Physica A 391 (2012) 5769–5778

Contents lists available at SciVerse ScienceDirect

Physica A

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

Uncovering missing links with cold ends

Yu-Xiao Zhu [a], Linyuan Lü [b,c], Qian-Ming Zhang [a], Tao Zhou [a,d,*]

[a] Web Sciences Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, PR China

[b] Institution of Information Economy, HangZhou Normal University, Hangzhou 310036, PR China

[c] Department of Physics, University of Fribourg, Chemin du Musee 3, CH-1700 Fribourg, Switzerland

[d] Beijing Computational Science Research Center, Beijing 100084, PR China

ARTICLE INFO

Article history:
Received 5 December 2011
Received in revised form 21 May 2012
Available online 9 June 2012

Keywords:
Complex networks
Link prediction
Cold ends
Sampling methods

ABSTRACT

To evaluate the performance of prediction of missing links, the known data are randomly divided into two parts, the training set and the probe set. We argue that this straightforward and standard method may lead to terrible bias, since in real biological and information networks, missing links are more likely to be links connecting low-degree nodes. We therefore study how to uncover missing links with low-degree nodes, namely links in the probe set are of lower degree products than a random sampling. Experimental analysis on ten local similarity indices and four disparate real networks reveals a surprising result that the Leicht–Holme–Newman index [E.A. Leicht, P. Holme, M.E.J. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2006) 026120] performs the best, although it was known to be one of the worst indices if the probe set is a random sampling of all links. We further propose an parameter-dependent index, which considerably improves the prediction accuracy. Finally, we show the relevance of the proposed index to three real sampling methods: acquaintance sampling, random-walk sampling and path-based sampling.

1. Introduction

Many social, biological, and information systems can be well described by networks, where nodes represent individuals and links denote the relations or interactions between nodes. The study of complex networks has therefore become a common focus of many branches of science. A fundamental tool for network analysis is the so-called *link prediction*, which attempts to estimate the likelihood of the existence of a link between two nodes, based on observed links and the attributes of nodes [1,2].

In many biological networks, such as food webs, protein–protein interaction networks and metabolic networks, the existence of a link must either be demonstrated by field and/or laboratory experiments, which are usually very costly. Our knowledge of these networks is very limited, for example, 80% of the molecular interactions in cells [3] and 99.7% of human [4,5] are still unknown. Instead of blindly checking all possible interactions, to predict based on known interactions and focus on those links most likely to exist can sharply reduce the experimental costs if the predictions are accurate enough. Social network analysis also comes up against the missing data problem [6,7], where link prediction algorithms may play a role. In addition, the data in constructing biological and social networks may contain inaccurate information, resulting in spurious links [8,9]. Link prediction algorithms can be applied to identify these spurious links [10,11]. Besides helping in analyzing networks with missing data, the link prediction algorithms can be used to predict the links that may appear

* Corresponding author at: Web Sciences Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, PR China.
E-mail address: zhutou@ustc.edu (T. Zhou).

0378-4371/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2012.06.003

in the future of evolving networks. For example, in online social networks, very likely but not yet existent links can be recommended as promising friendships, which can help users in finding new friends and thus enhance their loyalties to the web sites. Other applications of link prediction include the evaluation of network evolving models [12,13], the classification of partially labeled networks [14], and so on (see the review article [1] for the detailed discussion on real applications).

To evaluate the algorithmic performance, the data set is divided into two parts: the training set is treated as known information while the probe set is used to estimate the prediction accuracy. To our knowledge, the datasets are always divided completely randomly. This is the most straightforward way, and it seems a very fair method without any statistical bias. However, this straightforward and standard method may lead to terrible bias, since in real biological and information networks, missing links are more likely to be links connecting low-degree nodes. For example, the known structure of the World Wide Web is just a sampling, where the hyperlinks from popular web pages have higher probability to be uncovered. In contrast, hyperlinks from unbeknown web pages are probably lost. Actually, in common sense, interaction between two significant proteins, hyperlink between two well-known web pages and relationship between two famous persons are of less probability to be missed. Accordingly, in this article, we study how to uncover missing links with low-degree nodes. That is to say, we divide the data set into two parts and make the links in the probe set less popular (i.e., of less degree products) than the links in the training set. Experimental analysis on ten local similarity indices and four disparate real networks reveals a surprising result that the Leicht–Holme–Newman (LHN) index [15] performs the best, although it was known to be one of the worst indices if the probe set is a random sampling of all links [16]. We further propose an parameter-dependent index, which considerably improves the prediction accuracy. Finally, we show the relevance of the proposed index to three real sampling methods.

This article is organized as follows. In the next section, we will clearly define the problem of link prediction, describe the standard metric to evaluate the prediction accuracy, introduce the state-of-the-art local indices for node similarity and how to sample less popular links for the probe set. Experimental results for the traditional sampling method and the proposed method are presented in Section 3. In Section 4, we will propose an improved index which performs even better than the LHN index. In Section 5, we will introduce three mainstream sampling methods, and test the improved index on the corresponding sampled networks. Finally, we summarize our results in Section 6.

2. Problem description

2.1. Link prediction: problem and evaluation

Given an undirected network $G(V, E)$, where V and E are the sets of nodes and links respectively. The multiple links and self-connections are not allowed. Denote by U the universal set containing all $\frac{|V|(|V|-1)}{2}$ possible links, where $|V|$ denotes the number of elements in set V. Then, the set of nonexistent links is $U \setminus E$, in which there are some missing links (i.e., the existed yet unknown links). The task of link prediction is to uncover these links. Each node pair x and y will be assigned a score s_{xy} according to a given prediction algorithm. The higher the score is, the higher existence likelihood this link has. The score matrix S is symmetry for G is undirected. All the nonexistent links are sorted in descending order according to their scores, and the top-ranked links are most likely to exist.

To test the algorithmic accuracy, the observed links E are divided into two groups: the training set E^T is treated as known information, while the probe set E^P is used for testing and no information therein is allowed to be used for prediction. Clearly, $E = E^T \cup E^P$ and $E^T \cap E^P = \phi$. The accuracy of prediction is quantified by a standard metric called AUC (short for *area under the receiver operating characteristic curve*) [17]. Specifically, this metric can be interpreted as the probability that a randomly chosen missing link (links in E^P) has higher score than a randomly chosen nonexistent link (links in $U \setminus E$). In the implementation, among n times of independent comparisons, if there are n' times that the missing link has higher score and n'' times the missing link and nonexistent link have the same score, the AUC value is defined as

$$\text{AUC} = \frac{n' + 0.5n''}{n}. \tag{1}$$

If all the scores are generated from an independent and identical distribution, the accuracy should be about 0.5. Therefore, the extent to which the accuracy exceeds 0.5 indicates how much better the algorithm performs than pure chance.

2.2. Similarity indices

The simplest framework of link prediction is the similarity-based algorithm, where the aforementioned score s_{xy} is directly defined as the similarity between node x and y [1,18,19]. All non-observed links are ranked according to their scores, and the links connecting more similar nodes are supposed to be of higher existence likelihoods. Owning to its simplicity, the study on similarity-based algorithms is mainstream.

In this article, we adopt the simplest local similarity indices. Zhou et al. [16] have investigated the performances of ten local indices showing that the resource allocation index performs best, and the LHN and preferential attachment indices perform the worst. However, these results are obtained based on random division of the training and probe sets. We will compare the performances of these ten indices on predicting the missing links with low-degree nodes. The details of these indices are as follows [1,16].

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

$$s_{xy}^{\mathrm{CN}} = |\Gamma_x \cap \Gamma_y|. \tag{2}$$

Readers are encouraged to see a complicated variant of CN index based on a Bayesian model [20].

(2) Salton index [21]. It is defined as

$$s_{xy}^{\mathrm{Salton}} = \frac{|\Gamma_x \cap \Gamma_y|}{\sqrt{k_x \times k_y}}, \tag{3}$$

where k_x is the degree of node x. The Salton index is also called the cosine similarity in the literature.

(3) Jaccard index [22]. This index was proposed by Jaccard over one hundred years ago, defined as

$$s_{xy}^{\mathrm{Jaccard}} = \frac{|\Gamma_x \cap \Gamma_y|}{|\Gamma_x \cup \Gamma_y|}. \tag{4}$$

(4) Sørensen index [23]. This index is mainly used for ecological community data and is defined as

$$s_{xy}^{\mathrm{Sørensen}} = \frac{|\Gamma_x \cap \Gamma_y|}{k_x + k_y}. \tag{5}$$

(5) Hub Promoted index (HPI) [24]. This index is proposed for quantifying the topological overlap of pairs of substrates in metabolic networks, defined as

$$s_{xy}^{\mathrm{HPI}} = \frac{|\Gamma_x \cap \Gamma_y|}{\min\{k_x, k_y\}}. \tag{6}$$

Under this measurement, the links adjacent to hubs are likely to be assigned high scores since the denominator is determined by the lower degree only.

(6) Hub Depressed index (HDI) [16]. Analogously to the above index, Zhou et al. [16] considered a measurement with the opposite effect on hubs, defined as

$$s_{xy}^{\mathrm{HDI}} = \frac{|\Gamma_x \cap \Gamma_y|}{\max\{k_x, k_y\}}. \tag{7}$$

(7) Leicht–Holme–Newman index (LHN) [15]. This index assigns high similarity to node pairs that have many common neighbors compared not to the possible maximum, but to the expected number of such neighbors. It is defined as

$$s_{xy}^{\mathrm{LHN}} = \frac{|\Gamma_x \cap \Gamma_y|}{k_x \times k_y}, \tag{8}$$

where the denominator, $k_x \times k_y$, is proportional to the expected number of common neighbors of nodes x and y in the configuration model [25,26].

(8) Preferential Attachment (PA) [27]. The mechanism of preferential attachment can be used to generate evolving scale-free networks, where the probability that a new link is connected to the node x is proportional to k_x. Motivated by this mechanism, the corresponding similarity index can be defined as

$$s_{xy}^{\mathrm{PA}} = k_x \times k_y, \tag{9}$$

which has been widely used to quantify the functional significance of links subject to various network-based dynamics, such as percolation, synchronization and transportation. Note that this index does not require the information of the neighborhood of each node, as a consequence, it has the least computational complexity.

(9) Adamic–Adar (AA) index [28]. This index refines the simple counting of common neighbors by assigning the less-connected neighbors more weights, and is defined as

$$s_{xy}^{\mathrm{AA}} = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{\log k_z}. \tag{10}$$

(10) Resource Allocation (RA) index [16,29]. This index is motivated by the resource allocation dynamics on complex networks, and is defined as

$$s_{xy}^{\mathrm{RA}} = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{k_z} \tag{11}$$

where z runs over all common neighbors of x and y.

Note that the above indices except PA are all common-neighbor-based. Therein the Salton index, Jaccard index, Sørensen index, HPI, HDI and LHN are different in the dominators which take into account the degrees of the two endpoints of the predicted links, while AA and RA indices consider the effects of their common neighbors' degrees.

Table 1
The basic topological features of the giant components of the four example networks. NS, CE and PB are the abbreviations for NetScience, C.elegans and Political Blogs networks, respectively. $N = |V|$ and $M = |E|$ are the total number of nodes and links, respectively. C is the clustering coefficient [32] that is defined as the average ratios of the number of connected pairs of a node's neighbors to the possible maximum. r is the assortative coefficient [39], the Pearson correlation coefficient of degree between pairs of connected nodes. r lies between -1 and 1, and $r > 0$ indicates a positive correlation while $r < 0$ indicates a negative correlation. $\langle k\rangle$ is the average degree of the network, and $\langle d\rangle$ is the average shortest distance between node pairs. H denotes the degree heterogeneity defined as $H = \frac{\langle k^2\rangle}{\langle k\rangle^2}$.

Datasets	N	M	C	r	$\langle k\rangle$	$\langle d\rangle$	H
USAir	332	2 126	0.749	−0.208	12.81	2.46	3.46
NS	379	914	0.798	−0.082	4.82	4.93	1.66
CE	297	2 148	0.308	−0.163	14.46	2.46	1.80
PB	1222	16 717	0.361	−0.221	27.36	2.51	2.97

2.3. Sampling for probe set

Traditionally, the probe links are randomly selected from E, namely each link has an equal probability of being selected into the probe set (called *random sampling*), and the algorithmic accuracy measured by AUC is an average of the probe links. However, the links may have different predictabilities for their different roles in the network. Some algorithms may be good at predicting the links connecting the high-degree nodes, while some are adept in the links connecting the low-degree nodes. Therefore, in order to evaluate the performance of different algorithms on different links, the dataset should be divided with preference.

Motivated by evaluating the algorithm's performance on uncovering the links with low-degree nodes, in this article, we propose a preferential partition method according to the link *popularity* which is defined as:

$$\mathrm{pop}_{(x,y)} = (k_x - 1) \times (k_y - 1), \tag{12}$$

where k_x denotes the degree of node x. Clearly, links with high-degree endpoints have higher popularities than those with low-degree ends. Thus for a given network, the links whose popularities are higher than the average popularity $\langle \mathrm{pop}\rangle$ are called popular links, and those with lower popularities than $\langle \mathrm{pop}\rangle$ are unpopular links. The detailed partition steps are as follows: (i) Calculate the popularity score of each observed link according to Eq. (12), and rank these links in descending order based on their popularity scores. (ii) Uniformly divide this list from down to up into D groups respectively denoted by $E_1, E_2, \ldots, E_D$. Clearly, $E_1 \cup E_2 \cup \cdots \cup E_D = E$ and $E_i \cap E_j = \phi$, $(i, j = 1, 2, \ldots, D, i \neq j)$. The popularity of each link in E_i is no higher than that in E_j if $i < j$. (iii) For each subset E_i, we randomly choose half of the links therein to constitute the probe set labeled by E_i^P. Then the rest links (i.e., $E_i \setminus E_i^P$) constitute the corresponding training set labeled by E_i^T. Denote by $\langle \mathrm{pop}\rangle_i$ the average popularity of the links in probe set E_i^P, we have $\langle \mathrm{pop}\rangle_1 < \langle \mathrm{pop}\rangle_2 < \cdots < \langle \mathrm{pop}\rangle_D$. E_1^P consisting of the most unpopular links is called cold probe set in this article. We design this sampling method for the convenience of theoretical analysis, however, this method is far different from real sampling methods. We will therefore test the relevance and validity of our main results on real sampling methods in Section 5.

3. Experimental analysis

3.1. Data description

We consider four representative networks drawn from disparate fields: (i) USAir. The network of the US air transportation system, which contains 332 airports and 2126 airlines [30]. (ii) NetScience. A network of coauthorships between scientists who are themselves publishing on the topic of networks [31]. The network contains 1589 scientists, 128 of whom are isolated. Here we consider the largest component that contains only 379 scientists and 914 links. (iii) C.elegans. The neural network of the nematode worm C.elegans, in which an edge joins two neurons if they are connected by either a synapse or a gap junction [32]. This network contains 297 neurons and 2148 links. (iv) Political Blogs. The network of US political blogs [33]. The original links are directed, here we treat them as undirected links. Table 1 summarizes the basic topological features of these networks. Brief definitions of the monitored topological measures can be found in the table caption. For more details about the topological metrics, one can see the review articles [34–38].

3.2. Results for random sampling

As we have mentioned above, the mainstream method to prepare the probe set is random sampling, namely all the links in E^P are randomly chosen from E. For example, Zhou et al. compared ten local similarity indices on five real networks [16] with this randomly selected probe set, and gave an overall evaluation measured by AUC. Instead of obtaining a collective performance of the whole probe set, here we investigate the algorithm's performance on each link. The accuracy of one link is defined as the probability that this link has higher score than that of one randomly chosen nonexistent link. The dependence of four typical algorithms' accuracies on the popularity of links is shown in Fig. 1. Note that the popularity of each link in the probe set is calculated according to the original dataset, not the training set. Fig. 1(a)–(l) show that the AUC

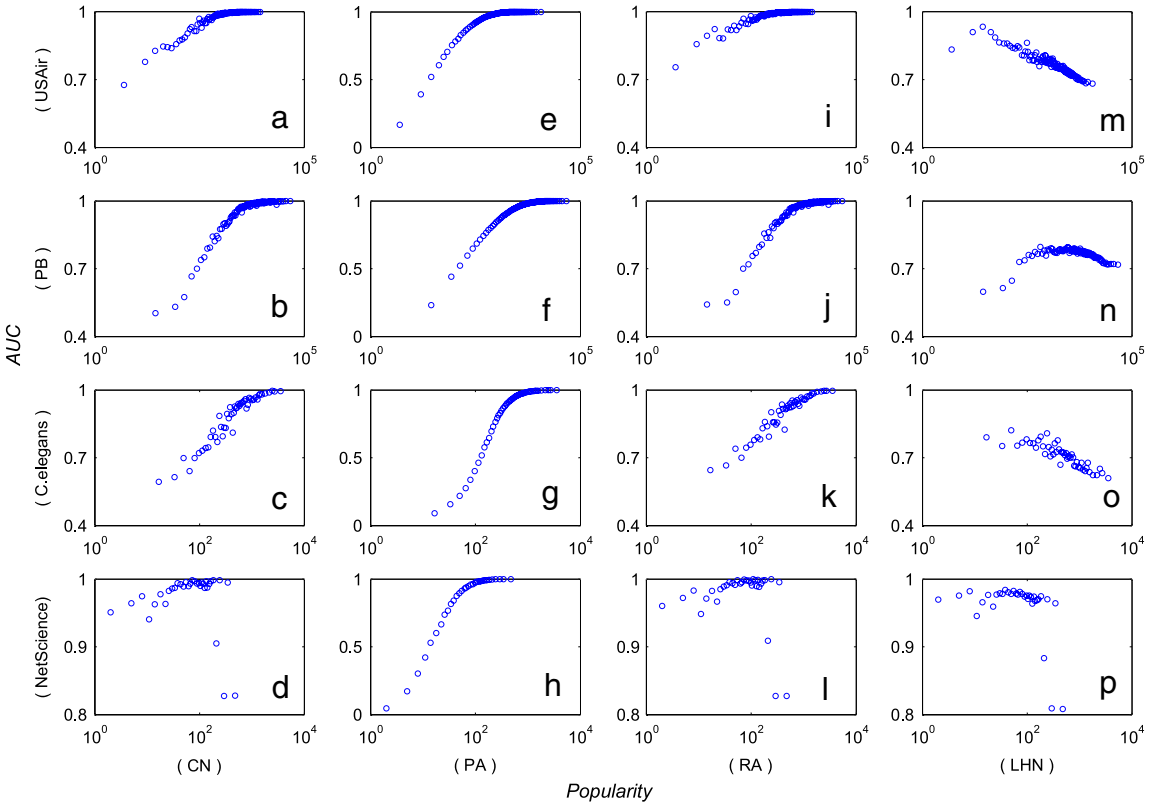

Fig. 1. The dependence of algorithmic accuracy on the popularity of links. The results for four typical indices, namely CN, PA, RA and LHN are shown respectively in four columns. Each subgraph is obtained by averaging over 100 implementations with independently random partitions of training set and probe set. The probe set contains 5% of observed links. Note that the y-coordinate denotes the average AUV value of links with the same popularity. The statistics are conducted with log-bin of popularity.

Table 2
Algorithmic novelty of all local similarity indices on C.elegans, measured by the top-L popularity. Sal, Jac and Sør are the abbreviations for Salton, Jaccard and Sørensen indices respectively. These abbreviations are also adopted in remaining sections. Each number is obtained by averaging over 100 implementations with independently random divisions of training set and probe set. The probe set contains 5% of observed links. Here $L = 100, 500, 1000, 5000, 10\,000, 20\,000$.

L	CN	Sal	Jac	Sør	HPI	HDI	LHN	PA	AA	RA
100	136.7	6.603	9.880	9.880	0.271	14.43	0.000	119.3	147.5	139.9
500	48.23	5.285	7.946	7.947	1.756	9.405	0.006	47.69	51.89	48.89
1 000	30.97	6.766	8.759	8.759	0.878	9.507	0.035	30.62	32.80	31.29
5 000	9.761	5.427	5.367	5.367	6.574	4.838	0.628	9.483	9.956	9.973
10 000	5.397	4.475	4.018	4.018	4.770	3.544	1.383	5.197	5.421	5.453
20 000	2.819	2.749	2.687	2.687	2.791	2.652	2.637	2.786	2.819	2.819

increases with the increasing of link popularity. This indicates that the CN, PA and RA algorithms tend to give higher accurate predictions on popular links, especially for USAir, PB and C.elegans networks. In comparison, the LHN index, which has been demonstrated to be a lowly accurate method in previous works, can give highly accurate prediction on the unpopular links. The reason is that LHN is likely to assign higher scores to the unpopular links by using $k_x \times k_y$ as its dominator to depress the scores of popular links.

We further investigate the average popularity of the top-L predicted links of different algorithms. In principle, a link prediction algorithm provides a descending ordered list of all non-observed links (i.e., $U \setminus E^P$ in our experiment) according to their scores, of which we only focus on the top-L predicted links. Then, the top-L popularity is defined as the average popularity of links in the top-L places. A low average top-L popularity indicates that the algorithm tends to rank the missing links connecting low-degree nodes at the top places. Table 2 shows the top-L popularity of ten local similarity indices on the C.elegans network. For CN, PA, AA and RA indices the top-100 popularities are extremely high, lager than 100, and the scores will decrease with the increasing of L. This indicates that these four indices tend to rank the popular links at the top places. For the other six indices, namely Salton, Jaccard, Sørensen, HPI, HDI and LHN, the top-100 popularity scores are very low. Especially the score of the LHN index is very small, approximated to zero, and will increase with the increasing of L,

Table 3
Algorithmic accuracies on C.elegans for different probe sets, measured by the AUC value. Each value is obtained by averaging over 100 implementations with independently divisions of training set and probe set using new partitioning method. The average popularity of these ten probe sets are shown in the brackets. The average popularity of the whole set is 523. The highest AUC value in each row is emphasized in black.

Probe sets	CN	Sal	Jac	Sør	HPI	HDI	LHN	PA	AA	RA
E_1^P (57)	0.615	0.724	0.723	0.722	0.713	0.710	**0.755**	0.247	0.634	0.653
E_2^P (110)	0.735	0.775	0.780	0.779	0.758	0.778	**0.787**	0.435	0.756	0.772
E_3^P (158)	0.754	0.754	0.759	0.759	0.745	0.758	0.748	0.584	0.777	**0.791**
E_4^P (211)	0.823	0.799	0.806	0.806	0.782	0.805	0.768	0.708	0.842	**0.849**
E_5^P (283)	0.829	0.777	0.780	0.781	0.773	0.782	0.732	0.806	0.854	**0.864**
E_6^P (372)	0.867	0.798	0.800	0.800	0.793	0.800	0.726	0.881	**0.884**	**0.884**
E_7^P (493)	0.910	0.813	0.807	0.807	0.819	0.797	0.707	**0.929**	0.921	0.916
E_8^P (650)	0.929	0.818	0.801	0.801	0.846	0.769	0.684	**0.956**	0.934	0.923
E_9^P (939)	0.943	0.821	0.800	0.800	0.857	0.771	0.669	**0.980**	0.948	0.939
E_{10}^P (1987)	0.947	0.814	0.797	0.797	0.848	0.777	0.649	**0.995**	0.959	0.965

suggesting that the LHN index is likely to assign higher scores to the links among whose two endpoints there is at least one node with degree being equal to 1. When L is large, the overlap of two ranking lists generated by two algorithms is very high, and thus leads to similar top-L popularity scores. This result further demonstrates that the LHN index is more competent to uncover the unpopular links, especially the links with very low-degree endpoints.

3.3. Results for cold probe set

We employ the new partition method to prepare the probe set for further experiments. Here we set $D = 10$, and thus, we obtain ten different probe sets E_i^P ($i = 1, \ldots, 10$). Clearly, each probe set contains 5% of observed links, and $\langle \text{pop} \rangle_1 < \langle \text{pop} \rangle_2 < \cdots < \langle \text{pop} \rangle_{10}$. The algorithmic performances on C.elegans network for different probe sets are shown in Table 3. The results for other three networks are similar.

Compared with other nine indices, LHN has the best performance for predicting the very unpopular links (the links in E_1^P and E_2^P), while has the worst performance on the links in the probe sets with $P \geq 5$, especially the popular links in E_8^P, E_9^P and E_{10}^P. On the contrary, the PA index gives very good predictions on the popular links, while extremely bad predictions on the links with low-degree nodes where the accuracy is even much lower than pure chance. In the middle region where the average popularity is close to that of the randomly selected probe set, the RA index outperforms others, which is in accordance with the conclusion in previous studies [16,40,41].

4. Improved index

To design a method that is able to effectively predict both popular and unpopular links, we propose a parameter-dependent index, which is defined as:

$$s_{xy} = \frac{|\Gamma_x \cap \Gamma_y|}{(k_x \times k_y)^\lambda}, \tag{13}$$

where λ is a free parameter. This index is also neighborhood-based and requires only the information of the nearest neighbors, and thus no extra calculational complexity arises. Clearly, when $\lambda = 0$, this index degenerates to CN, and for the cases $\lambda = 0.5$ and 1, this index respectively degenerates to the Salton index and LHN index. Given a network, one can tune λ to find its optimal value subject to the highest accuracy.

We apply the new index to respectively predict the links in E_i^P ($i = 1, \ldots, 10$). The prediction of results of four example networks in the (λ, i) plane are shown in Fig. 2 where we focus on the unpopular links ($i = 1, \ldots, 5$). The results show that the optimal λ is positive when i is small (i.e., $i = 1, 2$), while it becomes negative for large i. For NetScience λ^* becomes negative for $i = 6$. The dependence of optimal parameter λ^* on i is shown in Fig. 3. Overall speaking, the optimal parameters λ^* of four networks are negatively correlated with i. In other word, the index with higher (positive) λ gives better predictions on unpopular links, while the index with lower (negative) λ is good at predicting popular links. For example in the C.elegans network, when $i = 1$, namely the probe set is constituted with the most unpopular links, the optimal $\lambda^* = 2.2$, indicating that to depress the scores of popular links is a better choice, while for $i = 10$, namely the probe set is constituted with the most popular links, the optimal $\lambda^* \approx -3$, which indicates that we had better assign higher scores to the popular links.

The algorithmic accuracies of ten local similarity indices as well as the proposed index for predicting unpopular links in E_1^P are shown in Table 4. Among the investigated ten local similarity indices, LHN outperforms others for predicting unpopular links. However, the proposed index can improve the accuracy with a proper λ which are all positive for all these four networks. Especially, the improvements on PB and C.elegans networks are significant, respectively 10.8% and 8.6% compared with LHN.

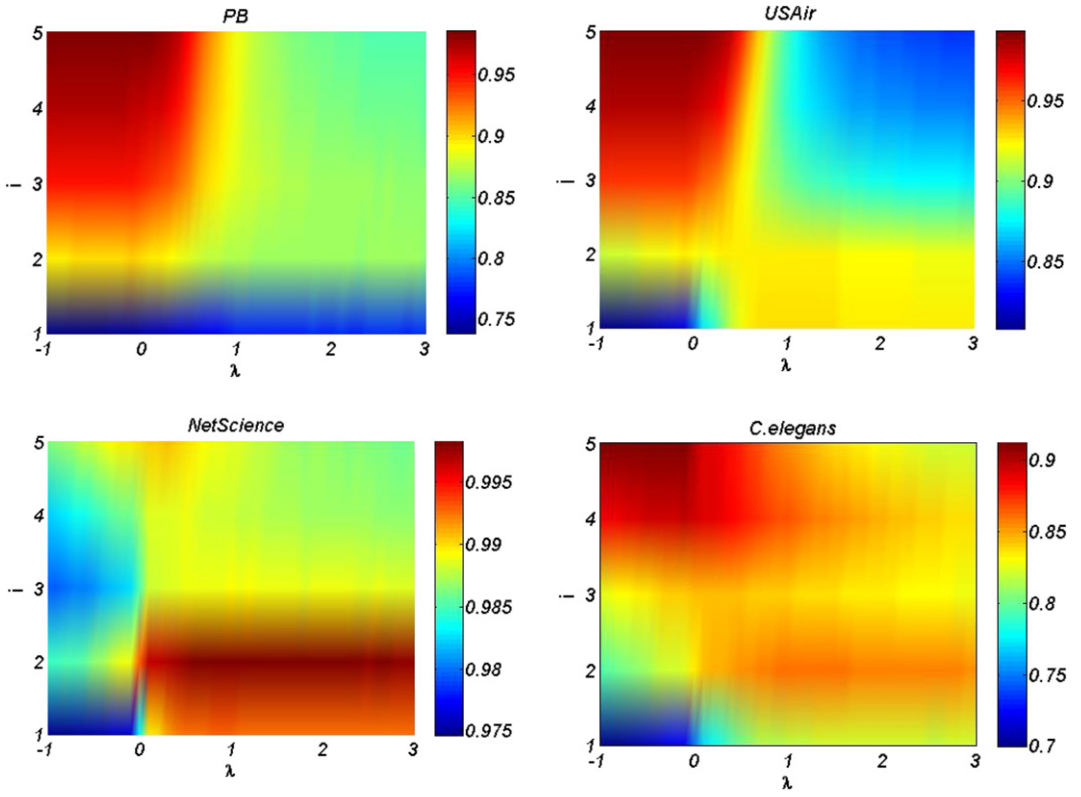

Fig. 2. The dependence of AUC on different λ and i. These results are obtained by averaging over 100 implementations with independently divisions of training set and probe set (by the new partitioning method introduced in Section 2.3). Here we set $D = 10$.

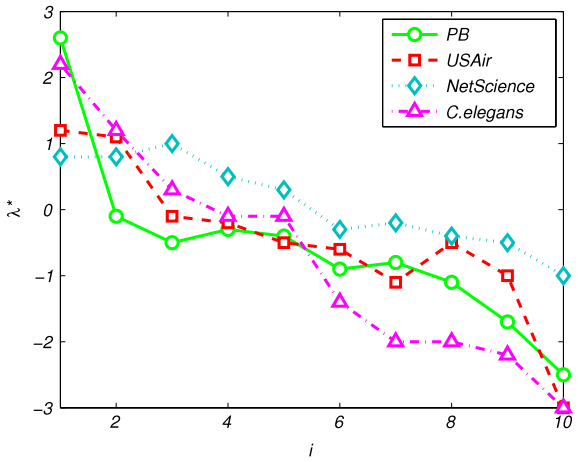

Fig. 3. The dependence of the optimal parameter λ^* on i. The probe set contains 5% of the known links. Each value is obtained by averaging the results over 100 independently implementations.

5. Experiments on real sampling methods

To connect our study to the real sampled networks, in this section, we will test the improved index on some real sampling methods. Firstly, we introduce four mainstream sampling methods as follows.

The first one is called *snowball sampling* (i.e., spider sampling or breadth first sampling, see Ref. [42]), which is a non-probability technique and gets widely used in the studies of World Wide Webs and large-scale social networks. In the beginning of this method, we randomly select one or a few nodes that consist of the initial sampled set, and then we crawl all the neighbors of the nodes in the sampled set, and put them into the sampled set. This process will carry on until a required number of nodes are sampled out. Obviously, it is not relevant for the link prediction problem because this method only leaves missing nodes rather than missing links.

Table 4
Comparison of algorithmic accuracy (AUC) on the cold probe set (i.e., $i = 1$). NS, CE and PB are the abbreviations for NetScience, C.elegans and Political Blogs, respectively. For each network, the probe set contains 5% of the known links. Each value is obtained by averaging over 100 independently implementations. The entries corresponding to the highest accuracies are emphasized in black. For the proposed index, the optional parameters for the highest AUC values are shown in the brackets.

Datasets	CN	Sal	Jac	Sør	HPI	HDI	LHN	PA	AA	RA	New
PB	0.649	0.674	0.664	0.664	0.690	0.662	0.701	0.584	0.656	0.664	**0.777** (**2.6**)
USAir	0.742	0.888	0.881	0.880	0.831	0.875	0.903	0.370	0.792	0.818	**0.928** (**1.2**)
NS	0.973	0.991	0.991	0.991	0.988	0.990	0.992	0.095	0.980	0.981	**0.993** (**0.8**)
CE	0.615	0.724	0.723	0.723	0.713	0.709	0.756	0.247	0.633	0.654	**0.821** (**2.2**)

Table 5
Average popularities of missing links corresponding to different sampling methods. 80% and 90% means the proportion $|E^T|/|E|$.

	PB		USAir		NS		CE	
	80%	90%	80%	90%	80%	90%	80%	90%
Acquaintance sampling	9674.7	11543.4	3115.2	3662.8	84.3	105.3	854.7	1030.9
Random-walk sampling	5361.1	5220.3	1658.2	1563.6	49.4	41.0	507.7	501.1
Path-based sampling	4505.1	4405.3	1158.4	958.1	18.7	18.0	344.4	296.6

The second one, called *acquaintance sampling*, is motivated by epidemic immunization with lack of information [43]. In this method, at each time step, a random link of a randomly selected node is sampled out (i.e., being put into the training set) until a required number of links have already been selected. Considering a link (x, y), if it is not yet sampled out, the probability it will be selected at this time step is $\frac{1}{N}(\frac{1}{k_x} + \frac{1}{k_y})$. Although a link with lower popularity is not necessarily with high $\frac{1}{k_x} + \frac{1}{k_y}$, statistically speaking, the probability $\frac{1}{N}(\frac{1}{k_x} + \frac{1}{k_y})$ is negatively correlated with the popularity $(k_x - 1)(k_y - 1)$. To our knowledge, this method is a very special method where unpopular links are more likely to be sampled out yet the popular links consist of the probe set.

The third one is named *random-walk sampling* [44]. A simple method adopted is as follows: (i) initialize a particle on a randomly selected node; (ii) this particle jumps to a randomly selected neighbor and the corresponding link will be added into the training set (i.e., sampled out); (iii) repeat (ii) until a certain number of links have been sampled out, and the rest links compose the probe set. It is well-known that the distribution of visiting frequency of a random walker on a connected network will soon converges to the degree distribution, namely the probability at a certain time step the random walker locate in a node x, say $\psi(x)$, is equal to $\frac{k_x}{2M}$, where $2M = \sum_y k_y$ serves as a normalization factor. Considering a link (x, y), if it is not yet sampled out, the probability it will be selected at this time step is $\psi(x)\frac{1}{k_x} + \psi(y)\frac{1}{k_y} = \frac{1}{M}$. That is to say, the average popularity of links in the probe set is approximately the same to that from the random sampling (we have checked it by simulation). However, the random-walk sampling is not the same as random sampling, for example, the sampled network from the former is always connected yet the one from the latter may contain several components.

The last one is called path-based sampling, which has been applied in extracting the topology of Internet at router level (http://www.routerviews.org). Indeed, this method tracks the transmission of information packets in the Internet, and a link passed by more packets has higher probability of being sampled out. To simulate this process, at each time step, we randomly select a starting point and an end, and assume that a packet will go from the starting point to the end through a randomly selected shortest path. After a sufficiently large number of time steps, a link with more than a threshold, N_T, packets will be put into the training set while others compose the probe set. Here for simplicity, we set $N_T = 1$. Under this method, the links with high betweenness centrality (betweenness centrality quantifies the traffic load of a link, depending on the routing strategy of packet transmission [45]) are favored. Since the popularity of a link is strongly positively correlated with its betweenness centrality, the average popularity of links in the probe set is lower than that of the random sampling.

The average popularities of missing links corresponding to different sampling methods are shown in Table 5. Agreeing with our analysis, the average popularities of the links in the probe set obeys the inequality $\langle \mathrm{pop} \rangle_{\text{acquaintance}} > \langle \mathrm{pop} \rangle_{\text{random-walk}} > \langle \mathrm{pop} \rangle_{\text{path-based}}$. Fig. 4 reports the algorithmic performance (measured by the AUC value) for different sampling methods and different λ. Very clearly, aiming to predict popular links (e.g., under acquaintance sampling) λ^* is negative while to predict cold links, λ^* is positive. In fact, there is strongly negative correlation between the average popularity of links in the probe set and the optimal value of lambda. As shown in Fig. 5, where we order the three sampling methods with decreasing average popularity of missing links and thus a positive correlation is observed.

6. Conclusion and discussion

To our knowledge, in the previous studies on link prediction [1], the data sets are always divided in a random manner. Inspired by the in-depth thought about the features of missing links, this article challenges such a straightforward method. Applying a simple measure on link popularity, we propose a method to sample unpopular links for the probe set.

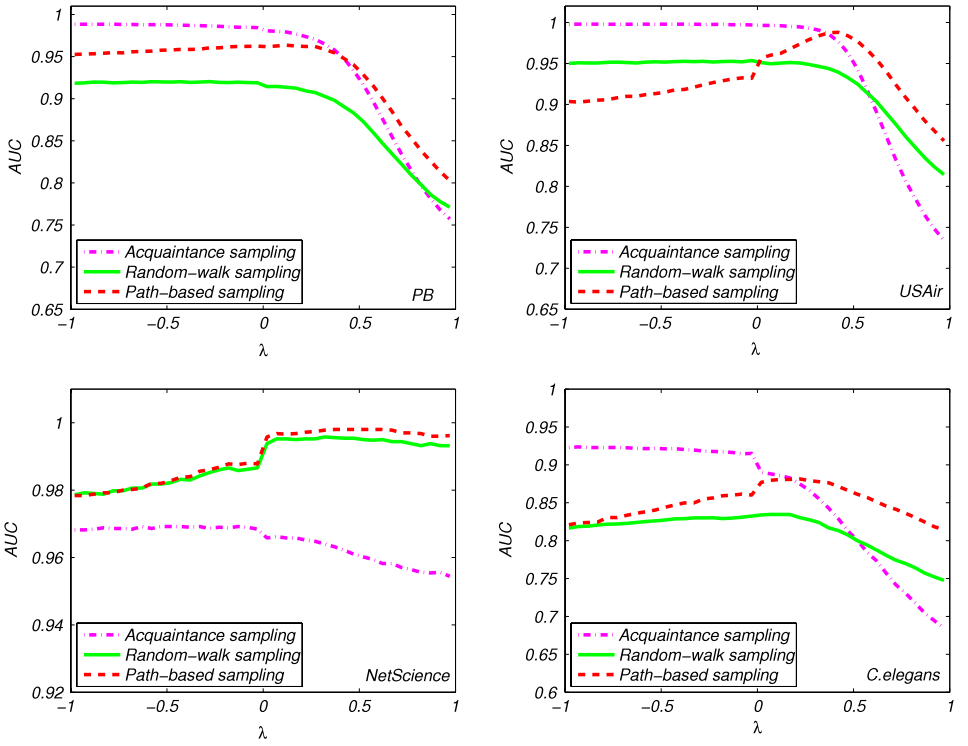

Fig. 4. Algorithmic performance of the improved index (see Eq. (13)) for the three real sampling methods with different values of λ. The training set contains 90% of observed links.

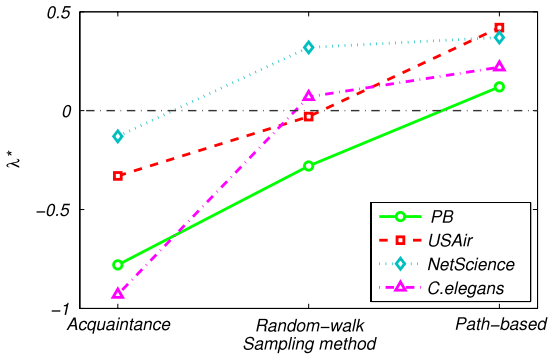

Fig. 5. The optimal values of λ for the three sampling methods. The dash line denotes $\lambda = 0$ as a visual guidance.

Experimental analysis shows a surprising result that the LHN index performs the best, although it was known to be one of the worst indices if the probe set is a random sampling of all links. We propose a similarity index with a free parameter λ, by tuning which this index can degenerate to the Common Neighbor index, the Salton index and the LHN index. The optimal value of λ monotonously depends on the average link popularity of the probe set. We further test this index on three real sampling methods. Agreeing with the main results from theoretical analysis, the optimal value of λ increases with the decreasing of the average popularity of links in the probe set. Again, the improved index in a well-tuned range can outperform others under real sampling methods.

Notice that the main contribution of this article does not lie in the proposed index. Instead, the significance of this work is to raise the serious question about how to properly determine the probe set. To us, this is a very important yet completely ignored problem in information filtering. The reconsideration of dataset division will largely change the understanding and thus the design of algorithms in information filtering (also relevant to the so-called recommendation problem [46]). As a starting point, we give a naive method and a preliminary analysis, which is of course far from a satisfactory answer to the question. In fact, we think an in-depth understanding of real sampling methods may shed light on this issue.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant Nos. 11075031 and 10635040, the Swiss National Science Foundation under Grant No. 200020-132253, and the Fundamental Research Funds for the Central Universities.

References

[1] L. Lü, T. Zhou, Link prediction in complex networks: a survey, Physica A 390 (2011) 1150.
[2] L. Getoor, C.P. Diehl, Link mining: a survey, ACM SIGKDD Explo. Newslett. 7 (2005) 3.
[3] H. Yu, et al., High-quality binary protein interaction map of the yeast interactome network, Science 322 (2008) 104.
[4] M.P.H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H.J. An, M. Lappe, C. Wiuf, Estimating the size of the human interactome, Proc. Natl. Acad. Sci. USA 105 (2008) 6959.
[5] L.A.N. Amaral, A truer measure of our ignorance, Proc. Natl. Acad. Sci. USA 105 (2008) 6795.
[6] G. Kossinets, Effects of missing data in social networks, Soc. Networks 28 (2006) 247.
[7] J.W. Neal, Kracking the missing data problem: applying Krackhardt's cognitive social structures to school-based social networks, Soc. Edu. 81 (2008) 140.
[8] C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Field, P. Bork, Comparative assessment of large-scale data sets of protein–protein interactions, Nature 417 (2002) 399.
[9] C.T. Butts, Network inference, error, and information (in)accuracy: a Bayesian approach, Soc. Networks 25 (2003) 103.
[10] R. Guimerà, M. Sales-Pardo, Missing and spurious interactions and the reconstruction of complex networks, Proc. Natl. Acad. Sci. USA 106 (2009) 22073.
[11] An Zeng, Giulio Cimini, Removing spurious interactions in complex networks, Phys. Rev. E 85 (2012) 036101.
[12] H.-K. Liu, L. Lü, T. Zhou, Empirical study of Chinese city airline network, Acta Phys. Sinica 56 (2007) 106.
[13] W.-Q. Wang, Q.-M. Zhang, T. Zhou, Evaluating network models: a likelihood analysis, Europhys. Lett. 98 (2012) 28004.
[14] Q.-M. Zhang, M.-S. Shang, L. Lü, Similarity-based classification in partially labeled networks, Internat. J. Modern Phys. C 21 (2010) 813.
[15] E.A. Leicht, P. Holme, M.E.J. Newman, Vertex similarity in networks, Phys. Rev. E 73 (2006) 026120.
[16] T. Zhou, L. Lü, Y.-C. Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (2009) 623.
[17] J.A. Hanely, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1980) 29.
[18] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. 58 (2007) 1019.
[19] L. Lü, C.-H. Jin, T. Zhou, Similarity index based on local paths for link prediction of complex networks, Phys. Rev. E 80 (2009) 046122.
[20] Z. Liu, Q.-M. Zhang, L. Lü, Tao Zhou, Link prediction in complex networks: a local naïve Bayes model, Europhys. Lett. 96 (2011) 48007.
[21] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Auckland, 1983.
[22] P. Jaccard, Étude comparative de la distribution florale dans une portion des alpes et des jura, Bull. Soc. Vaud. Sci. Nat. 37 (1901) 547.
[23] T. Sørensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, Biol. Skr. 5 (1948) 1.
[24] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barábasi, Community structure in social and biological networks, Science 297 (2002) 1553.
[25] M. Molloy, B. Reed, A critical point for random graphs with a given degree sequence, Random Structures Algorithms 6 (1995) 161.
[26] M.E.J. Newman, S.H. Strogatz, D.J. Watts, Random graphs with arbitrary degree distributions and their applications, Phys. Rev. E 64 (2001) 026118.
[27] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509.
[28] L.A. Adamic, E. Adar, Friends and neighbors on the web, Soc. Networks 25 (2003) 211.
[29] Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, B.-Q. Yin, Power-law strength-degree correlation from resource-allocation dynamics on weighted networks, Phys. Rev. E 75 (2007) 021102.
[30] V. Batageli, A. Mrvar, Pajek datasets. Available at: http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm.
[31] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104.
[32] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (1998) 440.
[33] R. Ackland, Mapping the us political blogosphere: are conservative bloggers more prominent, Presentation to BlogTalk Downunder, Sydney, 2005. Available at: http://incsub.org/blogtalk/images/robertackland.pdf.
[34] R. Albert, A.-L. Barábasi, Statistical mechanics of complex networks, Rev. Mod. Phys. 74 (2002) 1.
[35] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of networks, Adv. Phys. 51 (2002) 1079.
[36] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2003) 167.
[37] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Huang, Complex networks: structure and dynamics, Phys. Rep. 424 (2006) 175.
[38] L. da F. Costa, F.A. Rodrigues, G. Travieso, P.R.U. Boas, Characterization of complex networks: a survey of measurements, Adv. Phys. 56 (2007) 167.
[39] M.E.J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (2007) 208701.
[40] L. Lü, T. Zhou, Link prediction in weighted networks: the role of weak ties, Europhys. Lett. 89 (2010) 18001.
[41] W. Liu, L. Lü, Link prediction based on local random walk, Europhys. Lett. 89 (2010) 58007.
[42] P. Biernacki, D. Waldorf, Snowball sampling: problems and techniques of chain referral sampling, Sociol. Methods Res. 10 (1981) 141.
[43] R. Cohen, S. Havlin, D. ben-Avraham, Efficient immunization strategies for computer networks and populations, Phys. Rev. Lett. 91 (2003) 247901.
[44] L. Lovász, Random walks on graphs: a survey, Combinatorics 2 (1993) 1.
[45] G. Yan, T. Zhou, B. Hu, Z.-Q. Fu, B.-H. Wang, Efficient routing on complex networks, Phys. Rev. E 73 (2006) 046108.
[46] T. Zhou, Z. Kusckis, J.-G. Liu, M. Medo, J.R. Wakeling, Y.-C. Zhang, Solving the apparent diversity-accuracy dilemma of recommender systems, Proc. Natl. Acad. Sci. USA 107 (2010) 4511.