Predicting top-\(L\) missing links with node and link clustering information in large-scale networks

Zhihao Wu, Youfang Lin, Huaiyu Wan and Waleed Jamil

Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, People’s Republic of China
E-mail: zhwu@bjtu.edu.cn

Received 11 January 2016
Accepted for publication 20 June 2016
Published 12 August 2016

Abstract. Networks are mathematical structures that are universally used to describe a large variety of complex systems, such as social, biological, and technological systems. The prediction of missing links in incomplete complex networks aims to estimate the likelihood of the existence of a link between a pair of nodes. Various topological features of networks have been applied to develop link prediction methods. However, the exploration of features of links is still limited. In this paper, we demonstrate the power of node and link clustering information in predicting top-\(L\) missing links. In the existing literature, link prediction algorithms have only been tested on small-scale and middle-scale networks. The network scale factor has not attracted the same level of attention. In our experiments, we test the proposed method on three groups of networks. For small-scale networks, since the structures are not very complex, advanced methods cannot perform significantly better than classical methods. For middle-scale networks, the proposed index, combining both node and link clustering information, starts to demonstrate its advantages. In many networks, combining both node and link clustering information can improve the link prediction accuracy a great deal. Large-scale networks with more than 100 000 links have rarely been tested previously. Our experiments on three large-scale networks show that local clustering information based methods...
Predicting top-\(L\) missing links with node and link clustering information in large-scale networks

outperform other methods, and link clustering information can further improve the accuracy of node clustering information based methods, in particular for networks with a broad distribution of the link clustering coefficient.

**Keywords:** data mining (experiments), network dynamics, random graphs, networks

---

**Contents**

1. **Introduction** 2
2. **Methods** 4
   2.1. Definitions ........................................ 4
   2.2. Similarity indices based on node and link clustering information ....... 6
   2.3. Evaluation metrics .................................. 8
3. **Experiments and analysis** 9
   3.1. Test on small-scale networks ............................ 9
   3.2. Test on middle-scale networks ........................... 10
   3.3. Tests on large-scale networks ............................ 15
   3.4. Time costs ....................................... 18
4. **Conclusions** 19
**Reference** 20

---

**1. Introduction**

In the ‘big data’ era, more and more of the data of our society are collected by various kinds of information systems. These rapidly growing data sets contain a large amount of potentially useful information, but valuable information is becoming harder to extract. Complex social, biological, and technological networks have become the focus of research aiming to express and mine useful latent information from large-scale complex systems. The results of these studies are closely related to our daily lives and social structures. The technique of network analysis has become a popular tool in various research and application fields [1–5].

Link prediction in complex networks is an open issue in data mining and knowledge discovery, and has been widely studied by researchers from disparate scientific communities in recent years [6, 7]. As a branch of link mining, link prediction mainly focuses on how to estimate the likelihood of the existence of links between nodes based on available network information, and has many valuable applications in different fields. For example, predicting whether two users of a social networking site know
Predicting top-L missing links with node and link clustering information in large-scale networks

Each other can be used to recommend new friends. In E-commerce systems, it can also help to understand a user’s purchase preferences. In the field of biology, as our knowledge of many biological networks is very limited, using predicted top-L latent links to guide laboratory experiments, rather than blindly checking all possible interactions, will greatly reduce the experimental costs [8].

To fulfill the task of link prediction, two types of information are widely utilized, including the entity or node’s property information and the network’s topological information. Some machine learning based methods have achieved very good prediction results by employing property information. However, in many cases, property information is difficult to access because of privacy preservation or data quality issues. Therefore, in this paper we only focus on methods based on solely topological information.

Among various link prediction methods, the simplest framework is the set of similarity based algorithms, where each node pair is assigned a score to estimate the similarity between two nodes. A wide range of approaches has been developed, including methods based on local and global analysis, maximum likelihood estimation methods, etc [7].

The simplest local method is the preferential attachment (PA) index [9], which is defined as the product of the degrees of two seed nodes. The common-neighbor (CN) [10] method counts the number of CNs and the Jaccard index (JC) [11] is a normalization of the CN method. To obtain better resolution, the Adamic–Adar (AA) [12] and resource allocation (RA) [13] methods are defined by punishing high degree CN nodes, and obtain better results than the CN method. However, degree information is not enough to describe the role of a CN. Cannistraci et al designed a series of Cannistrai–Alanis–Ravai (CAR)-based similarity indices to improve classical similarity indices with links between CNs [8]. However, CAR-based methods are not efficient enough for large-scale networks. Based on Bayesian theory, Liu et al presented a local naive Bayes model to differentiate the role of neighboring nodes [14]. The final form of the naive Bayes model based index shows great relevance to node clustering coefficient. Recently, Wu et al designed a new index, called the clustering coefficient for link prediction (CCLP) index [15], with a direct node clustering coefficient. Their experimental results showed that the simple CN based method with a node clustering coefficient has good effectiveness in predicting missing links.

The above methods are all local measures and, to pursue higher precision, some global and quasi-local methods have also been proposed, such as Katz [16], SimRank [17], hitting time [18], average commute time [19], local path (LP) [20], transferring similarity [21], matrix forest index [22], and so on. Obviously, considering more information and features in prediction methods will require greater outlay of time and memory.

Some results indicate that community/cluster structures can improve the performance of link prediction accuracy [8, 23]. Therefore, several methods directly combine the communities detected by various community detection algorithms with existing similarity indices [24, 25], but how to choose a proper community detection algorithm is still not very clear.

Maximum likelihood estimation methods apply some presumed rules and parameters with the maximum probability of the known structure to predict missing links in a network. Clauset et al proposed an algorithm based on the hierarchical network structure, which gives good predictions for networks with hierarchical structures.
Guimera et al solved this problem using a stochastic block model [28]. Lü et al proposed a structural perturbation method and the concept of structural consistency, which can be used to estimate the link predictability of a network [29].

Until now, efficiency has remained a great challenge for link prediction in large-scale network data sets. Local methods are still a good applicable choice for solving link prediction problems. In this paper, we continue to work in the node-neighborhood based framework. The local na"ive Bayes common neighbor (LNBCN) [14] and CCLP [15] indices present better descriptions of the role of a CN than classical methods, while they still employ universal estimations, without considering the local network environment specific to a predicted pair of nodes. To solve the above problem, we shift our focus to local links and explore the power of link clustering information in predicting top-$L$ missing links. The experimental results on nine middle-scale and three large-scale networks show that link clustering information based indices outperform the other compared methods, estimated based on both the precision and area under precision curve (AUP) estimators.

2. Methods

2.1. Definitions

In this work, we focus on unweighted and undirected simple networks $G(V, E)$, where $V$ is the set of nodes and $E$ is the set of links. For each pair of nodes $x, y \in V$, we assign a score to the pair of seed nodes. All nonexistent links are sorted in decreasing order according to their scores, and the links at the top are most likely to exist. The commonly used framework always sets the similarity to the score, so a higher score means a higher similarity and vice versa.

To compare the link prediction performance of the proposed method with those of other methods in large-scale networks, we choose six efficient local similarity indices, including three classical and three recently proposed indices. As classical similarity indices, we choose CN, AA, and RA, which have been proved to perform better than other classical indices. LP, LNBCN, and CCLP are the three advanced methods. The LP index is defined with the LP structures of a pair of nodes. LNBCN is an index inferred from the na"ive Bayes theory and its final form demonstrates a close relation with the node clustering coefficient. The CCLP index purely estimates the contribution of CN nodes with a node clustering coefficient.

The definitions of these similarity indices are as follows.

1. CN index. For a pair of nodes, the CN method counts the number of CNs. As is common sense, more CNs indicate a larger probability of the formation/existence of a link between two nodes. The definition of a CN is given in equation (1), in which $\Gamma(x)$ denotes the set of neighbors of node $x$ and $|A|$ is the number of items in set $A$,

$$s_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|.$$  (1)
Predicting top-L missing links with node and link clustering information in large-scale networks

(2) AA index. The AA index refines the simple counting of CNs by assigning more weight to less-connected nodes, and is defined as

\[ s^{AA}_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log(k_z)}, \]  

where \( k_z \) is the degree of node \( z \).

(3) RA index. RA is also an index based on CNs, and the motivation comes from RA dynamics on complex systems. For a pair of unconnected nodes \( x \) and \( y \), the node \( x \) can send some resource to \( y \), with their CNs playing the role of transmitters. In the simplest case, assume that each transmitter has a unit of resource, and it will equally distribute that resource to all its neighbors. The similarity can be defined as given by equation (3), which measures the amount of resource that node \( y \) received from \( x \). Compared to AA, which simply replaces \( k_z \) by \( \log k_z \), this small difference only makes the results significantly different when the degrees of the CNs are comparatively high,

\[ s^{RA}_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}. \]

(4) LP index. The LP index considers the information regarding the next nearest neighbors, which can remarkably enhance the prediction accuracy. The definition is

\[ s^{LP}_{xy} = (A^2)_{xy} + \varepsilon (A^3)_{xy}, \]

where \( \varepsilon \) is a parameter. \((A^2)_{xy}\) and \((A^3)_{xy}\) are the number of different paths with lengths 2 and 3 connecting \( x \) and \( y \), respectively. In our experiments, \( \varepsilon \) is fixed as 0.001.

(5) LNBCN index. The LNBCN index is inferred according to the Bayes theory. The final definition of the LNBCN index is given in equation (5). It consists of two parts: the first part is a function of the node clustering coefficient (7) and the second part is a function of the whole network density as shown in (6). The authors of LNBCN also combined this index with other classical indices to generate a series of LNB-based indices, but other similarity indices can also be mixed with classical indices. For convenience and clarity, we only make comparisons to the basic and natural local Naïve Bayes index, i.e. LNBNC,

\[ s^{LNBCN}_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} (\log R_z + \log s) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \left( \log \left( \frac{C_z}{1 - C_z} \right) + \log \left( \frac{1 - \rho}{\rho} \right) \right), \]

where \( \rho \) is defined in equation (6).

\[ \rho = \frac{m}{n(n-1)/2} \]
The node clustering coefficient $C_z$ is,

$$C_z = \frac{t_z}{k_z(k_z - 1)/2},$$

where $t_z$ is the number of triangles passing through node $z$, and $k_z$ is the degree of node $z$.

(6) CCLP index. The CCLP index estimates the contribution of a CN node directly from the clustering coefficient. The definition is

$$s_{xy}^{\text{CCLP}} = \sum_{z\in \Gamma(x)\cap \Gamma(y)} C_z.$$  

### 2.2. Similarity indices based on node and link clustering information

Clearly, the problem of link prediction has a close relationship with the network evolving mechanism [30, 31]. One of the most important mechanisms, called triadic closure, makes simulated networks demonstrate a high clustering coefficient [32–34]. Therefore, considering clustering/triangular structure information in link prediction methods is very natural.

In most cases, when the clustering coefficient is mentioned, we usually think about the node clustering coefficient. The LNBCN index starts from the local naïve Bayes model, and its final expression demonstrates that node clustering coefficient plays an important role in estimating the contribution of a CN, although LNBCN also contains another part, which is a function of the whole network density (see its definition in equation (5)). The CCLP index simply describes the contribution of a CN with its clustering coefficient in the framework of the CN method. Although these methods are effective, the node clustering coefficient is not local enough for each predicted node pair. It is a local measure from the perspective of the whole network, but it is the same for different predicted pairs of nodes. This factor may limit the performance of predictors, since a CN should play different roles in the local networks of different pairs of nodes.

In this paper, we shift our focus from nodes to links and demonstrate the power of link clustering information in predicting top-$L$ missing links. In fact, the concept of link-community has been proposed before by Ahn et al [35]. From the aspect of links, more local or versatile information can be achieved, such as discovering overlapping circles. That means a node can play different roles in different local network environments, and this may convey specific local clustering information that we want.

Figure 1 shows the node and link clustering based similarity indices. As shown in figure 1(a), the CCLP index employs the node clustering coefficient to describe the contribution of a CN, and links between neighbors of a CN are used to calculate the number of triangles passing through the CN. In our new method, we focus on the clustering coefficient of links between seed nodes and the CN. For example, in figure 1(b), for the CN node $z$, we consider links $(x, z)$ and $(y, z)$, respectively. Link $(x, z)$ it has a probability of clustering node $y$ in the same triangle with node $x$ and $z$ through the CN node $z$ and vice versa.
We note that Wang et al gave a definition of the edge clustering coefficient (ECC) \[36\] as shown in equation \((9)\), in which \(CN(x, z)\) is the number of triangles passing through the link \((x, z)\) or the number of CNs between \(x\) and \(z\), and \(\min(k_{x} - 1, k_{z} - 1)\) gives the constraint of the maximal number of triangles. Here \(k_{x}\) is the degree of \(x\),

\[
ECC_{x,z} = \frac{CN(x, z)}{\min(k_{x} - 1, k_{z} - 1)}. \tag{9}
\]

Since we only care about the clustering ability of a link through the CN end node \(z\), the denominator of equation \((9)\) can be modified to \(k_{z} - 1\) in our methods.

Nodes and links are two perspectives from which to observe a complex system. Previous literature has demonstrated that a hybrid community structure of both nodes and links is superior in revealing network characteristics \[37, 38\]. Therefore, we conceive that combining both of the two factors in one similarity index may demonstrate some interesting results.

The similarity index based on both node and link clustering information, called the NLC index, is given by

\[
s^{NLC}_{xy} = \sum_{x \in \Gamma(x) \cap \Gamma(y)} \left( \frac{CN(x, z)}{k_{x} - 1} \times C_{z} + \frac{CN(y, z)}{k_{z} - 1} \times C_{z} \right). \tag{10}\]

In this definition, the multiplication of link clustering coefficient and node clustering coefficient can be regarded as a local clustering coefficient of the CN node specific to one seed node. The part of the link clustering coefficient of \((x, z)\) is the proportion of the number of CNs of node \(x\) and \(z\) to the number of neighbors of node \(z\). This ratio can be explained as the degree of acceptance of node \(x\) to the clustering coefficient of node \(z\).
node \( z \). Therefore, the NLC index can also be regarded as the summation of the local clustering coefficient of node \( z \) to node \( x \) and node \( y \), respectively. If the CN node \( z \) has a high clustering coefficient, but if there is no CN between \( x \) and \( z \), the high clustering property of node \( z \) may have a low correlation with node \( x \), and cannot contribute too much in clustering \( x \) and \( y \) together.

### 2.3. Evaluation metrics

To evaluate the accuracy of our method, here we employ two estimators: precision and AUP. The preparation for the two chosen estimators is the same. First, a network will be divided into two parts. Let us use \( E \) to indicate the whole link set, which is randomly divided into \( E_t \) and \( E_p \). The training set \( E_t \) is used as the known topological information, while the probe set \( E_p \) is used as the test set. We can obtain \( E_t \cap E_p = \text{null} \). To make sure that the training set contains true structure information, we only set 10\% of links as the test links, because removing too many links from the network may destroy the structure of the original network. The final results of our experiments are the average values of 1000 independent runs for small- and middle-scale networks, and 300 independent runs for large-scale networks.

After the calculation of the similarity matrix, we employ a commonly used way to estimate the effectiveness of the proposed method [7, 29]. First, all non-observed links are ranked in descending order. Then the precision is defined as the ratio of relevant items selected to the number of items selected [39]. This means if we take the top \( L \) links as the predicted ones, among which \( L_r \) links are right, then the precision can be defined as equation (11). Higher precision indicates higher prediction accuracy.

In this paper, we choose \( L \) as 20 or 100 for networks with fewer or more than 1000 links, respectively. In some of the literature, \( L \) can be chosen as the size of \( E_p \) [29]. It seems that such a choice is more reasonable, but in fact, we usually have no opportunity of accessing the number of missing links directly for a brand new real network.

In a practical link prediction application, the predicted links always need to be checked further. Thus, a proper \( L \) always needs to meet the user’s validation capabilities. For example, in biological experiments, verifying an interaction between a pair of proteins always costs a lot of money and time. Therefore, even though more than 90\% of interaction information is missing in some biological systems, it is meaningless to predict too many latent interactions, simply due to limited resources. For a specific application, if it is really necessary to investigate a proper \( L \), in our opinion link predictability analysis techniques [29] can bring some hints for solving this problem to some extent. A lower link predictability indicates more randomness in forming links and more difficulties in predicting precisely. To hit missing links, a relatively large value of \( L \) may be better, and vice versa,

\[
\text{precision} = \frac{L_r}{L}. \tag{11}
\]

AUP is defined as the area under the precision curve, and was proposed in [8]. Here, the precision curve is achieved by using ten different values of \( L \), which are 10, 20, ..., 100 for networks with more than 1000 links and 2, 4, ..., 20 for networks with fewer
Predicting top-$L$ missing links with node and link clustering information in large-scale networks

Table 1. The basic topological features of the six small networks.

| Nets   | $N$ | $M$ | $\langle k \rangle$ | $\langle d \rangle$ | $\langle C \rangle$ | $\langle ECC \rangle$ |
|--------|-----|-----|--------------------|-------------------|---------------------|---------------------|
| Karate | 34  | 78  | 4.59               | 2.41              | 0.57                | 0.27                |
| Dolphins | 62 | 159 | 5.13               | 3.36              | 0.26                | 0.12                |
| Football | 115 | 613 | 10.66              | 2.51              | 0.41                | 0.17                |
| Netsci | 379 | 914 | 4.82               | 6.04              | 0.74                | 0.69                |
| Mouse | 18  | 37  | 4.11               | 1.97              | 0.22                | 0.11                |
| Grassland | 75 | 113 | 3.01               | 3.88              | 0.34                | 0.23                |

Note: $N$ and $M$ are the total number of nodes and links, respectively. $\langle k \rangle$ is the average degree of a network. $\langle d \rangle$ is the average shortest distance between node pairs. $\langle C \rangle$ is the average node clustering coefficient. $\langle ECC \rangle$ is the average ECC.

than 1000 links. Since AUP considers the results for different values of $L$, it should be more robust and reliable.

3. Experiments and analysis

To investigate the effectiveness of the proposed NLC index in predicting missing links, we test the compared methods on three groups of networks: small-scale, middle-scale, and large-scale. In most related literature, link prediction algorithms are seldom tested on large-scale networks with more than 100 000 links. The network scale factor has not attracted enough attention in the past. Our experimental results demonstrate that the performance of link prediction algorithms is also affected by the factor of network scale.

3.1. Test on small-scale networks

In this section, we collect six small network data sets which are usually used in network analysis. Karate [40] and Dolphins [41] are club and dolphin social networks. Netsci [42] is a co-author network in network science. Mouse [43] is a neural network. Football [44] is a competition network about a football game. Grassland [45] is a food-web network. The basic properties and statistics of these networks are given in table 1.

The link prediction results for the six small-scale networks are given in tables 2 and 3. The best results are shown in bold face. It is not difficult to see that every index can only perform best on one or two networks at most. Since these networks are very small and the structures are not very complex, advanced similarity indices cannot demonstrate their advantages compared to classical indices.

The main reason for this is that most nodes in small networks have very small degrees. When these small-degree nodes are CNs, it is difficult for the NLC and CCLP indices to generate better resolution. As shown in figure 2, all kinds of situations of nodes with degrees equalling 2 and 3 are listed. For a CN of a pair of candidate nodes with degree equal to 2, the clustering coefficient is zero. Thus, common-neighbors with degree equal to 2 cannot contribute any value to the formation of the candidate link in both the CCLP and NLC indices. On one hand, there is a large probability that the clustering coefficient of a small-degree node is zero, so the link clustering coefficient can do nothing further in the NLC index; on the other hand, the possible situations are very
Predicting top-$L$ missing links with node and link clustering information in large-scale networks

### Table 2. Link prediction accuracy for small-scale networks estimated by precision.

| Precision | NLC   | CCLP  | LP    | LNBCN | CN   | AA   | RA   |
|-----------|-------|-------|-------|-------|------|------|------|
| Karate    | 0.103(55)| 0.095(52)| 0.100(54) | 0.097(53) | 0.104(57) | 0.102(53) | **0.107(52)** |
| Dolphins  | 0.119(60)| 0.108(59) | 0.136(64) | 0.118(61) | **0.124(62)** | 0.114(61) | 0.098(59) |
| Football  | 0.253(36)| 0.257(36) | 0.285(30) | 0.277(34) | **0.291(30)** | 0.285(30) | 0.284(30) |
| Netsci    | 0.943(38)| 0.898(37) | 0.846(56) | 0.885(41) | 0.869(43) | **0.973(17)** | **0.973(16)** |
| Mouse     | **0.062(39)** | 0.052(37) | 0.048(35) | 0.060(39) | 0.048(36) | 0.046(36) | 0.046(37) |
| Grassland | 0.103(56)| 0.102(57) | 0.065(50) | **0.136(64)** | 0.068(51) | 0.119(61) | 0.117(60) |

*Note: The numbers in brackets denote standard deviations. For example, 0.103(55) means that the precision value is 0.103 and the standard deviation is $55 \times 10^{-3}$."

### Table 3. Link prediction accuracy for small-scale networks estimated by AUP.

| AUP     | NLC   | CCLP  | LP    | LNBCN | CN   | AA   | RA   |
|---------|-------|-------|-------|-------|------|------|------|
| Karate  | 0.139(90) | 0.131(89) | 0.149(87) | 0.133(86) | **0.158(89)** | 0.137(85) | 0.131(85) |
| Dolphins| 0.145(92) | 0.146(96) | **0.188(111)** | 0.160(98) | 0.186(110) | 0.158(98) | 0.138(94) |
| Football| **0.334(46)** | 0.331(47) | 0.263(44) | 0.323(44) | 0.273(41) | 0.275(41) | 0.272(41) |
| Netsci  | 0.992(7) | 0.985(9) | 0.960(11) | 0.981(16) | 0.977(12) | **0.995(5)** | 0.990(10) |
| Mouse   | **0.103(67)** | 0.082(62) | 0.068(57) | 0.072(60) | 0.052(53) | 0.054(56) | 0.054(56) |
| Grassland| 0.111(69) | 0.122(74) | 0.073(73) | 0.138(86) | 0.051(59) | 0.167(93) | **0.170(94)** |

*Note: The numbers in brackets denote standard deviations. For example, 0.139(90) means that the AUP value is 0.139 and the standard deviation is $90 \times 10^{-3}$."

Figure 2. The situations of small-degree nodes as CNs. Nodes in red indicate CNs and those in black indicate predicted pairs of nodes. (a) Degree = 2. (b) Degree = 3. (c) Degree = 3. (d) Degree = 3.

Limited for a small-degree CN, thus the link clustering coefficient cannot contribute too much resolution. In the following sections we will show the advantages of node and link clustering information in relatively large networks.

doi:10.1088/1742-5468/2016/08/083202
Predicting top-L missing links with node and link clustering information in large-scale networks

The basic topological features of the nine middle-scale networks.

| Nets   | $N$  | $M$  | $\langle k \rangle$ | $\langle d \rangle$ | $\langle C \rangle$ | $\langle ECC \rangle$ |
|--------|------|------|---------------------|-------------------|------------------|-------------------|
| Jazz   | 198  | 2742 | 27.697              | 2.24              | 0.62             | 0.36              |
| Hamster| 1788 | 12476| 13.96               | 3.45              | 0.14             | 0.03              |
| PB     | 1222 | 16714| 27.36               | 2.74              | 0.32             | 0.10              |
| C elegans | 297  | 2148 | 14.47               | 2.46              | 0.29             | 0.08              |
| Macaque| 94   | 1515 | 32.23               | 1.77              | 0.77             | 0.58              |
| Yeast  | 2375 | 11693| 9.85                | 5.10              | 0.31             | 0.27              |
| USAir  | 332  | 2126 | 12.81               | 2.74              | 0.63             | 0.37              |
| Grid   | 4941 | 6594 | 2.67                | 15.87             | 0.08             | 0.06              |
| INT    | 5022 | 6258 | 2.49                | 5.99              | 0.012            | 0.01              |

Note: $N$ and $M$ are the total number of nodes and links, respectively. $\langle k \rangle$ is the average degree of a network. $\langle d \rangle$ is the average shortest distance between node pairs. $\langle C \rangle$ is the average node clustering coefficient. $\langle ECC \rangle$ is the average ECC.

The link prediction accuracy for the middle-scale networks estimated by precision.

| Precision | NLC    | CCLP   | LP     | LNBCN  | CN    | AA    | RA    |
|-----------|--------|--------|--------|--------|-------|-------|-------|
| Jazz      | 0.861(38) | 0.857(38) | 0.806(36) | 0.853(38) | 0.820(36) | 0.836(33) | 0.819(34) |
| Hamster   | 0.127(29) | 0.040(18) | 0.015(12) | 0.021(13) | 0.016(11) | 0.011(10) | 0.005(7)  |
| PB        | 0.467(44) | 0.408(41) | 0.425(40) | 0.410(39) | 0.420(39) | 0.379(38) | 0.249(38) |
| C elegans | 0.147(32) | 0.136(31) | 0.138(30) | 0.136(30) | 0.129(30) | 0.136(30) | 0.129(30) |
| Macaque   | 0.651(35) | 0.607(34) | 0.601(34) | 0.597(35) | 0.576(33) | 0.577(34) | 0.553(33) |
| Yeast     | 0.694(69) | 0.683(67) | 0.684(61) | 0.685(64) | 0.680(60) | 0.700(52) | 0.497(42) |
| USAir     | 0.638(46) | 0.622(47) | 0.603(47) | 0.613(47) | 0.604(47) | 0.622(47) | 0.637(37) |
| Grid      | 0.164(33) | 0.172(33) | 0.136(32) | 0.164(33) | 0.124(28) | 0.100(27) | 0.081(24) |
| INT       | 0.247(45) | 0.172(38) | 0.129(33) | 0.120(28) | 0.104(24) | 0.107(25) | 0.084(24) |

Note: The numbers in brackets denote standard deviations. For example, 0.861(38) means that the precision value is 0.861 and the standard deviation is $38 \times 10^{-3}$.

3.2. Test on middle-scale networks

In this section, we choose nine middle-scale networks with more than 1000 links from various fields. Jazz [46], Hamster\(^1\), and PB [47] are three social networks. C elegans [48], Macaque [49], and Yeast [50] are three biological networks. USAir\(^2\), Grid [48], and INT [51] are three technological networks. The basic properties and statistics of these networks are given in table 4.

The predicting results of these methods under the estimation of precision are given in table 5. It is clear compared to the other methods NLC performs the best. It attains the seven best results in the nine tested networks. After combining link clustering information, the NLC index can perform better than the CCLP index.

For the Hamster and INT networks, the link clustering information shows its great value. For the Hamster network, NLC index can improve the precision results of the other methods, which do not utilize link clustering information, by several times. When compared separately, NLC can beat AA, CCLP, and LNBCN on eight networks, and

---

\(^1\) Hamsterster friendships network dataset—KONECT, July 2016 http://konect.uni-koblenz.de/networks/petster-friendships-hamster.

\(^2\) A Pajek data sets http://vlado.fmf.uni-lj.si/pub/networks/data/mix/usair97.net.

doi:10.1088/1742-5468/2016/08/083202

11
Predicting top-L missing links with node and link clustering information in large-scale networks

Table 6. Link prediction accuracy for the middle-scale networks estimated by AUP.

|      | NLC | CCLP | LP  | LNBCN | CN  | AA  | RA  |
|------|-----|------|-----|-------|-----|-----|-----|
| Jazz | 0.940(16) | 0.927(17) | 0.875(25) | 0.920(17) | 0.886(23) | 0.891(22) | 0.882(24) |
| Hamster | 0.104(27) | 0.030(17) | 0.012(11) | 0.015(12) | 0.011(11) | 0.006(7) | 0.002(2) |
| PB   | 0.563(53) | 0.479(54) | 0.473(52) | 0.466(53) | 0.463(52) | 0.397(49) | 0.263(48) |
| Celegans | 0.167(43) | 0.153(42) | 0.157(40) | 0.155(41) | 0.152(40) | 0.157(41) | 0.150(41) |
| Macaque | 0.726(36) | 0.680(34) | 0.676(35) | 0.650(48) | 0.650(37) | 0.647(35) | 0.605(36) |
| Yeast | 0.740(28) | 0.733(29) | 0.723(28) | 0.725(28) | 0.721(28) | 0.710(27) | 0.416(35) |
| USAir | 0.792(37) | 0.784(37) | 0.777(39) | 0.781(38) | 0.775(39) | 0.785(38) | 0.754(38) |
| Grid | 0.245(54) | 0.251(55) | 0.219(51) | 0.246(53) | 0.218(48) | 0.105(36) | 0.086(33) |
| INT  | 0.284(52) | 0.218(50) | 0.142(34) | 0.132(35) | 0.127(34) | 0.106(31) | 0.077(26) |

Note: The numbers in brackets denote standard deviations. For example, 0.940(16) means that the AUP value is 0.940 and the standard deviation is $16 \times 10^{-3}$.

other the three methods on all nine tested networks. It demonstrates the great value of combining node and link clustering information in predicting top-L missing links.

Then, we compare these similarity indices under the estimation of AUP in table 6. The results are very similar with those estimated from precision. The NLC index attains eight bests. Comparing separately, the NLC index beats the CCLP and LNBCN indices on eight networks and the other four indices on all tested networks. We plot the precision curves on each network with various values of L in figure 3. It should be emphasized that the NLC index always performs better than the other methods on the nine tested networks. It will be more difficult to predict accurately when L becomes larger. On most networks, the precision accuracy decreases with the increase of L, except on the Hamster and Yeast networks. In particular, on the Hamster network, the improvements made by the NLC index become much larger with the increase of L.

To further compare the performance of the NLC and CCLP indices in predicting the top-100 missing links, we plot figure 4 for the nine middle-scale networks. In each figure, a black point (NLC–CCLP) indicates the average CCLP-rank of the hitting links at some position in the NLC-rank. For example, a black point with $L_1 = 50$ and $L_2 = 80$ indicates that for hitting links with NLC-rank equal to 50 in 1000 independent runs, their corresponding average CCLP-rank is 80. This means links that can be hit by NLC at a rank equal to 50 can also be hit by CCLP with a lower rank in most cases, because 80 is still fewer than 100 and the hitting links are still in the candidate list when the precision is calculated with $L = 100$. If $L_2$ is larger than 100, these hitting links in this position of the NLC-rank cannot be hit by CCLP.

We can observe that for the Jazz and USAir networks, the differences between NLC–CCLP and CCLP–NLC are very small, while on the Hamster, PB, Yeast, and INT networks the differences are obvious. For the Hamster network, the hitting links of the CCLP index can, not only, be hit by NLC, but also have even higher average rank. But when $L_1$ is larger than 40, the hitting links of NLC are more difficult for CCLP to hit. It can be easily understood that hitting links with lower rank are more difficult to predict accurately. Thus, figure 4 can be regarded as a comparison of the predicting ability of the two methods. In short, NLC has a better or at least the same predicting ability on the tested networks, except for the Grid network.

The above experiments are all about top-100 missing link prediction, which means we only observe the results of the first 100 candidates. Next, we compare the NLC and
CCLP indices from another perspective, i.e. the required length of candidates ($L$) to hit $K$ links. Figure 5 is plotted to show the results on nine tested middle-scale networks. Since the observed range of $K$ is from 1 to 100, the scope of the corresponding $L$ is largely extended, rather than being only in $[1, 100]$. This extension allows us to find the difference between the NLC and CCLP methods more clearly. Only in the Jazz network do the two compared methods show almost no difference, while on all the other networks the differences can be noted clearly. On the Hamster and INT networks, huge improvements can be observed, and in this case the improvement of NLC to CCLP in the INT network is larger than that in the Hamster network. We also find that when $K$ is larger than 70, NLC performs much better than CCLP on the Yeast network. CCLP can only perform a bit better than NLC on the Grid network. Since the Grid network has a very long average distance (see table 4), there are a lot of nodes degrees of only 2 or 3, which makes link clustering information generate only very low level resolution.

Figure 3. Precision curves of compared methods for the nine middle-scale networks with various values of $L$. 

doi:10.1088/1742-5468/2016/08/083202
The experimental results of figure 5 tell us that even though the value of $L$ may be chosen differently in different applications, the effectiveness of NLC is very stable. We also note that the larger the value of $L$ is, the bigger are the improvements made by NLC. All of the above experimental results show that NLC performs significantly better than CCLP on the Hamster and INT networks, but the reason for this is still unknown. We try to find some different features between Hamster, INT, and the other networks. Finally, we note that the average ECCs of the two networks are very small (see table 4). Further, we plot the distribution of the ECC for the nine tested networks in figure 6. It is surprising that the ECC distributions are very different. Some of them are similar to a normal distribution, such as Jazz and USAir; and some
are similar to an exponential distribution, such as Celegans and PB; while the ECC distributions of the Hamster and INT networks are more like power-law distributions. It looks like the results have a close relation with the distribution of ECC, and a broader distribution of ECC may be the key reason why NLC performs very well on those networks.

3.3. Tests on large-scale networks

Finally, we test the link prediction algorithms on three large-scale networks. Enron [52] is an email network of Enron between 1999 and 2003. Facebook [53] is an online social

doi:10.1088/1742-5468/2016/08/083202
network. Epinions [54] is a trust network from the online social network Epinions. The basic properties and statistics are given in table 7.

The prediction results of precision and AUP are shown in tables 8 and 9. For large-scale networks, NLC index outperforms the other compared methods. It should be noted that the local clustering information based methods, i.e. CCLP and NLC, perform much better than the other methods, and the combining link clustering information NLC index performs better, even when the precision values achieved by CCLP are already very high. Figure 7 shows the precision curves of the tested methods with
Predicting top-$L$ missing links with node and link clustering information in large-scale networks

Table 8. Link prediction accuracy for the large-scale networks estimated by precision.

| Precision | NLC     | CCLP    | LP      | LNBCN   | CN      | AA     | RA     |
|-----------|---------|---------|---------|---------|---------|--------|--------|
| Enron     | 0.877(14) | 0.853(18) | 0.749(35) | 0.781(31) | 0.766(33) | 0.780(25) | 0.679(26) |
| Facebook  | 0.869(26)  | 0.837(24) | 0.573(34) | 0.648(35) | 0.573(36) | 0.582(33) | 0.659(26) |
| Epinion   | 0.568(36)  | 0.338(37) | 0.228(36) | 0.232(35) | 0.222(34) | 0.196(35) | 0.255(43) |

Note: The numbers in brackets denote standard deviations. For example, 0.877(14) means that the precision value is 0.877 and the standard deviation is $14 \times 10^{-3}$.

Table 9. Link prediction accuracy for the large-scale networks estimated by AUP.

| AUP       | NLC     | CCLP    | LP      | LNBCN   | CN      | AA     | RA     |
|-----------|---------|---------|---------|---------|---------|--------|--------|
| Enron     | 0.871(16) | 0.832(18) | 0.776(28) | 0.785(24) | 0.782(26) | 0.774(23) | 0.674(33) |
| Facebook  | 0.881(21)  | 0.841(23) | 0.561(43) | 0.647(39) | 0.554(42) | 0.559(44) | 0.594(34) |
| Epinion   | 0.565(50)  | 0.322(43) | 0.232(48) | 0.225(47) | 0.221(48) | 0.204(49) | 0.273(50) |

Note: The numbers in brackets denote standard deviations. For example, 0.871(16) means that the AUP value is 0.871 and the standard deviation is $16 \times 10^{-3}$.

Figure 7. Precision curves of the compared methods for the three large-scale networks with various values of $L$.

different $L$. It can be observed that NLC always performs best and very stably with the increase of $L$.

Further, we plot figures 8 and 9 to compare the predicting ability of NLC and CCLP for large-scale networks. For the Enron and Epinions networks, most of the red points are lower than the black points in figure 8, which means NLC can always set a higher rank for links hit by CCLP. Since these three networks are very large, we set the range of $K$ from 1 to 300 in figure 9. For the Epinions network, the differences between CCLP and NLC are very clear starting from a very small $K$, while for the Enron network, the difference is difficult to observe until $K$ is larger than 150. That is why we cannot observe a huge improvement in the results of the precision and AUP with $L$ equal to 100. For the Facebook network, since NLC and CCLP perform better than each other in different ranges of $L$, the final accuracy of the two compared methods does not differ too much.

doi:10.1088/1742-5468/2016/08/083202
Finally, we also plot the ECC distribution for the three large-scale networks in figure 10. The distributions of the Enron and Facebook networks are similar to exponential distributions and that of Epinions is more like a power-law distribution. This is consistent with the conclusion made for middle-scale networks. When the ECC distribution is broader, the results of NLC are much better than those of CCLP. In the distributions in figure 10, there are some strange abnormal protuberances, which may be caused by some specific sampling strategy used to collect the data sets from their original systems.

3.4. Time costs

Table 10 demonstrates the time costs of these methods for the middle-scale and large-scale networks. The results for middle-scale and large-scale networks are average of 1000 and 300 independent runs, respectively. The differences between different independent runs mainly comes from randomly removing 10% of the test links for each.
Predicting top-$L$ missing links with node and link clustering information in large-scale networks

In this paper, we focus on the power of node and link clustering information in predicting top-$L$ missing links on complex networks. We propose a novel similarity index, called NLC, which employs both node and link clustering information. In our experiments, we test link prediction algorithms on three groups of networks, including small-scale, middle-scale, and large-scale networks. The network scale factor has not been

| Time  | NLC  | CCLP | LP   | LNBCN | CN   | AA  | RA   |
|-------|------|------|------|-------|------|-----|------|
| Jazz  | 0.008(1) | 0.006(0) | 0.010(1) | 0.011(1) | 0.003(0) | 0.003(0) | 0.003(0) |
| Hamster | 0.132(16) | 0.104(15) | 0.238(32) | 0.232(29) | 0.041(5) | 0.059(8) | 0.056(9) |
| PB    | 0.161(16) | 0.116(10) | 0.245(29) | 0.232(26) | 0.047(4) | 0.054(4) | 0.054(4) |
| C elegans | 0.011(1) | 0.008(1) | 0.013(1) | 0.014(1) | 0.003(0) | 0.004(1) | 0.004(1) |
| Macaque | 0.003(1) | 0.002(0) | 0.003(0) | 0.003(0) | 0.001(0) | 0.001(0) | 0.001(0) |
| Yeast | 0.119(9) | 0.107(11) | 0.096(12) | 0.124(13) | 0.038(6) | 0.071(11) | 0.072(12) |
| USAir | 0.011(6) | 0.007(3) | 0.014(7) | 0.015(7) | 0.003(2) | 0.005(3) | 0.005(3) |
| Grid | 0.308(61) | 0.304(174) | 0.197(50) | 0.301(52) | 0.189(42) | 0.298(50) | 0.309(129) |
| INT | 0.185(24) | 0.161(32) | 0.160(19) | 0.255(23) | 0.145(35) | 0.238(31) | 0.250(46) |

Table 10. Time costs of the compared methods for the nine middle-scale and three large-scale networks (in seconds).

Note: The numbers in brackets denote standard deviations. For the middle-scale network, 0.008(1) means the standard deviation is $1 \times 10^{-3}$, and for the large-scale network, 32.3(73) means the standard deviation is $73 \times 10^{-1}$.

The results for small-scale networks are not given, because the time costs are too small to be counted precisely. Apparently, the simplest index CN runs fastest. NLC has competitive efficiency with the other advanced methods. It is only a bit slower than CCLP index, but is still very efficient. The Matlab source codes used in our experiments can be downloaded from: http://inis.bjtu.edu.cn/nlc.html

Figure 10. ECC distribution for the three large-scale networks.

4. Conclusions

In this paper, we focus on the power of node and link clustering information in predicting top-$L$ missing links on complex networks. We propose a novel similarity index, called NLC, which employs both node and link clustering information. In our experiments, we test link prediction algorithms on three groups of networks, including small-scale, middle-scale, and large-scale networks. The network scale factor has not been
considered in the previous literature related to link prediction, and large-scale networks with more than 100 000 links have rarely been tested.

Our experimental results demonstrate that the performances of link prediction algorithms are different for networks with different scales. For small-scale networks, since the structure of these networks are not very complex and most of the nodes have very small degrees, classical similarity indices can perform competitively with advanced methods. With an increase in network scale, an abundant and complex network structures emerge, and advanced methods demonstrate their advantages gradually. For middle-scale and large-scale networks, the NLC index always performs very well. In particular, we find that the performance of NLC is closely related to the average ECC and the distribution of ECC.

The main contribution of this work is that we find that link clustering information, which is not usually considered in other methods, is very valuable in predicting missing links. On some networks, such as Hamster, INT, and Epinions, node clustering information based methods have made great improvements on classical methods. When combining link clustering information, the accuracy can be improved even more, even when the accuracy achieved by CCLP is already very high. We consider that the good results mainly profit from the local clustering information, which is specific to the local network environment of the predicted pair of nodes. When the ECC distribution approaches a power-law distribution, link clustering information can support a much better resolution to discriminate the contribution of a CN in different local networks. The broader the ECC distribution is, the better is the performance of NLC.

In addition, another advantage of NLC index is the fact that it is parameter free, which makes it more easy to apply to different kinds of networks, as choosing an appropriate parameter for a specific network always requires further information, which may be difficult to access.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants No. 61403023) and State Education Commission – China Mobile Research Fund (Grants No. MCM20150513).

Reference

[1] Albert R and Barabási A L 2002 Rev. Mod. Phys. 74 47
[2] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D U 2006 Phys. Rep. 424 175–308
[3] Costa L D F, Rodrigues F A, Travieso G and Villas Boas P R 2007 Adv. Phys. 56 167–242
[4] Bianconi G, Pin P and Marsili M 2009 Proc. Natl Acad. Sci. 106 11433–8
[5] Shen H W and Barabási A L 2014 Proc. Natl Acad. Sci. 111 12325–30
[6] Liben-Nowell D and Kleinberg J 2007 J. Am. Soc. Inf. Sci. Technol. 58 1019–31
[7] Lü L and Zhou T 2011 Physica A 390 1150–70
[8] Cannistraci C V, Alanis-Lobato G and Ravasi T 2013 Sci. Rep. 3 1613
[9] Newman M E 2001 Phys. Rev. E 64 025102
[10] Lorrain F and White H C 1971 J. Math. Sociol. 1 49–80
[11] Jaccard P 1901 Bulletin del la Société Vaudoises Sciences Naturelles 37 241–72
[12] Adamic L A and Adar E 2003 Soc. Netw. 25 211–30
[13] Zhou T, Lü L and Zhang Y C 2009 Eur. Phys. J. B 71 623–30
[14] Liu Z, Zhang Q M, Lü L and Zhou T 2011 Europhys. Lett. 96 48007

doi:10.1088/1742-5468/2016/08/083202
Predicting top-L missing links with node and link clustering information in large-scale networks

[15] Wu Z, Lin Y, Wang J and Gregory S 2016 Physica A 452 1–8
[16] Katz L 1953 Psychometrika 18 39–43
[17] Jeh G and Widom J 2002 SimRank: a measure of structural-context similarity Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (New York: ACM) pp 538–43
[18] Göbel F and Jagers A 1974 Stoch. Process. Appl. 2 311–36
[19] Fouss F, Pirotte A, Renders J M and Saerens M 2007 IEEE Trans. Knowl. Data Eng. 19 355–69
[20] Lu L, Jin C H and Zhou T 2009 Phys. Rev. E 80 046122
[21] Sun D, Zhou T, Liu J G, Liu R R, Jia C X and Wang B H 2009 Phys. Rev. E 80 017101
[22] Chebotarev P Y and Shamis E 1997 Avtom. Telemekh. 9 125–37
[23] Liu Z, He J L, Kapoor K and Srivastava J 2013 PloS One 8 e72908
[24] Yan B and Gregory S 2012 Phys. Rev. E 85 056112
[25] Soundarajan S and Hopcroft J 2012 Using community information to improve the precision of link prediction methods Proc. 21st Int. Conf. Companion on World Wide Web (New York: ACM) pp 607–8
[26] Clauset A, Moore C and Newman M E 2008 Nature 453 98–101
[27] Redner S 2008 Nature 453 47–8
[28] Guimerà R and Sales-Pardo M 2009 Proc. Natl Acad. Sci. 106 22073–8
[29] Lu L, Pan L, Zhou T, Zhang Y C and Stanley H E 2015 Proc. Natl Acad. Sci. 112 2325–30
[30] Wang W Q, Zhang Q M and Zhou T 2012 Europhys. Lett. 98 28004
[31] Zhang Q M, Xu X K, Zhu Y X and Zhou T 2015 Sci. Rep. 5 10350
[32] Newman M E and Girvan M 2004 Phys. Rev. E 69 026113
[33] Wu Z H, Menichetti G, Rahmede C and Bianconi G 2015 Sci. Rep. 5 10073
[34] Bianconi G, Darst R K, Iacovacci J and Fortunato S 2014 Phys. Rev. E 90 042806
[35] Ahn Y Y, Bagrow J P and Lehmann S 2010 Nature 466 761–4
[36] Wang J, Li M, Wang H and Pan Y 2012 IEEE/ACM Trans. Comput. Biol. Bioinform. 9 1070–80
[37] He D, Jin D, Chen Z and Zhang W 2015 Sci. Rep. 5 8638
[38] Jin D, Gabrys B and Dang J 2015 Sci. Rep. 5 8600
[39] Herlocker J L, Konstan J A, Terveen L G and Riedl J T 2004 ACM Trans. Inf. Syst. 22 5–53
[40] Zachary W W 1977 J. Anthropol. Res. 33 452–73
[41] Lusseau D, Schneider K, Boisseau O J, Haase P, Slooten E and Dawson S M 2003 Behav. Ecol. Sociobiol. 54 396–405
[42] Newman M E 2006 Phys. Rev. E 74 036104
[43] Bock D D, Lee W C A, Kerlin A M, Andlerrman M L, Hood G, Wetzel A W, Yurgensom S, Soucy E R, Kim H S and Reid R C 2011 Nature 471 177–82
[44] Girvan M and Newman M E 2002 Proc. Natl Acad. Sci. 99 7821–6
[45] Dawah H A, Hawkins B A and Claridge M F 1995 J. Animal Ecol. 70–82
[46] Gehrke P R and Danon L 2003 Adv. Complex Syst. 6 565–73
[47] Adamic L A and Glance N 2005 The political blogosphere and the 2004 US election: divided they blog Proc. 3rd Int. Workshop on Link Discovery (New York: ACM) pp 36–43
[48] Watts D J and Strogatz S H 1998 Nature 393 440–2
[49] Kötter R 2004 Neuroinformatics 2 127–44
[50] Bu D et al 2003 Nucl. Acids Res. 31 2443–50
[51] Spring N, Mahajan R and Wetherall D 2002 Measuring ISP topologies with rocketfuel ACM SIGCOMM Computer Communication Review vol 4 (New York: ACM) pp 133–45
[52] Klimt B and Yang Y 2004 The Enron corpus: a new dataset for email classification research Machine Learning: ECML 2004: 15th European Conf. on Machine Learning (Pisa, Italy, 20–24 September, 2004) (Berlin: Springer) pp 217–26
[53] Viswanath B, Mislove A, Cha M and Gummadi K P 2009 On the evolution of user interaction in facebook Proc. 2nd ACM Workshop on Online Social Networks (New York: ACM) pp 37–42
[54] Richardson M, Agrawal R and Domingos P 2003 Trust management for the semantic web The Semantic Web-ISWC 2003: 2nd Int. Semantic Web Conf. (Sanibel Island, FL, USA, 20–23 October 2003) (Berlin: Springer) pp 531–68

doi:10.1088/1742-5468/2016/08/083202