

SCIENTIFIC REPORTS



OPEN

Similarity-based future common neighbors model for link prediction in complex networks

Shibao Li, Junwei Huang, Zhigang Zhang, Jianhang Liu, Tingpei Huang & Haihua Chen

Link prediction aims to predict the existence of unknown links via the network information. However, most similarity-based algorithms only utilize the current common neighbor information and cannot get high enough prediction accuracy in evolving networks. So this paper firstly defines the future common neighbors that can turn into the common neighbors in the future. To analyse whether the future common neighbors contribute to the current link prediction, we propose the similarity-based future common neighbors (SFCN) model for link prediction, which accurately locate all the future common neighbors besides the current common neighbors in networks and effectively measure their contributions. We also design and observe three MATLAB simulation experiments. The first experiment, which adjusts two parameter weights in the SFCN model, reveals that the future common neighbors make more contributions than the current common neighbors in complex networks. And two more experiments, which compares the SFCN model with eight algorithms in five networks, demonstrate that the SFCN model has higher accuracy and better performance robustness.

Many social, biological, and food-chain systems can be well described by networks, where nodes denote individuals, biological elements, and so on, and links represent the relations between nodes. The complex networks has therefore become a popular focus of many branches of science. An attractive research topic is link prediction, whose purpose is to predict the possibility or necessity of forming links between unconnected node pairs via the information of complex networks^{1,2}. Thus Link prediction can predict the existing yet unknown links (the missing links) and the links that may appear in the future (the future links)^{3,4}. With the amount of data increasing nowadays, Link prediction plays a more crucial role in recommendation system, data mining, complex networks, and so on. For instance, in the protein-protein interaction network of Yeast, 80% of the molecular interactions are still unknown. Whether a link between two nodes exists must be demonstrated by field and/or laboratorial experiments⁵. However, if the accurate prediction results are applied into the laboratorial experiments instead of blindly checking all possible interactions, the costs will be sharply reduced⁶. In the scientists cooperation network, link prediction helps to find the potential cooperation between scientists^{7,8}. Besides, link prediction is also employed in recommending friends for online social networks^{9,10} and identifying spurious links in a noisy environment^{11,12}.

Until now, many indexes for link prediction have been proposed. Generally, they are classified into three main models: models based on Markov chains^{13–15}, models based on machine learning^{16,17} and models based on the similarity of topological structure^{1,18}. Though the first two have high prediction accuracy in many networks, they don't apply to the large-scale networks due to their high computational complexity. Nevertheless, models based on similarity can avoid such problems and easily obtain the information of networks. For instance, the Common Neighbor (CN) index, which is the most widely used index, just counts the number of common neighbors between node pair. Newman¹⁹ used this quantity in the study of collaboration networks, showing a positive correlation between the number of common neighbors and the probability that two scientists will collaborate in the future. By taking into account the common neighbors number and the degrees of two nodes, Salton et al. pointed out the Salton index²⁰; Leicht, Holme and Newman proposed the LHN index²¹. To characterize the topological similarity between reactants in the metabolic network, Ravasz E. and Somera A. L. et al. proposed the hub promoted index (HPI)²². HPI index insists that the links adjacent to hubs are likely to be assigned high scores since the denominator is determined by the common neighbors number and the lower degree. To measure with the opposite effect of hubs, Zhou and Lü et al. put forward the hub depressed index (HDI)²³. Furthermore,

China University of Petroleum, College of Computer and Communication Engineering, Qingdao, Shandong, 266580, China. Zhigang Zhang, Jianhang Liu, Tingpei Huang and Haihua Chen contributed equally. Correspondence and requests for materials should be addressed to S.L. (email: Lishibao@upc.edu.cn)

they proposed the resource allocation index (RA)²³. Motivated by the resource allocation dynamics on complex networks, the RA index can effectively improve the accuracy by restraining the contributions of large-degree common neighbors. Additionally, Liu et al. proposed a Local Naive Bayes (LNB) model²⁴, which insists that different common neighbors play different roles and make different contributions. Based on the LNB model, they improved the CN, RA and AA index. As the similarity indexes can predict links in networks, we can apply the similarity indexes to evaluate the evolving mechanisms for the evolving networks¹.

Obviously, similarity-based algorithms for link prediction can predict the future links by using the current common neighbors information²⁵. However, on above principles of the similarity indexes, some nodes, currently not common neighbors, can turn into the common neighbors in the future. More importantly, these nodes raise a series of new questions worth exploring. First of all, it is whether these nodes contribute to the current prediction between node pair. Although the previous algorithms have proved that the current common neighbors can promote two nodes to connect, people still doubt whether nodes, which are currently not a common neighbor but can become a common neighbor in the future, are also helpful in the current link prediction. Second, if they do make contribution, then how we can locate these nodes and measure their contribution, simultaneously. Previous algorithms can easily count the number of the current common neighbors by only analyzing the network topology. However, the nodes described above have not yet become the common neighbors, and they can get different topology structures when they are together with the node pairs and their surrounding nodes. These lead to the challenge of locating these nodes and measuring their contributions via a simple method.

To address the above problems clearly, firstly, we define nodes, which are currently not common neighbors but can turn into the common neighbors in the future, as the future common neighbors and divide them into three types according to their topology structure with other nodes. Second, we propose the similarity-based future common neighbors (SFCN) model for link prediction. The SFCN model accurately finds out all the future common neighbors, besides the current common neighbors. And simultaneously, it can also measure their contributions by only using the existing similarity indexes. We also design and observe three MATLAB simulation experiments. First, we conduct a priori experiment on α and β in FWFBN network. The results provide strong evidence that the future common neighbors have more positive contribution than the common neighbors in complex networks. Second, by comparing the SFCN model with eight similarity-based algorithms in five networks, we find that the SFCN model has higher prediction accuracy from the whole perspective. Third, the experiments, where we change the ratio of the training set to the probe set in five networks, also demonstrates that the SFCN model has better performance robustness. So, the proposed SFCN model has higher accuracy and performance robustness than popular algorithms, and the future common neighbors is necessary to be considered for link prediction in evolving networks.

Results

Network and problem description. A network can be represented by an undirected network $G(V, E)$ without self-connections and multiple links between node pair. In $G(V, E)$, V is the set of nodes, and E is the set of links. Then $|V|$ represents the quantity of nodes in V . Define the fully connected network as U that contains $(|V|(|V| - 1))/2$ links. So, $U - E$ is the set of the nonexistent links. To evaluate the prediction accuracy of algorithms, we divide the observed link set E into the training set E^T and the probe set E^P randomly. E^T is the known information while E^P is the unknown information. Obviously, $E = E^P \cup E^T$, and $\phi = E^P \cap E^T$. Accurately detecting the missing links or the future links from $U - E$ is the purpose of link prediction. Give the link between node pair (x, y) in U a score $(s_{x,y})$, which is calculated by the link prediction algorithm. All the nonexistent links are sorted in descending order according to their scores, and the links at the top are most likely to exist.

The future common neighbors. Most similarity-based algorithms for link prediction predict the future links by using the current common neighbors. However, on the prediction principles of the above similarity indexes, some nodes, which are currently not common neighbors, can turn into the common neighbors in the future. To analyze whether such nodes are factors that contribute to the current prediction between node pair, and in order to accurately locate these nodes to measure their contribution, we define them as the future common neighbors and propose the similarity-based future common neighbors model for link prediction in evolving networks.

The future common neighbors are nodes that are currently not the common neighbors but can turn into the common neighbors in the future on the principle of the similarity index. According to their topology with other nodes, the future common neighbors are divided into three types shown in Fig. 1, where x and y are the target node pair for link prediction. The first future common neighbor, like node i in Fig. 1(a), has a direct link with x while no direct link with y . Currently, the similarity score between i and y is $s_{i,y}$. According to the prediction principle of the similarity algorithm, i and y may form a link in the future (the greater the $s_{i,y}$, the greater the probability of forming a link). Therefore, i has connected with y and turn into the common neighbor between x and y in a future time. The second future common neighbor, like i in Fig. 1(b), has direct link with y while no direct link with x . The third future common neighbors does not connect with both x and y , seen node i in Fig. 1(c). According to the existing similarity indexes for link prediction, if $s_{x,i}$ and $s_{i,y}$ are great enough, i will form links with both x and y . Thus, the i in Fig. 1(c) are also a common neighbor between x and y in a future time.

Similarity-based future common neighbors model. Combining the future common neighbors topology with the similarity-based indexes, this paper designs the similarity-based future common neighbors model. The model is to accurately find out all the future common neighbors in complex networks and effectively measure their contributions.

Taking the chaotic network in Fig. 2 as an example, for node i ($i = 1, 2, 3, \dots, |V|$), we assume that $s_{x,i}^{C2}$ is a similarity score between x and i , and $s_{x,i}^{C2}$ is calculated by any classical similarity-based algorithms that we mark as C2.

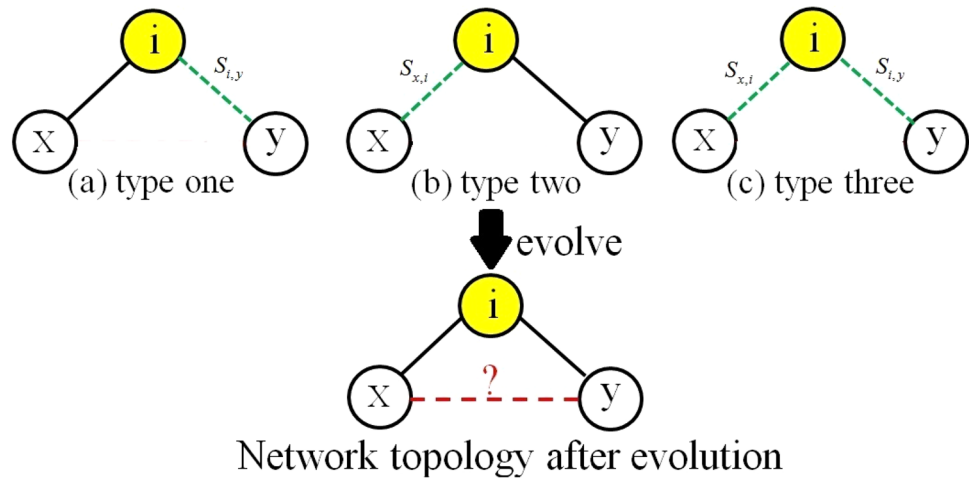


Figure 1. Three types of the future common neighbors.

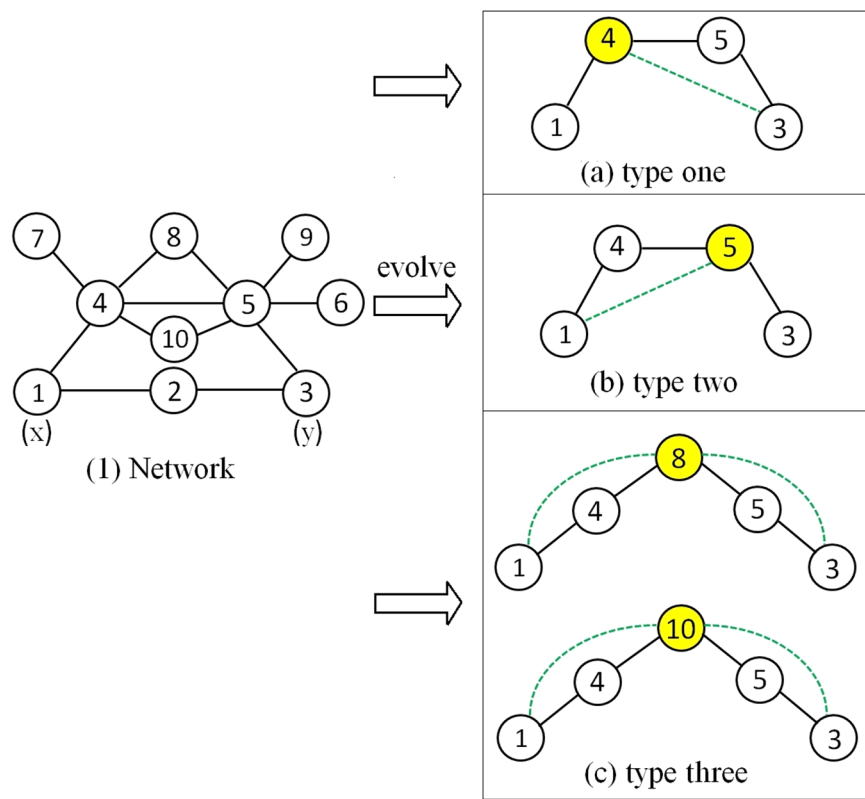


Figure 2. The process of identifying the future common neighbors from the chaotic network. The yellow nodes are the future common neighbors between x and y .

Therefore, $s_{x,i}^{C2}$ also symbolizes the possibility of forming a link between x and i on the principle of C2 when the network is evolving. We make $r_{i,y}$ indicate whether i and y are connected ($r_{i,y}=1$ if i and y are connected, otherwise $r_{i,y}=0$), which can be obtained from the observed networks. Similarly, $s_{i,y}^{C2}$ is a similarity score between i and y calculated by algorithms C2; $r_{x,i}$ represents whether x and i are connected.

The SFCN model identifies the above three types of the future common neighbors from chaotic network by employing their topological rules. (1) i is the first type of the future common neighbor only when $s_{x,i}^{C2} \cdot r_{i,y} \neq 0$. It is necessary to note that we set $s_{x,i}^{C2} = 0$ and $r_{i,y} = 0$ when $i = x$ or $i = y$ in order to keep non self-connections. (2) The rest rules can be deduced by analogy, i is the second type of the future common neighbor if $r_{x,i} \cdot s_{i,y}^{C2} \neq 0$. (3) And i is the third type of the future common neighbor if and only if $s_{x,i}^{C2} \cdot s_{i,y}^{C2} \neq 0$.

To accumulate the contributions of the future common neighbors, which meet the above rules, we construct four vectors for x and y in eqs 1, 2, 3, 4.

$$\Gamma_x = (r_{x,1} \ \cdots \ r_{x,x-1} \ 0 \ r_{x,x+1} \ \cdots \ 0), \tag{1}$$

$$(\Gamma_y)^T = (r_{1,y} \ \cdots \ r_{y-1,y} \ 0 \ r_{y+1,y} \ \cdots \ 0)^T, \tag{2}$$

$$S_x^{C2} = (s_{x,1}^{C2} \ \cdots \ s_{x,x-1}^{C2} \ 0 \ s_{x,x+1}^{C2} \ \cdots \ 0), \tag{3}$$

$$(S_y^{C2})^T = (s_{1,y}^{C2} \ \cdots \ s_{y-1,y}^{C2} \ 0 \ s_{y+1,y}^{C2} \ \cdots \ 0)^T, \tag{4}$$

where the superscript T denotes matrix transposition, and the black highlighted parts are the row vectors or column vectors. Γ_x stores the connections of x to all nodes. S_x^{C2} stores the similarity scores of x to all nodes. Similarly, $(\Gamma_y)^T$ stores the connections of y to all nodes. And $(S_y^{C2})^T$ stores the similarity scores between y and all other nodes.

Therefore, we get the similarity-based future common neighbors model as eq. 5:

$$\begin{aligned} s_{x,y}^{SFCN} &= \alpha \cdot s_{x,y}^{C1} + \beta \cdot \left(\Gamma_x \cdot (S_y^{C2})^T + S_x^{C2} \cdot (\Gamma_y)^T + S_x^{C2} \cdot (S_y^{C2})^T \right) \\ &= \alpha \cdot s_{x,y}^{C1} + \beta \cdot \sum_{i=1}^{|V|} (r_{x,i} \cdot s_{i,y}^{C2} + s_{x,i}^{C2} \cdot r_{i,y} + s_{x,i}^{C2} \cdot s_{i,y}^{C2}), \end{aligned} \tag{5}$$

where $s_{x,y}^{C1}$ is the similarity score between x and y , and $s_{x,y}^{C1}$ is calculated by any similarity algorithm that we temporarily mark as C1. C1 and C2 are two similarity algorithms, and they can be the same or different. The two free parameters, α and β , is to adjust the contributions of the current common neighbors and the future common neighbors, respectively. When $\alpha \neq 0$ and $\beta = 0$, the model only considers the current common neighbors contributions. When $\alpha = 0$ and $\beta \neq 0$, the model only considers the future common neighbors contributions. Both C1 and C2 are the. A special case, when both C1 and C2 are represented by CN algorithm and only the first class of the future common neighbors are considered, the model degenerates into the LP index. In a word, the SFCN model, employed in evolving networks, takes into account the contributions of the future common neighbors besides the current common neighbors.

Example 1. This section gives an example of how to find the future common neighbors between node pair and how to measure the contributions of three future common neighbors. Suppose C1 and C2 are the LHN and RA algorithms, respectively. Take the network in Fig. 2 as an example and treat nodes (1, 3) as the target node pair (x, y) . Then we can get four vectors:

$$\Gamma_1 = (0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0), \tag{6}$$

$$(\Gamma_3)^T = (0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0)^T, \tag{7}$$

$$S_1^{C2} = \left(0 \ 0 \ 0 \ 0 \ \frac{1}{5} \ 0 \ \frac{1}{5} \ \frac{1}{5} \ 0 \ \frac{1}{5} \right), \tag{8}$$

$$(S_3^{C2})^T = \left(0 \ 0 \ 0 \ \frac{1}{6} \ 0 \ \frac{1}{6} \ 0 \ \frac{1}{6} \ \frac{1}{6} \ \frac{1}{6} \right)^T. \tag{9}$$

From the calculation process (eq. 10), it is easy to observe that only node 4 is the first type of the future common neighbors. And the contribution of node 4 to (1, 3) is $\frac{1}{6} \cdot \beta$.

$$\Gamma_1 \cdot (S_3^{C2})^T = \sum_{i=1}^{|V|} (r_{1,i} \cdot s_{i,3}^{C2}) = r_{1,4} \cdot s_{4,3}^{C2} = \frac{1}{6}. \tag{10}$$

We can also observe that only node 5 is the second type of the future common neighbors from the eq. 11:

$$S_1^{C2} \cdot (\Gamma_3)^T = \sum_{i=1}^{|V|} (s_{1,i}^{C2} \cdot r_{i,3}) = s_{1,5}^{C2} \cdot s_{5,3}^{C2} = \frac{1}{5}. \tag{11}$$

At last, we can check out that 8 and 10 are the third type of the future common neighbors through eq. 12:

$$S_1^{C2} \cdot (S_3^{C2})^T = \sum_{i=1}^{|V|} (s_{1,i}^{C2} \cdot s_{i,3}^{C2}) = s_{1,8}^{C2} \cdot s_{8,3}^{C2} + s_{1,10}^{C2} \cdot s_{10,3}^{C2} = \frac{1}{15}. \tag{12}$$

Networks	$ V $	$ E $	$\langle d \rangle$	$\langle k \rangle$	$\langle H \rangle$	$\langle C \rangle$	$\langle r \rangle$
CE	297	2148	2.46	14.4646	1.8008	0.3079	-0.163
FWFB	128	2075	1.78	32.4219	1.2370	0.3346	-0.112
NS	379	914	4.93	4.82	1.66	0.798	-0.082
PB	1222	16714	2.74	27.3552	2.9707	0.3600	-0.221
Yeast	2375	11693	5.09	9.8467	3.4756	0.3883	0.454

Table 1. Details of networks. $|V|$ and $|E|$ are the number of nodes and links, respectively. $\langle d \rangle$ is the average shortest distance between node pairs. $\langle k \rangle$ is the average degree, and $\langle H \rangle$ denotes the degree heterogeneity. $\langle C \rangle$ represents the clustering coefficient. $\langle r \rangle$ is the assortative coefficient.

Therefore, the contribution of the future common neighbors is $\beta \cdot \left(\frac{1}{6} + \frac{1}{5} + \frac{1}{15}\right) = \frac{13}{30} \cdot \beta$.

Evaluation Metrics. In the Experiments, we introduce two standard metrics to quantify the prediction accuracy: the *AUC*²⁶ (area under the receiver operating characteristic curve) and *precision*²⁷. The AUC evaluate the algorithms performance according to the whole list. The AUC is comprehended as the probability that a link randomly chosen from set E^T has a much higher score than a link randomly chosen from nonexistent link $U-E$. In the n times independent comparisons, we select a link from E^T and $U-E$ respectively. Define their similarity scores as $S1$ and $S2$. When $S1 > S2$, set $n' = n' + 1$; when $S1 = S2$, set $n'' = n'' + 1$ (n' and n'' are initialed as 0, $n = n' + n''$). So, the *AUC* can be defined as eq. 13:

$$AUC = (n' + 0.5n'')/n. \quad (13)$$

Different from the *AUC*, precision focuses on the links with top ranks or highest scores. It is the ratio of correct links recovered out of the top L links in the candidate list generated by each link predictor. Assume L_r links are accurately predicted among the top- L links. Then the precision can be defined as eq. 14:

$$\text{Precision} = \frac{L_r}{L}. \quad (14)$$

Datasets of real networks. In order to compare the prediction accuracy of the SFCN model with the eight mainstream indexes mentioned in this paper, we do MATLAB simulation experiments in five real networks: the network of scientific communication (NS)²⁸, the US political blogs network (PB)²⁹, the protein interaction network (Yeast)³⁰, the neural network of *C.elegans* (CE)³¹, the food web network of florida bay (FWFB)³². All datasets of the five networks can be seen in the electronic supplementary material. The basic features of those networks are summarized in Table 1.

The metrics that characterize the networks can be seen in the caption of Table 1. We find that NS, PB and CE have similar characteristics, including the high clustering coefficient. Nevertheless, for FWFB network, the relation between predator and prey makes the network have a larger average degree and a shorter average distance between node pair.

Existing similarity indexes based on topological structure. Here, we introduce eight mainstream similarity indexes to compare with the SFCN model.

- CN. Let Γ_x be the set of neighbors of x . The CN index proposes that node pair (x, y) are more likely to connect if they have more common neighbors, namely:

$$s_{x,y}^{CN} = |\Gamma_x \cap \Gamma_y|. \quad (15)$$

- Salton²⁰. It is defined as:

$$s_{x,y}^{Salton} = \frac{|\Gamma_x \cap \Gamma_y|}{\sqrt{k_x k_y}}, \quad (16)$$

where k_x is the degree of node x .

- RA²³. The RA index assumes each transmitter has a unit of resource and will equally distributed to all its neighbors, concluded as:

$$s_{x,y}^{RA} = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{k_z}. \quad (17)$$

- HPI²². It is defined as:

$$s_{x,y}^{HPI} = \frac{|\Gamma_x \cap \Gamma_y|}{\min\{k_x, k_y\}}. \quad (18)$$

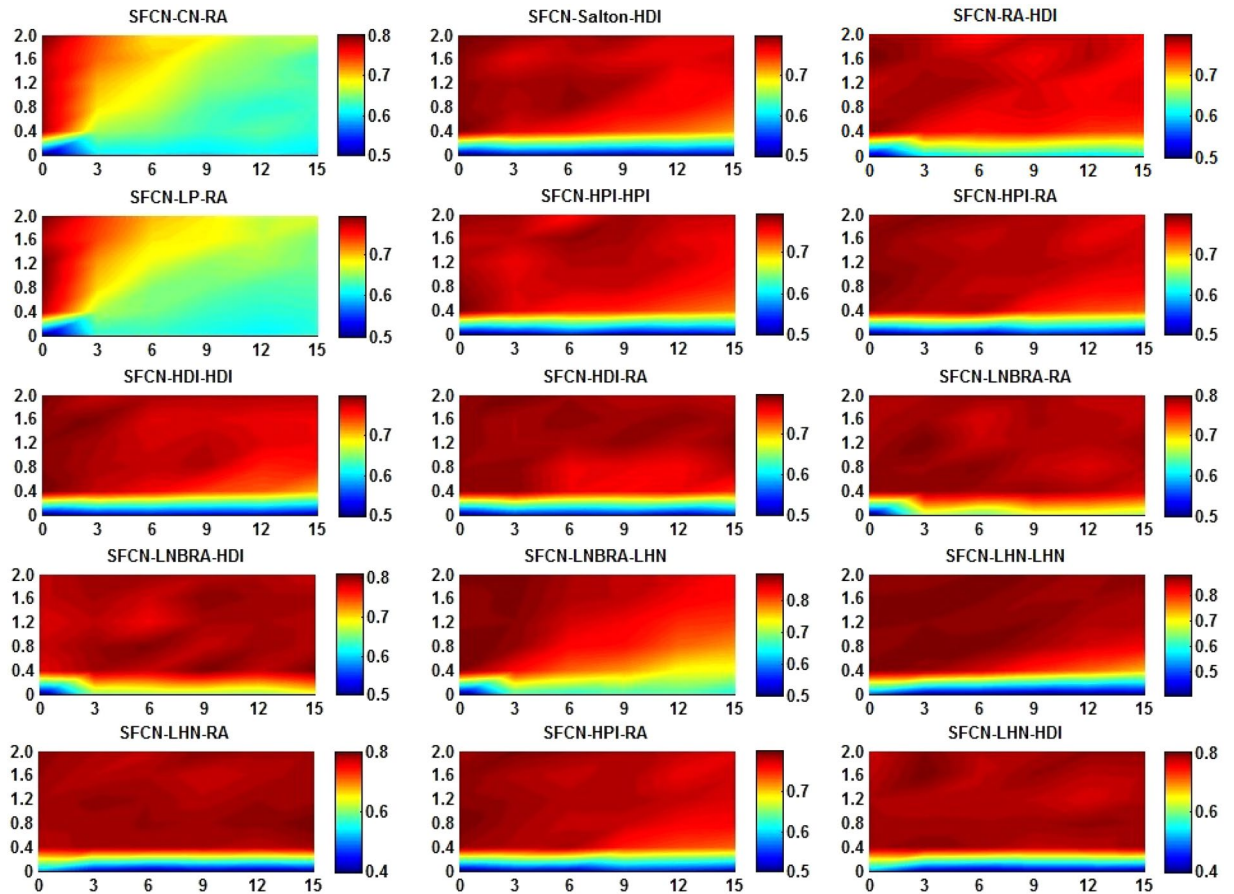


Figure 3. AUC sensitivity analysis of the SFCN model in FWFB network. X-axis is the α value that is taken from 0 to 15 at intervals of 3. Y-axis is the β value that is taken from 0 to 2 at intervals of 0.4.

- HDI²³. It is defined as:

$$s_{x,y}^{HDI} = \frac{|\Gamma_x \cap \Gamma_y|}{\max\{k_x, k_y\}}. \tag{19}$$

- Leicht-Holme-Newman index (LHN)²¹. The LHN index is defined as:

$$s_{x,y}^{LHN} = \frac{|\Gamma_x \cap \Gamma_y|}{k_x k_y}. \tag{20}$$

- LNBRA²⁴. The LNBRA index is an improvement in RA index based on the LNB model, defined as:

$$s_{x,y}^{LNBRA} = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{k_z} (\log_2 \eta + \log_2 R_z), \tag{21}$$

where η and R_z are defined as:

$$\eta = \frac{|V|(|V| - 1)}{2|E^T|} - 1, \tag{22}$$

$$R_z = \frac{N_{\Delta z} + 1}{N_{\nabla z} + 1}, \tag{23}$$

where $N_{\Delta z}$ and $N_{\nabla z}$ are respectively the numbers of connected and disconnected node pairs which have a common neighbor z .

- Local Path (LP)³³. This index considers the number of different orders, defined as:

AUC	CE	FWFB	NS	PB	Yeast
CN	0.846	0.616	0.989	0.925	0.917
SFCN-CN-RA	0.876	0.659	0.989	0.939	0.971
Salton	0.802	0.532	0.984	0.880	0.914
SFCN-Salton-HDI	0.851	0.793	0.984	0.937	0.974
RA	0.871	0.598	0.977	0.927	0.924
SFCN-RA-HDI	0.872	0.794	0.991	0.938	0.975
LP	0.861	0.633	0.980	0.938	0.971
SFCN-LP-RA	0.872	0.666	0.990	0.942	0.978
HPI	0.804	0.528	0.979	0.855	0.912
SFCN-HPI-HPI	0.811	0.762	0.985	0.901	0.972
SFCN-HPI-RA	0.865	0.790	0.989	0.945	0.974
HDI	0.775	0.527	0.980	0.873	0.914
SFCN-HDI-HDI	0.849	0.782	0.985	0.935	0.976
SFCN-HDI-RA	0.889	0.795	0.990	0.947	0.977
LNBRA	0.863	0.659	0.980	0.928	0.920
SFCN-LNBRA-RA	0.883	0.796	0.993	0.949	0.977
SFCN-LNBRA-HDI	0.881	0.809	0.992	0.942	0.975
SFCN-LNBRA-LHN	0.878	0.843	0.989	0.941	0.975
LHN	0.725	0.390	0.974	0.766	0.906
SFCN-LHN-LHN	0.810	0.891	0.983	0.891	0.974
SFCN-LHN-RA	0.876	0.797	0.992	0.947	0.977
SFCN-LHN-LP	0.806	0.704	0.974	0.928	0.961
SFCN-LHN-HDI	0.839	0.798	0.984	0.936	0.976

Table 2. There are the prediction accuracy results, measured by AUC, of classic indexes and corresponding algorithms based on the SFCN model in five real networks.

Precision	CE	FWFB	NS	PB	Yeast
CN	0.198	0.094	0.396	0.460	0.678
SFCN-CN-RA	0.202	0.108	0.404	0.464	0.766
Salton	0.012	0.008	0.290	0.000	0.024
SFCN-Salton-HDI	0.012	0.010	0.260	0.000	0.032
RA	0.124	0.094	0.564	0.256	0.520
SFCN-RA-HDI	0.130	0.098	0.566	0.278	0.440
LP	0.124	0.112	0.312	0.400	0.654
SFCN-LP-RA	0.140	0.112	0.324	0.410	0.696
HPI	0.026	0.068	0.556	0.224	0.860
SFCN-HPI-HPI	0.036	0.388	0.192	0.224	0.868
SFCN-HPI-RA	0.116	0.382	0.162	0.566	0.896
HDI	0.032	0.008	0.310	0.002	0.030
SFCN-HDI-HDI	0.086	0.360	0.320	0.516	0.900
SFCN-HDI-RA	0.106	0.360	0.294	0.590	0.882
LNBRA	0.131	0.162	0.544	0.252	0.586
SFCN-LNBRA-RA	0.136	0.166	0.554	0.250	0.580
SFCN-LNBRA-HDI	0.130	0.164	0.564	0.278	0.602
SFCN-LNBRA-LHN	0.132	0.154	0.580	0.260	0.586
LHN	0.000	0.014	0.138	0.000	0.010
SFCN-LHN-LHN	0.000	0.026	0.138	0.000	0.014
SFCN-LHN-RA	0.000	0.014	0.138	0.000	0.012
SFCN-LHN-LP	0.000	0.020	0.138	0.000	0.012
SFCN-LHN-HD	0.000	0.018	0.140	0.000	0.010

Table 3. There are the prediction accuracy results, measured by precision (top-100), of classic indexes and corresponding algorithms based on the SFCN model in five real networks.

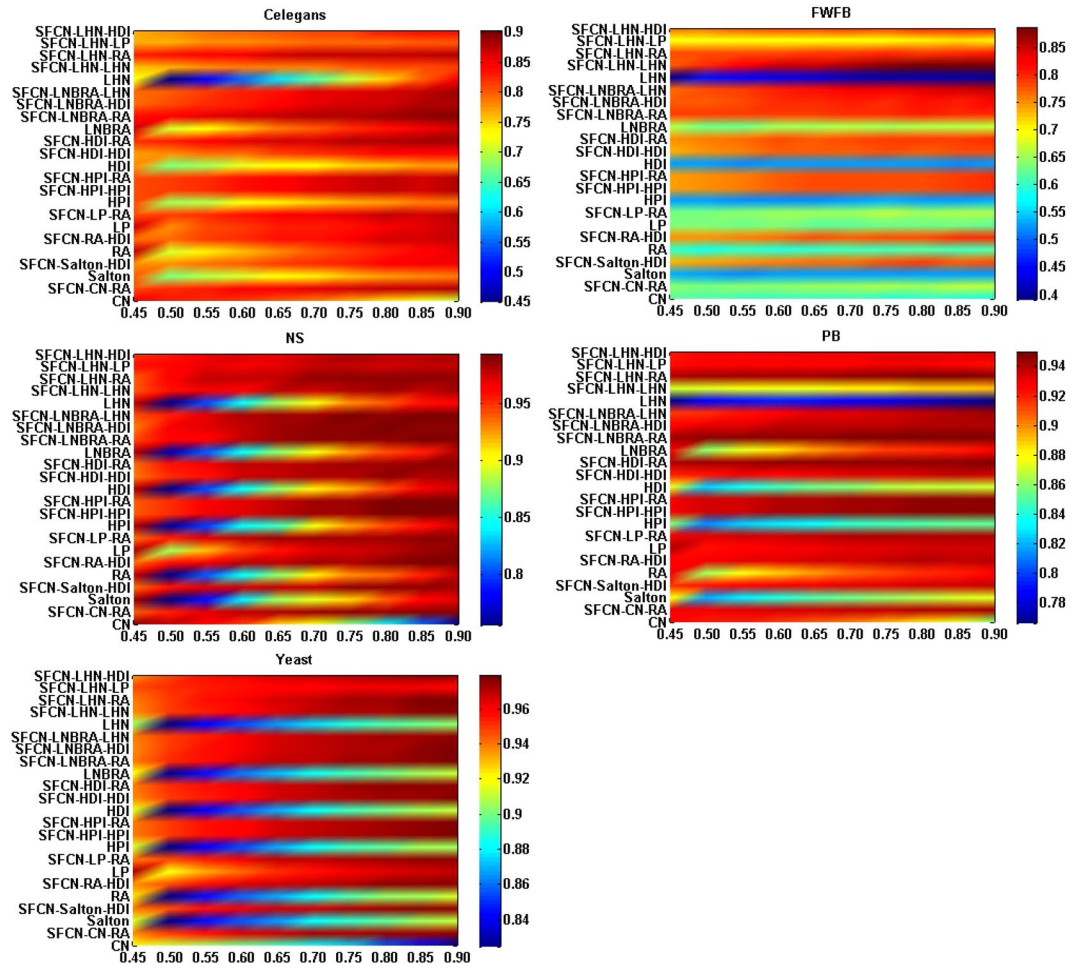


Figure 4. The AUC of different algorithms with different ratio of training sets to probe sets in real networks. X-axis is the ratio, and Y-axis is the each algorithm.

$$s_{x,y}^{LP} = (A^2)_{x,y} + \alpha \cdot (A^3)_{x,y}, \tag{24}$$

where α is an adjustable parameter and A is the adjacency matrix of network. $(A^i)_{x,y}$ represents the quantity that the order length is equal to i between x and y .

Experiments and performance analysis. In this section, we do three experiments and make corresponding analysis for three purposes. In the first and second experiments, the E^T contains 90% of links, while the remaining 10% of links constitute the E^P . In addition, all the following results are returned with the average over 100 independent experiments.

For the first experiment, to verify whether the contribution of the future common neighbors is necessary, we conducted priori experiments on α and β in FWFB network. Step 1, since the training set (E^T) is known, we divide E^T into the sub-training set (E^{T1}) and the sub-probe set (E^{P1}) to learn the values of α and β in step 2. Step 2, we apply the SFCN model to the sub-training set in order to obtain the similarity scores of the sub-probe set and get the AUC that varies with α and β . In this way, it is easy to select the numerical values of α and β with high AUC for the SFCN model. The experimental results are shown in Fig. 3. Before that, it is necessary to consider the two limit problems. When $\alpha = 0$, the current common neighbors in the SFCN model do not make any contribution, and only the future common neighbors make contributions. When $\beta = 0$, there are only contribution from the current common neighbors, and the future common neighbors do not make any contributions for link prediction. We can get two results from the Fig. 3. First, the AUC when $\beta = 0$ is much lower than that when $\beta \neq 0$. Second, the SFCN model can obtain highest AUC when α and β are adjust to a suitable value. For example, for the SFCN-CN-RA, SFCN-Salton-HDI, and SFCN-LNBRA-LHN algorithms, we should set α smaller and β larger to get a higher prediction accuracy in FWFB network. These two results illustrate the important contribution of the future common neighbors.

Therefore, in the second and the third experiments later, we set $\alpha = 9$ and $\beta = 1$, which meet the above condition.

The second experiment is to compare the SFCN model with other eight similarity-based indexes, including the CN, HDI, HPI, LP, RA, Salton, LNBRA and LHN index. The prediction results of AUC and precision are listed in Tables 2 and 3 for details, respectively. Most comparative experiments, in the Table 2, clearly demonstrate that the SFCN model has the best or close to the best AUC, especially in the FWFB and Yeast networks. Taking

the FWFB network as an example for analysis, we can see that there are 2075 links but only 128 nodes from the Table 1. And the average degree is as high as 32.422 while the average aggregation coefficient is low to 0.3346, which indicate that there are many random connections and high obscure similarity between the clusters in the FWFB network. These are the reasons why all nodes in the FWFB network have the tendency to gather and form some unknown clusters with the network evolving. The SFCN model takes into account the network evolution tendency via the principle of similarity index. In detail, the model has greatly improved the AUC in FWFB network by regarding the future common neighbors as the evolution direction. Moreover, Table 3 demonstrates that 90% of the precision results, predicted based on the SFCN model, are equal to or higher than that predicted based their original algorithm. For example, the precision results of the SFCN-HDI-RA algorithm are much higher than those of the original HDI algorithm in most network, because the contributions of the future common neighbors are taken into account.

Finally, in order to explore the robustness, we change the ratio of training set to probe set in the third experiment. The lower the ratio, the more links information that should be predicted³⁴. That is to say, there are less number of the known connected links and more number of the unknown links when the ratio is small. It is obviously to obtain two results from the Fig. 4. On the one hand, when the ratio is the same, the algorithms based on SFCN model have higher prediction accuracy results (measured by AUC) than their corresponding original algorithms. For instance, the SFCN-LHN-RA, SFCN-LHN-LP and SFCN-LHN-HDI algorithms have higher AUC compared with the original LHN algorithms when the ratio is the same. On the other hand, even when the ratio is low, the algorithms based on SFCN model still get high AUC, which indicates that the SFCN model has higher stability. Therefore, the SFCN model has better performance in prediction accuracy and stability even when there is few links information.

Discussion

Exploring what factors can provide a positive impact on link prediction is an important and challenging problem. In this paper, we firstly discover the existence of the future common neighbors, which are classified into three types according to their topological structure with other nodes. Then, to investigate whether the future common neighbors can make positive contribution for current link prediction, we propose the similarity-based future common neighbors model (SFCN), which accurately locates all the future common neighbors and effectively measure their contributions in complex networks, besides the current common neighbors.

We design three simulation experiments via the MATLAB for three different purposes. First, we conduct priori experiments on α and β in FWFB network. The results provide strong evidence that the future common neighbors can make great contribution than current common neighbors in complex networks. In the second experiments, we compare the SFCN model with eight algorithms in five networks, finding that the SFCN model has higher prediction accuracy, especially the AUC in the FWFB and Yeast networks. Third, in order to verify whether the SFCN model can get great accuracy when the known link information is little, we change the ratio of the training set to the probe set in five networks. And the experiment results show that the SFCN model has better performance robustness, even when the ratio is low to 0.45, compared with eight similarity-based algorithms. Therefore, the proposed SFCN model has higher accuracy and performance robustness than popular similarity-based algorithms, and the future common neighbors make more positive contribution than the current common neighbors that is widely used nowadays.

Some extensions of this work deserve further exploration. One is that we are limited to the current common neighbors and the future common neighbors in evolving networks. It is meaningful to research the contribution of the future nodes and the future links. For example, current path-based algorithms only consider the contribution of the existing paths currently, so it is significative to further exploit whether and how much the future paths, which are not existing currently but will exist after once prediction, can make a positive impact on current link prediction.

Methods

Algorithm of the SFCN model for link prediction. The adjacency matrix E is a sparse matrix of the complex network. And the pseudocode of the SFCN model is presented in algorithm 1.

Algorithm 1. Algorithm of the proposed SFCN framework.

Input: the adjacency matrix E of the complex network, parameters α and β .

Output: the similarity score matrix of all the nonexistent links, $\mathbf{S}_{x,y}^{\text{SFCN}}$;

- 1: divide E into E^T , E^P with a ratio;
 - 2: compute and store the similarity score $\alpha \cdot s_{x,y}^{C1}$ of the current common neighbors through method C1;
 - 3: compute the similarity score $(\mathbf{S}_y^{C2})^T$ and \mathbf{S}_x^{C2} through method C2;
 - 4: get the neighbors vector Γ_x of x , and get the neighbors vector $(\Gamma_y)^T$ of y ;
 - 5: calculate and store the similarity score of the future common neighbors through the following formula: $\beta \cdot (\Gamma_x \cdot (\mathbf{S}_y^{C2})^T + \mathbf{S}_x^{C2} \cdot (\Gamma_y)^T + \mathbf{S}_x^{C2} \cdot (\mathbf{S}_y^{C2})^T)$;
 - 6: compute the final similarity score matrix $\mathbf{S}_{x,y}^{\text{SFCN}}$ by adding up the contribution of the current common neighbor and the future common neighbors;
- return** $\mathbf{S}_{x,y}^{\text{SFCN}}$;
-

Complexity analysis. This part give a simple complexity analysis of the proposed SFCN model. The most time-consuming part occurs in computing the contribution of the future common neighbors. The time cost of (S_y^{C2}) is $O(|V||V|)$, and the time cost of (S_x^{C2}) is $O(|V||V|)$. Thus the total time cost of the future common neighbors is $3 \cdot O(|V||V|)$. Since complex network can be simplified as a sparse matrix, the final computational complexity is much less than $3 \cdot O(|V||V|)$.

References

- Lü, L. & Zhou, T. Link prediction in complex networks: A survey. *Phys. A Stat. Mech. Its Appl* **390**, 1150–1170 (2010).
- Lai, D., Shu, X. & Christine, N. Link prediction in complex networks via modularity-based belief propagation. *Chin. Phys. B* **26**, 604–614 (2017).
- Bai, M., Hu, K. & Tang, Y. Link prediction based on a semi-local similarity index. *Chin. Phys. B* **20**, 498–504 (2011).
- Butts, C. T. Network inference, error, and informant (in)accuracy: a bayesian approach. *Soc. Networks* **25**, 103–140 (2003).
- Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Sci* **322**, 104–110 (2008).
- Cannistraci, C. V., Alanislobato, G. & Ravasi, T. Erratum: From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Sci. Reports* **3**, 1613 (2013).
- Amaral, L. A. A truer measure of our ignorance. *Proc. Natl. Acad. Sci. United States Am* **105**, 6795–6796 (2008).
- Zhang, Q., Xu, X., Zhu, Y. & Zhou, T. Measuring multiple evolution mechanisms of complex networks. *Sci. Reports* **5**, 10350 (2015).
- Guimer, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. United States Am* **106**, 22073–22078 (2009).
- Zhang, Y. The research of information dissemination model on online social network. *Acta Phys. Sinica* **60**, 50501 (2011).
- Ma, C., Zhou, T. & Zhang, H. Playing the role of weak clique property in link prediction: A friend recommendation model. *Sci. Reports* **6**, 30098 (2016).
- Chongqing, H. & Jiang, W. Estimating topology of complex networks based on sparse bayesian learning. *Acta Phys. Sinica* **14**, 071 (2012).
- Bilgic, M., Namata, G. M. & Getoor, L. Combining collective classification and link prediction. In *IEEE International Conference on Data Mining Workshops, 2007. ICDM Workshops*, 381–386 (2007).
- Zhu, J., Hong, J. & Hughes, J. G. Using markov chains for link prediction in adaptive web sites. In *Soft-Ware 2002: Computing in an Imperfect World*, 60–73 (Springer, 2002).
- Sarukkai, R. R. Link prediction and path analysis using markov chains. *Comput. Networks* **33**, 377–386 (2000).
- Clauset, A., Moore, C. & Newman, M. E. Hierarchical structure and the prediction of missing links in networks. *Nat* **453**, 98–101 (2008).
- Popescu, A. & Ungar, L. H. Statistical relational learning for link prediction. *Work. on Learn. Stat. Model. from Relational Data at IJCAI-2003* 2003 (2003).
- Liben-Nowell, D. & Kleinberg, J. The link-prediction problem for social networks. *journal Assoc. for Inf. Sci. Technol.* **58**, 1019–1031 (2007).
- Newman, M. E. Clustering and preferential attachment in growing networks. *Phys. review E* **64**, 025102 (2001).
- Salton, G. & McGill, M. J. Introduction to modern information retrieval. *Auckland: McGraw-Hill* (1983).
- Leicht, E. A., Holme, P. & Newman, M. E. Vertex similarity in networks. *Phys. Rev. E* **73**, 026120 (2006).
- Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *science* **297**, 1551–1555 (2002).
- Zhou, T., Lü, L. Y. & Zhang, Y. C. Predicting missing links via local information. *Eur. Phys. J. B* **71**, 623–630 (2009).
- Liu, Z., Zhang, Q., Lü, L. & Zhou, T. Link prediction in complex networks: a local naïve bayes model. *Epl* **96**, 48007 (2011).
- Alain Barrat, M. B. & Vespignani, A. Dynamical processes on complex networks. *Camb. Univ. Press* (2008).
- Hanley, J. A. & Mcneil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiol* **143**, 29 (1982).
- Herlocker, J. L., Konstan, J. A., Terveen, L. & Riedl, J. T. Evaluating collaborative filtering recommender system. *ACM Transactions on Inf. Syst* **22**, 5–53 (2004).
- Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. review E* **74**, 036104 (2006).
- Adamic, L. A. & Glance, N. The political blogosphere and the 2004 u.s. election: divided they blog. *Proc. 3rd international workshop on Link discovery* 36–43 (2005).
- Von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nat* **417**, 399–403 (2002).
- Watts, D. J. & Strogatz, S. H. Collective dynamics of small-world networks. *nature* **393**, 440 (1998).
- Ulanowicz, R. E., Bondavalli, C. & Egnotovitch, M. S. Network analysis of trophic dynamics in south florida ecosystems, fy 97: The florida bay ecosystem [r/ol]. *Tech. report, CBL* **12**, 98–123 (1998).
- Lü, L., Jin, C. & Zhou, T. Similarity index based on local paths for link prediction of complex networks. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys* **80**, 046122 (2009).
- Shakibian, H. & Charkari, N. M. Mutual information model for link prediction in heterogeneous complex networks. *Sci. Reports* **7**, 44981 (1982).

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant Nos 61402433 and 61601519; The Fundamental Research Funds for the Central University, with No. 18CX02134A, No.18CX02137A, No.18CX02133A.

Author Contributions

S.L. and Z.Z. conceived the study; S.L., J.H. and Z.Z. designed the experiments and algorithms; J.H. prepared figures and tables; S.L. and Z.Z. collected data; S.L. and J.H. wrote the paper. J.L., T.H. and H.C. modified the pictures and paper. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-35423-2>.

Competing Interests: The authors declare no competing interests.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018