the considered networks are shown in Table 6, where each data is computed by averaging over 500 simulation runs. As an example, we show the in- and out-degree distributions for three networks in Fig. 4, including the OpenFlights, the UCsocial and the JM. The three networks are with $D(ALR) = 0.220, 0.311$ and 0.927, respectively. The differences between the in- and out-degree distributions for the three networks increase with the increasing of D.

Based on Table 6 and Fig. 4, on one hand, the LR depends on its degree distribution if in-degree and out-degree distribution tend to be similar. Notice that the Email, Jazz and NetSci are undirected networks, their in- and out-degree distributions are the same, therefore $D = 0$. We conjecture that the main reason of the consequence is that the network are undirected. However, it cannot explain why D is around zero for the OpenFlights, which is a directed network. To statistically measure whether two distributions are the same, we perform the two-sample KS test. We find that the two distributions for the OpenFlights are very similar (P-value $= 1$) (Fig. 4(a)), but they are very different for the other directed networks with non-zero D. It suggests that the LR algorithm depends on its degree distribution if its in- and out-degree distributions are

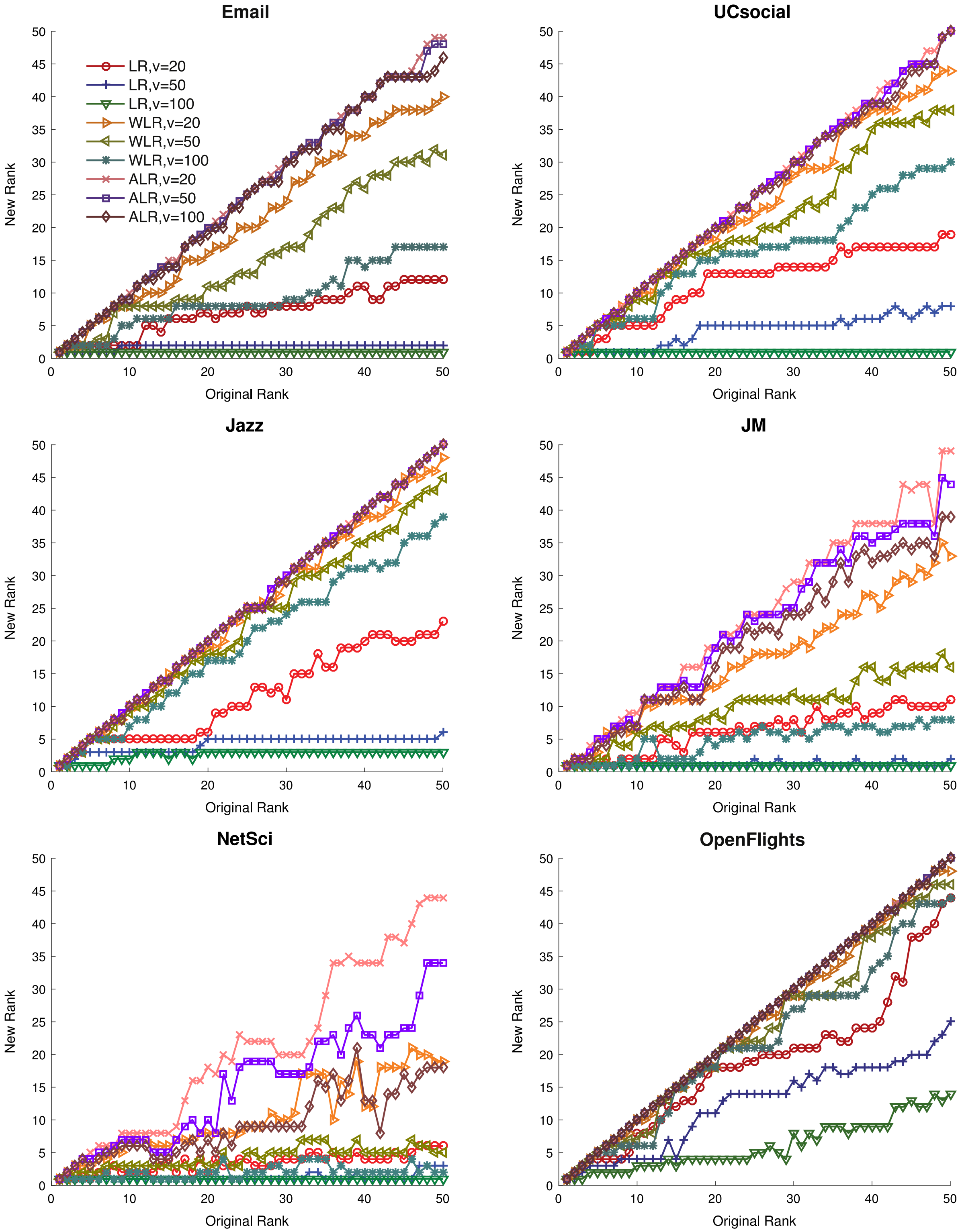


Fig. 3. Line charts of the original rank versus new rank after adding new nodes for the six networks.

similar. On the other hand, the LR-family depends on its degree distribution to different extent. The LR is with the greatest dependence, followed by the WLR and the ALR. Many networks is time-varying. For example, users may follow others during some periods, while unfollow others during some other periods in social network. Thus, degree distribution of a network

Table 6
Randomization results and KS test.

	$D(LR)$	$D(WLR)$	$D(ALR)$	P-value	KS-statistics
Email	0.000	0.068	0.248	1	0
UCsocial	0.274	0.286	0.311	$<10^{-5}$	0.2696
Jazz	0.000	0.216	0.481	1	0
JM	0.927	0.890	0.927	$<10^{-5}$	0.1225
NetSci	0.000	0.049	0.196	1	0
OpenFlights	0.010	0.120	0.220	1	0.002

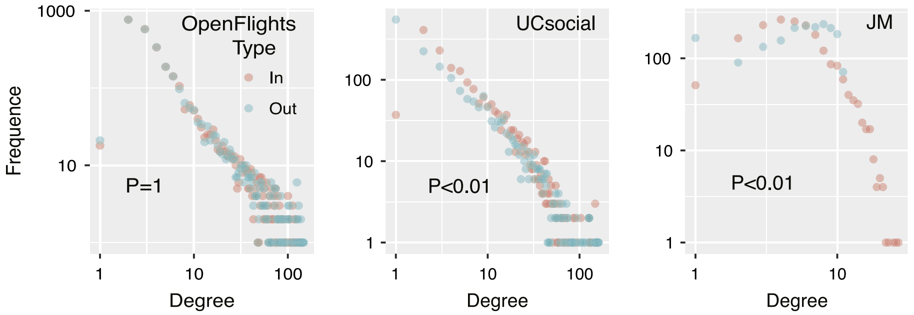


Fig. 4. Degree distribution. From left to right, the difference of in- and out-degree distributions increases.

may vary with time. The LR and the WLR are time-sensitive, since they depend on degree distribution very much. While, the ALR is more time-invariant. The result makes a significant step towards making possible applications to large-scale dynamical networks, because it guarantees that the identified rich-spreaders are effective even if the network evolves with time.

In fact, the process of the ALR algorithm can be used to explain the observed phenomenon. The LR is based on the standard random walk, where the transition probability totally depends on the degree of source node. The WLR and the ALR are based on the biased random walk, where the transition probability also relates to its in-degree and H-index, respectively. The H-index can be seen as the coarse-graining of degree, therefore, the WLR is more sensitive to degree.

5. Discussions and conclusions

On one hand, identification of influential node in complex networks is an open issue. Different types of networks have different features, it is generally difficult to develop an all-purpose measure to evaluate the importance of nodes in all complex networks. On the other hand, finding of vital spreaders in complex networks has fundamental importance in real-world applications, such as maximizing information spreading, controlling the spreading of rumours in social networks, finding drug targets in biological systems [25,26]. Therefore, it is interesting to design new measures for complex networks.

It is well known that real-world complex networks are dynamic and time-varying. Therefore, it is important yet interesting to design measures that can adapt to topological perturbation. Based on the LR and the H-index, we propose a new algorithm. The new algorithm is based on the biased random walk, instead of the standard random walk, where the transition probability relates to the H-index. Though the performance of the proposed algorithm on prediction accuracy is similar to the WLR (Fig. 2 and Table 3), detailed properties analysis of the new algorithm indicate it can further improve the LR and the WLR, and at the same time, with high resolution. More importantly, compared with the LR and the WLR, simulations suggest that the new algorithm can significantly promote the robustness against topological perturbations (Fig. 3 and Table 6), such as the adding of zombie fans and the change of degree distributions.

Network structure determines and affects node characters. Researches on the interesting inverse problem, identification of network topological structure is another interesting issue. Methods based on Granger causality, optimization or compressed sensing have been proposed [43–45]. But how to connect node importance and network structure reconstruction is still an open problem. However, the current work can help us to further understand the topological structure of complex networks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 61304151; The key scientific research projects in Colleges and Universities of Henan under Grants 17A120002. The Basal Research Fund of Henan University (yqpy20140049).

References

[1] R. Albert, A.L. Barabási, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (2002) art.no. 47.
[2] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2003) 167–256.
[3] A.L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[4] Y.Y. Liu, J.J. Slotine, A.L. Barabási, Controllability of complex networks, Nature 473 (2011) 167–173.
[5] J.D. Noh, H. Rieger, Random walks on complex networks, Phys. Rev. Lett. 92 (2004) art.no. 118701.
[6] L. Lü, T. Zhou, Link prediction in complex networks: A survey, Physica A 390 (2011) 1150–1170.
[7] X.F. Wang, Complex networks: topology, dynamics and synchronization, Internat. J. Bifur. Chaos 12 (2002) 885–916.
[8] S. Xu, P. Wang, Coarse graining of complex networks: a k-means clustering approach, in: Proc. 28th Chinese Control and Decision Conference, Yinchuan, May 28–30, 2016, pp. 4203–4208.
[9] R. Pastor-Satorras, A. Vespignani, Epidemic dynamics and endemic states in complex networks, Phys. Rev. E 63 (2001) art.no. 066117.
[10] M. Boguná, R. Pastor-Satorras, Epidemic spreading in correlated complex networks, Phys. Rev. E 66 (2002) art.no. 047104.
[11] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. 86 (2001) art.no. 3200.
[12] R. Pastor-Satorras, A. Vespignani, Immunization of complex networks, Phys. Rev. E 65 (2002) art.no. 036104.
[13] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, Epidemic processes in complex networks, Rev. Modern Phys. 87 (2015) art.no. 925.
[14] Y. Moreno, M. Nekovee, A.F. Pacheco, Dynamics of rumor spreading in complex networks, Phys. Rev. E 69 (2004) art.no. 066130.
[15] G. Sabidussi, The centrality index of a graph, Psychometrika 31 (1966) 581–603.
[16] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1977) 35–41.
[17] D.-B. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Physica A 391 (2012) 1777–1787.
[18] M. Kitsak, L.K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H.E. Stanley, H.A. Makse, Identification of influential spreaders in complex networks, Nat. Phys. 6 (2010) 888–893.
[19] J. Bae, S. Kim, Identifying and ranking influential spreaders in complex networks by neighborhood coreness, Physica A 395 (2014) 549–559.
[20] P. Wang, C. Tian, J. Lu, Identifying influential spreaders in artificial complex networks, J. Syst. Sci. Complex. 27 (2014) 650–665.
[21] L. Lü, T. Zhou, Q.-M. Zhang, H.E. Stanley, The H-index of a network node and its relation to degree and coreness, Nature Commun. 7 (2016) art.no. 10168.
[22] S. Brin, L. Page, Reprint of: The anatomy of a large-scale hypertextual web search engine, Comput. Netw. 56 (2012) 3825–3833.
[23] L. Lü, Y.-C. Zhang, C.H. Yeung, T. Zhou, Leaders in social networks, the delicious case, PLoS One 6 (2011) art.no. e21202.
[24] Q. Li, T. Zhou, L. Lü, D.-B. Chen, Identifying influential spreaders by weighted Leaderrank, Physica A 404 (2014) 47–55.
[25] P. Wang, J. Lü, X. Yu, Identification of important nodes in directed biological networks: A network motif approach, PLoS One 9 (2014) art.no. e106132.
[26] P. Wang, X. Yu, J. Lü, Identification and evolution of structurally dominant nodes in protein-protein interaction networks, IEEE Trans. Biomed. Circuits Syst. 8 (2014) 87–97.
[27] S.B. Seidman, Network structure and minimum degree, Soc. Networks 5 (1983) 269–287.
[28] D.J. Watts, S.H. Strogatz, Collective dynamics of small-worldnetworks, Nature 393 (1998) 440–442.
[29] M.E.J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (2002) art.no. 208701.
[30] M.E.J. Newman, Spread of epidemic disease on networks, Phys. Rev. E 66 (2002) art.no. 016128.
[31] R. Cohen, K. Erez, D. Ben-Avraham, S. Havlin, Resilience of the Internet to random breakdowns, Phys. Rev. Lett. 85 (2000) art.no. 4626.
[32] C. Castellano, R. Pastor-Satorras, Thresholds for epidemic spreading in networks, Phys. Rev. Lett. 105 (2010) art.no. 218701.
[33] J. Kunegis, The Koblenz Network Collection. URL: http://konect.uni-koblenz.de/ (accessed 16.04.23).
[34] 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 (2003) art.no. 065103.
[35] T. Opsahl, P. Panzarasa, Clustering in weighted networks, Soc. Networks 31 (2009) 155–163.
[36] P.M. Gleiser, L. Danon, Community structure in jazz, Adv. Complex Syst. 6 (2003) 565–573.
[37] J. Moody, Peer influence groups: Identifying dense clusters in large networks, Soc. Networks 23 (2001) 261–283.
[38] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) art.no. 036104.
[39] T. Opsahl, F. Agneessens, J. Skvoretz, Node centrality in weighted networks: Generalizing degree and shortest paths, Soc. Networks 3 (2010) 245–251.
[40] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.
[41] F.J. Massey Jr., The Kolmogorov-Smirnov test for goodness of fit, J. Amer. Statist. Assoc. 46 (1951) 68–78.
[42] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423.
[43] G. Mei, X. Wu, G. Chen, J. Lu, Identifying structures of continuously-varying weighted networks, Sci. Rep. 6 (2016) art.no. 26649.
[44] S. Zhang, X. Wu, J. Lu, H. Feng, J. Lü, Recovering structures of complex dynamical networks based on generalized outer synchronization, IEEE Trans. Circuits Syst.-I 61 (2014) 3216–3224.
[45] X. Wu, W. Wang, W. Zheng, Inferring topologies of complex networks with hidden variables, Phys. Rev. E 86 (2012) art.no. 046106.