HEM: An Improved Parametric Link Prediction Algorithm Based on Hybrid Network Evolution Mechanism

Dejing Ke[1] and Jiansu Pu[2](✉)

[1] University of Electronic Science and Technology of China, Sichuan 611731, China
[2] Big Data Visual Analysis Lab, University of Electronic Science and Technology of China, Sichuan 610000, China
jiansu.pu@uestc.edu.cn

Abstract. Link prediction plays an important role in the research of complex networks. Its task is to predict missing links or possible new links in the future via existing information in the network. In recent years, many powerful link prediction algorithms have emerged, which have good results in prediction accuracy and interpretability. However, the existing research still cannot clearly point out the relationship between the characteristics of the network and the mechanism of link generation, and the predictability of complex networks with different features remains to be further analyzed. In view of this, this article proposes the corresponding link prediction indices Reg, DFPA and LW on regular network, scale-free network and small-world network respectively, and studies their prediction properties on these three network models. At the same time, we propose a parametric hybrid index HEM and compare the prediction accuracy of HEM and many similarity-based indices on real-world networks. The experimental results show that HEM performs better than other indices. In addition, we study the factors that play a major role in the prediction of HEM and analyze their relationship with the characteristics of real-world networks. The results show that the predictive properties of factors are closely related to the features of networks.

Keywords: Link Prediction · Complex Networks · Network Evolution · Data Mining

1 Introduction

The network represents the relationship between entities in the form of connections, which is an effective and popular abstraction of the complex real world. Network science has been involved in biological, social, communication and economic fields and achieved fruitful achievement [1,2]. In network science, network evolution and link prediction are two most challenging and attractive directions.

Network evolution mechanism is one of the most important aspect of the research of complex networks. It aims to understand the root causes of changes in network structure and function. Currently there have been a lot of models to study network evolution mechanism. Such as ER, WS, BA and so on [3–6]. And

X. Yang et al. (Eds.): ADMA 2023, LNAI 14177, pp. 91–106, 2023.
https://doi.org/10.1007/978-3-031-46664-9_7

link prediction is an attracted and challenge task in complex network. Link prediction aims to predict missing links and new links in the network through existing structural information in the network. Link prediction can helps us to understand and infer the connection mechanism of complex networks. And Link prediction has been applied into all kinds of fields. There are a lot of efficient research of link prediction algorithms at present. No matter how a link prediction algorithm is expressed, it is essentially a guess of network evolution mechanism. A good link prediction algorithm can more accurately reveal the evolution behavior of a network [7].

The research of link prediction and complex networks is developing rapidly, but it also faces many challenges. Firstly, the existing similarity algorithms often perform well in the face of a few networks, but they are no longer effective when dealing with a wider range of real-world networks, including directed networks, weighted networks, heterogeneous edge networks and other complex situations [8–10]. Secondly, there is a strong correlation between the link prediction algorithm and the network structure characteristics and the link predictability of the network in theory [11,12]. However, how to describe and express the relationship between them is a challenging task. In addition, through link prediction, the evolution characteristics of the network can be reproduced to a certain extent, and the research on the evolution behavior of complex networks can be promoted, but the research on this aspect is still relatively lacking; on the other hand, link prediction needs to face large-scale real data at the application level, and our algorithm needs stronger adaptability and more efficient calculation [13].

Therefore, starting from these challenges, this paper attempts to study through the following aspects. Firstly, this paper studies the characteristics of regular networks, scale-free networks and small-world networks. According to these characteristics, we propose the corresponding link prediction indices Reg, DFPA and LW. Through these indices, we aim to verify: link prediction indices are often related to the characteristics of the network when predicting; a single index often cannot cope with many networks, and indices that fit a certain network characteristics will always be better for the network. After that, we propose a parametric hybrid index HEM. We hope that through this hybrid index, we can get a better generalization performance index that integrates the characteristics of different networks. This index has better adaptability and more accurate prediction effect on complex real-world networks.

In this article we first introduce some basic network evolution models, then introduce the evaluation metrics of link prediction and some representative similarity-based algorithms. Finally we introduce our proposed indices based on network evolution mechanism.

2 Related Work

At present, link prediction has been applied widely in recommendation systems [14,15], mining biological information [16,17], reconstructing network information [18,19], and evaluating network evolution models [20,21]. Current link prediction methods mainly include methods based on structural similarity, network embedding, matrix completion, ensemble learning and neural network methods, etc. [22–24].

Among all the link prediction algorithms, the similarity-based algorithms are favored in many fields because of its simplicity and good interpretability. The similarity-based algorithms compute the similarity of each pair of nodes. Then similarity is used for prediction. The similarity-based algorithms include local similarity-based and global similarity-based indices. The local ones often take "common neighbor" into mainly account, such as CN, Satlon, Jaccard, Sorensen, HPI, HDI, LHN1, etc. [25]. The global ones always take higher-order paths into consideration, like LP, Katz and LHN2 and LO [26–28]. And some indices predict links by randomly walking, like LRW and SRW [23,29]. And Some takes other global information [25]. The more information is considered, a better the performance there will be, but it also brings higher computational cost.

All the link prediction algorithms calculate the connection probability between nodes in the network and express the network connection mechanism to some extent. Through the study of network evolution mechanism, if we can deeply grasp the relationship between nodes in network evolution and deeply understand the basis of connections in the network, we are more likely to propose an excellent link prediction algorithm. Based on this idea, we proposes the link prediction algorithm via the evolution characteristics of the network.

So we firstly construct regular networks, scale-free networks and small-world networks and proposes our algorithms accordingly. We then perform link prediction on these networks to analyze the feature of indices.

Secondly, we propose an combined algorithm. The index sets two parameters for the prediction factors. We sample the parameters and perform predictions on some real-world networks. The results show that our index performs better than many classical similarity-based indices. We hope that through the combination of simple characteristic indices, we can conduct a more efficient and interpretable index.

Finally, we analyze the dominant factors of the hybrid index. Experiments show that the accuracy and the upper limit are determined by the main factors. In addition, we find that the main factors are always related to the characteristics of the network, which coincides with the prediction properties of individual index.

When performing predictions, we often pay attention to the best results, and parameter sampling should also be oriented to the upper limit of the index. Finding the main factor can help to optimize the sampling problem.

3 Network Model and Link Prediction

In this section, we will briefly introduce some network evolution models, link prediction evaluation metrics and similarity-based indices.

3.1 Network Evolution Model

The study of complex networks plays an increasingly important role in mathematics, statistical physics, computer science and other fields [30]. In order to study specific feature of networks, this article will focus on regular network, small world network and scale-free network. We choose them because they have the most common and basic characteristics of complex networks. And we hope to simulate the feature of complex network by their simple features.

(1) **Regular Network**. In the regular network each node has the same number of neighbors. Many crystal networks or protein networks in the field of chemistry can be regarded as regular networks.
(2) **Scale-Free Network**. Networks with power-law degree distribution are called scale-free networks [31]. The scale-free network always can be generate by preferential attachment, that is, new nodes tend to be connected to nodes with high degree.
(3) **Small-World Network**. The small-world network depicts the phenomenon of large clustering coefficient and small average short path length in the real world network. Social networks, protein networks, food chain networks, cultural networks and so on have been proved to have the characteristics of small-world networks. In small-world network the nodes tend to connect with their close neighbors.

3.2 Link Prediction Evaluation Metrics

Reference [23] proposed two methods to evaluate the accuracy of link prediction algorithms, namely AUC (area under the receiver operating characteristic curve) and Precision. The briefly review of them are below.

AUC. The AUC metric evaluates the accuracy of the algorithms by comparing the score of missing links and the nonexistent links. Suppose there are n independent comparisons in total. Among these comparisons, there are n1 times the missing link having a greater score and n2 times missing link and nonexistent link have the same score. Then the AUC value can be calculated as:

$$AUC = \frac{n_1 + 0.5n_2}{n} \tag{1}$$

When AUC is equal to 0.5, the prediction accuracy of the algorithm is equivalent to random prediction. The closer the AUC value is to 1, the better the prediction accuracy of the algorithm is.

Precision. The Precision metric sorts the scores of missing links and nonexistent links in descending order. We take the sorted top-L links as the predicted ones. Among these L links, N links belong to the test set. Then the Precision can be calculated as:

$$Precision = \frac{N}{L} \tag{2}$$

Compared with AUC, the Precision only focuses on whether the top L links are predicted accurately.

3.3 Link Prediction Similarity-Based Algorithms

The similarity-based algorithms for link prediction compute a similarity score S_{xy} for each pair of nodes x and y, which directly represent the link possibility between x and y. The algorithms can be classified into two categories: local similarity indices and global similarity indices. Here we choose some representative indices to introduce (These indices are similar to the indices proposed in this paper in terms of expression. So they are chosen to better analyze and explain the differences. We ignore some indices that are not comparable). The details are as follows.

3.4 Local Similarity Indices

(1) Common Neighbor (CN) [25]

$$S_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)| \tag{3}$$

$\Gamma(x)$ denotes the set of neighbors of the node x. In the CN index, the more common neighbors two nodes have, the more likely they are to connect.

(2) Salton Index [25]

$$S_{xy}^{Salton} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x \times k_y}} \tag{4}$$

k_x and k_y denote the degree of nodes x and y, respectively.

(3) Resource Allocation Index (RA) [25]

$$S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \tag{5}$$

The RA index defines the amount of resources x allocates to y.

(4) Cannistraci-Hebb index (CH) [32]

$$S_{xy}^{CH} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1 + k_z^i}{1 + k_z^e} \tag{6}$$

where k_z^i denotes the number of links of z with other common neighbors of x and y, and k_z^e denotes the number of links between z and nodes other than x and y or their common neighbors.

(5) Local Path Index (LP) [25]

$$S^{LP(n)} = A^2 + \epsilon A^3 + \epsilon^2 A^4 + \cdots + \epsilon^{n-2} A^n \tag{7}$$

where ϵ is a free parameter and n is the maximum order.

3.5 Global Similarity Indices

(1) Katz Index [26]

$$S_{xy}^{Katz} = (I - \beta A)^{-1} - I = \beta A_{xy} + \beta^2 A_{xy}^2 + \beta^3 A_{xy}^3 + \cdots \tag{8}$$

β is the free parameter. I is the identity matrix. The contribution of higher order path can be controlled by adjusting β. This index considers all path sets. It calculates all the paths and assigns less weight to long paths in an exponential decay.

(2) Linear Optimization index (LO) [28]

$$S^{LO} = \alpha A(\alpha A^T A + I)^{-1} A^T A = \alpha A^3 - \alpha^2 A^5 + \alpha^3 A^7 - \alpha^4 A^9 + \cdots \tag{9}$$

α is a free parameter. I is identity matrix and A is adjacency matrix. When α is small enough, LO degenerates to the index that calculates only the 3-hop paths A^3.

4 Link Prediction Based on Network Evolution Mechanism

According to the characteristics of regular networks, scale-free networks and small-world networks, this article proposes link prediction indices for these three networks, and proposes a hybrid indices for complex networks based on the three indices. Note that all the link prediction results in this article are obtained by using the 10-fold cross-validation method on test networks.

4.1 Index Based on Regular Networks

According to the characteristics of regular networks, this article proposes a link prediction index called Reg. Reg is expressed as follows:

$$S_{xy}^{Reg} = \frac{1}{\sqrt{k_x \times k_y}} \tag{10}$$

k_x and k_y represent the degree of nodes x and y, respectively. In the formula, the nodes with larger degree are less likely to be connected. Small nodes are more likely to generate connections. By suppressing the connection probability of large degree nodes and promoting the connection probability of small degree nodes, the degree balance is achieved to a certain extent.

In order to study the performance of Reg index, we compared the link prediction accuracies of Reg index, CN index and Salton index on random regular network (see results in Table 1).

We can see that the Reg index is significantly better than other indices. Due to the randomness of the regular network, the CN index has an AUC value of only 0.5, while the Satlon index shows random results even with the same computational factor (i.e., $\frac{1}{\sqrt{k_x \times k_y}}$) as Reg index. As the degree of each node increases, the prediction performance of Reg index will gradually decrease.

Table 1. Accuracies on regular networks

Network	Reg_3	Reg_8	Reg_13	Reg_18	Reg_23	Reg_28	Reg_33
Cn	0.500	0.497	0.493	0.493	0.494	0.492	0.489
Salton	0.500	0.498	0.493	0.496	0.500	0.503	0.506
Reg	**0.942**	**0.839**	**0.784**	**0.752**	**0.729**	**0.712**	**0.698**

Accuracies are measured by the AUC value. The number of nodes of the network are all 2000. The results are calculated on random regular network whose each node has 3, 8, 13, 18, 23, 28 and 33 neighbors, respectively. And these 7 regular networks are denoted as Reg_3, Reg_8, Reg_13, Reg_18, Reg_23, Reg_28 and Reg_33, respectively.

4.2 Index Based on Scale-Free Networks

In reference to the article [25], a link prediction index PA corresponding to the preferential attachment principle is proposed. The expression of PA is as follows.

$$S_{xy}^{PA} = k_x \times k_y \tag{11}$$

This article also proposes a link prediction algorithm called DFPA (Difference Preferential Attachment) for scale-free networks. The expression is as follows.

$$S_{xy}^{DFPA} = \frac{max(k_x, k_y)}{min(k_x, k_y)} \tag{12}$$

Compared with PA index, DFPA index pays more attention to the connection between nodes with large degree and nodes with small degree. Nodes with similar degree are more stable and less likely to connect with each other. Therefore, small degree nodes and large degree nodes develop faster according to DFPA index. Besides, the connection probability between nodes with large degree is smaller than PA.

We compare the link prediction accuracies of PA and DFPA on scale-free networks constructed by BA model. The results are shown in Fig. 1. Note that accuracies are measured by the AUC value. The number of nodes of the networks are all 2000. Based on the BA model, each time the new nodes generate 1, 2, 4, 8, 16, 32 and 64 links, respectively. Thus there are 7 kinds of scale-free networks.

According to the prediction results of PA and DFPA in these scale-free networks, DFPA performs better when the network is sparse. As the degree of each node increase, the performance of PA gradually becomes better, while that of DFPA shows a downward trend. However, DFPA has a higher upper limit than PA in prediction.

There is a definition of degree assortativity in article [33], when it is greater than 0, nodes with similar degrees tend to connect with each other. When it is less than 0, nodes with different degrees are more likely to connect with each other. DFPA considers the latter case. In theory, the DFPA index also predicts accurately on disassortative networks.

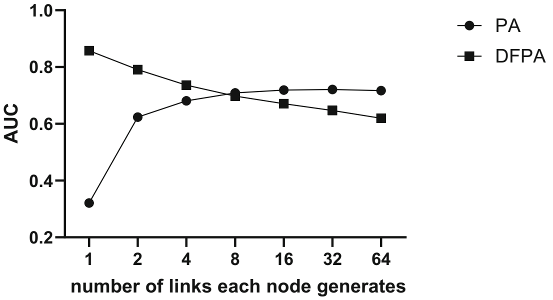

Fig. 1. Accuracies of PA and DFPA on scale-free networks

4.3 Index Based on Small-World Networks

In small-world network, each node is connected to the nearest k nodes. Based on that, this article proposes the LW (local world) index. The LW index considers that when two nodes have paths of length less than k or $k+1$, the two nodes are possible to have connection. The expression of LW index is as follows.

$$S^{LW} = A^k + A^{k+1} \tag{13}$$

k is the free parameter. A is the adjacency matrix of the network. A^k calculates the number of paths with length k between each pair of nodes. The paths calculated by A^k may go back and forth on some edges. So in order to consider both odd-order paths and even-order paths, LW calculates the sum of A^k and A^{k+1}.

k in LW represents the breadth and scope of information, which is similar to n in LP index. Compared with LP and Katz index, LW index does not consider that the lower order path has a higher weight. The weight of the path is related to the size of k and network structure. And the LW index has a small computational complexity.

To facilitate the comparison of LP and LW indices, we define the LPK index as:

$$S^{LPK} = A^2 + A^3 + \cdots + A^k + A^{k+1} \tag{14}$$

LPK is the case where the ϵ parameter of LP is set to 1 and the order n of LP is set to $k+1$.

For instance, we define LP2, LP4 and LP8 as the cases where the k value of LPK takes 2, 4 and 8 respectively. Similarly, define LW2, LW4, and LW8 as the cases where the k value of the LW index takes 2, 4 and 8, respectively.

We see that LPK and LW are basically equal. It is because $A^k + A^{k+1}$ are almost cover the information of A^i when i less than k.

4.4 Hybrid Index Based on Complex Network

Among the above three indices, Reg and DFPA are indices based on degree distribution, and LW is the index based on network topology. According to the three link prediction indices proposed by different network models, this article proposes a hybrid index called HEM (Hybrid Evolution Mechanism). The expression of HEM is as follows.

$$S_{xy}^{HEM} = {S_{xy}^{Reg}}^{\alpha} \times {S_{xy}^{DFPA}}^{1-\alpha} \times S_{xy}^{LW} \tag{15}$$

According to equation (10), (12) and (13), the above formula can be expanded as:

$$S_{xy}^{HEM} = \frac{1}{\sqrt{k_x \times k_y}}^{\alpha} \times \frac{max(k_x, k_y)}{min(k_x, k_y)}^{1-\alpha} \times (A^k + A^{k+1})_{xy} \tag{16}$$

There are two free parameters α and k in the HEM index. The α parameter is used to balance the degree distribution. The role of the k parameter is the same as in LW, representing the range of paths included.

By adjusting the α parameter, we can achieve the optimal balance of the HEM index in the link prediction on the mixed networks of regular networks and disassortative networks. When α is close to 1, the HEM index tends to predict on regular networks; when α is close to 0, the HEM index tends to predict on disassortative networks. The k parameter represents the path range considered in the prediction of LW index. If the k value is set too small, some high-order paths may not be taken into account for prediction. If it is too large, the paths that should not be considered will be involved. Therefore, the α and k parameters need to be adjusted simultaneously during the experiment.

In order to test the link prediction accuracy of the HEM index, this article selects the following network data sets (see in Table 2). The multiple edges are regarded as one single edge, and the directed edge is regarded as an undirected edge. The self-connections are not taken into account. In addition, we only consider the giant component when one network is not well connected.

Table 2. The features of 11 real-world networks

Network	N	M	K	Δ	D	C	ρ
PPI	2375	11693	9.85	118	15	0.306	0.454
NS	1461	2742	3.75	34	17	0.694	0.462
Grid	4941	6594	2.67	19	46	0.08	0.003
INT	5022	6258	2.49	106	15	0.012	−0.138
PB	1222	16714	27.36	351	8	0.32	−0.221
Yeast	2361	6646	5.63	64	11	0.13	−0.099
FBC	4039	88234	43.69	1045	8	0.606	0.064
HSS	1858	12534	13.49	272	14	0.141	−0.085
GrQc	5242	14484	5.53	81	17	0.53	0.659
AS	6474	12572	3.88	1458	9	0.252	−0.182
ER	1174	1417	2.41	10	62	0.017	0.127

Where N and M denote the number of nodes and edges of the network, respectively; K denotes the average degree; Δ denotes the maximum degree; D denotes the network diameter; C denotes the clustering coefficient; ρ denotes the degree assortativity. PPI is a protein-protein interaction network [34]. NS is a network of co-authorships in the area of network science [35]. Grid contains information about the power grid of the Western States of the United States of America [4]. INT represents the router-level topology of the Internet [36]. PB is a network of hyperlinks between political blogs about politics in the United States of America [37]. Yeast is a protein-protein interaction network in budding yeast [38]. FB consists of "friends lists" from Facebook, whose data was collected from survey participants using this Facebook app [39]. HSS represents the network of friendships between users of the website hamsterster.com [40]. GrQc is the collaboration network from the e-print arXiv and covers scientific collaborations between authors papers submitted to General Relativity and Quantum Cosmology category [40]. AS is the network of autonomous systems of the Internet connected with each other [40]. ER is the international E-road network, a road network located mostly in Europe [40].

There are many similarity indices in link prediction. This paper only selects some indexes that are similar to the indexes proposed in this paper in terms of expression. On the one hand, it is better to control variables and understand the factors that cause the difference in accuracy between indexes. On the other hand, some indices are quite different from the indicators in this paper in terms of predictive properties and computational performance, so that the predictive differences of the indicators cannot be accurately grasped, and the interpretability is also poor.

So this article compares the prediction accuracies of the HEM index and other similarity-based indices like CN, Salton, PA, RA, CH, LPK, Katz and LO on these networks. In these 11 networks, we calculate the AUC value and Precision value of these link prediction algorithms (see results in Table 3 and Table 4). Where The L value of Precision is 100. The parameter values in both Katz and LO indices are set to 0.01. The values of k parameter in LPK are selected as 2, 4 and 8, respectively. In the HEM index, we simultaneously sampled the α parameter and the k parameter. The values of α are selected as 0.0, 0, 25, 0.5, 0.75 and 1.0, respectively; and the values of k are selected as 2, 4 and 8, respectively. Among the 15 results obtained by combining the two parameters, we take the best result of the HEM index and record the α and k parameters when the AUC value is maximized.

Table 3. Algorithms' accuracy quantified by AUC

Network	PPI	Grid	INT	PB	Yeast	FB	HSS	GrQc	AS	NS	ER
Cn	0.893	0.589	0.559	0.919	0.706	0.992	0.805	0.922	0.696	0.943	0.526
Salton	0.892	0.588	0.559	0.875	0.705	0.992	0.789	0.922	0.676	0.944	0.526
PA	0.823	0.442	0.472	0.902	0.788	0.831	0.866	0.740	0.738	0.631	0.338
RA	0.894	0.589	0.559	0.923	0.706	**0.995**	0.809	0.923	0.700	0.944	0.526
CH	0.866	0.698	0.569	0.856	0.522	0.992	0.589	**0.938**	0.606	**0.988**	**0.713**
LP2	**0.939**	0.638	**0.633**	**0.932**	**0.839**	0.984	0.936	0.930	0.762	0.946	0.555
LP4	0.906	0.708	0.572	0.915	0.818	0.962	0.878	0.921	0.660	0.943	0.627
LP8	0.825	**0.772**	0.378	0.897	0.770	0.911	0.830	0.846	0.623	0.934	0.692
Katz	0.920	0.660	0.378	0.925	0.821	0.611	0.915	0.914	0.690	0.945	0.629
LO	0.935	0.560	0.623	0.929	0.813	0.986	**0.952**	0.846	**0.787**	0.852	0.486
α	*0.50*	*0.75*	*0.50*	*0.75*	*0.00*	*1.00*	*0.75*	*1.00*	*0.00*	*1.00*	*1.00*
k	*2*	*8*	*2*	*2*	*2*	*2*	*2*	*4*	*2*	*4*	*8*
HEM	**0.958**	**0.902**	**0.922**	**0.936**	**0.869**	**0.989**	**0.953**	**0.961**	**0.944**	**0.987**	**0.858**

Table 4. Algorithms' accuracy quantified by Precision

Network	PPI	Grid	INT	PB	Yeast	FB	HSS	GrQc	AS	NS	ER
Cn	0.474	0.000	0.008	0.078	0.003	0.040	0.003	0.354	0.059	0.200	0.000
Salton	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.011	0.000	0.046	0.000
PA	0.409	0.000	0.014	0.082	0.009	0.033	0.089	0.222	0.131	0.005	0.000
RA	0.002	0.000	0.000	0.028	0.001	0.041	0.000	0.000	0.016	0.004	0.000
CH	0.267	0.005	0.000	0.010	0.008	0.006	0.000	0.140	0.026	0.229	0.000
LP2	0.548	0.037	0.280	0.412	0.144	0.661	0.297	0.629	**0.253**	0.252	0.000
LP4	0.531	**0.046**	0.243	0.391	0.117	0.689	0.186	0.641	0.227	**0.253**	0.000
LP8	0.523	0.035	0.218	0.349	0.099	**0.694**	0.161	**0.644**	0.213	0.251	**0.001**
Katz	0.533	0.001	0.009	0.261	0.003	0.612	0.015	0.522	0.099	0.201	0.000
LO	**0.603**	**0.046**	**0.379**	**0.414**	**0.198**	0.037	**0.964**	0.301	0.185	0.230	**0.001**
α	*0.50*	*1.00*	*1.00*	*0.75*	*0.75*	*0.00*	*1.00*	*0.50*	*0.00*	*0.75*	*0.00*
k	*2*	*4*	*2*	*2*	*2*	*2*	*2*	*8*	*4*	*4*	*4*
HEM	**0.978**	**0.051**	**0.159**	**0.524**	**0.178**	**0.993**	**0.731**	**0.759**	**0.081**	**0.273**	**0.002**

According to the results of AUC, HEM performs much better than other indices in Grid, INT, AS and ER networks. In PPI, PB, Yeast, HSS, GrQc and NS networks, the prediction accuracies of HEM is also higher than other indices. For FB network, HEM and many other indices perform very well, the prediction accuracies are basically reaching 100%.

According to the results of Precision, the performance of HEM index on PPI, FB, HSS networks is much better than other indices, especially on PPI and FB networks, the Precision values of the HEM index are almost 1. HEM also has a better improvement on PB and GrQC networks compared to the classic indices. In contrast, in the AUC results, the HEM index outperforms in Grid, INT, and AS networks, but underperform in Precision compared to other indices, which indicates that most of correct predictions from the HEM index for these networks come from the second half of the lists of links.

Also in the tables we see that the parameters of HEM index differ when taking the maximum AUC and Precision values. Therefore, we need to study the role of parameters in the HEM index and their relationship with network characteristics.

5 Analysis of HEM Index

In order to understand the influence of different parameters, study which factor, including Reg, DFPA and LW, plays a major role in the prediction. Here we propose two methods.

(1) Calculate the prediction accuracies of different factors separately, and choose two factors with the highest accuracy.
(2) Sample α and k, then choose the top 5 combinations of α and k parameters from where the HEM index has the highest prediction accuracy. Where α takes the average value, and k takes the mode. If α is equal to 0.5, we only consider the k. Or when α is close to 0, take the factor DFPA; when it close to 1, take Reg.

The first method discusses the performance of individual factors, and the second method calculates the parameters that have a greater impact on the prediction. In practical considerations, The second method is used as the main reference, and the results obtained by the first method can make us have a better understanding of the characteristics of the network.

Here we discuss the situation when the prediction accuracy measured by the AUC value. The results of two methods may be different when it measured by the Precision value, but it has the same way. In this article we consider 5 factors, they are Reg, DFBA, LW2, LW4 and LW8.

We compare the main factors of the 11 networks obtained by the two methods, results are shown in Table 5.

Table 5. The main factors of 11 networks obtained by the method 1 and method 2

Method	PPI	Grid	INT	PB	Yeast	FB	HSS	GrQc	AS	NS	ER
1	LW2	LW8	DFPA, LW2	LW2	LW2	LW2	LW2	LW2	DFPA, LW2	LW2	REG, LW8
2	LW2	LW8	DFPA, LW2	LW2	DFPA, NW2	LW2	LW2	REG, LW4	DFPA, LW2	REG, LW4	REG, LW8

It can be seen that the results obtained by the two methods are basically the same except for the three networks of Yeast, GrQc and NS. In Yeast, the main factors calculated by method 2 has DFPA. While in method 1, DFPA in yeast performs better than Reg. In GrQc and NS networks, the main factors obtained by method 2 has Reg, while according to method 1, Reg factor performs worse than DFPA factor. Therefore, the influencing factors cannot be simply determined by the individual prediction accuracy.

Observe the several networks with high clustering coefficient: NS, FB, PB and GrQc, they have LW2 as their main factors based on the first method. LW2 performs very well on these networks, especially on FB. The FB network is the dense network with high clustering coefficient, and the prediction accuracies of LW indices basically reaches 1. So we guess that the LW index may be related to the clustering coefficient of the network. Besides, we can also observe that the density of the network also has a certain influence on the prediction of LW. For example, although the NS network has the highest clustering coefficient, the average degree of the network is only 3.75, far sparser than the FB network, and the LW2 and LW4 indices perform less well than on the FB network. Moreover, the main factors in the NS network obtained by the second method are Reg and LW4, indicating that due to the sparsity, a wider k in LW and additional consideration of regularity are needed to have a better prediction performance on the NS network. In addition, although the clustering coefficient of HSS network is low, the network is denser, then the performance of the LW index on the network is as good as that on the PPI and PB networks, whose clustering coefficient are much larger.

Both Grid and ER networks are sparse, and the diameter of the two networks is very large compared to other networks. Therefore, LW index needs to consider wider paths to predict the links. The main factors obtained in method 1 and method 2 are both LW8. The degree assortativity of the INT and AS networks is observed to be negative, indicating that the networks have the tendency of differential connection. Thus in these two networks, DFPA as their main factor performs the best among all the factors.

Moreover, the maximum degree of network AS is 1485, indicating that the degree distribution is very unbalanced, and the preferential attachment is more obvious. So the prediction performance of DFPA factor alone on AS network is also better. The maximum degree of GrQc, ER and NS networks is relatively small, indicating that the degree distribution of the network is relatively balanced. So on these 3 networks, the corresponding results obtained in the second method, Reg are their main factors. Though the Yeast network also has a small maximum degree, the degree assortativity is negative, indicating that connections on the network are still difference preferential. Correspondingly in the second method, DFPA is the main factor on Yeast network.

In summary, the Reg factor often acts on networks with relatively balanced degree distribution, that is, when the maximum degree is relatively small, we can take the Reg index into account to predict links. The DFPA index is usually more effective on networks with negative degree assortativity. The prediction performance of LW index is determined by clustering coefficient, average degree and network diameter. When clustering coefficient is higher and the network is denser, the link prediction of LW index is always more accurate. The size of the k of LW index depends largely on the diameter and average distance of the network.

By arranging the above results, we compare the prediction results(measured by the AUC value) of individual factors and hybrid index by tabular statistic (see in Table 6).

Table 6. Results of individual factors and hybrid index

	PPI	Grid	INT	PB	Yeast	FB	HSS	GrQc	AS	NS	ER
best of Factors	0.939	0.772	0.849	0.932	0.839	0.984	0.936	0.930	0.929	0.946	0.693
best of HEM	**0.958**	**0.902**	**0.922**	**0.936**	**0.869**	**0.989**	**0.953**	**0.961**	**0.944**	**0.987**	**0.858**

So we can see that the main factor largely determines the upper limit of the prediction accuracy of the hybrid index.

In general, the hybrid index always has a better prediction performance than the single index. The prediction performance is mainly determined by the main factor, and other factors may have some influence to the prediction, which will help to improve the overall result.

If we can determine the factors that have a greater impact in the link prediction of different networks, then we can save the sampling on the parameters of the HEM index that have little impact and reduce the computational complexity. Depending on the upper limit of the main factors, we can also have some idea of the upper limit of the HEM index. Determining the main factors can also give us some insight into the characteristics of the network.

6 Conclusion and Future Work

The link prediction indices proposed in this article, are based on the idea of simulating evolution mechanism through simple rules.

Thus, we firstly proposes corresponding link prediction algorithms on regular networks, scale-free networks and small-world networks respectively and studies their prediction properties on these three network models. Then we propose a parametric hybrid index, which has higher prediction accuracy than many similarity-based indices on real-world complex networks. Finally we studies the main predictors in the hybrid index, and analyzes and summarizes their relationship with network features.

In the future work, we will further refine the link prediction algorithms according to the network evolution mechanism. Firstly, we need to consider more details of topology structure. After all, path information is not sufficient to define the existence of links. Secondly, we only considers the mixed degree distribution of the regular network and the disassortativitive network. Therefore, it is necessary to consider the degree distribution more exactly in future research.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62272088 and U19A2078).

References

1. Newman, M.: Networks. Oxford University Press, New York (2018)
2. Barabási, A.-L.: “Network Science.” Network Science (2016)
3. Erdos, P.L., Rényi, A.: On the evolution of random graphs. Trans. Am. Math. Soc. **286**, 257–257 (1984)
4. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. Nature **393**(6684), 440–442 (1998)
5. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
6. Clauset, A., Moore, C., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. Nature **453**(7191), 98–101 (2008)
7. Wang, L., Shang, C.: Research on link prediction problem in scale-free network. Comput. Eng. **38**(3), 67–70 (2012)
8. Lü, L., Zhou, T.: Link prediction in weighted networks: the role of weak ties. EPL (Europhys. Lett.) **89**(1), 18001 (2010)
9. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World Wide Web (2010)
10. Murata, T., Moriyasu, S.: Link prediction of social networks based on weighted proximity measures. In: IEEE/WIC/ACM International Conference on Web Intelligence (WI 2007). IEEE (2007)
11. Lü, L., et al.: Toward link predictability of complex networks. Proc. Natl. Acad. Sci. **112**(8), 2325–2330 (2015)
12. Tan, S.Y., et al.: Link predictability of complex network from spectrum perspective. Acta Physica Sinica Chinese Edition **69**(8), 088901 (2020)
13. Lin-Yuan, L.: Link prediction on complex networks. J. Univ. Electron. Sci. Technol. China (2010)
14. Lü, L., et al.: Recommender systems. Phys. Rep. **519**(1), 1–49 (2012)
15. Bagci, H., Karagoz, P.: Context-aware friend recommendation for location based social networks using random walk. In: Proceedings of the 25th International Conference Companion on World Wide Web (2016)
16. Fakhraei, S., et al.: Network-based drug-target interaction prediction with probabilistic soft logic. IEEE/ACM Trans. Comput. Biol. Bioinf. **11**(5), 775–787 (2014)
17. Sridhar, D., Fakhraei, S., Getoor, L.: A probabilistic approach for collective similarity-based drug-drug interaction prediction. Bioinformatics **32**(20), 3175–3182 (2016)
18. Squartini, T., et al.: Reconstruction methods for networks: the case of economic and financial systems. Phys. Rep. **757**, 1–47 (2018)
19. Peixoto, T.P.: Reconstructing networks with unknown and heterogeneous errors. Phys. Rev. X **8**(4), 041011 (2018)
20. Wang, W.-Q., Zhang, Q.-M., Zhou, T.: Evaluating network models: a likelihood analysis. EPL (Europhys. Lett.) **98**(2), 28004 (2012)
21. Zhang, Q.-M., et al.: Measuring multiple evolution mechanisms of complex networks. Sci. Rep. **5**(1), 1–11 (2015)
22. Zhou, T.: Progresses and challenges in link prediction. Iscience **24**(11), 103217 (2021)
23. Lü, L., Zhou, T.: Link prediction in complex networks: a survey. Phys. A **390**(6), 1150–1170 (2011)

24. Mutlu, E.C., Oghaz, T.A.: Review on graph feature learning and feature extraction techniques for link prediction. arXiv preprint arXiv:1901.03425 (2019)
25. Zhou, T., Lü, L., Zhang, Y.-C.: Predicting missing links via local information. Eur. Phys. J. B **71**(4), 623–630 (2009)
26. Lü, L., Jin, C.-H., Zhou, T.: Similarity index based on local paths for link prediction of complex networks. Phys. Rev. E **80**(4), 046122 (2009)
27. Leicht, E.A., Petter, H., Newman, M.E.J.: Vertex similarity in networks. Phys. Rev. E **73**(2), 026120 (2006)
28. Pech, R., et al.: Link prediction via linear optimization. ArXiv abs/1804.00124 (2018): n. pag
29. Liu, W., Lü, L.: Link prediction based on local random walk. EPL (Europhys. Lett.) **89**(5), 58007 (2010)
30. Hou, L., et al.: Recent progress in controllability of complex network. Wuli Xuebao/Acta Physica Sinica **64**(18), 0188901 (2015)
31. Caldarelli, G.: Scale-Free Networks: Complex Webs in Nature and Technology. Oxford University Press, Oxford (2007)
32. Muscoloni, A., Abdelhamid, I., Cannistraci, C.V.: Local-community network automata modelling based on length-three-paths for prediction of complex network structures in protein interactomes, food webs and more. BioRxiv, 346916 (2018)
33. Newman, M.E.J.: Assortative mixing in networks. Phys. Rev. Lett. **89**(20), 208701 (2002)
34. von Mering, C., et al.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature **417**(6887), 399–403 (2002). https://doi.org/10.1038/nature750
35. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E Stat. Nonlinear Soft Mater. Phys. 74 3 Pt 2, 036104 (2006)
36. Mahajan, R., et al.: Inferring link weights using end-to-end measurements. In: International Memory Workshop (2002)
37. Adamic, L.A., Glance, N.S.: The political blogosphere and the 2004 U.S. election: divided they blog. In: LinkKDD 2005 (2005)
38. Jeong, H., et al.: Lethality and centrality in protein networks. Nature **411**(6833), 41–2 (2001). https://doi.org/10.1038/35075138
39. Mcauley, J.J., Leskovec, J.: Learning to discover social circles in ego networks. Neural Information Processing Systems Curran Associates Inc. (2012)
40. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: densification and shrinking diameters. ACM Trans. Knowl. Discov. Data **1**(1), 2 (2007)