Link Prediction Algorithm Based on Node Structure Similarity Measured by Relative Entropy

Meng Yuyu¹ and Guo Jing º

¹School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou, 730070, China
ºCorresponding author’s e-mail: mengyuyu@mail.lzjtu.cn

Abstract. To solve the problem that the link prediction method based on local information ignores the influence of neighbor structure information on the similarity measurement of nodes, a link prediction method based on relative entropy and local structure of nodes is proposed. Firstly, the second-order local network is introduced to describe the local structure of nodes; then, the structural similarity between nodes is described by redefining the relative entropy; finally, the structural similarity of nodes is measured based on relative entropy, and the structural similarity index of the node structure is proposed considering the structure information of the neighbor. Experimental results on 7 real network data sets show that the proposed method can achieve better results and can be applied to networks with a small average aggregation coefficient compared with other similarity indexes based on local and global information, and also have better performance on large-scale networks.

1. Introduction

In reality, many complex systems can be represented by complex networks, such as the Internet [1], World Wide Web [2], social network [3] and protein network [4]. Any of the networks can be shown by a graph composed of points and edges, which is defined as $G(V, E)$—$V$ is the set of vertices and $E$ is the set of connected edges. In the research and application of complex networks, link prediction refers to predicting the possibility of link between two nodes in a network that has not yet produced a link and inferring the missing edges in the network through the known network topology and network node attributes [5]. Some of extra applications include friend recommendation in social networks (such as Facebook) [6,7], prediction of protein interaction [8], news recommendation to users [9].

At present, the algorithm based on local information is to select neighbor nodes and make predictions according to the similarity between nodes. This algorithm can calculate similarity index very efficiently and can be well executed and suitable for large-scale network applications in many cases. For example, common neighbor index (CN) [10] focuses on whether two nodes are in the same environment; Jaccard similarity index [11] refers to introducing the node on the basis of the number of common neighbors between any two points to describe similarity; Adamic-Adar index (AA) [12] is that the contribution of common neighbor nodes with the small degree is greater than that of common neighbor nodes with the large degree; Resource allocation index (RA) [13] considers that nodes in the network that are not directly connected pass resources through common neighbors; The preferential attachment index (PA) [14] defines that the probability of a new link node is proportional to the product of two nodes degrees; Lu [15] found that all common neighbors had the same contribution to node pairs by studying several benchmark indicators and put forward the local naive Bayes model (LNB) on the basis of this assumption [16]. The algorithm based on global information is in accordance with the measure similarity
of the topological structure in the whole network, for instance, Katz index [17] takes into account the full path information of nodes. The prediction effect is improved, but the complexity is high, meanwhile, it does not perform well in large-scale networks. The method based on the quasi-local information considers more information than the local index, and discards the redundant information which has no contribution or little contribution to the accuracy of prediction. For example, the local path (LP) index [18] is not suitable for networks with the large average shortest distance. In the past 20 years, most alternative methods based on the maximum likelihood method have been proposed [19,20]. The maximum likelihood method computes the possibility of any unobserved link on the premise of some organization principles of network structure, such as random block model [21-25].

To solve these problems, this paper proposes link prediction method based on the relative entropy and the structure similarity of the local structure of the node. The algorithm is mainly based on local information and introduces the concept of local network, which measures the similarity between nodes by relative entropy under the joint action of first-order and second-order neighbor information. This paper will compare five algorithms based on local similarity with one algorithm based on global similarity on seven real network data sets.

2. Similarity index based on local network relative entropy

2.1. Local network

In the research of complex networks, local structure plays an important role. The local area network [27] is constructed by nodes and their neighbours, and each node itself is also included in the local network, as shown in Figure 1. Subfigure(a) shows the first-order local network of the node 1. subfigure(b) shows the second-order local network of node 1.

![Local network definition](image)

Fig.1 Local network definition

2.2. Similarity index based on second-order local network relative entropy

2.2.1. Redefinition of relative entropy. Relative entropy [26] can be used to measure the difference between two probability distributions. Assuming that p(x) and q(x) are two probability distributions of values in discrete random variable x, then the relative entropy of p to q is:

$$H(p, q) = \frac{1}{2}(d_{KL}(p||q) + d_{KL}(q||p))$$

where

$$d_{KL}(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

(1)

2.2.2. Node similarity measurement. The local networks of node x are represented by L_x (N, D). First, the structure information of the node will be obtained. The similarity of each pair of nodes is determined by calculating the local relative entropy [27] of each node. In the whole network, the maximum node degree of the network is defined as m=max(D). p (x, k) represents the set of probability sets of node x in descending order, and p (y, k) represents the set of probability sets of node y in descending order. Then p (x, k) is defined as:

$$p(x, k) = \begin{cases} \frac{D(k)}{\sum_{k=1}^{m} D(k)} & k \leq \text{Degree}(x) + 1 \\ 0 & k > \text{Degree}(x) + 1 \end{cases}$$

where D(k) is the degree set of node x.

The relative entropy of each pair of nodes is calculated on p (x, k) and p (y, k), and the similarity matrix
The local relative entropy of node $x$ and node $y$ is defined as:

$$D_{KL} (p(x)||p(y)) = \sum_{k} p(x,k) \log \frac{p(x,k)}{p(y,k)}$$

(3)

Next, local relative entropy is used to describe the similarity between nodes.

**Definition 1** (Similarity between nodes based on local relative entropy, RE) For a network $G (V,E)$, $x$ and $y$ are two unconnected nodes, whose similarity can define the difference between $x$ and $y$ for local relative entropy and unit volume, as shown in formula (4):

$$S_{xy} (RE)=1- \frac{\max(H)}{H(x,y)}$$

(4)

$H(x,y)=\frac{1}{2}(D_{KL}(x||y)+D_{KL}(y||x))$ denotes the local relative entropy between nodes $x$ and $y$. $\max(H)$ represents the maximum value of local relative entropy of all nodes in a network.

This paper focuses on more structural information of nodes and further considers the second-order local network through the expansion of the first-order local network. Based on the generation rule of node similarity in the first-order local network, formula (3) is used to calculate the local relative entropy of node pairs in the second-order local network, and the node similarity based on the second-order local network is obtained according to formula (4).

**Similarity index based on relative entropy of the local network.** This paper is according to the structure information of the node and the structure information of the node neighbours, the node similarity based on the local network will be redefined as:

$$S_{xy}^{(RE)}=S_{xy} (RE) + \alpha * S_{xy}^{(RE)}$$

(5)

$S_{xy} (RE)$ is the similarity matrix of the first-order local network. $\alpha$ adjustment parameter indicates how much the second-order neighbour node affects the node structure similarity and the most appropriate value can be selected according to the specific network. $S_{xy}^{(RE)}$ is the similarity matrix of the second-order local network.

### 2.3. Algorithmic steps

The steps of similarity algorithm based on the relative entropy of the local network are as follows:

- **Input:** Network $G (V,E)$ and all unknown link sets $V$ (the set element $V_{xy}$ represents a node pair composed of nodes $x$ and $y$).
- **Output:** The average prediction accuracy AUC of the algorithm.

1. Step 1. Divide edge set $E$ into $E_T$ and $E_P$, and $E=E_T+E_P$;
2. Step 2. For $i = 1$ to $V$, for $j = 1$ to $V$;
3. Step 3. Compute $D_{KL} (p(x)||p(y)) = \sum_{k} p(x,k) \log \frac{p(x,k)}{p(y,k)}$, $S_{xy} (RE)=1- \frac{H(x,y)}{\max(H)}$.
4. Step 4. Compute $S_{xy}^{(RE)}=S_{xy} (RE) + \alpha * S_{xy}^{(RE)}$.
5. Step 5. Compute $AUC=\frac{n^2+0.5*n*n}{n}$, $n=1000$;
6. Step 6. Cycle steps from 1 to 5 for 20 times to get the prediction accuracy AUC and average it and output it;
7. Step 7. Determine the best value of $\alpha$.

### 3. Experimental analysis

#### 3.1. Experimental data

The experiment uses the following seven open complex network datasets to test the algorithm: Jazz [28], Celegans [29], Router [28], Yeast [30], Net Science (NS) [31], Political blogs (PB) [29] and wiki-vote [32]. Table 1 is the summary of network topology characteristics.
Table 1. Topological characteristic parameters of the dataset

| Network        | Jazz | Celegans | Router | Yeast | NS | PB | wiki-vote |
|----------------|------|----------|--------|-------|----|----|-----------|
| |V| | 198 | 297 | 5022 | 2375 | 1536 | 1222 | 7066 |
| |E| | 2742 | 2148 | 6258 | 11693 | 2742 | 19021 | 103663 |
| <K> | 27.7 | 14.46 | 2.49 | 9.2 | 4.97 | 27.36 | 14.573 |
| C | 0.618 | 0.308 | 0.033 | 0.388 | 0.79 | 0.36 | 0.142 |
| D | 6 | 5 | 15 | 15 | 17 | 8 | 4.77 |

3.2. Evaluation indicator

The AUC evaluation indicator [25], which measures the accuracy of the algorithm as a whole, is based on the comparison between the similar value of the edges in the test set and the similar value of the non-existent edges.

\[
AUC = \frac{n' + 0.5n''}{n}
\]

\( n \) means independent comparison \((n=10000)\); \( n' \) means the similarity value of edges in the test set is greater than that of non-existent edges; \( n'' \) means the similarity value of edges is equal to nonexistent edges in the test set.

3.3. Experiment

3.3.1. Influence and determination of parameter \( \alpha \) on the algorithm. The similarity index based on relative entropy of local network proposed in this paper comprehensively considers the influence of first and second neighbour information on the similarity of nodes. For the similarity measurement method based on a semi-local path, it is believed that the longer path has limited space to improve prediction performance. Too much noise information in some cases will result in a decrease in prediction accuracy. Generally, only the second and third paths, at most the fourth path will be considered [25]. Therefore, this paper considers the third path of nodes and ignores the influence of neighbour information from the third path, because neighbour information from the second path has less influence on the similarity of nodes than neighbour information from the first path. The result of experiment show that \( \alpha \) has a high prediction performance in the range of \([0, 1]\), so \( \alpha \in (0,1) \) is determined. The value of \( \alpha \) directly affects the prediction accuracy of RE. In order to determine the appropriate value of \( \alpha \) in the range of \((0,1)\) with 0.001 as the step size, different \( \alpha \) values will be selected to predict the RE index and the average value of AUC will also be calculated.

Figure 2 shows the prediction performance of RE indicators in each dataset on the basis of 90% training set and 10% test set. The higher the prediction accuracy of AUC, the higher the accuracy and the better the algorithm. The first-order local network information is considered when \( \alpha = 0 \). With the increase of \( \alpha \), the information of the second-order local network becomes more important in the index, but the prediction performance of different networks changes to different extents with the continuous increase of \( \alpha \). For Router networks and wiki-vote networks, the average aggregation coefficient is so small that the first-order local network basically covers all the information, which causes little influence on node similarity. The importance of the second-order local network is determined by \( \alpha \), so it is necessary to determine the best \( \alpha \) value according to different datasets. A large number of parameter tuning experiments are performed on the real datasets to determine the best \( \alpha \) value for different networks. The specific values are shown in Table 2.
3.3.2. The comparison of algorithm performance. $\alpha$ is the value in Table 2 and the corresponding AUC prediction accuracy is calculated to measure the performance of this algorithm. Figure 3 shows the average AUC prediction accuracy of each dataset under different indicators. The RE algorithm in this paper is superior to other comparison algorithms in Celegans, Yeast, PB and wiki-vote datasets and is 4.6%, 2.5%, 2.4% and 10.1% higher than LNBRA, LNBCN, Jaccard and LNBAA algorithms. The RE index in Jazz network has increased by 9.1% compared with the PA index and the router network has increased by 9.2% compared with the PA index. The NS network has increased by 12.9%, 13%, 13% and 12.9% respectively compared with the algorithm in LNBRA, LNBCN, Jaccard and LNBAA. The link prediction performance achieved by RE algorithm in each network shows that more comprehensive network structure information can be obtained when predicting by using relative entropy theory and considering the similarity index of second-order local network node information, which effectively improve prediction accuracy. The algorithm presented in this paper is not suitable for networks with high average aggregation coefficient by observing the network characteristics of the selected dataset, such as Jazz network and NS network, but has very good performance for networks with low connection density, such as wiki-vote network, Celegans network, Yeas network and PB network. For a particular network Router, the accuracy of the algorithm in this paper also has a good performance, which shows that the algorithm is suitable for networks with low connection density and small aggregation coefficient and has good performance on large-scale networks. The average AUC prediction accuracy (90% training set) for each dataset is shown in Table 3.
In order to comprehensively measure the performance of the algorithm, the link information will be changed with the proportion of the training set $E_T$ from 90% to 99% and 80%. As the proportion of the training set decreases, good connectivity will be lost between nodes and more link information is missing in the prediction, so the proportion from a smaller training set is no longer considered. The above experimental process is repeated to obtain the AUC values of average prediction accuracy for each prediction algorithm. The test results are shown in Figure 4. Experiments have found that the AUC of all prediction algorithms decreases when the training set is reduced from 99% to 80%. The decrease of the proportion of training set causes less network information obtained during prediction. No matter how the proportion of the training set changes, the algorithm in this paper has good prediction accuracy in Celegans network, Yeast network, PB network and wiki-vote network, and it will be less affected as the proportion of the test set decreases.

Tables 4 and 5 show the specific values of AUC difference of each algorithm in different scales of the training set. In PB and Router networks, the AUC difference of this algorithm in this paper is smaller than that of LNBRA, LNBCN, LNBA and Jaccard algorithms, and the performance of this algorithm is more stable. In wiki-vote networks, the performance of this algorithm is more stable than that of Katz0.01 algorithm. However, the algorithm has little change in AUC difference compared with most other algorithms in the Jazz, Celegans, NS, and Yeast networks. This shows that the prediction performance of the algorithm in this paper doesn’t have any significant reduction when the training set is decreased, which can better adapt to complex environments.

### Table 4. AUC difference between 99% and 90% training set of each algorithm (%)

| Network | Jazz | Celegans | NS | PB | Router | Yeast | wiki-vote |
|---------|------|----------|----|----|--------|-------|-----------|
| LNBRA   | 0.420| 0.120    | 0.300| 6.310| 4.850  | 1.400 | 0.410     |
| LNBCN   | 0.820| 0.430    | 0.300| 6.690| 4.860  | 1.390 | 0.400     |
| Jaccard | 0.600| 1.600    | 0.300| 2.120| 4.820  | 1.390 | 0.430     |
| LNBA    | 0.620| 1.130    | 0.300| 6.420| 4.860  | 1.390 | 0.400     |
| PA      | 1.670| 1.500    | 0.630| 0.260| 0.580  | 1.560 | 0.760     |
| Katz0.01| 1.410| 0.600    | 0.450| 6.640| 1.860  | 0.570 | 21.540    |
| RE      | 0.040| 0.320    | 0.040| 0.590| 1.220  | 0.500 | 0.880     |

### Table 5. AUC difference of each algorithm in 90% and 80% training set (%)

| Network | Jazz | Celegans | NS | PB | Router | Yeast | wiki-vote |
|---------|------|----------|----|----|--------|-------|-----------|
| LNBRA   | 0.110| 2.910    | 1.100| 0.840| 4.420  | 0.960 | 0.700     |
| LNBCN   | 0.190| 2.570    | 1.090| 0.840| 4.430  | 0.950 | 0.680     |
| Jaccard | 0.190| 2.710    | 1.100| 0.650| 4.400  | 0.960 | 0.660     |
| LNBA    | 0.150| 2.600    | 1.100| 0.810| 4.430  | 0.950 | 0.690     |
| PA      | 1.000| 0.610    | 0.880| 0.280| 0.360  | 0.120 | 1.000     |
| Katz0.01| 0.400| 1.240    | 0.160| 0.160| 2.080  | 1.080 | 23.720    |
| RE      | 0.020| 0.6600   | 0.420| 0.350| 0.630  | 0.280 | 1.010     |
4. Conclusion
In order to solve the problem that the structural information of nodes in the real network has an impact on the similarity measurement, a method for link prediction of node structure similarity based on relative entropy and node local structure is proposed from the perspective of information theory. Compared with other similar indicators, experimental tests on seven actual network datasets show that the proposed method not only can achieve better results according to the AUC measurement standard but also show more stable performance with a smaller proportion of the total sample in the training samples. The algorithm presented in this paper is more suitable for networks with low connection density and small aggregation coefficient by observing the network characteristics of the selected dataset. The main contribution of this paper is to introduce information theory into the link prediction problem and take into account the structural information of node neighbours. Other similar indicators will be introduced for hybrid link prediction and related research to improve accuracy.

Acknowledgments
This paper is one of the phased achievements of the general project of the National Natural Science Foundation of China, research and application of correlation algorithm between chemical pollution and chronic diseases of local food in Gansu Province (NSFC).
References
[1] Michalis Faloutsos, Petros Faloutsos, Christos Faloutsos. On power-law relationships of the Internet topology[C]// Conference on Applications. ACM, 1999.
[2] Albert, Réka, Jeong, et al. Diameter of the World Wide Web[J]. Nature, 1999, 401(6):130--131.
[3] Jiang Y, Jiang J C. Understanding Social Networks From a Multiagent Perspective[J]. Parallel and Distributed Systems, IEEE Transactions on, 2014, 25(10):2743-2759.
[4] Wang P, Yu X, Lu J. Identification and Evolution of Structurally Dominant Nodes in Protein-Protein Interaction Networks[J]. IEEE transactions on biomedical circuits and systems, 2014, 8(1):87-97.
[5] Getoor L, Diehl CP. Link mining: a survey[J]. Acm Sigkdd Explorations Newsletter, 2005, 7(2): 3-12.
[6] Z. Wang, J. Liao, Q. Cao, H. Qi, Z. Wang, Friend book: A semantic-based friend recommendation system for social networks, IEEE Trans. Mob.Comput. 14 (3) (2015) 538–551.
[7] ZHANG X. Complex network link analysis and social media prediction [D]. National University of Defense Science and technology. 2013
[8] I.A. Kovács, K.Luck, K. Spirohn,Y. Wang, C. Pollis, S.Schlabach, W.Bian, D.-K.Kim, N. Kishore, T.Hao,M.A.Calderwood,M.Vidal,A.-L.Barabási, Network-based prediction of protein interactions, Nature Commun. (2019).
[9] D.Wei, T.Zhou, G.Cimini, P.Wu, W.Liu, Y.-C.Zhang, Effective mechanism for social recommendation of news, Physica A 390 (11) (2011)2117–2126.
[10] LORRAIN F, WHITE H C. Structural equivalence find individuals in social networks[J]. Social Networks, 1977, 1(1): 67-98.
[11] Jaccard P. Etude de la distribution florale dansune portion des Alpes et du Jura[J]. Bulletin De La Societe Vaudoise Des Sciences Naturelles, 1901, 37(142): 547-579.
[12] Adamic L A, Adar E. Friends and neighbors on the Web[J]. Social Networks, 2003,25(3):211-230.DOI:10.1016/S0378-8733(03)00009-1 .
[13] Zhou T, Lü L, Zhang Y C. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630. DOI:10.1140/epjb/e2009-00335-8.
[14] Barabási A L, Albert R. Emergence of scaling in random networks[J]. science, 1999, 286(5439): 509-512.
[15] LÜ Linyuan, ZHOU Tao. Link prediction in weighted networks: the role of weak ties[J]. EPL (neurophysics letters), 2010, 89(1): 18001.
[16] LIU Zhen, ZHANG Qianming, LÜ Linyuan, et al. Link prediction in complex networks: a local naive Bayes model[J]. EPL (neurophysics letters), 2011, 96(4): 48007.
[17] Víctor Martínez, Fernando Berzal, Juan-Carlos Cubero. A Survey of Link Prediction in Complex Networks. 2016, 49(4):1-33.
[18] T. Zhou, L. Lü, Y.C.Zhang, Predicting missing links via local information, Eur. Phys. J. B 71 (2009) 623.
[19] W. Liu, L. Lü, Link prediction based on local random walk, Europhys. Lett. 89 (2010) 58007.
[20] J. Neville, Statistical models and analysis techniques for learning in relational data, Ph.D. Thesis, 2006.
[21] H.C. White, S.A. Boorman, R.L. Breiger, Social structure from multiple networks I: block models of roles and positions, Am. J. Social. 81 (1976) 730.
[22] P.W. Holland, K.B. Laskey, S. Leinhardt, Stochastic block models: first steps, Soc. Networks 5 (1983) 109.
[23] P. Dorelan, V. Batagelj, A. Ferligoj, Generalized Block modeling, Cambridge University Press, Cambridge, UK, 2005.
[24] E.M. Airoldi, D.M. Blei, S.E. Fienberg, X.P. Xing, Mixed-membership stochastic block models, J. Mach. Learn. Res. 9 (2008) 1981.
[25] L. Lü, Zhou T. Link prediction in complex networks: A survey[J]. Physical A: Stutal Mechanics and its Applications, 2010, 390(6):1150-1170.
[26] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Stat. (1951) 79–86.
[27] Qi Zhang, Meizhu Li, Yong Deng. Measure the structure similarity of nodes in complex networks based on relative entropy[J]. Physical A: Statistical Mechanics and its Applications, 2018, 491.
[28] Router and Jazz data http://linkprediction.org/index.php/link/resource/data.
[29] C.elegans data http://www-personal.umich.edu/~mejn/netdata.
[30] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, P. Bork, Comparative assessment large-scale data sets of protein–protein interactions, Nature 417 (2002) 399–403.
[31] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E74 (2006) 036104.
[32] wiki-vote data https://snap.stanford.edu/data/wiki-Vote.html.