Missing link prediction and spurious link detection based on attractive force and community

Hui Qu¹, Wei Chen¹ and Kuo Chi²
¹School of Economics and Management, Harbin Engineering University, Harbin, China
²School of Information and Communication Engineering, Hainan University, Haikou, China

Abstract
With the rapid development of Internet and information technology, networks have become an important media of information diffusion in the global. In view of the increasing scale of network data, how to ensure the completeness and accuracy of the obtainable links from networks has been an urgent problem that needs to be solved. Different from most traditional link prediction methods only focus on the missing links, a novel link prediction approach is proposed in this paper to handle both the missing links and the spurious links in networks. At first, we define the attractive force for any pair of nodes to denote the strength of the relation between them. Then, all the nodes can be divided into some communities according to their degrees and the attractive force on them. Next, we define the connection probability for each pair of unconnected nodes to measure the possibility if they are connected, the missing links can be predicted by calculating and comparing the connection probabilities of all the pairs of unconnected nodes. Moreover, we define the break probability for each pair of connected nodes to measure the possibility if they are broken, the spurious links can also be detected by calculating and comparing the break probabilities of all the pairs of connected nodes. To verify the validity of the proposed approach, we conduct experiments on some real-world networks. The results show the proposed approach can achieve higher prediction accuracy and more stable performance compared with some existing methods.

Keywords
Missing link prediction, spurious link detection, social networks, attractive force, community
Introduction

In recent years, the rapid development of communication technology and Internet have greatly accelerated the sharing and spread of information, they have also provided more convenience and options for individuals. Some individuals are connected to each other to form a network, and the network will become more complex with the increasing participant individuals. Extracting the relations between individuals from many of these networks for analysis has been a crucial research issue at present.\(^1\)\(^–\)\(^3\) In a network, a node can be used to represent a participant individual, and a link between two connected nodes is often used to indicate the existence of the connecting relation between them. In addition, a link can be regarded as a transmission path of information between nodes. Therefore, link prediction plays an important role in research of recommendation system and information spread.\(^4\)\(^–\)\(^7\)

Link prediction can be normally used to predict the possibility if a link exists between two unconnected nodes based on the observable network structure.\(^8\)\(^–\)\(^10\) The problem can be divided into two categories: one is to detect the missing links that should exist in networks,\(^11\)\(^–\)\(^14\) these links may be missed in the process of data acquisition; the other one is to predict the potential links that will become real links in the future.\(^15\) In recent decades, various link prediction methods based on similarity metrics and machine learning models have been proposed.\(^16\)\(^–\)\(^18\) Similarity-based methods can achieve good prediction accuracy with less time cost due to they predict the possibility that a link should exist or not in the network by analyzing the network structure, such as the connections and paths between nodes. Learning-based methods predict the links by model building and repeated calculation to achieve better prediction accuracy but they also take more time. However, most of link prediction methods ignore an important problem that may occur in data processing: some real links may be spurious links that should not exist in networks, which can have a negative impact on accuracy of networks and information dissemination.

In order to solve the above problems, based on previous work on community detection,\(^19\) a novel similarity-based link prediction approach via community detection is proposed in this paper, which can not only predict the missing links in the network, but also detect the spurious links. In the beginning, we define the attractive force for any pair of nodes to denote the strength of the relation between them. Then, a community detection algorithm is proposed to divide all the node into some communities according to their degrees and the attractive forces on them. Next, the link prediction approach based on the community structures is proposed, including two algorithms LPCA-P and LPCA-R. LPCA-P can be used to predict the missing links by calculating and comparing the connection probabilities of the potential links in a network, the higher connection probability of a potential link between two unconnected nodes means the two nodes are more possible to be connected. LPCA-R can be applied to detect the spurious links by calculating and comparing the break probabilities of the real links, the higher break probability of a real link between two connected nodes means the link is more possible to be a
spurious link. Finally, experiments are conducted on some real-world networks and compared with some existing link prediction methods to verify the validity of the proposed algorithms.

The rest of this paper is organized as follows: In Section 2, we describe the current research work on link prediction. In Section 3, we define the attractive force for any pair of nodes in a network and propose the community detection algorithm. Then, we make detailed descriptions and explanations of our link prediction approach in Section 4, including the algorithm LPCA-P for missing link prediction and the algorithm LPCA-R for spurious link detection. In Section 5, we conduct experiments on some real-world networks and compare the results with some existing link prediction methods. Finally, we summarize the paper and forecast the future work.

Related work

Most of the existing link prediction methods focus on the prediction of the potential links between pairs of unconnected nodes according to the known network structure since they have been proposed. These methods can be basically divided into two categories, one is based on similarity metrics and the other one is based on machine learning.

The methods based on similarity metrics measure the importance of a potential link between two unconnected nodes by extracting local information or global information in a network. Local information may be the attributes of nodes such as age, interest or hobby that have been uploaded by users, or a few local characteristics of the network. Some typical methods are: the common neighbors (CN) index,20 Adamic and Adar21 (AA) index, Jaccard22 index, Katz23 index, and SimRank algorithm.24 Moreover, inspired by these methods, Zhou et al.12 defined the amount of resource for the nodes and calculated the amount of resource transmission of potential links by considering the common neighbors of unconnected nodes, they presented the resource allocation (RA) index. Dai et al.25 constructed the belief vector for different types of links by calculating the belief of each node and measured the influence between different types of relations by comparing the similarity of belief vectors. Rafiee et al.26 proposed a similarity-based link prediction algorithm CNDP to measure the similarity score of pairs of nodes by considering the clustering coefficient of nodes. These methods can generally achieve acceptable prediction results even if they are implementable, however, they perform unsatisfactorily on sparse networks. The similarity-metrics-based methods via global information mainly focus on the network topology structure, such as the paths between pairs of unconnected nodes, that have broader fitness and better robustness than the methods only used the attributes of nodes. Liu et al.27 considered the distance between nodes in different communities of social networks and introduced the Hash method to calculate the similarity between nodes in different communities, it could make dynamic link prediction by collecting the change of local nodes information. Akiba et al.28 put forward an algorithm for K shortest
route between any two nodes in networks, and the algorithm has been successfully used in link prediction. Liu et al.\textsuperscript{29} presented the method of superposed random walk (SRW) by superimposing T-step and preorder results based on the concept of local random walk. Zhang et al.\textsuperscript{30} measured the inter-similarity by employing the local diffusion processes and proposed a bi-directional hybrid diffusion method for identifying missing and spurious links in bipartite networks. These methods can normally achieve higher prediction accuracy than the methods using local information, however, they need to take more time.

The methods based on machine learning can predict the possibility of potential links becoming real links via building machine learning models. Scellato et al.\textsuperscript{31} defined some new features based on the original network features and proposed a supervised learning framework to predict new links of friends-friends and places-friends. Zhou and Jia\textsuperscript{32} characterized the similarity of any pairs of unconnected nodes by considering the knowledge quantity of the nodes on the paths between them and proposed the knowledge-dissemination-based link prediction (KDLP) algorithm for link prediction based on knowledge dissemination mechanism. Pech et al.\textsuperscript{33} introduced robust principal component analysis (robust PCA) for link prediction on dense networks by completing the adjacent matrixes of networks. Williamson\textsuperscript{34} presented a Bayesian nonparametric approach for link prediction on sparse networks that combined structure explanation with predictive performance. Muniz et al.\textsuperscript{7} proposed weighted criteria by combining contextual, temporal and topological information to improve the link prediction accuracy. Singh et al.\textsuperscript{35} presented a community detection algorithm to divide the network into clusters and further proposed a missing link prediction approach based on information diffusion. These methods need to train the data and deploy parameters in advance, and the trained models have great influence on the final prediction accuracy.

Moreover, a few link prediction approaches have been proposed specifically on detecting spurious links in network.\textsuperscript{36–41} Pan et al.\textsuperscript{38} defined a structural Hamiltonian for a network and calculated the conditional probability to detect spurious links. Samei and Jalili\textsuperscript{40} used two novel similarity metrics based on the hyperbolic distance of node pairs to predict missing and spurious links in multiplex networks. They further combined intralayer similarity with interlayer relevance to rank the links to predict whether there are spurious links.\textsuperscript{41} These works are effectively promoting the research of spurious link detection.

**Community detection based on the attractive force between nodes**

**Problem description**

For an undirected network $\mathcal{N} = (\mathcal{V}, \mathcal{E})$, $\mathcal{V}$ is the set of nodes, and $\mathcal{E}$ is the set of real links that exist between nodes. In addition, an adjacency matrix $A = [a_{ij}]_{n \times n}$ can be used to represent this network, $n$ is the number of nodes, $a_{ij} = 1$ if the link $(i,j)$ between node $i$ and node $j$ exists in the network. Otherwise, $a_{ij} = 0$, and it is regarded as a potential link.
In view of the deviation in the process of network data acquisition, especially on some large-scale networks, some links that should exist in the network may be ignored (missing links) and some links that should not exist may be misdeemed as real links (spurious links). To address this problem, we define connection probability \( p_c \) for the potential link and define break probability \( p_b \) for the real link as Definition 1 and Definition 2, respectively.

**Definition 1 (Connection probability)** For a potential link \( \langle i, j \rangle \) between two unconnected nodes \( i \) and \( j \) in network \( N \), \( \langle i, j \rangle \notin E \), its connection probability \( p_c \) denotes the possibility of being a real link. \( p_c \) is within the closed interval \([0, 1]\), and the greater value means the higher possibility as a missing link.

**Definition 2 (Break probability)** For a real link \( \langle i, j \rangle \) between two connected nodes \( i \) and \( j \) in network \( N \), \( \langle i, j \rangle \in E \), its break probability \( p_b \) denotes the possibility of the connection breaking between the two nodes. \( p_b \) is also within the closed interval \([0, 1]\), and the greater value means the higher possibility as a spurious link.

Table 1 shows the symbols used in this paper and their meaning.

Before calculating the probabilities for missing link prediction and spurious link detection, we need to divide all the nodes into some communities based on the attractive force between nodes. Details will be described in the following.

**The attractive force between nodes**

In order to measure the strength of the relation between any two nodes in network \( N \), we define the attractive force for each pair of nodes by referring to the formula of gravitation. The attractive force between node \( i \) and node \( j \) in \( N \) can be defined as Definition 3.

**Definition 3 (The attractive force between nodes)** For any two nodes \( i \) and \( j \) in network \( N \), the attractive force between them can be calculated by equation (1).

| Symbol | Meaning               | Symbol | Meaning               |
|--------|-----------------------|--------|-----------------------|
| \( N \) | Network               | \( V \)  | Sets of nodes in \( N \) |
| \( E \) | Sets of links in \( N \) | \( \langle i, j \rangle \) | The link between node \( i \) and node \( j \) |
| \( n \) | Number of nodes in \( N \) | \( m \)  | Number of links in \( N \) |
| \( a_{ij} \) | The indication of link type | \( A \)  | The adjacent matrix of \( N \), \( A = [a_{ij}]_{n \times n} \) |
| \( p_c \) | Connection probability of potential links | \( p_b \) | Break probability of real links |
| \( d_i \) | Degree of node \( i \) | \( R_{ij} \) | Shortest path distance between \( i \) and \( j \) |
| \( f_{ij} \) | Attractive force between \( i \) and \( j \) | \( C \)  | Sets of communities in \( N \) |
| \( t \)  | Number of communities in \( N \) | \( f_{i,c_i} \) | Attractive force between \( i \) and \( j \) |
where $d_i$ denotes the degree of node $i$, $R_{ij}$ denotes the shortest topological distance between the two nodes in $N$.

The numerator in the right side of equation (1) should be the product of the mass of node $i$ and node $j$. Consider that nodes can be regarded as the information promulgators in a network, and links can be used as the propagation paths of information as well as the media of influence diffusion. Although it is difficult to evaluate the influence of each node based on a same attribute in different networks, the number of links between a node and its neighbors can reflect the ability of the node to spread influence. Therefore, we take the degree of each node as its mass. In addition, since the actual geographical distance has been no longer limited in networks, two nodes can connect with each other no matter how far away they are. We only need to consider the path distance between two nodes and choose the shortest path distance here.

**Community detection**

Due to there are $m$ real links in the network $N = (V, E)$, the number of potential links is $n(n - 1)/2 - m$, it may be huge when the scale of the network is large. Moreover, finding all the shortest distance paths between pairs of nodes may waste a lot of time, especially on the social networks with large diameters. Therefore, we need to divide the nodes into some communities. Like the structure of galaxies in the universe, the nodes with large degree are also easily to affect and attract the nodes with small degree. Accordingly, we select the nodes with local maximum degree as the core of each community in the global network (a node owns local maximum degree if it has greater degree than all its neighbors). It is noteworthy that the core of a community may be just one node in general, however, if multiple connected nodes all own local maximum degree, they will be the core of a community together. The core of a community can be deemed as the initial structure of the community. Then, we divide other nodes into corresponding communities by considering the attractive forces between nodes and communities. Definition 4 describes the attractive force between a node and a community.

**Definition 4 (The attractive force between a node and a community)** For node $i$ and community $c_s$ in network $N$, the attractive force between them can be calculated by summing the attractive forces between node $i$ and all the nodes connected to $i$ in $c_s$. It can be expressed by equation (2).

$$f_{i,c_s} = \sum_{j \in c_s, k \in a_i} f_{ij} = f_{i,c_s}$$ (2)
For a non-core node, if it is attracted by only one community, it can be divided directly into this community, if it is attracted by multiple communities, it should be divided into the community with greater attractive force. Repeat the above process until all the nodes in the network are divided into communities. Algorithm 1 describes the details of community detection.

Algorithm 1: Community detection based on the attractive force between nodes

**Input:** $N = (V, E)$ and $A = [a_{ij}]_{n \times n}$

**Output:** $C = \{c_1, c_2, \ldots, c_t\}$; ($t$ is the number of communities)

1. initialize $t = 1$, $c_t = \emptyset$,
2. for each $i \in V$,
3. $m_i = d_i = \sum_{j=1}^{n} a_{ij}$;
4. if $m_i \geq m_{\text{neighbors}(i)}$
5. $c_i \leftarrow (i \text{ and its neighbors have the same degree})$, $t = t + 1, c_t = \emptyset$;
6. end if
7. end for
8. for each $i \in V \& i \notin C$ and $c_t \in C$
9. $f_i, c_t = \sum_{j \in c_t, k, a_{ij} = 1} f_j$ (only considering the attractive force between connected nodes here);
10. if $f_i, c_t \geq \text{ the attractive force between node } i \text{ and any other community}$
11. $c_i \leftarrow i$;
12. end if
13. end for
14. repeat 8-13 until $|C| = |V|$;
**Return** $C = \{c_1, c_2, \ldots, c_t\}$

For a non-core node, if it is attracted by only one community, it can be divided directly into this community, if it is attracted by multiple communities, it should be divided into the community with greater attractive force. Repeat the above process until all the nodes in the network are divided into communities. Algorithm 1 describes the details of community detection.

The time complexity of Algorithm 1 is $O(nm)$. Figure 1 shows Karate club social network as a sample. In the network, all the nodes can be divided into two communities.

![Figure 1. A sample network-Karate club social network.](image_url)
communities, node 1 and node 34 are the core of the two communities, respectively. In the community which takes node 1 as the core, node 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 18, 22 and node 17 are included. And the other community consists of node 34, 9, 10, 14, 15, 16, 19, 20, 21, 23, 24, 27, 28, 29, 30, 31, 32, 33, 25 and node 26.

**The approach for missing link prediction and spurious link detection**

The proposed approach will be described in two parts, one part is used for predicting the missing links from all the potential links and the other one is applied to detecting the spurious links from all the real links.

**Predicting the missing links from potential links**

For a potential link \( (i,j) \) in the network \( N \), obviously, it does not belong to \( E \), and the distance between the two nodes is at least 2. In order to judge whether it is a missing link, we need to calculate the connection probability of this potential link.

We first consider the case that the distance between node \( i \) and node \( j \) is 2. When the two nodes are at different communities, the attractive forces from the nodes they connect to in their respective communities can make the connection between the two nodes difficult to form, and the connection probability may be tiny. Therefore, we decide to ignore this case to simplify the calculation of our approach and set the connection probability between two nodes at different communities to 0.

When node \( i \) and node \( j \) are at the same community, their common neighbors can help them connect with each other. In addition, other nodes they respectively connect to may not prevent them from connecting, otherwise the community may be split. Therefore, the connection probability of the potential link \( (i,j) \) in this case can be defined as equation (3).

\[
p_c(i,j) = \frac{f_{ij} + \sum_k f_{ik} + \sum_k f_{jk}}{F_i + F_j}
\]

where \( k \) is a common neighbor of node \( i \) and node \( j \), \( F_i \) denotes the sum of attractive forces between node \( i \) and all its neighbors.

Such as node 7 and node 11 of Karate club social network, they belong to the same community, and \( R_{7,11} = 2 \), the connection probability of the potential link \( <7,11> \) can be calculated according to equation (3): \( p_c(7,11) = \frac{(f_{77} + f_{57} + f_{67}) + (f_{11,11} + f_{5,11} + f_{6,11})}{(f_{1,7} + f_{5,7} + f_{6,7} + f_{17,7}) + (f_{1,11} + f_{5,11} + f_{6,11})} = \frac{(3 + 92 + 69)}{(100 + 69)} = 0.970. \) Another such as node 25 and node 34 of the network, similarly, \( p_c(25,34) = \frac{(12.75 + 30 + 340)}{(39 + 1105)} = 0.335. \)

If the distance between node \( i \) and node \( j \) is more than 2. It can be regarded as the extend of the case that \( R_{ij} = 2 \). We need to find a shortest path between the two nodes at first, then selecting the nodes \( \{k_1, k_2, \ldots, k_l\} \) from the path that divide the
shortest path into several parts of length 2. In the final, we can obtain the connection probability of this potential link by multiplying the connection probability of each part together.

After the connection probabilities of all the potential links are calculated, we can select some potential links with great connection probabilities as the missing links in the network. Algorithm 2 describes the details of missing link prediction (LPCA-P).

Algorithm 2: Link prediction approach based on community structure and the attractive force between nodes for potential links (LPCA-P)

Input: $N = (V, E)$ and $C = \{c_1, c_2, \ldots, c_t\}$

Output: $P_c = \{pc\}_{n(n-1)/2-|E|}$

1. for each $i \in V, j \in V, (i, j) \not\in E$ and $R_{ij} = 2$
   1. if $i \in c_r, j \in c_s, r, s = 1, \ldots, t, r \neq s$
   2. $p_c(i, j) = 0$
   3. end if
   4. end for
   5. if $i, j \in c_t$
   6. for $k \in V, (i, k) \in E$, and $(j, k) \in E$
   7. $p_c(i, j) = f_{ij} + \sum_k f_{ik} + \sum_k f_{jk}$
   8. end for
   9. end if
10. end for
11. for each $i \in V, j \in V, (i, j) \not\in E$ and $R_{ij} > 2$
12. find a shortest path between $i$ and $j$;
13. select nodes $\{k_1, k_2, \ldots k_l\}$ that make $R_{ik_1} = R_{k_1 k_2} = \ldots = R_{k_l j} = 2$ (or $R_{jk_l} = 1$);
14. calculate $R_{ik_1}, R_{k_1 k_2}, \ldots$ according to 2-10;
15. $p_c(i, j) = p_c(i, k_1)p_c(k_1, k_2) \ldots p_c(k_l, j)$;
16. end for

Return $P_c = \{pc\}_{n(n-1)/2-|E|}$

Detecting the spurious links from real links

For a real link $(i, j)$ in $N$, that is, $(i, j) \in E$ and the distance between node $i$ and node $j$ is 1. In order to judge whether it is a spurious link, we need to calculate the break probability of this real link.

When node $i$ and node $j$ are at the same community, their connection is difficult to break because the attractive forces from the nodes outside the community could be very weak. Similar to the potential links prediction between two nodes at different communities, we can ignore the break probability in this case to simplify the calculation of our method and set the break probability between two nodes at the same communities to 0.

When node $i$ and node $j$ are at different communities, the attractive forces between each of them and each neighbor in their respective communities may
prevent them from connecting, however, their common neighbors can help them maintain the connection. Therefore, the break probability of the real link \( <i,j> \) in this case can be defined as equation (4).

\[
p_{b(i,j)|R_i} = \frac{F_i - \sum_k f_{ik} - f_{ij} + F_j - \sum_k f_{jk} - f_{ij} - (f_{ij} + \sum_k f_{ik} + \sum_k f_{jk})}{F_i + F_j - f_{ij}}
\]

\[= \frac{F_i + F_j - (3f_{ij} + 2\sum_k f_{ik} + 2\sum_k f_{jk})}{F_i + F_j - f_{ij}} \tag{4}\]

Such as real link \( <3,10> \) of Karate club social network, the two nodes belong to different communities. The break probability can be calculated according to equation (4): \( p_b(3,10) = (660 + 54 - 90)/(660 + 54 - 30) = 0.912 \).

After the break probabilities of all the real links are calculated, we can select some real links with great break probabilities as the spurious links in the network. Algorithm 3 describes the details of real links prediction (LPCA-R).

The time complexity of Algorithm 3 is \( O(n^3/t) \). Like Algorithm 2, we only consider the real links that the connected nodes are at different communities, Algorithm 3 can also be applied to distributed processing to reduce the time complexity.

**Experiments**

In this paper, we use some real-world networks for experiments to verify the proposed approach. In order to evaluate the prediction accuracy, we select AUC as the evaluation indicator and compare the results with some existing algorithms.

---

**Algorithm 3:** Link prediction approach based on community structure and the attractive force between nodes for real links (LPCA-R)

**Input:** \( N = (V, E) \) and \( C = \{c_1, c_2, \ldots, c_t\} \)

**Output:** \( P_b = \{p_b\}_E \)

1. for each \( i \in V, j \in V, <i,j> \in E \)
2. if \( i, j \in c_s \)
3. \( p_b(i,j) = 0; \)
4. end if
5. if \( i \in c_r, j \in c_s, r, s = 1, \ldots, t, r \neq s \)
6. for \( k \in V, <i,k> \in E, \text{ and } <j,k> \in E \)
7. \( p_b(i,j) = \frac{f_{i,k} + f_{j,k} - (3f_{ij} + 2\sum_i f_{ik} + 2\sum_j f_{jk})}{f_{i,k} + f_{j,k} - f_{ij}} \)
8. end for
9. end if
10. end for

**Return** \( P_b = \{p_b\}_E \)
Experimental setup

Seven real-world networks are used to conduct experiments to verify the proposed approach, including Dolphins network, American College Football Network, Neural network, Eu-core network, Political blogs network, Co-authorships network in network science, and P2P Gnutella 08 network. Table 2 shows the basic information of these networks.

AUC (Area Under the Receive Operating Characteristics Curve) is a common evaluation index of link prediction. It can be used to evaluate the accuracy of link prediction by measuring the times that the probability of a link in the test set is larger than the probability of a link randomly selected from the validation set after many times comparisons. The formula is defined as equation (5):

$$AUC = \frac{\sum_{i=1}^{n} T_i}{T}$$

where $T$ denotes the times of comparisons. If the probability of a link in the test set is larger than the probability of a link randomly selected from the validation set in the $i$th comparison, $T_i = 1$; If the link probabilities are the same, $T_i = 0.5$; otherwise, $T_i = 0$. The result of AUC is in the range of $[0,1]$, and the greater value means the higher prediction accuracy of the algorithm.

In addition, to verify the prediction accuracy of our method, we select some classic algorithms and newly proposed algorithms as comparison algorithms, including CN index, AA index, RA index, CRA algorithm, RWR algorithm, and KDLP algorithm. The detail descriptions of these comparison algorithms are in following:

(1) CN index evaluates a potential link between two unconnected nodes by counting their common neighbors, it can be defined as equation (6):

$$S_{ij}^{CN} = |\Gamma(i) \cap \Gamma(j)|$$

where $\Gamma(i)$ denotes the neighbors set of node $i$.  

| Network          | Number of nodes | Number of real links | Number of potential links |
|------------------|-----------------|----------------------|---------------------------|
| Dolphins         | 62              | 159                  | 1732                      |
| Football         | 115             | 616                  | 5939                      |
| Neural           | 307             | 2656                 | 44,315                    |
| Eu-core          | 1005            | 25,571               | 478,939                   |
| Political blogs  | 1491            | 19,025               | 1,091,770                 |
| Co-authorships   | 1589            | 2742                 | 1,258,924                 |
| P2P Gnutella 08  | 6301            | 20,777               | 19,827,373                |

Table 2. The basic information of experimental data sets.
(2) AA index can be regarded as a variant of CN index and it endows the common neighbors with more weight if they have lower degree. It can be defined as equation (7).

$$S_{ij}^{AA} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log|\Gamma(k)|}$$ (7)

(3) RA index is similar to AA and it can be defined as equation (8).

$$S_{ij}^{RA} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{|\Gamma(k)|}$$ (8)

(4) CRA algorithm considers both the common neighbors between two unconnected nodes and the common neighbors between the three. It can be defined as equation (9).

$$S_{ij}^{CRA} = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{|\Phi(k)|}{|\Gamma(k)|}$$ (9)

where $\Phi(k)$ is the set of the common neighbors of node $i$, node $j$ and node $k$.

(5) RWR algorithm is based on random walk, it assumes that some particles can walk randomly between any two nodes, and the importance of a link can be measured by counting the number of times that particles pass in it at a certain number of rounds. It can be defined as equation (10).

$$S_{ij}^{RWR} = q_{ij} + q_{ji}$$ (10)

where $q_{ij}$ denotes the probability that particles walk randomly via the real link $(i,j)$ from node $i$ to node $j$ at a certain number of steps.

(6) KDLP algorithm defines the knowledge quantity for each node by calculating its H-index and weights the links in a network based on knowledge dissemination between nodes. It is defined in equation (11).

$$KDLP(i,j) = \sum_{i=1}^{n} \beta^n(WA)^n_{ij}$$ (11)

where $\beta$ is a free parameter, $WA$ denotes the weighted adjacency matrix of a network and $n$ denotes the length of a path between node $i$ and node $j$.

Table 3 shows the time complexity of algorithms. Compared with the comparison algorithms, the approach proposed in this paper has less time complexity.
Experimental results of missing link prediction

In the experiments of missing link prediction, we divide all the potential links of a network into two parts, one is the validation set which account for \( t(0 < t < 1) \) of the potential links, and the other one that contains the remaining potential links is the test set. For the potential links in the validation set, they will be regarded as real links in the experiments. Table 4 shows the AUC results of missing link prediction on the seven networks by LPCA-P and the comparison algorithms when \( t \) is set to 5%.

According to the AUC results shown by Table 4, the proposed algorithm LPCA-P can achieve top two prediction accuracy on most of the experimental networks except for Eu-core network and Political blogs network when \( t \) = 5%. In the two networks, LPCA-P can also achieve the prediction accuracy close to the highest value.

In order to verify the performance of algorithms when more misinformation exists in a network, we further increase the value of \( t \) that means more potential links in the network are identified as real links. Figure 2 shows the changes of the AUC results in missing link prediction on these networks by LPCA-P and the comparison algorithms when \( t \) is increasing from 10% to 30% by 5%.

According to the results shown in Figure 2, the proposed algorithm LPCA-P can achieve and maintain high prediction accuracy with the increase of \( t \) on these networks. In addition, LPCA-P, CN, and KDLP perform better than AA, RA, and CRA. RWR algorithm performs unstable because of the uncertainty of the random walk. On the whole, the proposed algorithm LPCA-P can achieve...
satisfactory and stable prediction accuracy compared to the comparison algorithms on the seven experimental networks.

**Experimental results of spurious link detection**

Like the experiments of missing link prediction, we select some potential links which account for \( t(0 < t < 1) \) of all the potential links in a network randomly and treat them as real links. These selected potential links form the validation set, they and all the real links form the test set. Table 5 shows the AUC results of spurious links prediction on the experimental networks by the proposed algorithm LPCA-R and the comparison algorithms when \( t \) is set to 5%.

According to the AUC results shown by Table 5, it is obviously that the proposed algorithm LPCA-R can achieve top two prediction accuracy on all the seven experimental networks.

![Figure 2](image_url)  
Figure 2. Changes of the AUC results in missing link prediction on experimental data sets by different algorithms with the increasing \( t \) from 10% to 30% by 5%.
When \( t \) is 5\%. Then, like the experiments of missing link prediction, we further increase the value of \( t \) to verify the performance of algorithms when more spurious links exist in these networks. Figure 3 shows the changes of the AUC results in spurious link detection on the seven networks by LPCA-R and the comparison algorithms when \( t \) is increasing from 10\% to 30\% by 5\%.

It can be observed from Figure 3 that the proposed algorithm LPCA-R can achieve higher prediction accuracy than the comparison algorithms, except for KDLP algorithm, in real links prediction on the seven networks. However, KDLP algorithm performs best on Eu-core network, Political blogs network, and Co-authorships in network science. According to the results, the proposed algorithm LPCA-R and KDLP algorithm can achieve satisfactory and stable prediction accuracy on large-scale network, such as Political blogs network, Co-authorships in network science and P2P Gnutella 08 network, other algorithms underperform on these networks.

### Conclusions

In this paper, a link prediction approach based on community structure and the attractive force between nodes is proposed to predict the missing links and detect the spurious links in social networks. Firstly, the attractive force between any two nodes is defined to measure the strengths of the relation between nodes in a network. In addition, the nodes with local maximum degree are chosen as the core of communities and other nodes can be divided into these communities depending on the attractive force between them and communities. Then, the approach is described, which contains two algorithms LPCA-P and LPCA-R. LPCA-P calculates and compares the connection probability of each potential link to predict whether it is a missing link, and LPCA-R calculates and compares the break probability of each real link to detect whether it is a spurious link. Finally, we conduct experiments on seven real-world networks to verify the validity of the proposed approach and compare with some existing algorithms. The experimental results demonstrate that the proposed approach can achieve satisfactory and stable

| Network                | CN    | AA    | RA    | CRA   | RWR   | KDLP  | LPCA-R |
|------------------------|-------|-------|-------|-------|-------|-------|--------|
| Dolphins               | 0.8571| 0.7857| 0.9286| 0.6429| 0.7857| 0.8571| 1.0000 |
| Football               | 0.7167| 0.6667| 0.6500| 0.5500| 0.6667| 0.7000| 1.0000 |
| Neural                 | 0.7803| 0.8068| 0.7917| 0.6061| 0.7348| 0.9053| 0.9924 |
| Eu-core                | 0.6628| 0.7070| 0.7598| 0.6835| 0.7461| 0.9378| 0.7884 |
| Political blogs        | 0.6435| 0.6430| 0.6362| 0.5689| 0.6351| 0.9627| 0.9159 |
| Co-authorships network | 0.5085| 0.5086| 0.5191| 0.5066| 0.5073| 0.9440| 0.8666 |
| P2P Gnutella 08        | 0.5087| 0.5092| 0.5082| 0.5005| 0.5096| 0.9697| 0.9595 |
prediction accuracy compared to the comparison algorithms in missing link prediction and spurious link detection. In the future, we will try to further improve the performance of our approach on large-scale networks and plan to use distributed computing to reduce time consumption.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the National Social Science Foundation of China under No. 14BGL007, the Fundamental Research Funds for the
Central Universities under No.3072020CFW0910, the Postdoctoral Science Foundation of Heilongjiang Province under No. LBH-Z18052, and the Social Science Fund of Heilongjiang Province under No. 19GLC160.

**ORCID iD**

Hui Qu https://orcid.org/0000-0002-0244-7047

**References**

1. Aiello LM, Barrat A, Schifanella R, et al. Friendship prediction and homophily in social media. *ACM Trans Web* 2012; 62: 9.
2. Li Z, Fang X and Sheng ORL. A survey of link recommendation for social networks: methods, theoretical foundations, and future research directions. *ACM Trans Manage Inf Syst* 2018; 9(1): 1.
3. Kumar A, Singh SS, Singh K, et al. Link prediction techniques, applications, and performance: a survey. *Phys A* 2020; 553: 124289.
4. Li X and Chen H. Recommendation as link prediction in bipartite graphs: a graph kernel-based machine learning approach. *Decis Support Syst* 2018; 54(2): 880–890.
5. Gündoğan E and Kaya B. A link prediction approach for drug recommendation in disease-drug bipartite network. In: *Proceedings of 2017 international artificial intelligence and data processing symposium (IDAP)*, Malatya, Turkey, 16–17 September 2017, pp.1–4. New York: IEEE.
6. Li D, Zhang T, Xu Z, et al. Exploiting information diffusion feature for link prediction in Sina Weibo. *Sci Rep* 2016; 6: 20058.
7. Muniz CP, Goldschmidt R and Choren R. Combining contextual, temporal and topological information for unsupervised link prediction in social networks. *Knowl Based Syst* 2018; 156: 129–137.
8. Lü L, Pan L, Zhou T, et al. Toward link predictability of complex networks. *PNAS* 2014; 112(8): 2325–2330.
9. Lü L and Zhou T. Link prediction in complex networks: a survey. *Phys A* 2011; 390(6): 1150–1170.
10. Yao Y, Zhang R, Yang F, et al. Link prediction in complex networks based on the interactions among paths. *Phys A* 2018; 510: 52–67.
11. Tasami Y and Safaei F. A novel multilayer model for missing link prediction and future link forecasting in dynamic complex networks. *Phys A* 2018; 492: 2166–2197.
12. Zhou T, Lü L and Zhang Y-C. Predicting missing links via local information. *Eur Phys J B* 2009; 71(4): 623–630.
13. Clauset A, Moore C and Newman MEJ. Hierarchical structure and the prediction of missing links in networks. *Nature* 2008; 453: 98–101.
14. Ding J, Jiao L, Wu J, et al. Prediction of missing links based on community relevance and ruler inference. *Knowl Based Syst* 2016; 98: 200–215.
15. Yasami Y and Safaei F. A novel multilayer model for missing link prediction and future link forecasting in dynamic complex networks. *Phys A* 2018; 492: 2166–2197.
16. Wu Z, Lin Y, Wang J, et al. Link prediction with node clustering coefficient. *Phys A* 2016; 452: 1–8.
17. Hasan MA, Chaoji V, Salem S, et al. Link prediction using supervised learning. In: *Proceedings of SDM 06 workshop on link analysis, counterterrorism and security*, April 2006, pp. 798–805. Bethesda, Maryland, USA: SIAM.

18. Doppa JR, Yu J, Tadepalli P, et al. Learning algorithms for link prediction based on chance constraints. In: *Proceedings of the 2010 European conference on machine learning and knowledge discovery in databases*, Span, September 2010, pp. 344–360. Barcelona: Springer.

19. Yin G, Chi K, Dong Y, et al. An approach of community evolution based on gravitational relationship refactoring in dynamic networks. *Phys Lett A* 2017; 381(16): 1349–1355.

20. Lorrain F and White HC. Structural equivalence of individuals in social networks. *Soc Netw* 1971; 1(1): 67–98.

21. Adamic LA and Adar E. Friends and neighbors on the Web. *Soc Netw* 2003; 25(3): 211–230.

22. Jaccard P. Etude de la distribution florale dans une portion des Alpes et du Jura. *Bull De La Soc Vaud Des Sci Nat* 1901; 37(142): 547–579.

23. Katz L. A new status index derived from sociometric analysis. *Psychometrika* 1953; 18(1): 39–43.

24. Fouss F, Pirotte A, Renders J-M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Trans Knowl Data Eng* 2007; 19(3): 355–369.

25. Dai C, Chen L, Li B, et al. Link prediction in multi-relational networks based on relational similarity. *Inf Sci* 2017; 394–395: 198–216.

26. Rafiee S, Salavati C and Abdollahpouri A. CNDP: link prediction based on common neighbors degree penalization. *Phys A* 2020; 539: 122950.

27. Liu Z, Dong W and Fu Y. Local degree blocking model for link prediction in complex networks. *Chaos* 2015; 25(1): 013115.

28. Akiba T, Hayashi T, Nori N, et al. Efficient top-k shortest-path distance queries on large networks by pruned landmark labeling. In: *Proceedings of the twenty-ninth AAAI conference on artificial intelligence*, January 2015, pp. 2–8. Austin, Texas, USA: AAAI Press.

29. Liu W and Lü L. Link prediction based on local random walk. *Europhys Lett* 2010; 89(5): 58007.

30. Zhang P, Zeng A and Fan Y. Identifying missing and spurious connections via the bi-directional diffusion on bipartite networks. *Phys Lett A* 2014; 378(32–33): 2350–2354.

31. Scellato S, Noulas A and Mascolo C. Exploiting place features in link prediction on location-based social networks. In: *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, August 2011, pp. 1046–1054. San Diego, California, USA: ACM.

32. Zhou W and Jia Y. Predicting links based on knowledge dissemination in complex network. *Phys A* 2017; 471: 561–568.

33. Pech R, Hao D, Pan L, et al. Link prediction via matrix completion. *Europhy Lett* 2017; 117(3): 38002.

34. Williamson SA. Nonparametric network models for link prediction. *J Mach Learn Res* 2016; 17(1): 1–21.

35. Singh SS, Mishra S, Kumar A, et al. CLP-ID: community-based link prediction using information diffusion. *Inf Sci* 2020; 514: 402–433.
36. Guimera R and Sales-Pardo M. Missing and spurious interactions and the reconstruction of complex networks. *PNAS* 2009; 106(52): 22073–22078.

37. Zeng A and Cimini G. Removing spurious interactions in complex networks. *Phys Rev E* 2012; 85(3): 036101.

38. Pan L, Zhou T, Lü L, et al. Predicting missing links and identifying spurious links via likelihood analysis. *Sci Rep* 2016; 6: 22955.

39. Zhang P, Qiu D, Zeng A, et al. A comprehensive comparison of network similarities for link prediction and spurious link elimination. *Phys A* 2018; 500: 97–105.

40. Samei Z and Jalili M. Application of hyperbolic geometry in link prediction of multiplex networks. *Sci Rep* 2019; 9: 12604.

41. Samei Z and Jalili M. Discovering spurious links in multiplex networks based on interlayer relevance. *J Complex Netw* 2019; 7(5): 641–658.

42. Lusseau D, Schneider K, Boisseau OJ, et al. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: can geographic isolation explain this unique trait? *Behav Ecol and Sociobiol* 2003; 54(4): 396–405.

43. Girvan M and Newman MEJ. Community structure in social and biological networks. *PNAS* 2002; 99: 7821–7826.

44. Watts DJ and Strogatz SH. Collective dynamics of ‘small-world’ networks. *Nature* 1998; 393: 440–442.

45. Stanford Large Network Dataset Collection, http://snap.stanford.edu/data/(accessed July 2019).

46. Adamic LA and Glance N. The political blogosphere and the 2004 U.S. election: divided they blog. In: *Proceedings of the 3rd international workshop on Link discovery*, August 2005, pp. 36–43. Chicago, Illinois, USA: ACM.

47. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E* 2006; 74: 036104.

48. Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143(1): 7063747.

49. Cannistraci CV, Alanis-Lobato G and Ravasi T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Sci Rep* 2013; 3: 2613.

50. Tang H, Faloutsos C and Pan J. Fast random walk with restart and its applications. In: *Proceedings of the sixth international conference on data mining*, December 2006, pp. 613–622. Hong Kong, China: IEEE.

**Author biographies**

Hui Qu received the BS degree from College of Science and technology software, Shenyang Normal University, China, in 2010. Since 2015, she has been a doctoral candidate in School of Economics and Management at Harbin Engineering University, Harbin, China. Her research interests include technological innovation, cooperation of industry-university-research.

Wei Chen received the BS degree from School of Management, Jilin University, China, in 1982. Then he received MS degree from School of Economics and Management, Harbin Engineering University, China, in 1989. And he received PhD degree from School of Economics and Management, Harbin Engineering University, China, in 2003. Since 2006,
he has been a professor at School of Economics and Management, Harbin Engineering University, Harbin, China. His research interests include intelligent property protection and management strategy, technological innovation, cooperation of industry-university-research.

**Kuo Chi** received the BS degree from School of Mathematics and Statistics, Shandong University, China, in 2012. Then he received MS degree from College of Computer Science and Technology, Harbin Engineering University, China, in 2014. From 2017 to 2018, he studied as a visiting student at Department of Computer and Information Science, Temple University, Philadelphia, USA. In 2019, he received Ph.D. degree from College of Computer Science and Technology, Harbin Engineering University, China. Since 2020, he has been a lecturer at School of Information and Communication Engineering, Hainan University, Haikou, China. His research interests include social network analysis, intelligent information processing and wireless network.