Evaluating link significance in maintaining network connectivity based on link prediction

Cite as: Chaos **29**, 083120 (2019); https://doi.org/10.1063/1.5091608
Submitted: 04 February 2019 . Accepted: 26 July 2019 . Published Online: 20 August 2019

Mingze Qi, Suoyi Tan, Hongzhong Deng, and Jun Wu

Evaluating link significance in maintaining network connectivity based on link prediction

Cite as: Chaos **29**, 083120 (2019); doi: 10.1063/1.5091608
Submitted: 4 February 2019 · Accepted: 26 July 2019 ·
Published Online: 20 August 2019

Mingze Qi,[1] Suoyi Tan,[1,a)] Hongzhong Deng,[1] and Jun Wu[2,b)]

AFFILIATIONS

[1]College of Systems Engineering, National University of Defense Technology, Changsha, Hunan 410073, People's Republic of China
[2]International Academic Center of Complex Systems, Beijing Normal University, Zhuhai, Guangdong 519087, People's Republic of China

[a)]**Electronic mail:** tansuoyi_cn@outlook.com
[b)]**Electronic mail:** wujunpla@hotmail.com

ABSTRACT

Evaluating the significance of nodes or links has always been an important issue in complex networks, and the definition of significance varies with different perspectives. The significance of nodes or links in maintaining the network connectivity is widely discussed due to its application in targeted attacks and immunization. In this paper, inspired by the weak tie phenomenon, we define the links' significance by the dissimilarity of their endpoints. Some link prediction algorithms are introduced to define the dissimilarity of nodes based solely on the network topology. Experiments in synthetic and real networks demonstrate that the method is especially effective in the networks with higher clustering coefficients.

Various real-world systems can be represented by networks of nodes connected by links. The structure and function of complex networks attracted a huge amount of attention from many branches of science. In particular, how the significance of nodes and links is evaluated is an important issue because of its theoretical significance and application value. The essential purpose of this work is to evaluate link significance in maintaining the connectivity.

I. INTRODUCTION

A network is a set of nodes with connections between them, which we call links.[1] The complex networks graphically represent the interactions between the system's components. Examples include the cellular network,[2] social network,[3] power grid,[4] and many others. Evaluating the nodes' significance with some certain structural or functional objectives has always been a significant issue in studying complex networked systems.[5,6]

Due to the wide meaning of significance, many methods have been proposed from different aspects and have high application values.[7] Based on the nodes' capacity to impact the behavior of their surrounding neighbors, the degree centrality,[8] LocalRank,[9] ClusterRank,[10] Coreness, and H-index[11] are presented in succession. These centralities could be used to measure the academic impacts of researchers or journals based on their publication and citations. From the viewpoint of information dissemination, the eccentricity,[12] closeness centrality,[13] Katz centrality,[14] information index,[15] betweenness centrality,[16] and subgraph centrality are given based on paths in networks. They could help us optimize the use of limited resources to facilitate information propagation. There are also many iterative refinement centralities in which both the number of a node's neighbors and the influence of its neighbors are considered, such as eigenvector centrality,[17] PageRank,[18] LeaderRank,[19] and HITs.[20] These algorithms have pervasive applications, such as ranking web pages.

In addition to the above applications, the significant nodes also mean a lot to the immunization of a network against the epidemic spreading and targeted destruction of networks by targeted attacks. The main task in these cases is to find the nodes that play significant roles in maintaining the network connectivity. Many heuristic algorithms, such as collective influence,[21] BPD,[22] and explosive immunization algorithms,[23] have also been proposed to identify a set of critical nodes whose removal is most efficient in destroying

the network connectivity. However, in many situations, the nodes' removal is infeasible, while the link removal is a better choice. For instance, for many infectious diseases like HIV, a vaccine is not available.[24] We could only control the spread of the disease by modifying the network connectivity. Moreover, for networks in which links are the entity, such as road networks connecting cities,[25] the target of attack or control should be the links rather than nodes. Therefore, evaluating links' significance in maintaining connectivity is also an important issue.

There are several methods presented to quantify the link significance in maintaining the network connectivity. Based on the endpoints' number of neighbors, the degree product index[26] considers those who connect "hub" nodes important. The edge betweenness centrality index[27] assumes that the links passed by many of the shortest paths are of great significance. The bridgeness index[28] considers that the links connecting the larger cliques are of great importance. It is worth mentioning that the bridgeness is inspired by the weak tie phenomenon,[29] which refers to the fact that the links with weaker strength may play significant roles in maintaining the global connectivity.

Generally, the tie strength is characterized through the weight of the links, which is given by the attribute information on links or nodes, for instance, the time spent on communication in mobile communication networks or the content similarity in document networks. However, the information on the nodes is usually hard to obtain. Therefore, we try to define the tie strength based solely on the network topology. In sociology, another important phenomenon discussed widely with the weak tie theory is homophily, which shows that a strong tie tends to form between nodes with similar attributes in social networks. Examples include that individuals with similar interest and social status are more likely to establish close interactive relationships. Therefore, in contrast, the weak tie could be identified by the dissimilarity of endpoints.

Actually, the issue about the similarity or dissimilarity of nodes has been extensively discussed in another field, named link prediction.[30] The main goal of the link prediction problem is to estimate the existence likelihood of nonobserved links based on the known topology. Many similarity-based algorithms have been proposed according to the hypothesis that nodes tend to form links with other similar nodes. Conversely, these algorithms could be used to define the dissimilarity of two nodes. In this paper, we introduce these methods to evaluate link significance and study their effectiveness from the perspective of maintaining the network connectivity.

The article is organized as follows. In Sec. II, we define the dissimilarity of endpoints using link prediction methods and give a discussion about the selection of algorithms. In Sec. III, we compare some representative algorithms with other existing methods in both synthetic networks and real networks. Finally, the conclusion and discussion are given in Sec. IV.

II. EVALUATING LINK SIGNIFICANCE IN MAINTAINING THE NETWORK CONNECTIVITY USING LINK PREDICTION METHODS

As we discussed above, because of the importance of weak tie in maintaining the network connectivity, we present a way to determine the tie strength based solely on the network structure to evaluate the link significance in maintaining the network connectivity. Considering the homogeneity of nodes connected by strong ties, we could define the tie strength by the dissimilarity of the endpoints. Therefore, we could define the link significance in maintaining the network connectivity based on the dissimilarity of endpoints. In this section, we first introduce some link prediction methods to define the dissimilarity of the endpoints. Moreover, we give a discussion about the selection of the algorithms.

A. The link significance in maintaining the network connectivity based on the dissimilarity of endpoints

A network can be presented by a simple undirected graph $G(V, E)$, where V is the set of nodes, and E is the set of links. Multiplex links and self-loops are not allowed. According to the above discussion, we define s_{xy} as the scores of the existing links between nodes x and y. The dissimilarity between endpoints becomes higher as the value of s_{xy} decreases, and the link between them is more significant in maintaining the network connectivity.

The issue about the measure of the similarity or dissimilarity of nodes has been widely studied in link prediction. As the simplest and most effective framework of the link prediction problem, the similarity-based algorithms are grounded in the empirical evidence that two entities are more likely to interact if they are similar, which is identical to the homogeneity principle mentioned above. For improving the precision of prediction, various indices have been presented to measure the similarity of nodes based on the structural topology. On the other hand, these indices also give different definitions of the dissimilarity of nodes.

We mainly introduce nine link prediction methods to define the scores s_{xy} in this paper, and more indices could be found in a review article.[30] These indices could be divided into three categories. The common neighbors index (CN), the Adamic-Adar index (AA), and the resource allocation index (RA) are local approaches, which use the node neighborhood-related structural information to compute the similarity of nodes. The average commuter time (ACT), Cosine based on L^+ index (Cos+), and the random walk with restart index (RWR) are global approaches that use the whole network topological information to score each link. The local path index (LP), the local random walks (LRW), and the superposed random walks (SRW) are quasilocal approaches that consider both local and global information. The definitions of several representative methods are listed below, and others could be found in Appendixes A–C.

(i) Common neighbors index (CN): For a node x, let Γ_x denote the set of neighbors of x. In common sense, two nodes x and y are more likely to have a link if they have many common neighbors. The CN index is the direct count of its neighborhood overlap,

$$s_{xy}^{CN} = \left|\Gamma_x \cap \Gamma_y\right|. \tag{1}$$

(ii) Local path index (LP): The LP index is an index that takes consideration of local paths, with a wider horizon than CN. It is defined as

$$s^{LP} = A^2 + \epsilon A^3, \tag{2}$$

where ϵ is a free parameter. The $(A^2)_{xy}$ is the number of common neighbors between nodes x and y (second-order path). Clearly, this measure degenerates to CN when $\epsilon = 0$. If x and y are not

directly connected, $(A^3)_{xy}$ is equal to the number of different paths with a length of 3 connecting x and y. The parameter ϵ is used to control the proportion of the numbers of second- or third-order paths, and we generally choose a small ϵ to make the impact of the second-order path higher than the third-order path. In this paper, we make the $\epsilon = 10^{-5}$, when the network scale does not exceed 10^5, which guarantees that the elements in ϵA^3_{xy} are always smaller than 1.

(iii) Average commute time index (ACT): Denoted by r_{xy}, the average number of steps required by a random walker starting from node x to reach node y, the average commute time between x and y is

$$n_{xy} = r_{xy} + r_{yx}, \tag{3}$$

which can be obtained in terms of the pseudoinverse of the Laplacian matrix L^+,

$$n_{xy} = M(l^+_{xx} + l^+_{yy} - 2l^+_{xy}), \tag{4}$$

where l^+_{xy} denotes the corresponding entry in L^+ and M is a constant factor. Assuming two nodes are more similar if they have a smaller average commute time, then the similarity between the nodes x and y can be defined as the reciprocal of n_{xy}, namely,

$$s^{ACT}_{xy} = \frac{1}{l^+_{xx} + l^+_{yy} - 2l^+_{xy}}. \tag{5}$$

Notably, in the link prediction issue, the algorithms mentioned above are initially designed for predicting missing links in networks, and the scores given by similarity-based algorithms are calculated in the nodes pair without links connecting. Therefore, we should know whether these definitions are still effective for assessing the dissimilarity of endpoints. So in Sec. II B, we give a discussion about the selection of the link prediction algorithms.

B. The selection of the link prediction algorithms

In the initial design process, the above algorithms were calculated between two nodes without connecting links. Thus, the existing links may mislead the calculation of their endpoints' similarity scores. For instance, in Fig. 1, we use the LP algorithm to calculate the similarity scores in a sample network. The score of the link connecting nodes b and c decreases from 0.09 to 0.01 after removing link (b, c), while other links' similarity scores are slightly affected. According to the principle of the LP algorithm, we could find that the existing link between nodes would mislead the calculation of the number of the third-order path. To make use of the link prediction methods without mistake, we need to remove the links between node pairs before we calculate the similarity score. However, we could not directly compare the scores given in different network topologies. Thus, we introduce an iterative calculation method to select the link prediction algorithms below. This method is also named the ranking score in the link prediction issue.

Giving a network $G(V, E)$ and a link prediction algorithm, for every link $e \in E$, we make the set $E^T = e$. Then, we calculate the score s_{xy} of all the link in set $H = U - E + E^T$, where U is the universal set containing all possible links. The set H concludes all the nodes pair without links connecting in the network after removing the link e.

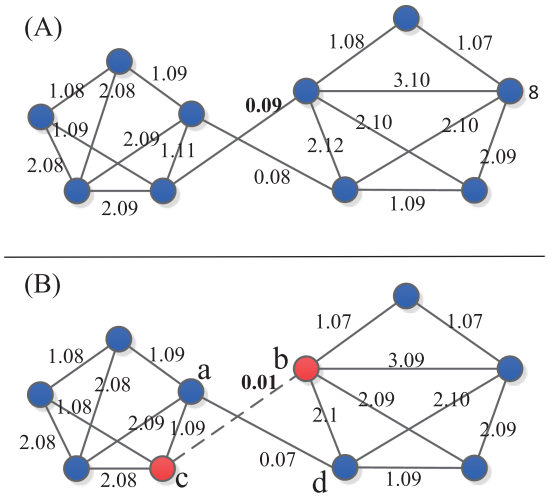

FIG. 1. The illustration of a mistake caused by the existing links. The scores of links calculated by the LP algorithm are shown in (a). After removing link (b, c), the scores are shown in (b). Because the size of the sample network is only 10, we make $\epsilon = 0.01$.

Then, we sort these scores in a descending order and record the rank of link e as r_e. We define the ranking score of the link e as

$$RS_e = \frac{r_e}{|H|}. \tag{6}$$

Note that these scores are calculated between node pairs without link connecting. So, the misleading phenomenon mentioned above would not appear in this calculation process. We sort all the links in a descending order of their ranking score and have rank A. Meanwhile, we calculate the scores directly in the same network and sort them in an ascending order to get rank B. Then, we could judge the effectiveness of the link prediction algorithm by comparing the difference of these two ranks. We introduce the Kendall rank correlation coefficient (commonly referred to as Kendall's tau coefficient) to measure the ordinal association between them. This coefficient depends on the number of inversions of pairs of objects which would be needed to transform one rank order into the other.

For the pair of links i and j, the x_i and y_i are the sequence numbers of i in a different order. If both $x_i < x_j$ and $y_i < y_j$, or if both $x_i > x_j$ and $y_i > y_j$, they are said to be concordant. Otherwise, they are said to be discordant. The Kendall τ coefficient is defined as

$$\tau = \frac{N_1 - N_2}{\frac{1}{2}N(N-1)}, \tag{7}$$

where N_1 is the number of concordant link pairs, N_2 is the number of discordant pairs, and N is the total number of the links. Obviously, when there is a positive correlation between the two ranks, $0 < \tau < 1$. Otherwise, $-1 < \tau < 0$. The value of $|\tau|$ represents the degree of correlation. The larger the $|\tau|$, the stronger the positive or negative correlation of the two ranks is.

The results are shown in Fig. 2. The network we used is a scale-free network given by the configuration model.[31] We produce random scale-free networks through the degree distribution $p(k)$

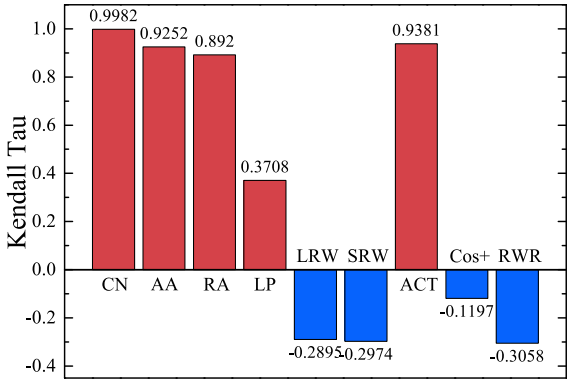

FIG. 2. Kendall's tau coefficient between ranks A and B with 9 different algorithms. The network used in the experiment is a scale-free network with $N = 1000$, $\lambda = 2.5$, $k_{\min} = 4$, and average degree $\langle k \rangle = 10.952$.

$= (\lambda - 1)k_{\min}^{\lambda-1}k^{-\lambda}$, where λ is the power-law exponent and $k_{\min}$ is the smallest degree. As Fig. 2 shows, in the three local indices (CN, AA, and RA), as well as ACT, the ranks A and B have a strong positive correlation. In LP, they have a lower positive correlation. However, in other algorithms, they are almost completely unrelated or even negatively correlated. The influence caused by the existing links might make these indices meaningless.

By analyzing the principle of the algorithm, it is not difficult to find the reasons behind the phenomenon. The three local indices are calculated by the information on two endpoints' common neighbors, which is not affected by the existing links between the nodes. The LP algorithm is determined by the number of paths with length 2 or 3. Just as Fig. 1 shows, when there is a link connecting two nodes, the number of the third-order path would include those who round trip to the neighbors first and then pass this link to arrive at another endpoint. The number of extra computations is related to the degree of the endpoints. As for the ACT index, it is the reciprocal of the average commute time given by the average number of steps required by a random walker starting from one node to another. Thus, the effect caused by the link between endpoints is insignificant.

In the rest of the article, we mainly use the three representative indices to calculate the scores s_{xy} of links in different categories: CN (local index), LP (global index), and ACT (quasilocal index). Their details have been given in Sec. II A.

III. EXPERIMENTAL ANALYSIS

In this section, we aim to verify the effectiveness of the method based on the dissimilarity of endpoints through experimental analysis in artificial and real networks.

A. The evaluation indices in the link percolation process

In general, if a significant order of links is effective in maintaining the network connectivity, the network will disintegrate much faster when we remove the links successively in the descending order of links' significance. In other words, in our method, the links are removed in an ascending order of the scores. This process is also named as the link percolation process.

1. The giant components

To evaluate the impact of removing links, we use the relative size of giant components C, which is the fraction of nodes contained in the giant components, to characterize the network connectivity. If we remove the links with high significance in maintaining the network connectivity, C will decrease faster. The parameter q is defined as the proportion of links removed,

$$q = \frac{m_r}{m}, \tag{8}$$

where m is the number of all links and m_r is the number of removing links. We record the function of the realized size of the giant components C over the proportion of q. We also call this link removal process the link percolation process and the change curve $C(q)$ the percolation curve.

2. The area under the percolation curve

In addition to visually observing the change in the relative size of giant components, to demonstrate the impact of the ranking of critical links on the networks in depth, we also consider the measure of the network R' to characterize the overall effect of evaluation methods,[32]

$$R' = \frac{1}{m+1}\sum_{m_r=0}^{m}\hat{C}(m_r) \approx \int_0^1 C(q)dq, \tag{9}$$

where $\hat{C}(m_r)$ is the relative size of the giant component after removing m_r links. The R' is calculated by the average of the $\hat{C}$ when the first m_r ($m_r = 0, 1, \ldots, m$) links are removed. C can be seen approximately as a continuous function when the total number of links m is large; therefore, R' could also be defined as the area under the percolation curve that corresponds to the integral of the curve $C(q)$. The value of R' can comprehensively evaluate the order of the edge significance in general. The smaller the R' value is, the more effective the method is. In Sec. II B, we have discussed the selection of link prediction algorithms. Next, we would compare the representative indices (whose principles have been given in Sec. II A) with other traditional indices.

B. Other indices for evaluating link significance

In this article, we introduce the three most common methods. Degree centrality and betweenness centrality are widely used in evaluating nodes' significance in maintaining the network connectivity, and similar principles have also been used to define the significance of links.

(i) Degree product: Based on the degree centrality of the links' endpoints, the degree product has been used in the studies of biased (degree-dependent) percolation. The degree product is defined as

$$Dp_e = (k_x k_y)^{\alpha}, \tag{10}$$

where x and y are two endpoints of link e. k_x is the degree of node x, whose value is the number of adjacent links of node x.

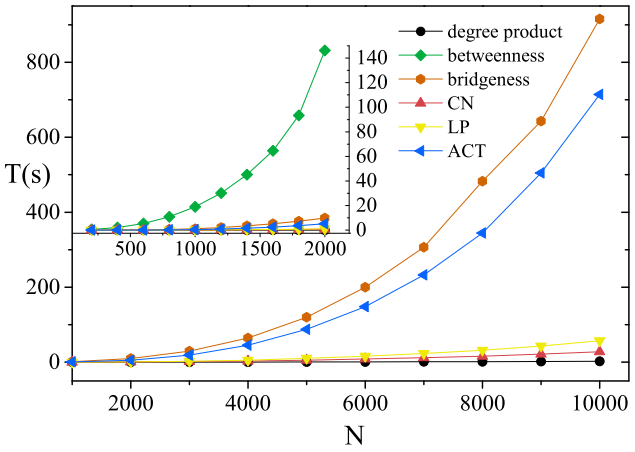

FIG. 3. The computation time of different methods vs network size N. The networks used in the experiments are scale-free networks with $\lambda = 2.5$ and $k_{min} = 4$.

α is a free parameter. Note that we only care about the order of the links and thus we choose $\alpha = 1$. The degree product index assumes that links connecting two nodes with a high degree ("hub" nodes) are of great significance. To better compare the results, we give link percolation processes with removing links in both ascending order (degree as) and descending (degree des) order of degree product.

(ii) Edge betweenness centrality: Betweenness was first extended from a graph theoretical notion to both connected and unconnected networks by Freeman in 1977. Then, Girvan generalized betweenness centrality to links and defined the edge betweenness centrality of a link as the number of shortest paths between pairs of nodes that run along it. The betweenness centrality is defined as

$$Bc_e = \sum_{x \neq y} \frac{\sigma_{xy}(e)}{\sigma_{xy}}, \tag{11}$$

where $\sigma_{xy}(e)$ is the number of shortest paths from x to y that pass through link e, and σ_{xy} is the number of shortest paths from node x to node y. Generally, if a network contains communities or groups that are only loosely connected by a few intergroup links, all shortest paths between different communities must go along one of these few links. Thus, the links connecting communities will have high edge betweenness.

Except for the above two indices expended from nodes' centrality, inspired by a weak tie phenomenon in document networks, a local index called bridgeness has also been proposed for evaluating link significance in maintaining the network global connectivity.

(iii) Bridgeness: A clique of size k is a fully connected subgraph with k nodes, and the clique size of a node x or an edge e is defined as the size of the maximum clique that contains this node or this edge. The bridgeness is defined as

$$Br_e = \left(\frac{\sqrt{S_x S_y}}{S_e} \right)^{\gamma}, \tag{12}$$

where S_x and S_y are the clique sizes of nodes x and y, respectively. S_e is the clique size of edge e. The bridgeness shows that the links between cliques may play a more significant role in maintaining the global connectivity.

The time complexity of our methods is determined by the principle of the link prediction algorithms. The algorithmic complexities of CN, LP, and ACT are approximately $O(Nk_{\max}^3)$, $O(3N^2k_{\max})$, and $O(N^3)$, where N is the number of nodes and $k_{\max}$ is the maximum degree of a node. We further give the time cost of the methods above in Fig. 3. The results show that the time cost of the betweenness is significantly higher than that of other methods. The time complexity of link prediction algorithms is relatively lower and can be adapted in the calculation of larger scale networks.

C. Experiments in synthetic networks

Due to the ubiquity in the real-world, we first focus on scale-free networks in this study.

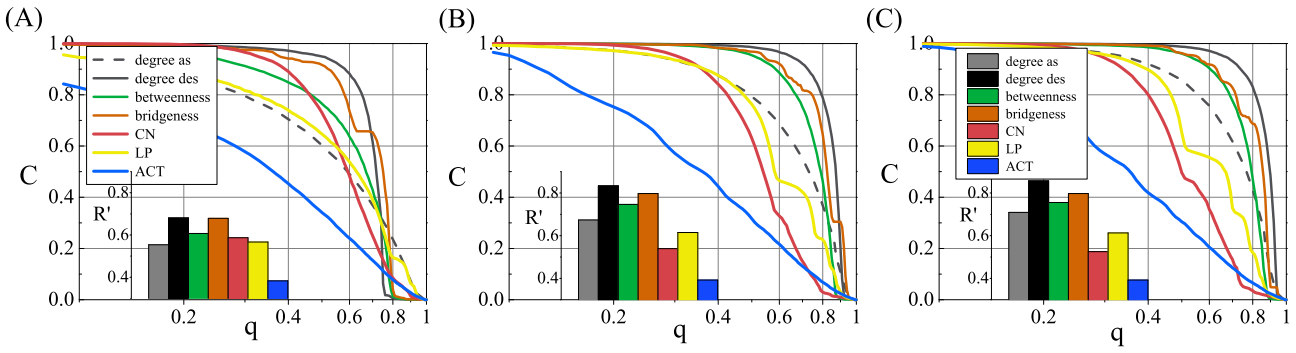

FIG. 4. Comparison of methods in random graphs. The size of networks $N = 10\,000$ and the connection probability $P = 0.0003$ (a), 0.0006 (b), and 0.0009 (c), respectively. The results are displayed in logarithmic coordinates.

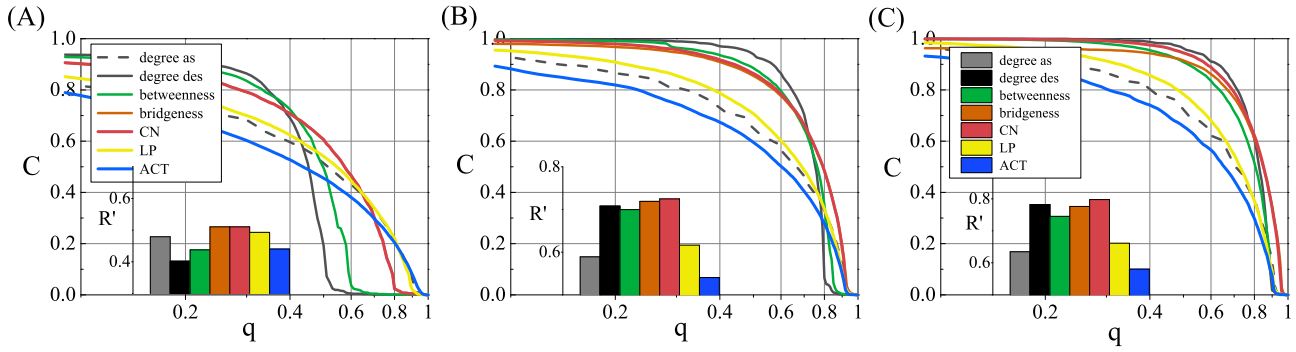

FIG. 5. Comparison of methods in scale-free networks. Their link percolation processes are given in scale-free networks with the same $N = 10\,000$, $\lambda = 2.5$ but different k_{min}. The k_{min} of the networks are 3 (a), 4 (b), and 5 (c), respectively. The average degree $\langle k \rangle$ of these three networks is 5.74, 11.53, and 14.41, respectively. The results of link percolation processes are displayed in logarithmic coordinates.

1. Scale-free networks

Networks with power-law degree distribution have been the focus of a great deal of attention in the literature. These networks sometimes are referred to as scale-free networks. Similarly, we generate a scale-free network by a configuration model, which has been presented in Sec. II B. In Fig. 5, we present the link percolation processes of three typical indices in scale-free networks with different average degree $\langle k \rangle$ and compare them with the other three evaluation indices, degree product (with both ascending and descending orders), betweenness and bridgeness. R' value of each link percolation process is given in the bar graph inside. From the results in Fig. 5, we find that the link prediction methods generally outperform the three existing methods. This is especially true when the networks have a higher average degree, just as Figs. 5(b) and 5(c) show. However, we could find that the percolation curve of LP index is similar to that of degree product with ascending order (degree as).

To explore the causes of these results, we further experiment in the other two kinds of special synthetic networks: random graphs and community networks.

2. Random graphs

The random graphs are generated by connecting nodes randomly.[33] Each link is included in the network with a probability P, independently from every other link. The average degree $\langle k \rangle$ of a random network is $N \times P$ approximately. The link percolation processes are given in Fig. 4. We noticed that the number of common neighbors is an important factor used to determine whether the endpoints of a link are similar in our methods. However, the nodes in random graphs have little common neighbors, which can be seen in the poor property of network transitivity (or clustering).

We generally use the clustering coefficient Cl proposed by Watts and Strogatz to measure the clustering property of networks.[34] A local value Cl_x is given to define the clustering coefficient of node x as

$$Cl_x = \frac{2E_x}{k_x(k_x - 1)}, \tag{13}$$

where E_x is the number of actual edges between the neighbors of node x and k_x is the degree of node x. For nodes with degree 0

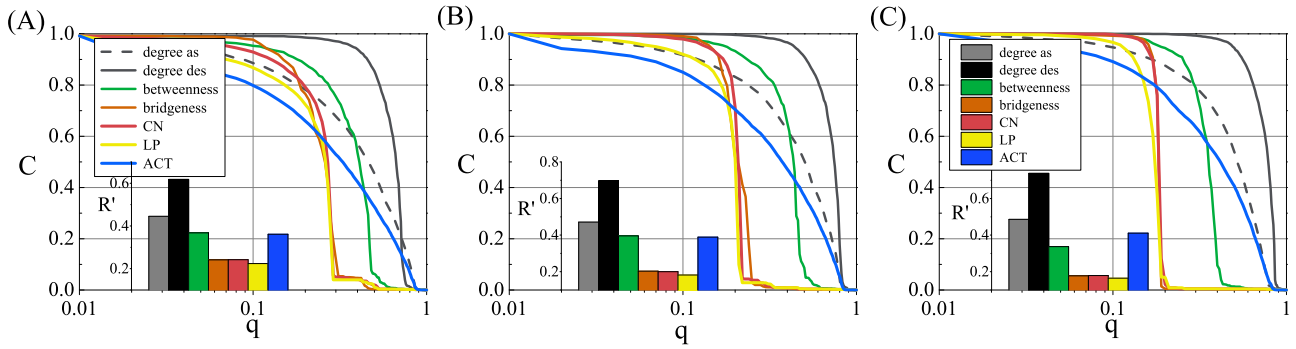

FIG. 6. Comparison of methods in community networks. The size of networks $N = 10\,000$ and the average degree $\langle k \rangle = 5.543$ (a), 7.915 (b), and 11.95 (c), respectively. The results are displayed in logarithmic coordinates.

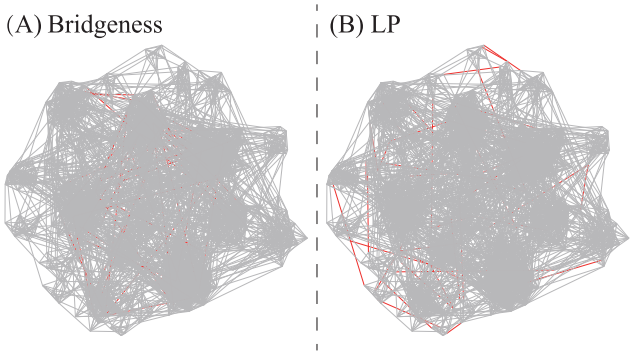

FIG. 7. The illustration to the top 1% links determined by bridgeness and LP in a community network with size $N = 500$.

or 1, we put $Cl_x = 0$. Then, the clustering coefficient for the whole network is the average

$$Cl = \frac{1}{N}\sum_x Cl_x. \tag{14}$$

The clustering coefficients Cl of the three scale-free networks in Fig. 5 are 0.0402, 0.0510, and 0.0615, respectively, while the Cl of the above three random graphs are only 0.00036, 0.00076, and 0.0011, respectively, which explain the poor performance of our method in the random graphs.

3. Community networks

We further experiment in artificial networks with clear community structures. The networks are obtained by a method proposed by Lancichinetti, which was initially used to generate benchmark graphs for testing community detection algorithms.[35,36] The model first creates unconnected communities and then chooses randomly internal links that are reconnected outside the community. The experimental results in the community network are shown in Fig. 6. The clustering coefficient Cl of the three community networks are 0.4471, 0.4454, and 0.4499, respectively. We find that bridgeness, CN and LP methods have a similar effect in the community networks. The removal of the edges determined by these methods would cause the network to collapse rapidly. Differently, the bridgeness index is biased to those links between large communities, while our method based on link prediction has no such bias. Even due to the special principle of link prediction algorithms, like LP, the links between small communities or nodes with a lower degree have advantages. We can see the difference between them visually through the visualization in Fig. 7.

Meanwhile, we found that compared with the above three algorithms, the ACT algorithm is not performing well in the community networks, which indicates that the average commute time (ACT) of random walker between two nodes is not sensitive to the community structure of the network. Although ACT has an excellent performance in many other networks, CN and LP are more dominant from identifying links between communities.

D. Experiments in real networks

The artificial networks facilitate our quantitative analysis but ignoring many situations that exist in real datasets. Therefore, we further experiment in eight real networks below. All these networks are undirected and unweighted.

(1) Euroroad is a road network located mostly in Europe.[37] Nodes represent cities and, a link between two nodes denotes that they are connected by an E-road.
(2) Adolescent health is a network of friendship of adolescent students.[38]
(3) Email is a network generated using email data. Nodes of the network are email addresses and if an address is sent at least one email to another address, an undirected link is created.
(4) Hamsterster is a network containing friendships between users of the website hamsterster.com.
(5) Chess is a network bout chess games. Each node is a chess player, and a link represents a game between two players.
(6) Cfinder google is a hyperlink network from pages within Google's own sites (on google.com).
(7) Astroph is a collaboration network of authors of scientific papers from the arXiv's Astrophysics (astro-ph) section. A link between two authors represents a common publication.
(8) Politic Blog is a network of US political blogs. A node represents a blog, and a link represents a hyperlink between two blogs.

TABLE I. The basic data for the eight real networks. The first column identifies the names of the networks. For each, we report in the following columns: the number of nodes n, the total number of links m, the average degree $\langle k\rangle$, the maximum degree k_{max} and the clustering coefficient Cl. We arrange the networks according to their average degree.

Network	n	m	$\langle k\rangle$	k_{max}	Cl
Euroroad	1174	1417	2.41	10	0.017
Adolescent_health	2539	10 455	8.24	27	0.147
Email	1133	5451	9.62	71	0.22
Hamsterster	1858	12 534	13.49	272	0.141
Chess	7301	55 899	15.31	181	0.177
Cfinder_google	15 763	149 456	18.96	11 401	0.526
Astroph	18 771	198 050	21.10	504	0.631
Politic_Blog	1222	16 714	27.36	351	0.32

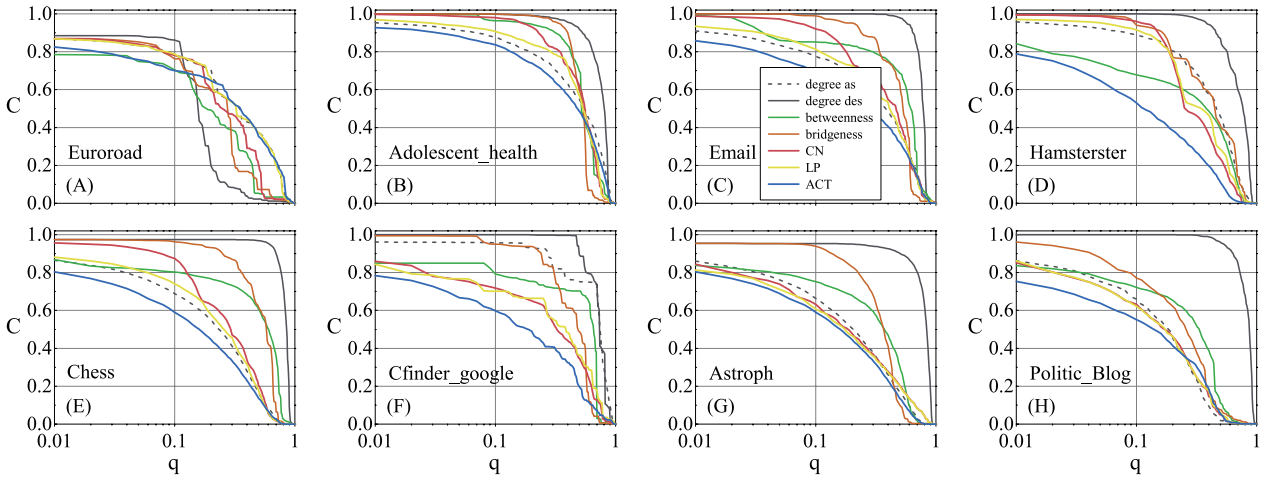

FIG. 8. Comparison of methods in eight real networks. The results are displayed in logarithmic coordinates.

These networks' datasets can also be downloaded from KONECT.[39] The basic data for these networks are shown in Table I.

The link percolation processes of different methods in real networks are shown in Fig. 8. The values of R' are presented in Fig. 9. We found that our dissimilarity based method is effective in most of the real networks, especially in those large social networks with a higher clustering coefficient. Meanwhile, we found that in the Euroroad network composed of the main roads between the cities of Europe [Fig. 8(a)], the clustering coefficient is only 0.017. Therefore, it is hard to identify the dissimilarity of nodes based solely on the topology information and our methods have poor performance in this network.

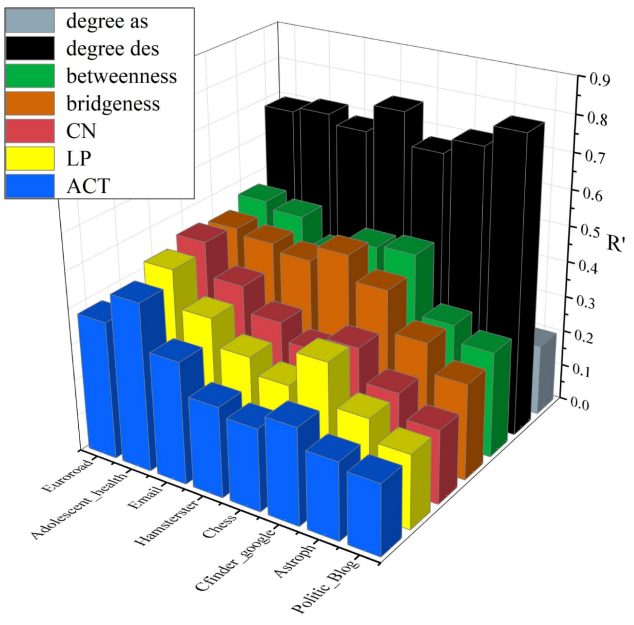

FIG. 9. The R' in real networks.

IV. CONCLUSION AND DISCUSSION

In this paper, we introduce some indices from the link prediction issue and use the dissimilarity of endpoints to define the significance of the links. Considering the mislead that may be caused by the existing link between node pairs, we give a discussion about the selection of link prediction algorithms. Experiments in both artificial and real networks show that removing links with the significance determined by our method leads to faster collapse of the networks. Meanwhile, we prove that our method is suitable for networks with higher clustering coefficients.

Different from the traditional methods which are based on endpoints' centrality or clique information, our method provide a new way to evaluating link significance in maintaining the network connectivity. According to the algorithmic principle, these dissimilarities are always defined by the number of accessible paths. The values of CN and LP algorithms are determined by the number of paths connecting nodes. In the ACT algorithm, the average commute time of a random walker is also significantly affected by the number of accessible paths. If there are few reachable paths between two nodes, the link between them is less replaceable. Therefore, the network connectivity would be seriously affected by those links' removal.

In this paper, we only tested some typical algorithms and conducted a preliminary analysis, which is far from providing a satisfactory answer. In the following work, we could test more existing link prediction algorithms or design some new definitions of dissimilarity. It is also a good idea to improve the performance of indices by combining some existing methods. At the same time, aside from their roles in maintaining connectivity, the performance

of similar methods in the network dynamics and other fields is also an important topic.

ACKNOWLEDGMENTS

Hongzhong Deng acknowledges the Natural Science Foundation of China under Grant Nos. 71771214 and 71690233. Jun Wu acknowledges the Natural Science Foundation of China under Grant Nos. 71371185 and 71871217 and the Hunan Provincial Natural Science Foundation of China under Grant No. 2019JJ20019. We thank Ziqiang Cao and Ye Deng for their assistant in designing figures. We are particularly grateful to the reviewers for their detailed suggestions which contributed greatly to the improvement of this paper.

APPENDIX A: LOCAL INDICES

1. Adamic-Adar index (AA)

This similarity measure refines the simple counting of common neighbors by assigning the less-connected neighbors more weights. For two nodes x and y, the AA index can be written as

$$s_{xy}^{AA} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log|\Gamma(z)|}, \tag{A1}$$

where $\Gamma(x)$ is the set of neighbors of x.

2. Resource allocation index (RA)

This index is motivated by the resource allocation process that take place in complex networks. It models the transmission of units of resources between two nodes, x and y, which are not directly connected. The similarity between x and y can be defined as the amount of resource y received from x, which is

$$s_{xy}^{RA} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{|\Gamma(z)|}. \tag{A2}$$

APPENDIX B: GLOBAL INDICES

1. Cosine based on L^+(Cos+)

This index is an inner-product-based measure. In the Euclidean space spanned by $v_x = \Lambda^{\frac{1}{2}} U^T \vec{e}_x$, where U is an orthonormal matrix made of the eigenvectors of L^+ ordered in the decreasing order of the corresponding eigenvalue λ_x, $\Lambda = \mathrm{diag}(\lambda_x)$, $\vec{e}_x$ is an $N \times 1$ vector with the xth element equal to 1 and others all equal to 0, and T is the matrix transposition, the pseudoinverse of the Laplacian matrix are the inner products of the node vectors, $l_{xy}^+ = v_x^T v_y$. Accordingly, the cosine similarity is defined as the cosine of the node vectors

$$s_{xy}^{\cos^+} = \cos(x,y)^+ = \frac{v_x^T v_y}{|v_x|\cdot|v_y|} = \frac{l_{xy}^+}{\sqrt{l_{xx}^+ \cdot l_{yy}^+}}. \tag{B1}$$

2. Random walk with restart (RWR)

This index is a direct application of the PageRank algorithm. Consider a random walker starting from node x, who will iteratively moves to a random neighbor with probability c and return to node x with probability $1-c$. Denote by q_{xy} the probability this random walker locates at node y in the steady state, we have

$$\vec{q}_x = cP^T\vec{q}_x + (1-c)\vec{e}_x, \tag{B2}$$

where P is the transition matrix with $P_xy = \frac{1}{k_x}$ if x and y are connected, and $P_{xy} = 0$ otherwise. The solution is straightforward, as

$$\vec{q}_x = (1-c)(I - cP^T)^{-1}\vec{e}_x. \tag{B3}$$

The RWR index is thus defined as

$$s_{xy}^{RWR} = q_{xy} + q_{yx}, \tag{B4}$$

where q_{xy} is the yth element of the vector $\vec{q}_x$. In this paper, we have $c = 0.9$.

APPENDIX C: QUASILOCAL INDICES

1. Local random walks (LRW)

A random walker is initially put on node x and thus the initial density vector $\vec{\pi}_x(0) = \vec{e}_x$. This density vector evolves as $\vec{\pi}_x(t+1) = P^T\vec{\pi}_x(t)$ for $t \geqslant 0$. The LRW index at time step t is thus defined as

$$s_{xy}^{LRW}(t) = q_x\pi_{xy}(t) + q_y\pi_{yx}(t), \tag{C1}$$

where q is the initial configuration function. In this paper, we have $q_x = \frac{k_x}{M}$ and $\pi_{xy} = \frac{k_y}{M}$.

2. Superposed random walks (SRW)

In the SRW index, the random walker is continuously released at the starting node, resulting in a higher similarity between the target node and the nodes nearby. The mathematical expression reads

$$s_{xy}^{SRW}(t) = \sum_{\tau=1}^{t} s_{xy}^{LRW}(\tau) = \sum_{\tau=1}^{t}[q_x\pi_{xy}(\tau) + q_x\pi_{yx}(\tau)], \tag{C2}$$

where t denotes the time steps. In this paper, we have $t = 3$.

REFERENCES

[1] M. E. J. Newman, "The structure and function of complex networks," SIAM Rev. **45**, 167–256 (2003).
[2] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," IEEE Trans. Wirel. Commun. **10**, 2752–2763 (2011).
[3] D. J. Watts, P. S. Dodds, and M. E. J. Newman, "Identity and search in social networks," Science **296**, 1302–1305 (2002).
[4] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid—The new and improved power grid: A survey," IEEE Commun. Surv. Tuts. **14**, 944–980 (2012).
[5] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," Phys. Rep. **650**, 1–63 (2016).
[6] S. Gao, J. Ma, Z. Chen, G. Wang, and C. Xing, "Ranking the spreading ability of nodes in complex networks based on local structure," Physica A **403**, 130–147 (2014).
[7] P. Holme, "Three faces of node importance in network epidemiology: Exact results for small graphs," Phys. Rev. E **96**, 062305 (2017).
[8] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," Nat. Phys. **6**, 888–893 (2010).
[9] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," Physica A **391**, 1777–1787 (2012).

[10]D. B. Chen, H. Gao, L. Lü, and T. Zhou, "Identifying influential nodes in large-scale directed networks: The role of clustering," PLoS One **8**, e77455 (2013).
[11]L. Lü, T. Zhou, Q.-M. Zhang, and H. E. Stanley, "The h-index of a network node and its relation to degree and coreness," Nat. Commun. **7**, 10168–10168 (2016).
[12]P. Hage and F. Harary, "Eccentricity and centrality in networks," Soc. Netw. **17**, 57–63 (1995).
[13]G. Sabidussi, "The centrality index of a graph," Psychometrika **31**, 581–603 (1966).
[14]L. Katz, "A new status index derived from sociometric analysis," Psychometrika **18**, 39–43 (1953).
[15]K. Stephenson and M. Zelen, "Rethinking centrality: Methods and examples," Soc. Netw. **11**, 1–37 (1989).
[16]L. C. Freeman, "A set of measures of centrality based on betweenness," Sociometry **40**, 35–41 (1977).
[17]P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," J. Math. Sociol. **2**, 113–120 (1972).
[18]S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," Web Conf. **30**, 107–117 (1998).
[19]L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the *delicious* case," PLoS One **6**, e21202 (2011).
[20]J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," J. ACM **46**, 604–632 (1999).
[21]F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," Nature **524**, 65–68 (2015).
[22]S. Mugisha and H.-J. Zhou, "Identifying optimal targets of network attack by belief propagation," Phys. Rev. E **94**, 12305 (2016).
[23]P. Clusella, P. Grassberger, F. J. Perez-Reche, and A. Politi, "Immunization and targeted destruction of networks using explosive percolation," Phys. Rev. Lett. **117**, 208301 (2016).
[24]E. A. Enns and M. L. Brandeau, "Link removal for the control of stochastically evolving epidemics over networks: A comparison of approaches," J. Theor. Biol. **371**, 154–165 (2015).
[25]H. Bast, S. Funke, P. Sanders, and D. Schultes, "Fast routing in road networks with transit nodes," Science **316**, 566–566 (2007).
[26]C. Giuraniuc, J. Hatchett, J. Indekeu, M. Leone, I. Castillo, B. V. Schaeybroeck, and C. Vanderzande, "Trading interactions for topology in scale-free networks," Phys. Rev. Lett. **95**, 98701 (2005).
[27]M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proc. Natl. Acad. Sci. U.S.A. **99**, 7821–7826 (2002).
[28]X.-Q. Cheng, F.-X. Ren, H.-W. Shen, Z.-K. Zhang, and T. Zhou, "Bridgeness: A local index on edge significance in maintaining global connectivity," J. Stat. Mech. Theory Exp. **2010**, 10011 (2010).
[29]J.-P. Onnela, J. Saramäki, J. Hyvönen, G. I. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," Proc. Natl. Acad. Sci. U.S.A. **104**, 7332–7336 (2007).
[30]V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," ACM Comput. Surv. **49**, 69 (2017).
[31]A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," Science **286**, 509–512 (1999).
[32]X.-L. Ren, N. Gleinig, D. Tolic, and N. Antulov-Fantulin, "Underestimated cost of targeted attacks on complex networks," Complexity **2018**, 1–15 (2018).
[33]P. Erd and A. R Nyi, "On the evolution of random graphs," Trans. Am. Math. Soc. **286**, 257–274 (2011).
[34]D. J. Watts and S. H. Strogatz, "Collective dynamics of "small-world" networks," Nature **393**, 440–442 (1998).
[35]A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," Phys. Rev. E **78**, 46110 (2008).
[36]P. Jensen, M. Morini, M. Karsai, T. Venturini, A. Vespignani, M. Jacomy, J.-P. Cointet, P. Mercklé, and E. Fleury, "Detecting global bridges in networks," J. Complex Netw. **4**, 319–329 (2016).
[37]L. Šubelj and M. Bajec, "Robust network community detection using balanced propagation," Eur. Phys. J. B **81**, 353–362 (2011).
[38]J. Moody, "Peer influence groups: Identifying dense clusters in large networks," Soc. Netw. **23**, 261–283 (2001).
[39]See http://konect.uni-koblenz.de/downloads/ for "Konect."