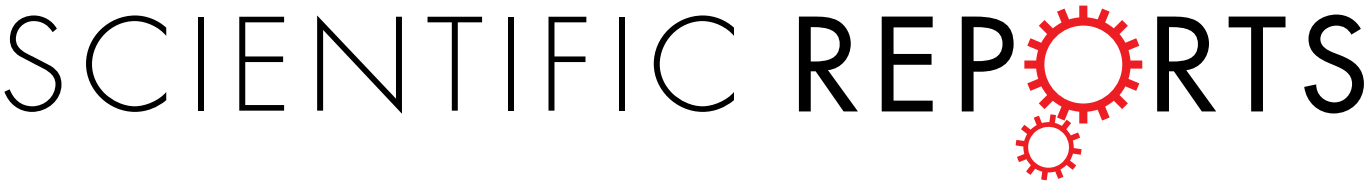

OPEN

Link Prediction in Evolving Networks Based on Popularity of Nodes

Received: 9 March 2017
Accepted: 26 June 2017
Published online: 02 August 2017

Tong Wang[1], Xing-Sheng He[1], Ming-Yang Zhou[2,3] & Zhong-Qian Fu[1]

Link prediction aims to uncover the underlying relationship behind networks, which could be utilized to predict missing edges or identify the spurious edges. The key issue of link prediction is to estimate the likelihood of potential links in networks. Most classical static-structure based methods ignore the temporal aspects of networks, limited by the time-varying features, such approaches perform poorly in evolving networks. In this paper, we propose a hypothesis that the ability of each node to attract links depends not only on its structural importance, but also on its current popularity (activeness), since active nodes have much more probability to attract future links. Then a novel approach named popularity based structural perturbation method (PBSPM) and its fast algorithm are proposed to characterize the likelihood of an edge from both existing connectivity structure and current popularity of its two endpoints. Experiments on six evolving networks show that the proposed methods outperform state-of-the-art methods in accuracy and robustness. Besides, visual results and statistical analysis reveal that the proposed methods are inclined to predict future edges between active nodes, rather than edges between inactive nodes.

Networks are effective descriptions of complex systems in society and nature[1,2], with entities denoted as nodes and relations as links, respectively. The organization of real networks evolve under the influence of certain patterns and irregular factors, in principle, only the former can be modeled with physical methodologies. A significant concern about complex networks is link prediction that conduces to explanations of these models and revelations of the hidden driving-mechanisms. Therefore, link prediction has drawn numerous attentions from various fields covering biology, sociology and others[3–6]. For example, in protein-protein interaction experiments in cells, only strong relations between proteins could be detected by limited precision of equipments. It is prohibitive to measure every interaction between all pair proteins due to sharply increasing experimental costs with the size of proteins[7,8], an appropriate approach is to evaluate the likelihood of potential relations and specifically test non-existing relations with the high likelihood. Also, in social contexts, two persons would build friendship in the near future with a high probability if they have many common friends or attributes, which could be utilized to uncover lost friends or predict future friends[9–11]. Besides, further extensive applications also include personalized recommendations in e-commerce[12,13] and aircraft route planning study[14], etc.

The crux of link prediction is to evaluate the likelihood of potential edges, based on which we can rank the potential edges in descending order and edges in the top of ranking list are predicted as underlying or future edges[15,16]. The similarity based approaches, which equate likelihood with similarity, are the most common frameworks that argue the prospective edges may exist between similar nodes. To achieve this, traditional attribute based methods measure the likelihood of links by learning how many common features (e.g. common hobbies, ages, tastes, geographical locations) the two endpoints share[17]. Many researches on social networks have shown that the pervasive homophily promotes ties between similar humans[18,19]. However this kind of methods suffer from the inaccessible and unreliable information of nodes due to the privacy policy in real scenario[20]. Luckily, the development of the complex network theory provides a new path in which only network topological structure is required regardless of privacy information to solve the problem. When evaluating the similarity between nodes, according to the structure differences, structure based methods could be classified into three categories: local

[1]Department of Electronic Science and Technology, University of Science and Technology of China, Hefei, 230027, P. R. China. [2]Guangdong Province Key Laboratory of Popular High Performance Computers, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, P. R. China. [3]Physics Department, University of Fribourg, Chemin du Musée 3, Fribourg, CH-1700, Switzerland. Correspondence and requests for materials should be addressed to M.-Y.Z. (email: zhoumy2010@gmail.com)

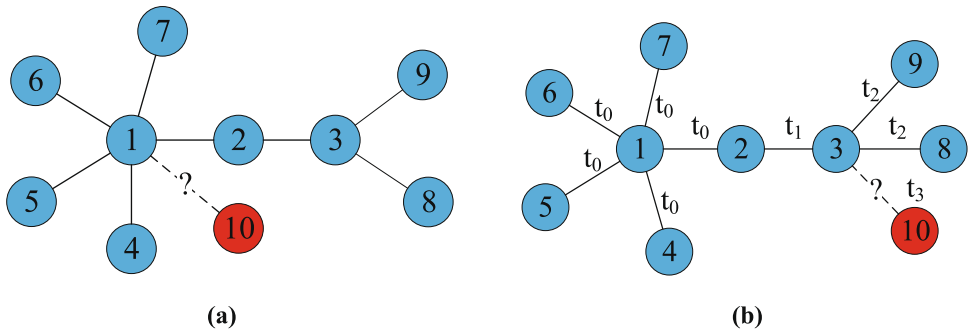

Figure 1. Illustration of the popularity. The fresh link and node 10 will be added into the existing networks at the next time t_3. In panel (a), attractiveness of nodes are determined by static features. According to the preferential attachment, node 10 prefers to connect with node 1 due to the largest degree. In panel (b), temporal effects are considered. The currently popular node 3 may become attractive and connect with node 10 at time t_3.

methods, global methods and Quasi-global methods. Local similarity is mainly based on common neighbors, such as the most well-known Common Neighbor (CN) index that counts the number of common neighbor nodes[21], Adamic-Adar (AA) index and Resource Allocation (RA) index that depress the large-degree neighbor nodes[22,23]. For large networks, Cui *et al*. proposed a fast algorithm for calculating the number of common neighbors[24]. Global similarity emphasizes the global topology information of network, such as Katz index that counts all of the paths between two nodes[25]. Quasi-global similarity is a well trade-off of local similarity methods and global similarity methods, such as Local Path (LP) index that only considers the short paths in Katz index[23], Local Random Walk (LRW) index that focuses on the limited random walk in local area[26]. Beyond that, some algorithms based on maximum likelihood methods and other exquisite models have been proposed. Clauset *et al*. proposed a Hierarchical Structure Model which presents well performance in hierarchical networks by using a dendrogram[27]. Lü *et al*. proposed a Structural Perturbation Method that approximates the observed networks by randomly repeated perturbations. This method outperforms state-of-the-art methods in accuracy and robustness[28]. In terms of information theory, Xu *et al*. proposed the Path Entropy index that considers the information entropies of shortest paths and penalizes the long paths[29]. Tan *et al*. proposed a Mutual Information (MI) method with the high accuracy and reasonable computation time, which considers the feature of common neighbors and denotes the likelihood of one link as the the conditional self-information of this link existing between the node pair when their common neighbors are given[30]. Zhu *et al*. generalized the MI index into Neighbor Set Information that is applicable to multiple structural features to enhance the accuracy[31].

Real networks are highly dynamic with the come-and-go of nodes and edges[32]. However, the aforementioned algorithms unexceptionally ignore the temporal aspects of real networks, in particular, the trend of nodes: yesterday active nodes that contacted numerous neighbors may be unpopular today. Inspired by this, we propose a hypothesis that the emergence of future links are not only determined by existing network structure, but also are affected by the popularity of endpoints. For instance, Fig. 1 illustrates the effects of popularity. The red node will enter in the network and connect with one of the existing nodes. In Fig. 1(a), according to the static analysis, node 10 prefers to connect with the large-degree node 1. While the birth time of each edge is given in Fig. 1(b), we can easily know that node 3 is of high popularity because only it attracts edges at the present time t_2. In practice, the fresh edge will be more likely to occur between node 10 and the active node 3 at the next period t_3. To comply with this scenario, unlike previous works that predict potential links mostly based on static networks, we propose a popularity based structural perturbation method (PBSPM) and its fast algorithm that integrate popularity of nodes and observed network topology to predict future edges. Experimental results on real-world networks show that the proposed methods outperform the other traditional approaches in accuracy and robustness.

Results

Popularity metrics. The definition of popularity is related to the concepts of temporal trend of nodes that could be obtained through the statistics and analysis of relevant historical information. For two nodes with the same degree, one may connect with its neighbors at early stage and not form any connections later, while the other one develops most of its connections at late stage. Intuitively the latter node would attract more fresh edges with high probability in the near future. Given this, a straightforward approach to evaluate the popularity of a node is counting the edges it recently attracts.

Given an undirected and unweighted network $G(V, E)$ where V and E represent the set of nodes and links, respectively, each link has a time-stamp that represents the entering time. In this work, multi-links and self-loops are not allowed. $k_i(t)$ denotes the degree of nodes i at time t. In the next time span T, node i would attract $\Delta k_i(t, T)$ new edges,

$$\Delta k_i(t, T) = k_i(t + T) - k_i(t). \tag{1}$$

Note that $\Delta k_i(t, T)$ in Eq. (1) determined by both t and T cannot reflect the relative popularity of node i, since even large degree nodes become inactive, they still attract more fresh edges than nodes of small degree due to the preferential attachment mechanism. To solve this issue, for a dataset spans starting from t_a to t_c, we divide its

edges into the fresh set and the old set according to a boundary $t_b \in (t_a, t_c)$. If an edge was constructed in (t_a, t_b), it belongs to the old set otherwise the fresh set. The fraction of old edges and fresh edges are denoted as p_{older} and $p_{fresher}$. The $p_{fresher}$ can be comprehended as the observation length of historical information. Then, the popularity of node i is

$$s_i = \frac{\Delta k_i(t_b, t_c - t_b)}{\Delta k_i(t_a, t_c - t_a)} = \frac{k_{i,fresher}}{k_{i,all}}, \tag{2}$$

where $k_{i,all}$ and $k_{i,fresher}$ indicate the whole degree and fresher degree of node i. Equation (2) improves the drawbacks of simply counting the new edges and quantifies the popularity in the normalized range. Clearly, if all links of node i locate in the fresh set, $s_i = 1$. For another case that all links of node i locate in the old set, node i becomes dormant, $s_i = 0$. Therefore $s_i \in [0, 1]$ and a higher s_i means a higher popularity.

Popularity based structural perturbation method. In this section, we propose a hypothesis that the observed network is determined by some latent attractors (e.g. similar hobbies, ages, gender, location) that independently influence the structural properties. For an attractor $x_k = [x_{k,1}, x_{k,2}, \ldots, x_{k,n}]^T$, $x_{k,i}$ represents the attractiveness of node i for the latent attractor x_k. Inspired by configuration model, the probability p_{ij} that an edge exists between two node i and j is proportional to $x_{k,i}x_{k,j}$. Supposing that there are m kinds of attractors, probability p_{ij} is defined as the weighted influence of each attractor,

$$p_{ij} = \sum_{k=1}^{m} w_k x_{k,i} x_{k,j}, \tag{3}$$

where w_k is a tunable parameter to balance the relative influence of each attractor x_k. The problem is how to seek the optimal w_k and $x_{k,i}$ that make p_{ij} approximate a_{ij} at most. Considering a network G with adjacent matrix $A = (a_{ij})_{n\times n}$, a special case is that $p_{ij} = 1$ if $a_{ij} = 1$, otherwise $p_{ij} = 0$. For optimal w_k and x_k,

$$A_p = (p_{ij})_{n\times n} = \sum_{k=1}^{m} w_k x_k x_k^T. \tag{4}$$

If $m = n$ in Eq. (4), where n is the size of the network, then Eq. (4) could be comprehended as the matrix decomposition, with w_k and x_k representing eigenvalues and eigenvectors respectively. In practice, many random connections exist in networks, L*ü et al.* proposed the structural perturbation method (SPM) that can reduce the influence of randomness[28]. In SPM, a small fraction p^H of edges ΔA is removed from the network, adjacent matrix A^R of the remaining network is decomposed into

$$A^R = \sum_{k=1}^{n} \lambda_k x_k x_k^T, \tag{5}$$

where λ_k and x_k are the eigenvalues and eigenvectors of $A^R, |x_k| = 1$. We could use A^R to evaluate A with

$$\tilde{A} = \sum_{k=1}^{n} (\lambda_k + \Delta\lambda_k)\, x_k x_k^T, \tag{6}$$

where $\Delta\lambda_k \approx \frac{x_k^T \Delta A x_k}{x_k^T x_k}$ is the coupling influence of x_k on λ_k. $\tilde{A}$ actually is a special case of A_p, $(\lambda_k + \Delta\lambda_k)$ and elements of eigenvector x_k represent weight difference and the attractiveness for attractor x_k separately.

As we have argued, the ability for node i to attract new edges is determined by both latent attractors and its current popularity. To better meet practice, an advanced attractiveness $x'_{k,i}$ is proposed as

$$x'_{k,i} = x_{k,i}(1 + \alpha s_i), \tag{7}$$

where α indicates the degree of temporal popularity. Equation (7), a combination of the static attractiveness and popularity, tightly captures both the static features and the temporal information of the evolving pattern. Later in Eq. (6), substituting x_k with x'_k to predict future links,

$$\tilde{A'} = \sum_{k=1}^{n} (\lambda_k + \Delta\lambda_k) x'_k x'^T_k. \tag{8}$$

Since Eq. (5) degenerates into Eq. (4) if the size m of attractors is less than n. Supposing that $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$, we substitute w_k and x_k in Eq. (4) with λ_k and x_k in Eq. (5). Similar to the same transition from Eq. (5) to Eq. (8), we obtain

$$A' = (p_{ij})_{n\times n} = \sum_{k=1}^{m} (\lambda_k + \Delta\lambda_k) x'_k x'^T_k, \tag{9}$$

which reduces into Eq. (8) if $m = n$. In the following experiments, we firstly measure the performance of Eq. (8), then show that we could reduce the calculation complexity by using only a few eigenvalues and eigenvectors, that is $m \ll n$ in Eq. (9).

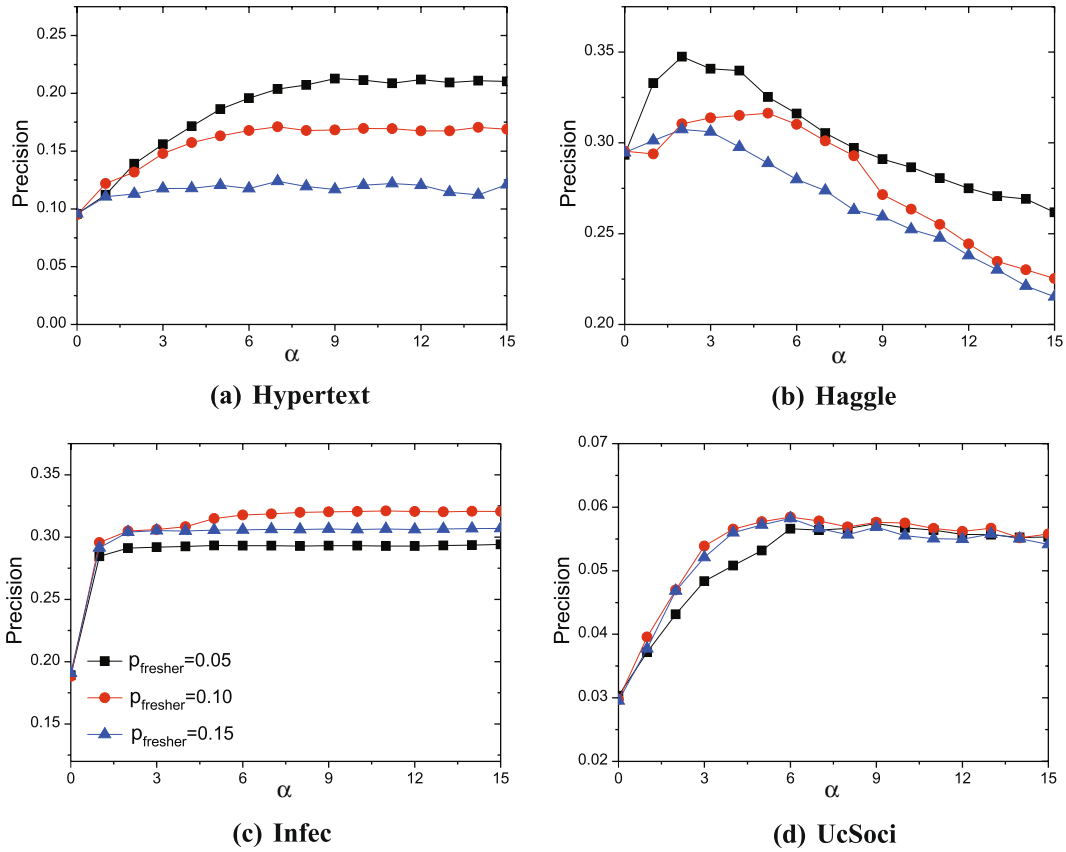

Figure 2. Precision versus α obtained by PBSPM. The experiments are performed on 90% training set and 10% probe set. Each data point is averaged over 10 independent realizations. The values of $p_{fresher}$ and α corresponding to the optimal precision reported in Table 1 vary for different networks: 0.05 and 9 for Hypertext, 0.05 and 2 for Haggle, 0.10 and 11 for Infec, 0.10 and 7 for UcSoci.

Precision	CN	AA	RA	Katz	SRW	SPM	PBSPM	Fast PBPSM
Hypertext	0.0959	0.1050	0.1005	0.0959	0.1187	0.0984	**0.2128**	**0.2194**
Haggle	0.1786	0.1888	0.1939	0.2041	0.2194	0.2928	**0.3475**	**0.3760**
Infec	0.0233	0.1163	0.1814	0.0233	0.3023	0.1949	**0.3210**	**0.3070**
UcSoci	0.0138	0.0153	0.0138	0.0138	0.0046	0.0298	**0.0584**	**0.0574**

Table 1. Precision of different methods for four networks. All the results are calculated under the optimal cases by adjusting parameters if any. The data in bold face are averaged over ten realizations with the same $p_{fresher}$ and α.

Experiments on real networks. The proposed method PBSPM, integrating the attractiveness $x_{k,i}$ and popularity s_i, reduces into the original SPM when $\alpha = 0$. With the increase of α, PBSPM prefers to predict links between popular nodes. Figure 2 gives the performance of PBSPM in contrast to SPM ($\alpha = 0$) under different $p_{fresher}$. The precision values tend to be stable or achieve the best when α brings the static attractiveness and popularity into balance. Clearly, the optimal value of α varies for different networks. For Hypertext, Infec and UcScoci, future links have high likelihood to exist between the active nodes. However, for the Haggle dataset, the temporal trend of nodes are less obvious. Hence, the precision curve is optimized at $\alpha = 2$, contrast to the other three networks of which the curves finally stabilize when α increases. Overall, when $\alpha \in [3, 5]$, PBSPM achieves improved performance compared with SPM in the four networks. Moreover, given the different length of historical information $p_{fresher}$, all the curves present different levels of superiority in precision, suggesting a general and robust range of $p_{fresher}$. Actually, it is difficult to choose the optimal value, which should follow the principle of keeping the balance between the length of historical information and future information (probe set). With regard to 10% probe set in this experiment, $P_{fresher} = 0.1$ is the balanced option because the corresponding curves all show the great improvements.

Reducing the number of eigenvectors could reduce the computation complexity. To address the high computation complexity, we propose the fast PBSPM that takes into account a few eigenvectors with only some large eigenvalues, which can well reflect the backbone structure of networks[33]. In practical networks, a huge gap exists in the eigenvalue space. Some eigenvectors with large eigenvalues play more important roles than those with small

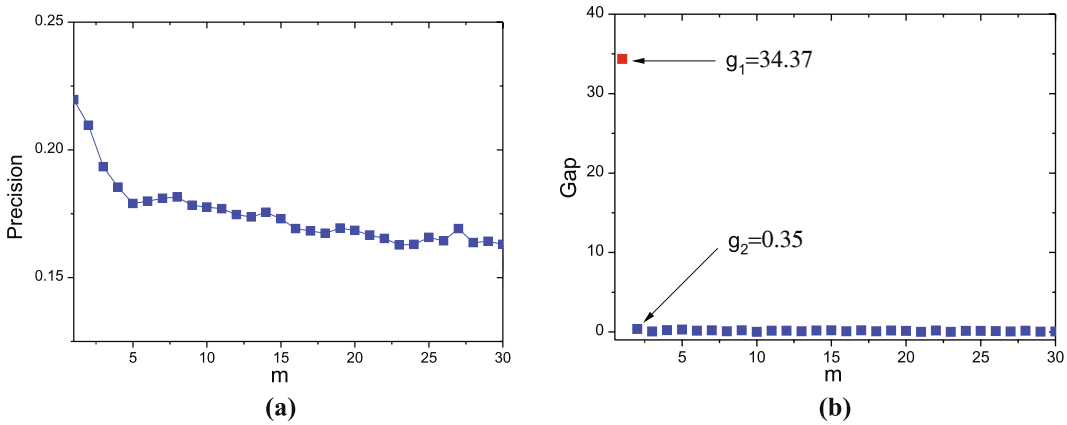

Figure 3. Precision versus m and gap $g_m = |\lambda_m| - |\lambda_{m+1}|$ for Hypertext. λ_m is the eigenvalue of adjacent matrix A^T. Panel (a) shows the performance of Eq. (9) on various $m \in [1, 30]$ with fixed $p_{fresher} = 0.05$ and $\alpha = 9$. Each data point is obtained over ten simulations. Panel (b) shows the difference g_m between $|\lambda_m|$ and $|\lambda_{m+1}|$. $g_1 = 34.37$ is distinct and the others are all close to 0.

Time(*ms*)	CN	AA	RA	Katz	SRW	SPM	PBSPM	Fast PBPSM
Hypertext	1.02	1.12	1.08	1.51	1.95	20.3	**20.51**	**15.73**
Haggle	2.25	2.58	2.54	3.08	4.78	50.45	**51.62**	**28.86**
Infec	5.62	6.22	5.95	7.39	11.4	175.71	**179.15**	**92.8**
UcSoci	204.27	239.23	228.62	272.06	856.08	15200.76	**15902.53**	**1122.95**

Table 2. Computation time of different methods for four networks. All the results are averaged over ten runs on AMD R7 computer with MATLAB R2016b and 8GB RAM.

eigenvalues. Taking Hypertext as example, Fig. 3(a) plots the precision for various m in Eq. (9). Compared with SPM, the curve presents significant improvements and achieves the best at $m = 1$, meeting the effectiveness of Eq. (9). Figure 3(b) gives the differences between two adjacent eigenvalues $g_m = |\lambda_m| - |\lambda_{m+1}|$ ($|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$). The distinct g_1 indicates a huge gap between $|\lambda_1|$ and $|\lambda_2|$, while the other gaps ($m \geq 2$) are all close to 0, suggesting that the huge gap g_1 induces the decline of precision when $m > 1$. Then, we choose $m = 1$ as the optimal value for Hypertext, analogously, the values for Haggle, Infec and UcSoci are respectively determined as $m = 2,19,2$ after which the g_m approaches to 0 approximately. In consequence, it only requires $O(n^2)$ time to calculate the top-m eigenvalues and corresponding eigenvectors, and the reconstruction of similarity matrix (Eq. 9) needs $O(m \times n^2)$ time. To reduce the randomness, the fast PBSPM repeats the random perturbation for ten times and obtains the averaged similarity matrix with $O(10 \times (mn^2 + n^2))$ time. Hence, with $m \ll n$ and the increase in size n, the time complexity of fast PBSPM is $O(n^2)$ in contrast with the time complexity $O(n^3)$ of PBSPM and SPM, where the decomposition and reconstruction consume $O(n^3)$ time. Besides, the time complexity is $O(n^2)$ for local similarity based methods, such as CN, RA, AA, and $O(n^3)$ for Katz and SRW.

Table 1 and Table 2 list the precision values and computation time of different link prediction algorithms. Obviously, the proposed methods achieve remarkable improvements, at most 84.84% for Hypertext, 28.42% for Haggle, 6.19% for Infec, 95.97% for UcSoci. In spite of this, PBSPM suffers from the huge computational cost that limits its extensive applications. Fast PBSPM, a well trade-off of computation complexity and accuracy, has the reasonable computational cost and the high accuracy. Due to the repeated steps in experimental procedures, the fast algorithm still consumes more time than some traditional predictors with the same time complexity. Additionally, the attractors ignored by the fast algorithm contain some secondary information that may either improve the accuracy as useful information or deteriorate the performance as network noise, hence, the precision slightly fluctuates around that of PBSPM. In general, the proposed methods show the high robustness because of the well performance for disparate networks, while other baselines give poor predictions for some networks. Apart from precision improvements, we also try to quantify the physical difference between the age of links selected by various methods, which can be comprehended as the average popularity of endpoints $\bar{s} = \sum \frac{s_i + s_j}{2_*|E^P|}$ if edge e_{ij} is selected by a certain predictor. According to Table 3, links selected by the proposed methods are much older than the others; that is, the potential links prefer to form between the active nodes in the earlier future.

In the following, we mainly focus on the performance of SPM and PBSPM to explore underlying reasons of the improvements. To figure out the effect of popularity, four typical nodes from the training set of Hypertext, the large-degree node 1 and 3 ($k_{1,training} = 78, s_1 = 0.051; k_{3,training} = 93, s_3 = 0.032$), and the active node 91 and 113 ($k_{91,training} = 29, s_{91} = 0.289; k_{113,training} = 14, s_{113} = 1$) are chosen to analyse their predicted connections and corresponding variation of attractiveness. Figure 4 plots the predicted future links attached to selected nodes by SPM and PBSPM when $p_{fresher} = 0.05$ and $\alpha = 9$. After that, the principal eigenvector x_1 of A^R and the advanced x_1' under

Precision	CN	AA	RA	Katz	SRW	SPM	PBSPM	Fast PBPSM
Hypertext	0.0411	0.0420	0.0445	0.0407	0.0509	0.0420	**0.2115**	**0.2243**
Haggle	0.0508	0.0513	0.0518	0.0488	0.0489	0.0609	**0.1313**	**0.1265**
Infec	0.0275	0.1228	0.2180	0.0275	0.4511	0.2148	**0.7901**	**0.8145**
UcSoci	0.0611	0.0666	0.0861	0.0617	0.1498	0.0672	**0.4407**	**0.4051**

Table 3. Average age of links selected by predictors. The bold data are averaged over ten runs and obtained under the optimal parameters.

Networks	Hypertext	Haggle	Infec	UcSoci
ΔCC	0.28	0.0582	0.3092	0.1155
$\Delta\lambda_1$	4.1822	4.79	1.86	4.2183

Table 4. Variation of correlation coefficient ΔCC and coupling influence $\Delta\lambda_1$. Each data is averaged over ten perturbations.

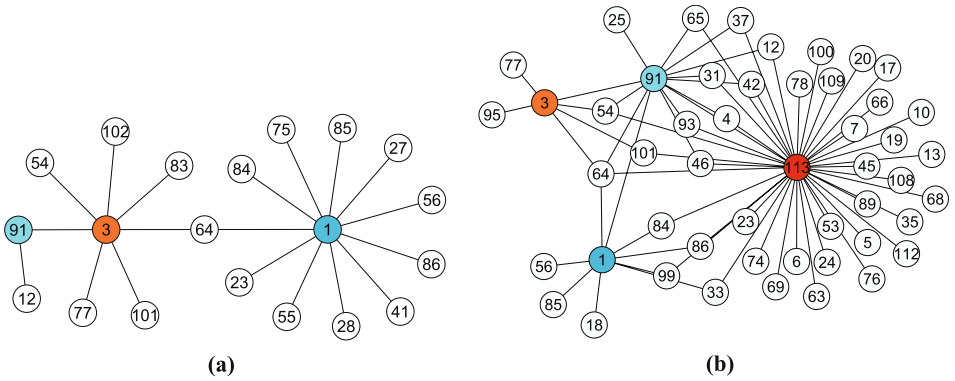

Figure 4. Predicted connections of large-degree node 1, 3 and active node 91, 113 in Hypertext. Only the selected nodes and their neighbors are plotted, and the connections are the subset of the top-$|E^P|$ predicted links. Panel (a) shows the connections predicted by SPM. Node 1 and 3 are much more attractive than node 91, and node 113 is not presented because of no connections. Panel (b) shows the connections predicted by PBSPM. The active node 91 and 113 attract numerous nodes, which gives rise to the explosive growth of edges.

the optimal case are calculated to quantify the attractiveness for the most weighted attractor. In addition, the principal eigenvector also characterizes the ranking of nodes, i.e. the importance[34,35]. In Fig. 4(a), node 1 and 3 ($x_{1,1}=0.1715, x_{1,3}=0.1899$) with the high importance are much more attractive than node 91 and 113 ($x_{1,91}=0.0648, x_{1,113}=0.0329$), especially, node 113 with the lowest importance has no connections at all. Contrastingly, the high popularity enhances the active nodes ($x'_{1,91}=0.1158, x'_{1,113}=0.1923$) and results in the burst of links connecting to the them in Fig. 4(b), notably the most active node 113. In summary, nodes with the higher popularity are emphasized by PBSPM to attract much more links, whereas the inactive despite their importance are weakened to reduce connections.

The above figures conduce to the understanding of how popularity imposes effects on several typical nodes, but note that, it is a rational speculation that the improvements must result from the advanced attractiveness of all nodes. As above argued, principal eigenvector denotes the attractiveness for the most weighted attractor. Because $(\lambda_1+\Delta\lambda_1)(x_1x_1^T)$ occupies the main body of $\tilde{A}$, neglecting constant term $\Delta\lambda_1+\Delta\lambda_1$, similarity $\tilde{a}_{i,j}$ is mainly determined by eigenvector x_1. The Pearson correlation coefficient (CC) between principal eigenvector and degree in the probe set, holistically reflecting the extent to which the attractiveness $x_{1,i}$ coincides with real degree increment $k_{i,probe}$, is computed as follows,

$$cc=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{1,i}-\overline{x}_{1,i}}{\delta_{x_{1,i}}}\right)\left(\frac{k_{i,probe}-\overline{k}_{i,probe}}{\delta_{k_{i,probe}}}\right), \tag{10}$$

where $\overline{x}_{1,i}$ and $\overline{k}_{i,probe}$ are the means of $x_{1,i}$ and $k_{i,probe}$. The CC between advanced x'_1 and degree in the probe set is obtained similarly. Table 4 lists the variation of CC after the addition of popularity and the coupling influence $\Delta\lambda_1$ averaged over ten independent perturbations. The positive ΔCC of four networks suggest attractiveness of some nodes are corrected to meet the degree increment in the future. Furthermore the positive $\Delta\lambda_1$ also strengthens the improvements of correlations. As a result, the popular nodes are assigned more connecting opportunities to promote the precision.

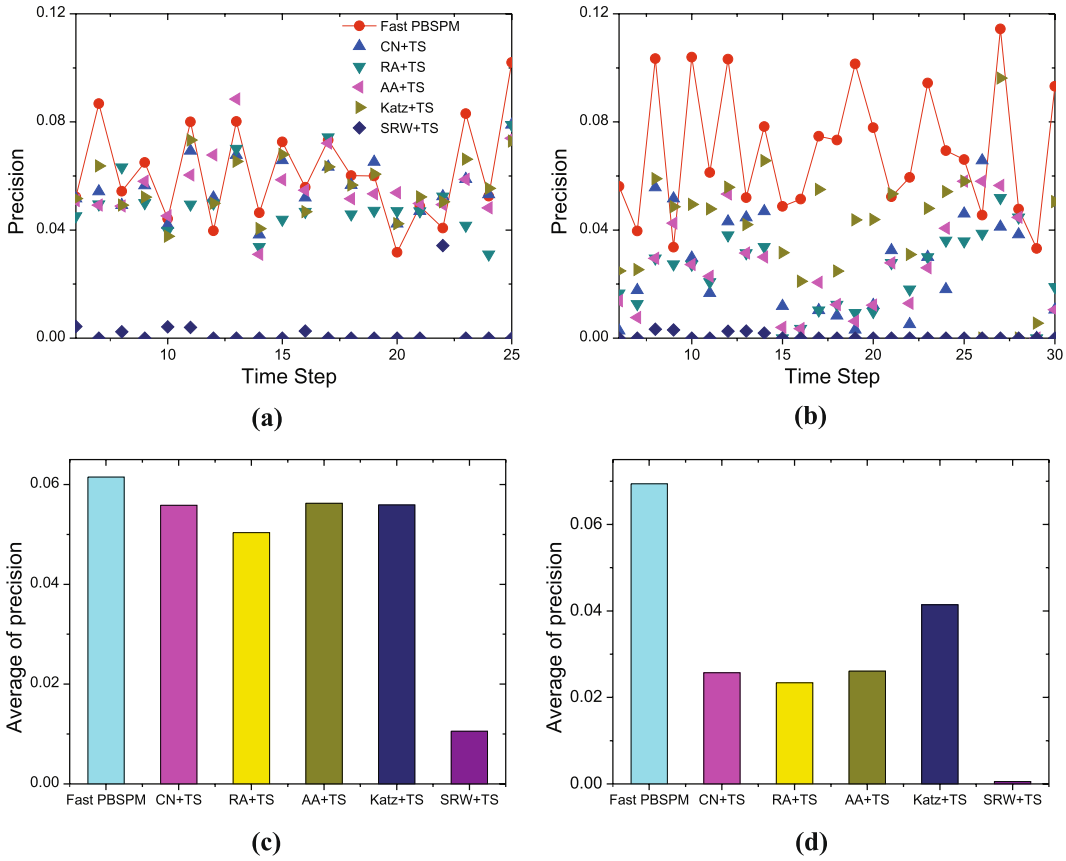

Figure 5. Precision at different time steps and their average values. $G_{t\sim t+T-1}(G_t, G_{t+1}, \ldots, G_{t+T-1})$ and G_{t+T} play the role of the training set E^T and probe set E^P. Panel (a) and panel (b) show the precision values at different time steps for LKMLR and Wiki. The red curves are respectively obtained by the fast PBSPM with $\alpha = 2$ and 5, $m = 2$ and 2. The other results are obtained under optimal cases by different forecasting models. Panel (c) and panel (d) give the average precision values of different methods for the two networks.

Eventually, to demonstrate the feasibility of the proposed methods in practical applications, we compare the fast PBSPM with time series (TS) based methods on continuous temporal networks, which have been effectively applied to the temporal link prediction[36–38]. For each network, the dataset is divided into T_N snapshots $(G_1, G_2, \ldots, G_{T_N})$ with the length of time period $P_{length} = 7$ days. Setting a specified time window $T = 5$, we use the graph series $(G_t, G_{t+1}, \ldots, G_{t+T-1})$ and its reduced static graph $G_{t\sim t+T-1}$ to predict the links that will occur in G_{t+T} $(t = 1, 2, \ldots, T_N - T)$. Then the popularity of each node is calculated as:

$$s_i = \frac{k_{i,G_{t+T-1}}}{k_{i,G_{t\sim t+T-1}}} = \frac{k_{i,fresher}}{k_{i,all}}. \tag{11}$$

During the evolution, certain mechanisms drive the network organization regularly and the structural features keep relatively stable. Hence, we obtain the optimal α and m by the known networks observed between the time period $1 \le t \le 6$ ($G_{1\sim5}$ as the training set, G_6 as the probe set) and apply them to the subsequent predictions. Figure 5 shows the precision at continuous time steps and the average accuracy of different methods. For LKMLR, though the fast PBSPM falls behind sometimes, its average value shows a slight advantage in precision (Fig. 5(a) (c)). For Wiki, not only does the fast PBSPM gain the upper hand at any time, but it achieves much higher average accuracy compared with TS based methods (Fig. 5(b) (d)). These experimental results demonstrate that the fast PBSPM has prospective applications in evolving networks.

Discussion

In this paper, we propose the PBSPM and its fast algorithm to predict future links. The main contribution is to investigate the popularity (activeness) of nodes in real-world evolving networks and apply it to link prediction. Unlike previous works that calculate temporal effects with complex theories, we infer the popularity of each node by its recently active edges. Then we propose a hypothesis that the future network is influenced by both existing structure and popularity of nodes. By introducing popularity into perturbation method, PBSPM could distinguish active and inactive historical important nodes, and prefer to predict new edges attached to active nodes. Subsequently, the fast method is proposed to get rid of the high computation complexity. Experimental

results on real-world evolving networks reveal that compared with traditional methods, the proposed methods achieve better performance in precision and robustness. Besides, further experiments are conducted to uncover the underlying reasons of the improvements.

Definitely, the performance of proposed methods largely depend on the popularity of each node. In other words, the popularity based methods are more applicable for the networks with obvious temporal effects, where the popularity metric can effectively quantify the popularity of each node. Hence, another important issue is that improving popularity performance would enhance the precision of link prediction, which is the future work. Since our work mainly explores prediction in evolving networks, it has possible applications in traffic prediction, airline control, recommendation of social network, and so on.

Methods

Experimental procedures. To predict the future links of evolving networks with PBSPM, there are five detailed steps to follow:

Step 1: We firstly divide the network into the training set E^T and the probe set E^P based on the birth time of each edge, the corresponding adjacent matrix are denoted by A^T and A^P.
Step 2: The training set is further divided into the old set and the fresh set to calculate the popularity via Eq. (2) or Eq. (11).
Step 3: We perturb the training set by randomly removing a small fraction $p^H = 0.1$ of edges ΔA, obviously, $A^T = A^R + \Delta A$.
Step 4: We decompose the matrix A^R and obtain the $\tilde{A}'$ via Eq. (7) and Eq. (8).
Step 5: Repeat step 3 and step 4 for ten times. In other words, we implement the perturbations for ten times to obtain the averaged $\langle \tilde{A}' \rangle$ where the score $\langle \tilde{a}'_{ij} \rangle$ represents the existent likelihood of the link between node i and j. Finally, non-observed edges with the top-$|E^P|$ scores are chosen as potential future edges.

Data description. In this work, six datasets are considered to evaluate the performance of algorithms. (1) Hypertext 2009 (Hypertext): a network of face-to-face contacts of the attendees of the ACM Hypertext 2009 conference from June 30 to July 1, 2009, including 113 nodes and 2196 unique links[39]. (2) Haggle: an undirected network representing contacts between people measured by carried wireless devices[40], including 188 nodes and 1947 unique links. The time span is 4 days. (3) Infectious (Infec): a network describing the face-to-face behavior of people during the exhibition INFECTIOUS: STAY AWAY in 2009[39], including 301 nodes and 2145 unique links. The time span is 8 hours. (4) UC Irvine messages (UcSoci): a directed network of messages between the users of an online community of students from the University of California, Irvine[41], including 1692 nodes and 13037 unique links. The dataset spans from April 15 to October 25, 2004. (5) Linux kernel mailing list replies (LKMLR): a communication network of the Linux kernel mailing list. The data considered in experiments is from January to June, 2013, including 2907 nodes and 78955 links. (6) Wikipedia elections (Wiki): a network of users from the English Wikipedia that voted for and against each other in admin elections. The data considered in experiments spans from October, 2005 to April, 2006, including 2309 nodes and 23707 links[42].

To simplified the problem, we ignore the direction and weighted of links, and remove the isolated nodes. What is more, the networks are divided into historical training set and future probe set only according to the timestamps that attach to edges.

Evaluation metric. AUC (Area Under the receiver operating characteristic Curve) and Precision are two standard metrics used to measure the link prediction algorithm[43,44]. The former randomly compares the score of a missing link with a non-existent link to evaluate the performance. The latter focuses on the links with top-L scores. When dealing with highly skewed datasets, the precision always gives a more informative picture of algorithms' performance[45]. Hence, We choose Precision index as the metric to evaluate the accuracy of the proposed method and other baselines. Precision is defined as the ratio of links predicted accurately to all links selected. Namely if we select top-L links in the all ranked non-observed links and only L_r links are predicted correctly in the probe set E^P, then the accuracy of predictor follows

$$Precision = \frac{L_r}{L}. \tag{12}$$

In our experiments, we select $L = |E^P|$ and count how many of top-$|E^P|$ links really exist in the probe set.

Baselines. For comparison, we briefly introduce five traditional algorithms based on all three kinds of structural similarity.

(1) Common Neighbors (CN), related to the concepts of the triadic closure, is the most well-known method with an assumption that two target points tend to connect with each other if the new connection may produce much more triangles in the graph.

$$s_{xy}^{CN} = |\Gamma(x) \bigcap \Gamma(y)|, \tag{13}$$

where $\Gamma(x)$ is the set of neighbors of node x and $|\Gamma(x) \cap \Gamma(y)|$ represents the set of common neighbors of x and y.

(2) Adamin-Adar (AA), advanced from CN, restricts the contributions of common neighbors by introducing a penalty factor, i.e., the logarithm of reciprocal of their degree.

$$s_{xy}^{AA} = \sum_{z \in \Gamma(x) \bigcap \Gamma(y)} \frac{1}{\log k_z}, \tag{14}$$

where k_z denotes the degree of common neighbor z.

(3) Resource Allocation (RA), motivated by transferring resource between two unconnected nodes, views the common neighbor as the intermediary of which the transfer capability equals to the reciprocal of degree of common neighbors.

$$s_{xy}^{RA} = \sum_{z \in \Gamma(x) \bigcap \Gamma(y)} \frac{1}{k_z}. \tag{15}$$

(4) Katz index, based on global information of network, counts all the paths connecting two endpoints with weakening the contributions of longer paths exponentially:

$$s_{xy}^{Katz} = \sum_{l=1}^{\infty} \alpha^l \cdot \left| paths_{x,y}^{\langle l \rangle} \right|. \tag{16}$$

When $|\alpha| < 1/\lambda_{\max}$, it can be rewritten as:

$$S = (I - \alpha \cdot A)^{-1} - I, \tag{17}$$

where I is the identity matrix, $\alpha > 0$ is the tunable parameter, $\lambda_{\max}$ is the largest eigenvalue of the adjacent matrix A.

(5) Superposed Random Walk (SRW) considers the summation of local random walks within t steps and degree of two endpoints to emphasize the local properties in real networks[26].

$$s_{xy}^{SRW}(t) = \sum_{\tau=1}^{t} [q_x \pi_{xy}(\tau) + q_y \pi_{xy}(\tau)], \tag{18}$$

where $q_x = \frac{k_x}{2|E|}$ denotes the initial distribution of resources and $\pi_{xy}(\tau)$ represents the transfer probability from x to y.

(6) Time series based methods explore the evolution of topological metrics to predict the future links[37]. It follows the steps below:

Step1: Choose a static-structure method (e.g. CN, RA, Katz, etc);
Step2: Establish the time series by calculating the similarity between unconnected nodes in each time period;
Step3: Compute the final score of unconnected nodes with a forecasting model (e.g. Moving Average, Liner Regression, Simple Exponential Smoothing, etc);
Step4: Measure the algorithms with future links in the next time period.

References

1. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47 (2002).
2. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
3. Barzel, B. & Barabási, A.-L. Network link prediction by global silencing of indirect correlations. *Nat. Biotechnol.* **31**, 720–725 (2013).
4. Hulovatyy, Y., Solava, R. W. & Milenković, T. Revealing missing parts of the interactome via link prediction. *Plos One* **9**, e90073 (2014).
5. Ermiş, B., Acar, E. & Cemgil, A. T. Link prediction in heterogeneous data via generalized coupled tensor factorization. *Data Mining and Knowledge Discovery* **290**, 203–236 (2015).
6. Stanfield, Z., Coşkun, M. & Koyutürk, M. Drug response prediction as a link prediction problem. *Sci. Rep.* **7**, 40321 (2017).
7. Mamitsuka, H. Mining from protein–protein interactions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**, 400–410 (2012).
8. Cannistraci, C. V., Alanis-Lobato, G. & Ravasi, T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Sci. Rep.* **3**, 1613 (2013).
9. Gong, N. Z. *et al*. Joint link prediction and attribute inference using a social-attribute network. *ACM Trans. Intell. Syst. Technol.* **5**, 27:1–27:20 (2014).
10. He, Y.-L., Liu, J. N., Hu, Y.-X. & Wang, X.-Z. Owa operator based link prediction ensemble for social network. *Expert Syst. Appl.* **42**, 21–50 (2015).
11. Tang, J., Chang, S., Aggarwal, C. & Liu, H. Negative link prediction in social media. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 87–96 (ACM, 2015).
12. Daminelli, S., Thomas, J. M., Durán, C. & Cannistraci, C. V. Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks. *New J. Phys.* **17**, 113037 (2015).
13. He, X.-S., Zhou, M.-Y., Zhuo, Z., Fu, Z.-Q. & Liu, J.-G. Predicting online ratings based on the opinion spreading process. *Physica A* **436**, 658–664 (2015).
14. Guimerà, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *P. Natl. Acad. Sci. USA* **106**, 22073–22078 (2009).
15. Liben-Nowell, D. & Kleinberg, J. The link-prediction problem for social networks. *J. AM. Soc. Inf. Sci. Technol.* **58**, 1019–1031 (2007).
16. Wang, P., Xu, B., Wu, Y. & Zhou, X. Link prediction in social networks: the state-of-the-art. *Sci. China Inform. Sci.* **58**, 1–38 (2015).
17. Lin, D. An information-theoretic definition of similarity. *Proceedings of the Fifteenth International Conference on Machine Learning*, 296–304 (Morgan Kaufmann Publishers Inc., 1998).

18. Yuan, G., Murukannaiah, P. K., Zhang, Z. & Singh, M. P. Exploiting sentiment homophily for link prediction. *Proceedings of the 8th ACM Conference on Recommender Systems*, 17–24 (ACM, 2014).
19. Hâncean, M.-G. & Perc, M. Homophily in coauthorship networks of east european sociologists. *Sci. Rep.* **6**, 36152 (2016).
20. Lü, L. & Zhou, T. Link prediction in complex networks: A survey. *Physica A* **390**, 1150–1170 (2011).
21. Newman, M. E. Clustering and preferential attachment in growing networks. *Phys.Rew. E* **64**, 025102 (2001).
22. Adamic, L. A. & Adar, E. Friends and neighbors on the web. *Social Networks* **25**, 211–230 (2003).
23. Zhou, T., Lü, L. & Zhang, Y.-C. Predicting missing links via local information. *Eur. Phys. J. B* **71**, 623–630 (2009).
24. Cui, W. *et al*. Bounded link prediction in very large networks. *Physica A* **457**, 202–214 (2016).
25. Katz, L. A new status index derived from sociometric analysis. *Psychometrika* **18**, 39–43 (1953).
26. Liu, W. & Lü, L. Link prediction based on local random walk. *Europhys. Lett.* **89**, 58007 (2010).
27. Clauset, A., Moore, C. & Newman, M. E. Hierarchical structure and the prediction of missing links in networks. *Nature* **453**, 98–101 (2008).
28. Lü, L., Pan, L., Zhou, T., Zhang, Y.-C. & Stanley, H. E. Toward link predictability of complex networks. *P. Natl. Acad. Sci. USA* **112**, 2325–2330 (2015).
29. Xu, Z., Pu, C. & Yang, J. Link prediction based on path entropy. *Physica A* **456**, 294–301 (2016).
30. Tan, F., Xia, Y. & Zhu, B. Link Prediction in Complex Networks: A Mutual Information Perspective. *Plos One* **9**, e107056 (2014).
31. Zhu, B. & Xia, Y. An information-theoretic model for link prediction in complex networks. *Sci. Rep.* **5**, 13037 (2015).
32. Kim, H. & Anderson, R. Temporal node centrality in complex networks. *Phys. Rev. E* **85**, 026107 (2012).
33. Godsil, C. & Royle, G. *Algebraic graph theory* (Springer-Verlag, New York, 2001).
34. Estrada, E. & Rodrguez-Velázquez, J. A. Subgraph centrality in complex networks. *Phys. Rev. E* **71**, 056103 (2005).
35. Borgatti, S. P. Centrality and network flow. *Social Networks* **27**, 55–71 (2005).
36. Huang, Z. & Lin, D. K. The time-series link prediction problem with applications in communication surveillance. *INFORMS J. Comput.* **21**, 286–303 (2009).
37. Soares, P. R. d. S. & Prudêncio, R. B. C. Time Series Based Link Prediction. *The 2012 International Joint Conference on Neural Networks*, 1–7 (IEEE, 2012).
38. Güneş, İ., Gündüz-Öğüdücü, Ş. & Çataltepe, Z. Link prediction using time series of neighborhood-based node similarity scores. *Data Mining and Knowledge Discovery* **30**, 147–180 (2016).
39. Isella, L. *et al*. What's in a crowd? Analysis of face-to-face behavioral networks. *J. Theor. Biol.* **271**, 166–180 (2011).
40. Chaintreau, A. *et al*. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Transactions on Mobile Computing* **6**, 606–620 (2007).
41. Opsahl, T. & Panzarasa, P. Clustering in weighted networks. *Social Networks* **31**, 155–163 (2009).
42. Leskovec, J., Huttenlocher, D. & Kleinberg, J. Predicting positive and negative links in online social networks. *Proceedings of the 19th international conference on World wide web*, 641–650 (ACM, 2010).
43. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
44. Herlocker, J. L., Konstan, J. A., Terveen, L. G. & Riedl, J. T. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* **22**, 5–53 (2004).
45. Davis, J. & Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. *Proceedings of the 23rd international conference on Machine learning*, 233–240 (ACM, 2006).

Acknowledgements

This work is jointly supported by the National Natural Science Foundation of China (Nos 61471243, 11547040 and U1301252), Science and Technology Innovation Commission of Shenzhen (Nos JCYJ20160520162743717, JCYJ20150625101524056, JCYJ20140418095735561, JCYJ20150731160834611 and SGLH20131010163759789), the PhD Start-up Fund of Natural Science Foundation of Guangdong Province (2017A030310374), the Young Teachers Start-up Fund of Natural Science Foundation of Shenzhen University.

Author Contributions

T.W. and M.Z. conceived and designed the experiments. T.W. and Z.F. performed the experiments. X.H. and Z.F. analysed the data and improved the methods. T.W., X.H. and M.Z. wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.