Link prediction for tree-like networks

Cite as: Chaos **29**, 061103 (2019); https://doi.org/10.1063/1.5107440
Submitted: 29 April 2019 . Accepted: 23 May 2019 . Published Online: 25 June 2019

Ke-ke Shang, Tong-chen Li, Michael Small, David Burton, and Yan Wang

Link prediction for tree-like networks

Cite as: Chaos **29**, 061103 (2019); doi: 10.1063/1.5107440
Submitted: 29 April 2019 · Accepted: 23 May 2019 ·
Published Online: 25 June 2019

Ke-ke Shang,[1,2,a)] Tong-chen Li,[1] Michael Small,[3,4,b)] David Burton,[5] and Yan Wang[1]

AFFILIATIONS

[1] Computational Communication Collaboratory, Nanjing University, Nanjing 210093, People's Republic of China
[2] College of Big Data and Intelligent Engineering, Yangtze Normal University, Chongqing 408100, People's Republic of China
[3] Complex Systems Group, Department of Mathematics and Statistics, The University of Western Australia, Crawley, Western Australia 6009, Australia
[4] Mineral Resources, CSIRO, Kensington, Western Australia 6151, Australia
[5] Western Australia Water Corporation, Leederville, Western Australia 6007, Australia

[a)] **Electronic addresses:** kekeshang@nju.edu.cn and keke.shang.1989@gmail.com
[b)] **URL:** http://school.maths.uwa.edu.au/~small/

ABSTRACT

Link prediction is the problem of predicting the location of either unknown or fake links from uncertain structural information of a network. Link prediction algorithms are useful in gaining insight into different network structures from partial observations of exemplars. However, existing link prediction algorithms only focus on regular complex networks and are overly dependent on either the closed triangular structure of networks or the so-called preferential attachment phenomenon. The performance of these algorithms on highly sparse or treelike networks is poor. In this letter, we proposed a method that is based on the network heterogeneity. We test our algorithms for three real large sparse networks: a metropolitan water distribution network, a Twitter network, and a sexual contact network. We find that our method is effective and performs better than traditional algorithms, especially for the Twitter network. We further argue that heterogeneity is the most obvious defining pattern for complex networks, while other statistical properties failed to be predicted. Moreover, preferential attachment based link prediction performed poorly and hence we infer that preferential attachment is not a plausible model for the genesis of many networks. We also suggest that heterogeneity is an important mechanism for online information propagation.

Traditional link prediction algorithms aim to predict the likely existence and location of unobserved links or edges in regular complex networks. These algorithms are strongly dependent on assumptions concerning the structure of networks—either the closed triangular structure or the so-called preferential attachment phenomenon (the new nodes preferential link to popular or high-degree nodes). However, many real-world systems are naturally represented as treelike networks with almost no closed triangular structure. Conversely, we note that the degrees or popularity of network nodes are often heterogeneous—so-called network heterogeneity. We propose a method that is based on network heterogeneity, and we show that our method is effective and performs better than traditional algorithms in three real large sparse networks. The technique appears to be particularly powerful for our archetypal social network—derived from twitter interactions. We further argue that heterogeneity is the prime rule for all kinds of complex networks, while other statistical properties failed to be predicted. Moreover, preferential attachment based link prediction performed poorly for all treelike networks and almost all regular networks. Our results not only broaden the scope of the link prediction problem, but also further demonstrate that preferential attachment is not sufficient as an explanatory model for many complex networks.

I. INTRODUCTION

Link prediction is the problem of predicting the location of either (unknown) unobserved or fake links[1] based on the statistical properties of the network structure (see Appendix A 1), that is, either observed links that should not be or unobserved links that should. Predicting which acquaintances would themselves be friends is a natural problem both in human interaction and also in expert recommender systems. It is at the very core of the famous six-degrees of separation experiment of Stanley Milgram.[2,3] Mathematically, however, the problem statement is incomplete and, therefore, intractable. In link prediction, what really should be asked is: given that a partially observed network is presumed to be consistent with some clearly articulated "prior,"[4] which link modifications would make that network more typical of that hypothesis?

Nonetheless, the existing solutions to the link prediction problem rely either on the assumption that friendship is assortative (my friends are more likely to be friends themselves) or the supremacy of the Barabasi-Albert preferential attachment (*BA*) model[5] (and its essential consequence that hubs gather in a central rich club[6,7]). Coupled to this, it is a simple statistical fact that link prediction becomes increasingly straightforward when links are more prevalent. Link prediction on denser networks is easier. Existing link prediction algorithms, therefore, adopt within their prior either a prevalence of triangles in the network (the assortativity of friendship) or a strong rich-club core and mild disassortativity on the leaves (a consequence of preferential attachment realized on a finite graph). The algorithms work best when the networks are densest.

In this paper, we propose a solution for when triads of connections or preferential attachment are insufficient. Moreover, we find that this method works extremely well in predicting missing links for the difficult case when the network is sparse (and links are rare). We apply this to real world networks and show that our solution out-performs existing methods when the underlying network structure deviates from the simplistic ideal. Of course, the limiting case of an ultimately sparse network is a tree. Ideal trees are not naturally amenable to link prediction—one would need to predict both links and "nodes." Nonetheless, we show that for a large class of treelike networks (utility distribution or transportation networks being a natural example), the link prediction schemes proposed here continue to perform very well.

In addition to finding hidden structure or identifying false connections, our method solves a second dual problem. The prior described above can also be formulated as a null hypothesis. The effectiveness of the link prediction performed on judiciously selected subsets of the entire network then becomes a statistical test of the appropriateness of that null hypothesis to describe the observed data—much as surrogate data methods have found popularity in nonlinear time series analysis.[8] When we do so, we are able to provide direct statistical evidence that preferential attachment is (or perhaps is not) a poor model of many real systems which otherwise appear to be scale-free. Of course, a scale-free degree distribution need not imply preferential attachment as a generative mechanism. What our results show is that while scale-free degree distributions may still be common, often the cause is not preferential attachment—or at least not preferential attachment alone.

In practical applications, link prediction may be applied as a useful expedient to assist in targeting an otherwise expensive experimental search. In biology, researchers allocate significant expense to recovering unknown interactions.[9,10] Fortunately, link prediction algorithms can help us identify unknown "potential" interactions to reduce the cost of experiment.[11–13] In social fields, link prediction algorithms may also aid in analyzing the network evolution.[14] On the other hand, link prediction algorithms are helpful for the algorithm design of recommender systems[15] and the spurious links detection problem.[16] In our recent studies, we used link prediction methods to analyze the triangle structure in several different kinds of networks.[17,18]

Moreover, recent studies focus on the so-called novel link prediction structures that are also based on the triangle structure.[19] Unfortunately, almost all previous methods rarely consider a special but substantial and important case—treelike networks, which are built by the open triangular structure. As depicted in Fig. 1, there are two typical treelike networks in the natural world—the engineering network and the human propagation network. The water distribution network is the typical treelike or tree network, and relate to human livelihood. We employ a metropolitan fresh water distribution network of the state utility provider for Perth, Western Australia,[20] to study the link prediction of the treelike network. Conversely, it is especially exciting, though challenging, to predict the behavior of systems as complex as humans, especially when the humans are interacting with each other in complex and unobserved ways.[14,21] Here, we also employ the open social treelike networks—a Twitter network from the Stanford Large Network Dataset Collection[22] and a well-known sexual contact network from Bearman *et al.*[23]—for our study. In addition, following our previous studies,[17,18] we also use the null model[24] to better understand the network characteristics.

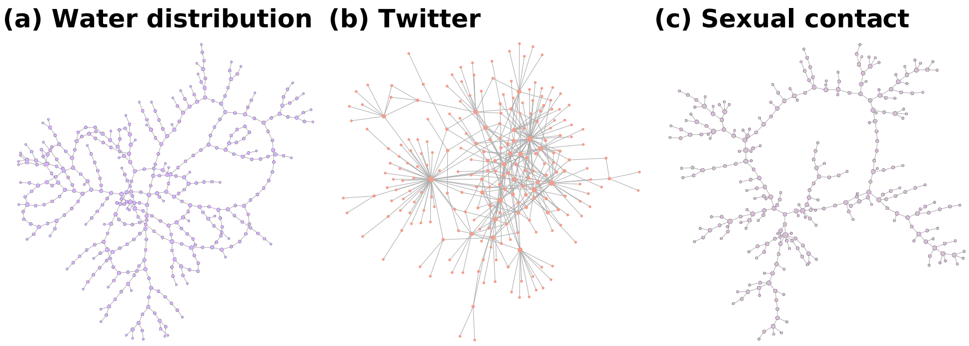

FIG. 1. Local structures of the engineering treelike network—Water distribution network (a) and the human propagation networks—Twitter network (b). The network structure of the Sexual Contact network (c). Almost no closed triangular structure in these networks. For the water distribution network: A node represents one pumping station, one residence, or one business building, and a link indicates that there is a pipe between them. Water distribution has 330 721 nodes and 343 721 links. For the Twitter network: A node represents one Twitter user. If one node has retweeted another node, there is a link between them. Twitter network has 256 491 nodes and 328 132 links. And for the sexual contact network: A node represents one student, a link indicates that there is a sexual contact between them. Sexual contact network has 288 nodes and 291 links.

In this paper, we analyze the AUC (Area Under Receiver Operating Characteristic Curve)[25] performance of the local link prediction algorithms, the quasilocal link prediction algorithms and the global link prediction algorithms[26] for the treelike networks. As we anticipate, the local link prediction algorithm and the quasilocal link prediction algorithm which are based on the closed triangular structure or on the preferential attachment principle both fail to predict the link existence of treelike networks, their performance remains around 0.5—that is, very similar to the random selection method. But for the global link prediction algorithm which is based on the global network structure,[27,28] we can achieve a better accuracy. These results show that the traditional link prediction algorithm which are based on the closed triangle structure have no effect on the treelike networks. The global link prediction algorithm which are based on the whole network information can achieve a definite effect.

However, the efficiency of global link prediction is accompanied by high computational complexity, and the so-called universal rules such as the preferential attachment principle[5] and the motifs of complex networks[29] both fail to predict. Hence, in turn, we pay attention to adopt the more universal principle—heterogeneity—to predict the sparse and treelike networks. Then, we propose a novel local link prediction algorithm to achieve a higher efficiency. We find that our heterogeneity algorithm can achieve the best performance in all treelike networks, especially for the information propagation network (Twitter). Finally, we use the randomized algorithm to test robust of the heterogeneity algorithm, and the results show that our algorithm can assure the robustness of accuracy.

II. METHODS

A. Metric for link prediction algorithm

A graph G can be described by a vertex set V, and an edge set E: $G = (V, E)$. Elements of the edge set are unordered pairs of elements of the vertex set: $e = (v_i, v_j) \in E$, where $v_i, v_j \in V$. The pair (v_i, v_j) occurs in at most one edge $e \in E$. The standard link prediction problem can be formulated as follows. The edge set E is divided into two parts: E^T and E^P, where $E^T \cup E^P = E$ and $E^T \cap E^P = \emptyset$. The division into E^P (typically including 10% of the observed links) and E^T (typically 90% of the observed links) is arbitrary and will be used for scoring purposes. That is, all the links in $E = E^P \cup E^T$ have been observed and are known, however, links in E^T will form a "training set" and are used to implement a link prediction score, the efficacy of which will be evaluated over the "probe set" E^P.

The Area under Receiver Operating Characteristic Curve (AUC),[30] originally applied to evaluate communication schemes, has since been widely applied to measure prediction accuracy[25] in a wide variety of settings. We use AUC as a link prediction accuracy measure for networks. Only the information of E^T is allowed to be used to compute the performance score $Score_{xy}$, we compare the prediction scores of m pairs of nodes from E^P and $\bar{E}$ randomly, if there are m' times that the score measured from E^P is bigger than the score measured from $\bar{E}$ and m'' times that the two scores are equal, then, $AUC = (m' + 0.5m'')/m$ (see Appendix A 2). Here, an AUC value closer to 1 means that the link prediction method is more efficient. Moreover, the AUC values are determined by the relationship between the network structure and the link prediction algorithm.

B. Link prediction algorithms for treelike networks

Obviously, the higher score that is computed by AUC means a better link prediction algorithm. Network scientists have proposed around 30 traditional algorithms,[26] which are based on various attributes of the network structure. Almost all link prediction algorithms are related to the closed triangular structure. In this paper, we employ 13 traditional link prediction algorithms as a baseline for comparison. We then compare these to our proposed heterogeneity algorithms in view of the weaknesses of the traditional algorithms. For the networks with more degree heterogeneity and lesser degree homogeneity, we propose the heterogeneity index (HEI),

$$S_{ij}^{HEI} = |k(i) - k(j)|^{\alpha}, \tag{1}$$

where $k(i)$ is the degree of node i and $k(j)$ is the degree of node j, α is a free heterogeneity exponent. Inversely, for the networks with more degree homogeneity and lesser degree heterogeneity, we propose the homogeneity index (HOI),

$$S_{ij}^{HOI} = \frac{1}{|k(i) - k(j)|^{\alpha}}. \tag{2}$$

In addition, for the more complex networks, we can further combine HEI and HOI algorithms, and we named it as heterogeneity adaption index (HAI),

$$S_{ij}^{HAI} = \alpha|k(i) - k(j)| + (1 - \alpha)\frac{1}{|k(i) - k(j)|}, \tag{3}$$

where $0 \leq \alpha \leq 1$. In this letter, we only applied HEI for treelike networks to outline our principle.

C. Traditional link prediction algorithms

The friend of our friend is our friend also, as is figured by a closed triangular structure. This common intuition is the basis of all local link prediction algorithms except the preferential attachment index. Newman *et al.* firstly use this rule to study the cooperation behaviors of scientists,[31] then provide a foundation for link prediction problem.[21] Based on these studies, the well-known common neighbors index (CN) has been proposed,

$$S_{ij}^{CN} = |\Gamma(i, j)|, \tag{4}$$

where $\Gamma(i, j)$ denotes the set of common neighbors of the nodes i and j. The algorithm indicates that if you have a common friend with another person, there is a possible relationship between you. That is to say the friend (common neighbors) of friend is our friend.

Resource allocation index (RA) is based on the principle of the resource allocation.[32,33] which also use the closed triangular structure (common neighbors),

$$S_{ij}^{RA} = \sum_{z \in \Gamma(i,j)} \frac{1}{k(z)}, \tag{5}$$

where $k(z)$ is the degree of the node z.

Adamic-Adar index (AA)[14] improves the effect of the lower-degree common neighbors, which is defined as

$$S_{ij}^{AA} = \sum_{z\in\Gamma(i,j)} \frac{1}{\log k(z)}. \tag{6}$$

Jaccard index[34] is the earliest local link prediction algorithm, which is proposed by Jaccard in 1901, and also based on the role of common neighbors,

$$S_{ij}^{Jaccard} = \frac{|\Gamma(i,j)|}{|\Gamma(i)\cup\Gamma(j)|}, \tag{7}$$

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

Sørensen index[35] has been widely applied in the ecological community data, which is also use the role of common neighbors,

$$S_{ij}^{Sørensen} = \frac{2|\Gamma(i,j)|}{k(i)+k(j)}. \tag{8}$$

Hub depressed index (HDI)[36] is associated with the hub promoted index (HPI),[37] both of them are based on the role of common neighbors. The hub promoted index aims at the improving of hub nodes effects,

$$S_{ij}^{HPI} = \frac{|\Gamma(i,j)|}{\min\{k(i)+k(j)\}}. \tag{9}$$

On the contrary, the hub depressed index aims at the decreasing of hub nodes effects,

$$S_{ij}^{HDI} = \frac{|\Gamma(i,j)|}{\max\{k(i)+k(j)\}}. \tag{10}$$

Leicht-Holme-Newman index ($LHN1$)[38] directly compare the number of common neighbors and the possible maximum number of that

$$S_{ij}^{HDI} = \frac{|\Gamma(i,j)|}{k(i)\times k(j)}. \tag{11}$$

Preferential attachment index (PAI)[5,36] is based on the preferential attachment principle, the probability of a new link connect to node i is proportional to k_i,[5] and the probability of a new link connect to node j is proportional to k_j. Hence, the probability that there is a link between i and j is proportional to $k_i \times k_j$, then,

$$S_{ij}^{PAI} = k(i)\times k(j). \tag{12}$$

Local path index (LP)[32,39] is a typical quasilocal link prediction algorithm, which not only adopts the paths between node i and node j with 2 steps (CN) and further adopts that with 3 steps,

$$S_{ij}^{LP} = A^2 + \alpha A^3, \tag{13}$$

where A is the adjacency matrix of the network. $(A^2)_{ij}$ is equal to the number of all paths within 2 steps that connect i and j. Obviously, this algorithm will be simplified into CN when $\alpha = 0$.

The Katz index[40] is one of the earliest global link prediction algorithm, which combine the role of common neighbors and global information,

$$S_{ij}^{Katz} = \sum_{l=1}^{\infty} \alpha^l \cdot |path_{ij}^{\langle l\rangle}| = \alpha A_{ij} + \alpha^2 (A^2)_{ij} + \alpha^3 (A^3)_{ij} + \cdots, \tag{14}$$

where α is a free exponent of path weights, and $path_{ij}^{\langle l\rangle}$ denotes all paths with length l that connect i and j. When α is lower than the reciprocal of the matrix (A) largest eigenvalue, the equation can be simplified as

$$S_{ij}^{Katz} = (I-\alpha A)^{-1} - I. \tag{15}$$

Average commute time index (ACT)[27,28] adopt the global information and the principle of random walk,

$$n(x,y) = m(i,j) + m(j,i), \tag{16}$$

where $m(,i,j)$ means the average number of steps from node i to node j via random walk. Here, we can use D which denotes the degree matrix of the network, then the Laplacian matrix $L = D - A$, and L^+ denotes the pseudoinverse of L. Previous studies[27,28] deduced that

$$n(x,y) = M(l_{ii}^+ + l_{jj}^+ + 2l_{ij}^+), \tag{17}$$

where l_{ij}^+ denotes the entry of L^+, and the entry position is ij. Hence, the equation can be simplified as

$$S_{ij}^{ACT} = \frac{1}{l_{ii}^+ + l_{jj}^+ - 2l_{ij}^+}. \tag{18}$$

III. RESULTS

A. The failure of traditional algorithms

Research into the accuracy of the link prediction algorithm has recently attracted increasing attention.[13,18,19,21] However, none directly explore the definition, resolution, and boundary of the link prediction problem. Recently, research on network structures has provided a new perspective on the nature of the link prediction problem.[17] For real world systems, sparse and treelike networks are ubiquitous. In this paper, we begin by testing the traditional local algorithms—with lower computation complexity—for the treelike networks.

1. Local indices

Obviously, the most famous link prediction algorithm—common neighbors algorithm CN[21]—is based on the closed triangle structure. On the other hand, the earliest link prediction algorithm "Jaccard" index[34] is also based on the closed triangle structure. Actually, all later local link prediction algorithms for the static network—evolutionary algorithms of CN or Jaccard—are also based on the closed triangle structure. Hence, as we expected, all traditional link prediction algorithms fail to predict the water distribution network and the sexual contact network (Table I), due to their AUC performance that is similar to that of the randomly chosen method (around 0.5). Especially for the PA index[26] which is based on the principle of the preferential attachment, the AUC performance is exceptionally poor. Furthermore, all traditional link prediction algorithms have a bit prediction effect for the Twitter network except the PA index. The PA index also has the worst performance in other complex networks.[26] We argue

TABLE I. For all networks, the performance of the traditional local link prediction algorithms with the metric of *AUC*. Bold numbers are the worst performance of all algorithms. In this letter, we compute the *AUC* of all algorithms 100 times independently.

AUC	*CN*	*RA*	*AA*	*Jaccard*	*Salton*	Sørensen	*HPI*	*HDI*	*LHN*1	*PA*
Water distribution	0.5011	0.5020	0.5010	0.5012	0.5000	0.5030	0.5021	0.5020	0.5041	**0.1400**
Twitter	0.5382	0.5291	0.5401	0.5400	0.5332	0.5381	0.5410	0.5210	0.5441	**0.4548**
Sexual contact	0.4942	0.4984	0.4943	0.4909	0.4978	0.4952	0.4948	0.4960	0.4980	**0.3392**

that the so-called preferential attachment is not the most popular rule as we recognized.

2. Quasilocal and global indices

Next, we test the traditional quasilocal and global algorithms—algorithms with higher computation complexity—for sparse and treelike networks. In particular, global algorithms which are based on the global structure information are extremely computationally expensive. As shown in Table II, the *LP*[32,39] index and the *Katz*[40] which also associate with the common neighbors (closed triangular structure) cannot predict the all treelike networks well, but the *ACT*[27,28] index which only associates with the global structure is significantly better than other traditional algorithms and gets the best performance for all treelike networks.

B. The heterogeneity index

We now propose to test our novel local algorithms for the treelike network. Our three novel methods have similar best performance, hence we only show the results of *HEI* to state our algorithm principle—the heterogeneity network structure. As depicted in Fig. 2, with the variation of the heterogeneity parameter α, *HEI* can achieve a better performance than the traditional link prediction algorithms, in general, especially for the Twitter network. Here, we infer that the high performance for the Twitter network is due to a number of hub nodes attracting most of the links, which then forms a strong heterogeneous network. For the water distribution network and the sexual contact network, the *ACT* performance is similar to that of *HEI*. However, the *ACT* is infeasible for the large datasets. *ACT* takes too much time to get the result, hence we only can get the *ACT* approximate performance by the sampled data of the water distribution network. To this end, we suggest that our algorithm is more suitable for the treelike network than traditional link prediction algorithms.

TABLE II. For all networks, the performance of the traditional quasilocal and global link prediction algorithms with the metric of *AUC*. Bold numbers are the best performance of all algorithms. Here, we randomly choose a sample set for the *Katz* and the *ACT* every time (see Appendix B 2). For *LP* and *Katz*, we only show the best performance of them.

AUC	*LP*	*Katz*	*ACT*
Water distribution	0.5049	0.5026	**0.5509**
Twitter	0.6440	0.4551	**0.8118**
Sexual contact	0.5031	0.4370	**0.6450**

To evaluate the reliability of our algorithm, we introduce the randomized edges algorithm (null model)[24] to destroy the network structure. The structure of null model is totally random, while the node degree is maintained. As depicted in Fig. 3, the *HEI* still has the prediction effect for the null model, though the performance of it decrease slightly. Our study suggests that the heterogeneity is still the popular rule for complex networks while other rules loose efficacy for the structure prediction. Especially for the so-called preferential attachment principle, the heterogeneity will be emerged with the appearing of that rule, but not vice versa.

C. The limitation of the heterogeneity index

Compare the performance of our methods for three real-world treelike networks, we find that our methods are more suitable for more heterogeneous treelike network (Twitter network). Moreover, previous study proposed a scale-free network model with different power law degree distribution exponents[41] that also can create networks with different heterogeneities. On the other hand, Xulvi-Brunet and Sokolov's algorithm[42] can help us build null models with different heterogeneities.[43] Inspired by previous studies, we test our algorithms in three networks with different heterogeneities to further evaluate the reliability of our algorithm. Here, we built the *BA* network[5] with 100 nodes and 294 links, then adopt simple null model methods to create the corresponding homogeneity *BA* network and the heterogeneity *BA* network (see Appendix B 1). As depicted in Table III, we find our algorithms only can achieve a good performance for the networks with the higher heterogeneity (*BA* and heterogeneity *BA*) and achieve the best performance in the heterogeneity *BA* network. In addition, our algorithms have the worst performance in the homogeneity *BA* network though are effective ($AUC > 0.5$) for the homogeneity *BA* network. That is to say the performance of all our algorithms are in positive correlation with the heterogeneity. In other words, our methods are more suitable for the regular treelike networks with the higher heterogeneity.

TABLE III. For the homogeneity *BA* network, the *BA* network, and the heterogeneity *BA* network, the best performance of our link prediction algorithms with the metric of *AUC*. Boldface denotes the best performing network for each link prediction algorithm.

	Homogeneity *BA*	*BA*	Heterogeneity *BA*
HEI	0.5956	0.7340	**0.8414**
HOI	0.5980	0.7379	**0.8404**
AHI	0.6090	0.7369	**0.8331**

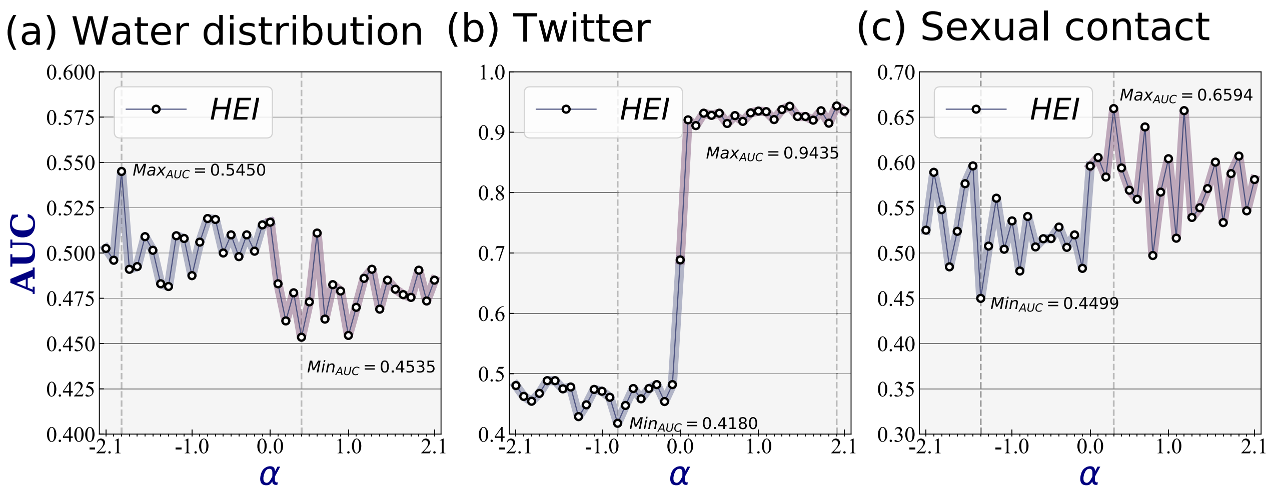

FIG. 2. Under the metric of *AUC*, the performance of *HEI* for all networks. The heterogeneity will play the main role when $\alpha > 0$ (b) and (c), and the homogeneity will play the main role when $\alpha > 0$ (a).

IV. DISCUSSION

Our study not only broadens the scope of the link prediction problem, but also argues that heterogeneity is the prime rule for many complex networks. The network structures of the real world are ever-changing and abundant due to one key factor—heterogeneity. Moreover, the treelike networks are exceedingly common in real systems. Consequently, we find that many real-world networks with complex structure and scaling exhibit structure and features that are atypical of what is provided by the preferential attachment alone. In this letter, we show an example to adopt the network to successfully predict the link existence. In social networks, automated link suggestions strongly influence the experience of social network users. Better prediction algorithms can improve ones' experience of social networks by enhancing interaction with the network. Our link prediction algorithm is especially suitable as a model for the online information propagation network. Hence, we suggest that heterogeneity is the driving force for the formation of the news propagation network, and our work will also benefit understanding and propagation of information in mass communication systems—the computational communication. Conversely, all algorithms that are based on or associated with the closed triangular structure will fail to predict the link for sparse or treelike networks, and the so-called

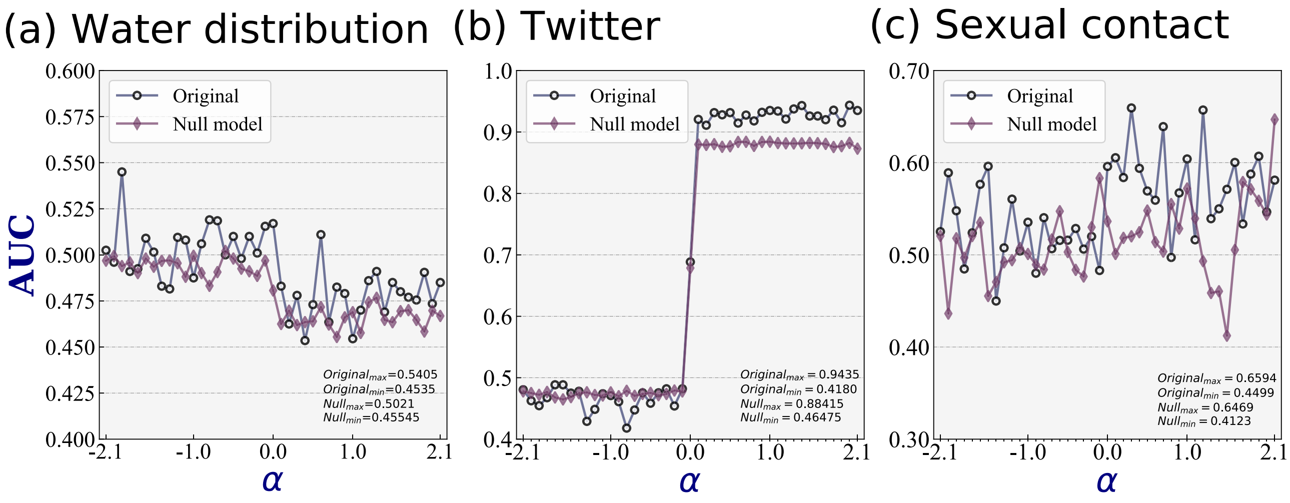

FIG. 3. Under the metric of *AUC*, the performance of *HEI* for all original networks and corresponding null models. The bottom right boxes show the best and worst performance for the original network and the corresponding null model.

preferential attachment rule can even act against favorable link prediction.

In addition, traditional link prediction algorithms in complex networks and our algorithms are only based on the observed network structure and suitable for the single layered network. On the other hand, the algorithm of community-based recommender systems can also help us analyze the similarity between two users,[44] that is, a special link prediction algorithm. Moreover, techniques such as Matrix factorization and Singular value decomposition can help us solve the data sparsity problem.[45] However, algorithms of community-based recommender system usually rely on the relevant features of users and analyze the relationships between users and items. We suggest that the algorithm of community-based recommender system is more suitable for the double layer network under the premise that users' public profiles are reliable.

ACKNOWLEDGMENTS

Ke-ke Shang is supported by the National Natural Science Foundation of China (NNSFC) (Grant No. 61803047) and the Tencent Research Institute. Michael Small is supported by ARC Discovery Project No. DP180100718. We would like to thank the data support of the Water Corporation, Western Australia's largest water utility. We are especially grateful to Professor Peter Bearman for sharing the sexual contact data.

APPENDIX A: LINK PREDICTION

1. Link prediction problem

A graph G can be described by a vertex set V, and an edge set E: $G = (V, E)$. Elements of the edge set are unordered pairs of elements of the vertex set: $e = (v_i, v_j) \in E$, where $v_i, v_j \in V$. The pair (v_i, v_j) occurs in at most one edge $e \in E$.

The standard link prediction problem can be formulated as follows. The edge set E is divided into two parts E^T and E^P, where $E^T \cup E^P = E$ and $E^T \cap E^P = \emptyset$. The division into E^P (typically including 10% of the observed links) and E^T (typically 90% of the observed links) is arbitrary and will be used for scoring purposes. That is, all the links in $E = E^P \cup E^T$ have been observed and are known, however, links in E^T will form a "training set" and are used to implement a link prediction score, the efficacy of which will be evaluated over the "probe set" E^P.

With sets E^T and E^P, the "static" link prediction problem is then applied where we consider an augmentation of G. Suppose that the information encapsulated in graph G provides an incomplete picture of a larger graph $G' = (V, E')$—the "truth." This larger graph G' may include both edges in E and also additional edges $\bar{E} = E' \backslash E$. In the real-world, these are additional edges that have not been observed. Let $U := \{(v_i, v_j) | v_i, v_j \in V,\ v_i \neq v_j\}$ be the universal set of all possible links (links are bidirectional and hence technically U should contain each link only once, so we will impose some ordering on the elements of V and insist—for example—that $i < j$).

For each link in U, we define and compute a prediction measure S_{ij} which measures—based only on link information contained in E^T—how close are nodes v_i and v_j. That is, the probability of nodes v_i and v_j has a link connecting each other is assessed as S_{ij} using information inferred from the links in E^T. Using the additional links in E^P, we can compute a performance score for the prediction measure (one can imagine many possible alternative prediction measures)—that is, based on the known links of E^P ("known structure")—how well correlated are the scores S_{ij} with edges $(v_i, v_j, w(i, j)) \in E^T$: are high values of S_{ij} associated with membership of E^P for edges in $E'' = U \backslash E^T$? This step (evaluating the score on E^T) is not strictly necessary, but provides a method to test how well our algorithm performs before moving to the unseen data in $\bar{E}$. Finally, our predictions of the unknown links in $\bar{E}$ can be obtained by ranking the scores S_{ij} for all $(v_i, v_j, w(i, j)) \in \bar{E}(U \backslash E)$. Links with highly ranked scores are those predicted to most likely exist—we expect that these highly ranked links will probably occur in $\bar{E}$, and other few lowly ranked links will probably occur in E^P.

The static link prediction problem can now be stated: given E^T and E^P (and also V), predict $\bar{E}$ and some fake links in E^P. That is, if we know some of the links of a network—those links being partitioned into the training set E^T and the probe set E^P—which we have observed, is it possible to predict the existence (or otherwise) of unobserved or fake links. The unobserved links are members of E'' and may be said to either *exist* (if they are also members of $\bar{E}$) or be nonexistent [if they are instead members of $U \backslash (E \cup \bar{E})$]. Of course, in general, the link prediction problem is ill-posed. If the links (network structure) are random and uncorrelated the information in E tells us nothing about any of the remaining possible pairs E'' and whether they are in $\bar{E}$. However, many real-world networks exhibit correlation among the links, on the contrary, the link prediction performance can reflect the network structural properties. In addition, we can use the link prediction performance to measure the variation of network structure.

For a graph $G = (V, E)$, the vertex set is stationary and does not evolve. Elements of the edge set are unordered pairs of elements of the vertex set: $e = (v_i, v_j) \in E$, where $v_i, v_j \in V$. The pair (v_i, v_j) occurs in at most one edge $e \in E$. The edge set E is divided into two parts E^T and E^P, where $E^T \cup E^P = E$ and $E^T \cap E^P = \emptyset$. The division into E^P (typically including 10% of the observed links in Ref. 36) and E^T (typically 90% of the observed links) is arbitrary and will be used for scoring purposes. The static[46] link prediction problem can be stated: given the training link set E^T and the probe link set E^P (and also V), $E'' = U \backslash E$, then predict a small part of unobserved links in E'' and that of fake links in E^P. That is, if we know some of the links of a network—those links being partitioned into the *training set* E^T and the probe set E^P—which we have observed, is it possible to predict the existence (or otherwise) of unobserved or fake links. The unobserved links are members of E'' and may be said to either exist or be nonexistent. Generally, the link prediction scores of existence links are bigger than that of fake links or nonexistent links.

2. A simple example for AUC

As shown in Fig. 4, as our previous study,[43] we describe a simple example to explain the *AUC* schematic. Here, we adopt the link prediction algorithm *PA* index and only compute the *AUC* once ($m = 1$): First, the observed link set (a) can be divided into a training set (b) and a probe set (c). Next, we can choose a pair of nodes *AB* from the probe set (c) and a pair of nodes from the nonobserved links set (d). Only the training set (b) can be used to compute the node degree. If we choose the pair of nodes *AD* from the nonobserved links set (d), $Score_{AB} = k_A \times k_B = 0$, $Score_{AD} = k_A \times k_D = 1$, namely, $Score_{AB}$

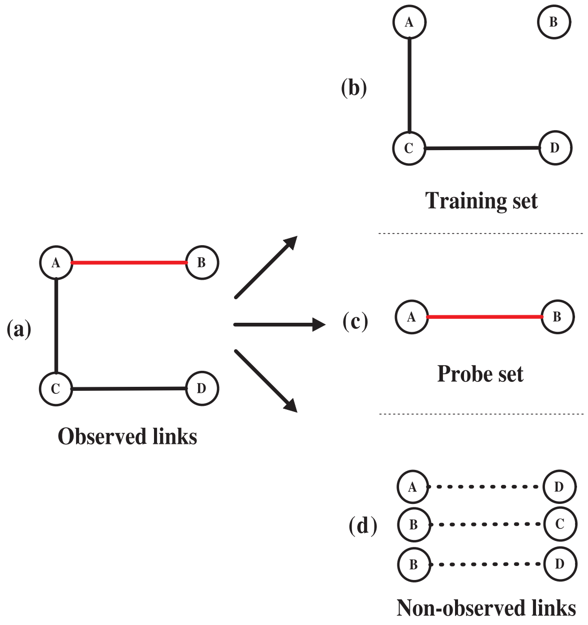

FIG. 4. The simple examples of the observed links set E, the nonobserved links set $\bar{E}$, the training set E^T, and the probe set E^P. Reproduced with permission from K. Shang *et al.*, Phys. A Stat. Mech. Appl. **474**, 49–60 (2017). Copyright 2017 Elsevier B.V.

$< Score_{AD}$, hence $m' = 0, m'' = 0$, then $AUC = (0 + 0.5 \times 0)/1 = 0$. If we choose the pair of nodes BD from the nonobserved links set (d), $Score_{AB} = k_A \times k_B = 0, Score_{BD} = k_B \times k_D = 0$, namely, $Score_{AB} = Score_{BD}$, hence $m' = 0, m'' = 1$, then $AUC = (0 + 0.5 \times 1)/1 = 0.5$.

APPENDIX B: NULL MODEL AND SAMPLING METHOD

1. Null models

Our aim with the null model is to destroy the original network structure, while maintaining the node degree. We can use the randomized edges algorithm (*RE*)[24] to rewire links of the original network. As shown in Fig. 5,[18] nodes *A* and *B* and nodes *C* and *D* are connected, while nodes *A* and *D* and nodes *B* and *C* are not connected. Then, cut the links *AB* and *CD*, and connect nodes *A* and *D* and nodes B and C, respectively. *RE* changes the structure of the network and the degree of each node is kept. *RE* is helpful for us to analyze the role of structure for link prediction. This is a standard edge switch, which will ensure that the network degree sequence remains unchanged. In this letter, the number of switches is equal to the number of nodes.

Furthermore, based on Xulvi-Brunet and Sokolov's algorithm,[42] as shown in Fig. 6,[43] we can use the disassortative *RE* method (*DARE*) and the assortative *RE* method (*ARE*) to change the structure and the assortativity or heterogeneity of the original *BA* networks. Obviously, the *DARE* will improve the heterogeneity of the original *BA* networks, and the *ARE* will play the contrary effect. The number of switches is also equal to the number of nodes.

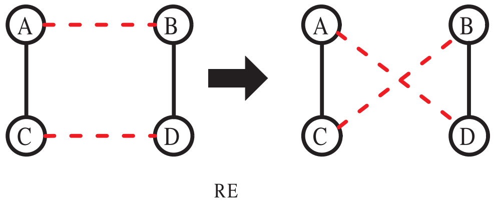

FIG. 5. Nodes *A* and *B* and nodes *C* and *D* are connected, while nodes *A* and *D* and nodes *B* and *C* are not connected. Then, cut the links *AB* and *CD*, and connect nodes *A* and *D* and nodes *B* and *C*, respectively. Reproduced with permission from K. Shang *et al.*, Europhys. Lett. **117**, 28002 (2017). Copyright 2017 EPLA.

2. Sampling method

Compared to other two kinds of link prediction algorithms, the global link prediction algorithm has the highest computational complexity and almost infeasible without a large workstation. Hence, to reduce computational load, we choose a sample set E_{sample} from original networks. Here, to select a suitable sample which has the proper density and number of nodes, we choose a node randomly, then only select some of its neighbors and neighbors of neighbors, and the links between these nodes as the sample set E_{sample}.[18] In this letter, the number of nodes in E_{sample} is 500, and the E_{sample} is only used for *Katz* and *ATC*.

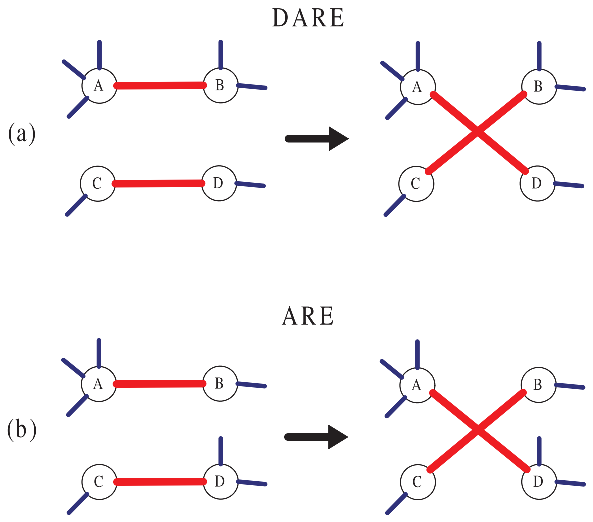

FIG. 6. (a) Nodes *A* and *B* and nodes *C* and *D* are connected, while nodes *A* and *D* and nodes B and C are not connected. In addition, $|k_A - k_B| < |k_A - k_D|$ and $|k_C - k_D| \leq |k_C - k_B|$, where k_i is the degree of node *i*. Then, cut the links *AB* and *CD*, and connect the nodes *A* and *D* and the nodes *B* and *C*, respectively. (b) Nodes *A* and *B* and nodes *C* and *D* are connected, while nodes *A* and *D* and nodes *B* and *C* are not connected. In addition, $|k_A - k_B| > |k_A - k_D|$ and $|k_C - k_D| \geq |k_C - k_B|$, where k_i is the degree of node *i*. Then, cut the links *AB* and *CD*, and connect nodes *A* and *D* and nodes *B* and *C*, respectively. Reproduced with permission from K. Shang *et al.*, Phys. A Stat. Mech. Appl. **474**, 49–60 (2017). Copyright 2017 Elsevier B.V.

REFERENCES

[1] For example, many users who follow each other in online social networks not only do not know each other in the offline world, but also have no communication in their online life. Such observed links can be stated as the fake link.

[2] S. Milgram, "The small world problem," Psychol. Today. **2**, 60–67 (1967).

[3] J. Travers and S. Milgram, "An experimental study of the small world problem," Sociometry **32**, 425–443 (1969).

[4] We disclaim at the outset that, despite adopting the language of Bayesians, the method itself does not at any point require a Bayesian framework. The discussion here, however, does suggests an alternative —purely Bayesian—solution to the same problem.

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

[6] M. Small, Y. Li, T. Stemler, and K. Judd, "Growing optimal scale-free networks via likelihood," Phys. Rev. E **91**, 042801 (2015).

[7] X. Xu, J. Zhang, and M. Small, "Rich-club connectivity dominates assortativity and transitivity of complex networks," Phys. Rev. E **82**, 046117 (2010).

[8] S. Michael, *Applied Nonlinear Time Series Analysis: Applications in Physics, Physiology and Finance* (World Scientific, 2005), Vol. 52.

[9] S.-Y. Takemura, A. Bharioke, Z. Lu, A. Nern, S. Vitaladevuni, P. K. Rivlin, W. T. Katz, D. J. Olbris, S. M. Plaza, and P. Winston *et al.*, "A visual motion detection circuit suggested by *Drosophila* connectomics," Nature **500**, 175–181 (2013).

[10] D. D. Bock, W.-C. A. Lee, A. M. Kerlin, M. L. Andermann, G. Hood, A. W. Wetzel, S. Yurgenson, E. R. Soucy, H. S. Kim, and R. C. Reid, "Network anatomy and *in vivo* physiology of visual cortical neurons," Nature **471**, 177–182 (2011).

[11] S. Redner, "Networks: Teasing out the missing links," Nature **453**, 47–48 (2008).

[12] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," Nature **453**, 98–101 (2008).

[13] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley, "Toward link predictability of complex networks," Proc. Natl. Acad. Sci. U.S.A. **112**, 2325–2330 (2015).

[14] L. A. Adamic and E. Adar, "Friends and neighbors on the web," Soc. Netw. **25**, 211–230 (2003).

[15] Z.-K. Zhang, C. Liu, Y.-C. Zhang, and T. Zhou, "Solving the cold-start problem in recommender systems with social tags," Europhys. Lett. **92**, 28002 (2010).

[16] A. Zeng and G. Cimini, "Removing spurious interactions in complex networks," Phys. Rev. E **85**, 036101 (2012).

[17] K. Shang, W. S. Yan, and M. Small, "Evolving networks—Using past structure to predict the future," Phys. A Stat. Mech. Appl. **455**, 120–135 (2016).

[18] K. Shang, M. Small, X. K. Xu, and W. S. Yan, "The role of direct links for link prediction in evolving networks," Europhys. Lett. **117**, 28002 (2017).

[19] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, "Simplicial closure and higher-order link prediction," Proc. Natl. Acad. Sci. U.S.A. **115**, E11221–E11230 (2018).

[20] A. Ballantyne, N. Lawrance, M. Small, M. Hodkiewicz, and D. Burton, "Fault prediction and modelling in transport networks," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2018), pp. 1–5.

[21] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," J. Am. Soc. Inf. Sci. Technol. **58**, 1019–1031 (2007).

[22] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," 2014.

[23] P. S. Bearman, J. Moody, and K. Stovel, "Chains of affection: The structure of adolescent romantic and sexual networks," Am. J. Sociol. **110**, 44–91 (2004).

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

[25] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology **143**, 29 (1982).

[26] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," Phys. A Stat. Mech. Appl. **390**, 1150–1170 (2011).

[27] D. J. Klein and M. Randić, "Resistance distance," J. Math. Chem. **12**, 81–95 (1993).

[28] F. Fouss, A. Pirotte, J. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," IEEE Trans. Knowl. Data. Eng. **19**, 355–369 (2007).

[29] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," Science **303**, 1538–1542 (2004).

[30] The abscissa stands for the false positive rate and the ordinate stands for the true positive rate, then we can draw a Receiver Operating Characteristic Curve (*ROC*). Statistically, the area under the ROC should be between 0.5 and 1. If the area is greater than 0.5, we can suggest that our method is effective. If the area equals 0.5, then our method is invalid. The case that the area is less than 0.5, is unrealistic.

[31] M. E. Newman, "Clustering and preferential attachment in growing networks," Phys. Rev. E **64**, 025102 (2001).

[32] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," Eur. Phys. J. B **71**, 623–630 (2009).

[33] Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, and B.-Q. Yin, "Power-law strength-degree correlation from resource-allocation dynamics on weighted networks," Phys. Rev. E **75**, 021102 (2007).

[34] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," Bull. Soc. Vaudoise Sci. Nat. **37**, 547–579 (1901).

[35] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons," Biol. Skr **5**, 1–34 (1948).

[36] L. Lü and T. Zhou, "Link prediction in weighted networks: The role of weak ties," EPL **89**, 18001 (2010).

[37] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," Science **297**, 1551–1555 (2002).

[38] E. A. Leicht, P. Holme, and M. E. Newman, "Vertex similarity in networks," Phys. Rev. E **73**, 026120 (2006).

[39] L. Lü, C.-H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks," Phys. Rev. E **80**, 046122 (2009).

[40] L. Katz, "A new status index derived from sociometric analysis," Psychometrika **18**, 39–43 (1953).

[41] F. Chung, L. Lu, and V. Vu, "Eigenvalues of random power law graphs," Ann. Comb. **7**, 21–33 (2003).

[42] R. Xulvi-Brunet and I. M. Sokolov, "Changing correlations in networks: Assortativity and dissortativity," Acta Phys. Polonica B **36**, 1431 (2005).

[43] K. Shang, M. Small, and W.-S. Yan, "Fitness networks for real world systems via modified preferential attachment," Phys. A Stat. Mech. Appl. **474**, 49–60 (2017).

[44] F. Gasparetti, A. Micarelli, and G. Sansonetti, "Community detection and recommender systems," in *Encyclopedia of Social Network Analysis and Mining*, edited by R. Alhajj and J. Rokne (Springer, New York, NY, 2017), pp. 1–14.

[45] G. Zhao, M. L. Lee, W. Hsu, W. Chen, and H. Hu, "Community-based user recommendation in uni-directional social networks," in *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management, CIKM'13* (ACM, New York, NY, 2013), pp. 189–198.

[46] And this is all that we consider here.