Predicting future links with new nodes in temporal academic networks

Yijun Ran1,∗, Si-Yuan Liu1, Xiaoyao Yu1, Ke-Ke Shang2,∗ and Tao Jia1,∗

1 College of Computer and Information Science, Southwest University, Chongqing 400715, People’s Republic of China
2 Computational Communication Collaboratory, Nanjing University, Nanjing 210023, People’s Republic of China
∗ Authors to whom any correspondence should be addressed.
E-mail: ranyij288@gmail.com, kekeshang@nju.edu.cn and tjia@swu.edu.cn

Keywords: link prediction, temporal academic networks, nodal attributes, new nodes

Abstract
Most real-world systems evolve over time in which entities and the interactions between entities are added and removed—new entities or relationships appear and old entities or relationships vanish. While most network evolutionary models can provide an iterative process for constructing global properties, they cannot capture the evolutionary mechanisms of real systems. Link prediction is hence proposed to predict future links which also can help us understand the evolution law of real systems. The aim of link prediction is to uncover missing links from known parts of the network or quantify the likelihood of the emergence of future links from current structures of the network. However, almost all existing studies ignored that old nodes tend to disappear and new nodes appear over time in real networks, especially in social networks. It is more challenging for link prediction since the new nodes do not have pre-existing structure information. To solve the temporal link prediction problems with new nodes, here we take into account nodal attribute similarity and the shortest path length, namely, ASSPL, to predict future links with new nodes. The results tested on scholar social network and academic funding networks show that it is highly effective and applicable for ASSPL in funding networks with time-evolving. Meanwhile, we make full use of an efficient parameter to exploit how network structure or nodal attribute has an impact on the performance of temporal link prediction. Finally, we find that nodal attributes and network structure complement each other well for predicting future links with new nodes in funding networks.

1. Introduction

Complex networks have emerged as a powerful modeling tool for understanding and describing the real-world systems from the traffic system [1, 2] to the power system [3, 4], from the neural system [5, 6] to the ecosystem [7], from the economic system [8, 9] to the social system [10, 11], in which nodes are the entities of a system and links are the interactions or connections between entities [12, 13]. Understanding the evolutionary mechanisms of real systems as to which generates a pattern governing the evolution is a problem of fundamental importance for network science [14]. Most network evolutionary models such as Watts–Strogatz small-world model [15], and Barabási–Albert preferential attachment model [16] provide an iterative process for constructing global properties, but cannot capture the evolutionary mechanisms of real systems. Although real system may seem to be found regularly, capturing directly patterns on the whole process of evolution is not trivial. Hence, over the last few years, network scientists devoted a large effort to study the properties of local structures such as link prediction, community detection and network reconstruction, which can help us understand the real system’s evolution law [17–19].

Link prediction has long been a question of great interest in network mining, which can be dated back to the first time by Liben-Nowell and Kleinberg [20]. The aim of link prediction is to uncover missing links from known parts of the network or quantify the likelihood of the emergence of future links from current structures of the network. Up to now, a considerable number of studies have demonstrated that link prediction not
only has a wide range of practical value but also has an important theoretical significance [21–24]. In applications, the results of link prediction are helpful to understand what structural damage causes a certain disease by analyzing the molecular network in the cell [21]. Link prediction techniques can also help pharmaceutical experts to develop therapeutic drugs without obvious side effects through studying the medical network [22]. In theoretical aspects, with the help of link prediction algorithms, one can understand the evolutionary mechanisms of real systems and further generate evolution model [25–27]. While some researches have been carried out on link prediction, there is still very little scientific understanding of link prediction in temporal networks.

Recently, several methods have been proposed to predict future links in temporal or time-evolving networks [28]. Shang et al. pointed out that if a pair of nodes are linked to each other in the current network, they will have a higher probability to link the same nodes in future networks [29]. Güneş et al. adopted the neighborhood-based node similarity scores with different timestamps as time series, and then utilized the time series model to predict new or recurring links in evolving networks [30]. Wang et al. showed that each node has the ability to attract other nodes where the ability of node depends not only on its structural importance, but also on its current popularity—active nodes have much more probability to connect future links [31]. Chi et al. proposed the similarity-based predictor, DLPA, that is associated with the attraction force between nodes to detect missing links and to predict whether potential links will become real links in the future [32]. Büttün et al. took advantage of both local and global topological structures to predict the citation count of scientists by employing the temporal link prediction metric [33]. The widely existing methods of link prediction almost focus on predicting future links in time-evolving networks with only links appearing or vanishing. However, most real systems evolve over time in which entities and the interactions between entities are added and removed—new entities or relationships appear and old entities or relationships vanish. Therefore, it is necessary to design an efficient approach to predict future links in temporal networks with nodes appearing or vanishing.

In real systems, two users would establish a friendship with a high chance in the near future if they have many common friends or interests, in contrast, the common friends or interests can be utilized to uncover lost friends or predict future friends [18, 34, 35]. This shows that network structure information and nodal attributes are both important for link prediction. However, users may suffer from unavailable and unreliable attribute information due to the privacy policy in most scenarios. Fortunately, the attributes of users in real systems such as scholar social system are easy to obtain and are relatively reliable. This is because a new scholar need to fill in some basic attribute information, such as affiliation, research direction, academic field, and so on, when he or she registers in an academic social system like ResearchGate, SCHOLAT. Hence, it is of great significance to take advantage of the attribute information to study the interaction between new users and old users.

In this study, we extend the structural equivalence and shortest path length hypotheses to predict future links between new nodes and old nodes in long-path networks with time-evolving [36]. This work generalizes our previous work [36] for an extension of long-path static networks into long-path networks with time-evolving. To solve the temporal link prediction with new nodes, here we integrate the network structure and external information (e.g., nodal attribute) to predict future links. Especially, we take into account nodal attribute similarity and shortest path length, namely, ASSPL to predict future links when the networks have new nodes. The results tested on scholar social network and academic funding networks show that ASSPL is highly effective and applicable for funding networks with time-evolving.

Our contributions are as following: (1) we propose a new similarity-based framework that integrates the network structure and nodal attribute for the temporal link prediction with new nodes. For this, we make full use of an efficient parameter to exploit how network structure or nodal attribute has an impact on the performance of temporal link prediction. (2) Extensive experiments are conducted on several scholar social networks and academic funding networks, and ASSPL predictor we proposed is shown to outperform state-of-the-art methods for temporal link prediction with new nodes in funding networks with time-evolving.

The remainder of the paper is organized as follows. We give a brief description of the temporal link prediction task and real network data in section 2. In section 3, we introduce static link prediction methods and propose temporal similarity methods. We report the main results and findings in section 4. Finally, section 5 is the conclusion and discussion.

2. Problem definition and data description

2.1. Problem definition

Given time periods, \( t = 1, \ldots, T \), where \( T \) is the window size, we consider an undirected time-evolving network (snapshot) \( G_t \) composed of a set of nodes \( V_t \) and a set of links \( L_t \), in which a node cannot connect to itself...
Figure 1. An example demonstrating the task of temporal link prediction. In this network, the old nodes or links may vanish and the new nodes or links may appear over time. We take advantage of the network structures at time \( t = T \) to predict future links at time \( t = T + 1 \). Here, the attributes of node 1 is \( A_1 = \{a\} \), the attributes of new node 7 is \( A_7 = \{c, d, e\} \), and so on. More importantly, from the beginning, every node in the network has attributes that will not change over time.

(no self-loops) nor share more than one link with another node (no repeated links). Each node \( i \) has a set of associated attributes denoted as \( A_i \) which do not change over time. Here, some new nodes or links in undirected time-evolving network will appear and some old nodes or links in undirected time-evolving network will vanish within time period \( t \). Given a time-evolving network \( G_t \) at time \( t \), we formulate the temporal link prediction problem as follows.

The problem of temporal link prediction is to predict the occurrence probabilities of future links by utilizing past network structures. Here, for a time-evolving network, we define the \( L \) links of the network at time \( t \) as the training set \( L^T_t \). The \( L_{t+1} \) links of the network at time \( t + 1 \) are considered as future links set \( L^P_{t+1} \) and need to be inferred from network topology given by the links in \( L^T_t \). To test the performance of temporal link prediction, another probe set \( L^N_{t+1} \) is used as the control group of \( L^P_{t+1} \), which is composed of randomly chosen nonexistent links from the network \( G_{t+1} \) usually with the same size of \( L^P_{t+1} \). And this sampling process must make sure that the randomly chosen nonexistent links also do not exist in \( G_t \). Figure 1 illustrates this procedure of temporal link prediction where the dot line in the time-evolving network at time \( t = T + 1 \) is future links set \( L^P_{t+1} \), the solid line in the time-evolving network at time \( t = T \) is training set \( L^T_t \).

In the temporal link prediction problem, a predictor takes node attributes and some features of the network, and then assigns a score \( S_{ab} \) to each pair of nodes \( a \) and \( b \), which is proportional to the probability that nodes \( a \) and \( b \) will be connected. As \( G_t \) is undirected, the score is symmetric, i.e., \( S_{ab} = S_{ba} \). The scores for node pairs that are not currently connected are sorted in descending order and the top candidates are likely future links. The prediction quality is measured by comparing the scores of predicted links in \( L^P_{t+1} \) and \( L^N_{t+1} \). Here, we randomly pick up a link from \( L^P_{t+1} \) and \( L^N_{t+1} \) respectively, and then compare the score of the two links each time. After \( n \) times of independent comparisons, if there are \( n' \) times that the score of future link in \( L^P_{t+1} \) is higher than that of the nonexistent link in \( L^N_{t+1} \), and \( n'' \) times that the future link and the nonexistent link have the same score [34, 37], hence the result of AUC can be calculated as

\[
AUC = \frac{n' + 0.5n''}{n}.
\] (1)

2.2. Data description and processing

2.2.1. Data description

In this work, two kinds of large temporal networks are taken to evaluate the performance of algorithms. One is scholar social network which is from the SCHOLAT website. The SCHOLAT is an emerging social system specifically designed and built for scholars, learners and course instructors in China. And the SCHOLAT website allows users around academic and learning discourse to strengthen social interaction and collaboration among scholar communities. This social network is an open-access dataset from SCHOLAT Lab which contains 10607 nodes and 168540 links. This dataset has been divided into the training set \( L^T_t \) and the future links set \( L^P_{t+1} \). According our precious study, the scholar social network is short-path network as shown in figure 2(a). More complete statistics of the social dataset are given in table 1.

Another one is academic funding network (co-occurrence network of funds) which denotes the relationship between two funds that appear in the same paper. We collect the fund datasets from the National Natural
Science Foundation of China (NSFC). There are 320,226 funds in the datasets we collected. In addition, for academic papers, we use the publication data of the Web of Science (WoS) from year 2008 to 2016. To mine the co-occurrence relationship which represents the fact that two fundings jointly fund a research paper, we perform link prediction analysis on academic funding network. Since a fund has its own survival time, here we take advantage of the datasets by a two-year window to construct a time-evolving network to predict the future links in the next year. Hence, we construct 7 academic funding networks utilizing the datasets we collected. For example, we denote Funding0809 in table 1 as the network which is constructed by using the data of year 2008 and 2009, and then we study the likelihood of the emergence of links within year 2010 on Funding0809 network. According to our precious study [36], the 7 academic funding networks are all long-path networks, such as Funding0809 network shown in figure 2(b). More complete statistics of the funding dataset are given in table 1.

Here, the scholar social network and academic funding network are typical time-evolving networks in which new nodes or links will appear and old nodes or links will vanish over time. Hence, the performance of traditional temporal approaches in those networks is poor. In this study, we propose a new similarity-based framework that integrates the network structure and nodal attribute for the temporal link prediction with new nodes.

2.2.2. Data processing
Regardless of whether a scholar social network or academic funding network is considered, the attribute information we collected is unstructured Chinese text data such as research interests, discipline, fund title and so on. Indeed, Chinese text data is totally different from English text data. We firstly carry out word segmentation for Chinese text data, and then remove stop words (stopwords) that do not provide any useful information. To better study the temporal link prediction task with new nodes, here we apply the latent Dirichlet allocation
Figure 3. An example of the attribute augmentation for a node without attributes. Here, the node E has no attributes, hence we take E's neighbors (A, B, C, and D) in the network $G_t$ to obtain the similar attributes. That is to say, we apply the union set of the keyword sets of the four neighbors to obtain the augmented keyword set of E.

...(LDA) topic model \cite{39} to build the topic vector of each node and measure the similarity between nodes by cosine similarity \cite{40}. The process of topic vector processing is as follows:

- **Word segmentation.** To segment the proper nouns well, for example, do not segment 'Southwest University' into 'Southwest' and 'University', here we first use the 'modify dictionary' mode in the Jieba tool \cite{41} to adjust the frequency of a single word which can segment the proper nouns well. Then we apply the 'keyword extraction' mode based on TF-IDF algorithm in the Jieba tool to re-segment the node's attributes and extract the top five weighted keywords of a node. For the node who has no keyword, in this study we take advantage of the node's neighbors in the network $G_t$ to obtain the union set of keywords. This attribute augmentation process is depicted in figure 3.

- **Keywords vector.** After getting the keywords list for all nodes, we build a dictionary for each keyword with a unique index. Then, we utilize the bag-of-words model (Doc2BOW in Gensim tool \cite{42}) to process each node's keywords into a vector according to the frequency of occurrence. We further use TF-IDF model to convert the result of the bag-of-words model from the frequency vector of the topics to the importance vector of keywords.

- **LDA topic model.** In this step, we use the importance vector of keywords calculated by TF-IDF to build LDA topic model, in which we set the count of topics as 10 (the dimension of the topic space). The 10 dimensions here are chosen after extensive testing. The LDA model can project a node's keywords into the topic space, which helps us to calculate the similarity between nodes by cosine similarity.

3. **Similarity-based methods**

Among many similarity-based predictors, Cao et al systematically compared common neighbors (CN) with other well-known similarity predictors and further demonstrated that CN predictor performs best in short-path static networks \cite{43}. Zhou et al proposed resource allocation (RA) predictor and showed that the predictive performance of RA is higher than CN, and Adamic–Adar (AA) index in short-path static networks \cite{44}. In addition, to study the role of weak ties, Lü and Zhou extended the CN, AA and RA into weighted static networks by introducing a free parameter that controls the contributions of weak ties \cite{45}. Hence, to study temporal link prediction in short-path networks, we employ the approach of Lü and Zhou to extend the CN, AA and RA as temporal similarity methods by integrating the network structure and nodal attribute. In our recent study, we proposed SESPL predictor which is associated with the principles of structural equivalence and shortest path length. The importance and originality of our recent study are that it systematically explored link prediction in short-path and long-path networks and showed that SESPL is a state-of-the-art predictor in long-path static networks. Therefore, in this study, we also extend SESPL predictor into long-path networks with time-evolving.

3.1. **Static similarity methods**

The static similarity methods mostly take advantage of network topology to predict future links in static networks. Here, we concentrate on CN, AA, RA and SESPL predictor, whose definitions are as following.

(1) Common neighbors (CN)
The CN directly counts the number of CN two nodes share [46]. The idea behind is that two nodes that share the same neighborhood node are likely to share other common features hence are likely to have a link. It is defined as

\[ S_{ab}^{CN} = |n(a) \cap n(b)|, \]  

where \( n(a) \) and \( n(b) \) denote the set of neighbors nodes that node \( a \) and \( b \) have, respectively.

(2) Adamic–Adar index (AA)

AA was proposed by Adamic and Adar and is used to compute the similarity between two web pages [47]. The AA emphasizes less-connected CN. It is defined as

\[ S_{ab}^{AA} = \sum_{c \in (n(a) \cap n(b))} \frac{1}{\log k(c)}, \]  

where \( k(c) = |n(c)| \) denotes the degree of node \( c \).

(3) Resource allocation index (RA)

Zhou et al. [44] motivated from the physical processes of RA, hence proposed the RA that puts penalties to large degree nodes. Yielding

\[ S_{ab}^{RA} = \sum_{c \in (n(a) \cap n(b))} \frac{1}{k(c)}, \]  

(4) SESPL

Our previous work proposed a novel similarity measure, SESPL, which is associated with the principles of structural equivalence and shortest path length [36]. The experimental results showed that the SESPL predictor works better than other classical similarity-based predictors in long-path networks. SESPL is defined as

\[ S_{ab}^{SESPL} = \frac{\text{SE}(a, b)}{d_{ab} - 1}, \]  

where \( d_{ab} \) denotes the shortest path length between node \( a \) and \( b \), and \( \text{SE}(a, b) \) is a measure that quantifies the structural equivalence between the local structures centered on \( a \) and \( b \), respectively. Here we quantify the structural equivalence of two nodes by taking into account local structures consisting of the first-order and second-order neighbors. The \( \text{SE}(a, b) \) is defined as

\[ \text{SE}(a, b) = 1 - J(P_a, P_b), \]  

where \( J(P_a, P_b) \) represents the Jensen–Shannon divergence between the local structure centered on nodes \( a \) and \( b \), respectively. And \( P_i \) is a vector in which the elements are the degree centrality of each node \( i \) in the local structure centered on nodes \( a \).

### 3.2. Temporal similarity methods

Most real systems evolve over time in which entities and the interactions between entities are added and removed—new entities or relationships will appear and old entities or relationships will vanish. This indicates that a systematic understanding of how a new node links to another node is the most meaningful work in time-evolving systems. However, because the new node does not have pre-existing structure information, it is very difficult to predict the future links with new node for static similarity methods. To deeply solve the temporal link prediction with new nodes, here we integrate the network structure and external information (e.g., nodal attribute) to predict future links. The topic vector of each nodal attribute was processed by the ways in section 2.2.2. Here, we first utilize the topic vector to define the similarity between two nodes by the cosine similarity [40], which is defined as

\[ \cos(a, b) = \frac{\sum_{i=1}^{m} V(a)_i V(b)_i}{\sqrt{\sum_{i=1}^{m} V(a)_i^2} \times \sqrt{\sum_{i=1}^{m} V(b)_i^2}} \]  

where \( V(a) \) is the topic vector of node \( a \) in which \( m \) represents the dimension of the topic space in LDA model.

We then extend static CN, AA and RA predictor into temporal networks by employing the approach of Lü and Zhou [45]. For CN predictor, here we name TCN in temporal networks. It is defined as

\[ S_{ab}^{TCN} = \begin{cases} \sum_{c \in (n(a) \cap n(b))} \cos(a, c)^\beta + \cos(c, b)^\beta & \text{if } a \text{ and } b \text{ are not new node,} \\ \cos(a, b) & \text{otherwise,} \end{cases} \]  

here \( \beta \) is a free parameter. Likewise, TAA and TRA are defined as
Table 2. The performance of static similarity predictors measured by AUC. In this study, all the results are based on the average over 1000 independent runs of simulation.

| Datasets   | CN  | AA  | RA  | SESPL |
|------------|-----|-----|-----|-------|
| Funding0809| 0.519| 0.519| 0.519| 0.499 |
| Funding0910| 0.522| 0.522| 0.523| 0.502 |
| Funding1011| 0.520| 0.520| 0.520| 0.499 |
| Funding112 | 0.524| 0.524| 0.524| 0.502 |
| Funding123 | 0.529| 0.529| 0.529| 0.509 |
| Funding1314| 0.530| 0.530| 0.530| 0.505 |
| Funding1415| 0.531| 0.531| 0.531| 0.507 |
| SCHOLAT    | 0.994| 0.995| 0.996| 0.989 |

\[
S_{\text{TAA}}^{ab} = \begin{cases} 
\sum_{c \in n(a) \cap n(b)} \frac{\cos(a,c)^{\beta} + \cos(c,b)^{\beta}}{\log(1 + s(c))} & \text{if } a \text{ and } b \text{ are not new node,} \\
\cos(a,b) & \text{otherwise,}
\end{cases} \tag{9}
\]

\[
S_{\text{TRA}}^{ab} = \begin{cases} 
\sum_{c \in n(a) \cap n(b)} \frac{\cos(a,c)^{\beta} + \cos(c,b)^{\beta}}{s(c)} & \text{if } a \text{ and } b \text{ are not new node,} \\
\cos(a,b) & \text{otherwise,}
\end{cases} \tag{10}
\]

where \(s(a) = \sum_{c \in n(a)} \cos(a,c)^{\beta}\).

What is interesting about the contributions in our previous work is that we take advantage of the principles of the structural equivalence and the shortest path length to mine missing links in long-path static networks. While the SESPL we proposed in the previous work cannot predict future links with new nodes in time-evolving networks, SESPL has efficient scalability. The single most striking feature to emerge from the SESPL predictor prompts us to further improve the algorithm so that it can be applied to temporal link prediction research. Therefore, we can make full use of nodal attribute similarity and shortest path length, namely, ASSPL, to predict future links with having new nodes. Because the more similar identities of two people, the more likely it is for them to promote trust and cooperation. Here, we define ASSPL as following:

\[
S_{\text{ASSPL}}^{ab} = \begin{cases} 
\beta \cdot \text{SE}(a,b) + \frac{\cos(a,b)}{d_{ab} - 1} & \text{if } a \text{ and } b \text{ are not new node,} \\
\cos(a,b) & \text{otherwise.}
\end{cases} \tag{11}
\]

4. Results

4.1. The failure of static similarity methods

To examine the temporal link prediction task, here we first apply static similarity methods which are the best-known algorithms for mining missing links to predict future links with new nodes in time-evolving networks. For the unconnected node pairs including new nodes, the static similarity methods score their similarity zero. As we expected, all static similarity algorithms fail to predict future links with new nodes in time-evolving funding networks (long-path networks), shown in table 2. This is because it is similar to the performance of the randomly chosen method for that of static predictors measured by AUC. Especially, for SESPL predictor which is associated with the principles of structural equivalence and shortest path length, it exhibits extremely poor accuracy. However, the result to emerge from table 2 is that all static similarity methods can achieve almost the best predictive performance in SCHOLAT network (short-path network). This may be because future links in SCHOLAT network have fewer new nodes than those in funding networks. Perhaps, this is also because link prediction may be fundamentally easier in social networks [48]. Together, static similarity algorithms cannot deal with future links with new nodes in time-evolving networks.

4.2. The highlight of temporal similarity methods

In order to mine future links with new nodes in time-evolving networks, we perform the predictive ability of our temporal similarity methods. In the beginning, we only consider nodal attributes for future links between old nodes, and set the score of future links involving one or two new nodes as zero. We find that each predictor in table 3 has a gain on the predictive accuracy comparing to the results in table 2. This shows that the attributes of the node can promote the performance of missing links mining.

Next, to further predict the future links with new nodes, we adopt nodal attributes to compute the similarity of future links with new nodes. Table 4 is quite revealing in several ways. First, compared to table 2, we can
Table 3. The performance of temporal similarity predictors measured by AUC. Here, we compute the score of future links only with old nodes in the next year by using network structure and nodal attributes. Hence, the score of future links with new nodes is zero. In this table, the parameter $\beta$ in all temporal similarity methods is 1.

| Datasets     | TCN  | TAA  | TRA  | ASSPL |
|-------------|------|------|------|-------|
| Funding0809 | 0.520| 0.520| 0.519| 0.501 |
| Funding0910 | 0.522| 0.522| 0.522| 0.503 |
| Funding1011 | 0.520| 0.520| 0.520| 0.498 |
| Funding1112 | 0.524| 0.524| 0.524| 0.503 |
| Funding1213 | 0.530| 0.529| 0.529| 0.511 |
| Funding1314 | 0.530| 0.530| 0.530| 0.506 |
| Funding1415 | 0.530| 0.531| 0.531| 0.509 |
| SCHOLAT     | 0.994| 0.995| 0.996| 0.989 |

Table 4. The performance of temporal similarity predictors measured by AUC. Here, we compute the score of all future links of next year by using network structure and nodal attributes. Bold numbers are the best performance of all algorithms. In this table, the corresponding optimal parameters $\beta$ in all temporal similarity methods are shown in table 5. Moreover, the AUC performance of ASSPL is also the best and can beyond 0.7 for funding networks, if we only consider the earliest 10% new nodes of next year.

| Datasets     | TCN  | TAA  | TRA  | ASSPL |
|-------------|------|------|------|-------|
| Funding0809 | 0.644| 0.642| 0.623| 0.653 |
| Funding0910 | 0.643| 0.643| 0.619| 0.654 |
| Funding1011 | 0.656| 0.653| 0.632| 0.666 |
| Funding1112 | 0.654| 0.652| 0.627| 0.670 |
| Funding1213 | 0.633| 0.632| 0.600| 0.654 |
| Funding1314 | 0.633| 0.633| 0.600| 0.657 |
| Funding1415 | 0.625| 0.621| 0.591| 0.650 |
| SCHOLAT     | 0.997| 0.998| 0.997| 0.993 |

Table 5. The optimal parameters $\beta$ in all temporal similarity methods subject to the highest AUC.

| Datasets     | TCN  | TAA  | TRA  | ASSPL |
|-------------|------|------|------|-------|
| Funding0809 | 1.6  | −1.0 | 1.0  | 3.0   |
| Funding0910 | −0.6 | −1.0 | 0.8  | 2.8   |
| Funding1011 | 1.0  | −0.6 | 3.0  | 3.0   |
| Funding1112 | 0.6  | 0.8  | 1.8  | 3.0   |
| Funding1213 | 1.6  | −1.0 | 1.8  | 2.6   |
| Funding1314 | 0.4  | −0.6 | 2.8  | 3.0   |
| Funding1415 | 0.4  | −0.8 | 3.0  | 3.0   |
| SCHOLAT     | 0.6  | 2.4  | 2.0  | 1.6   |

see that temporal similarity methods can achieve a better performance than that of static similarity predictors. Moreover, the average performance of ASSPL is 30.86% (0.653–0.499%) higher than that of SESPL. Second, compared to table 3, we find that the reasonable use of the similarity of new nodes can totally improve the performance of link prediction in time-varying networks. Third, the state-of-the-art predictor for predicting future links with new nodes comes from ASSPL in academic funding networks. This means that ASSPL we extended is also highly effective and efficient in long-path networks with time-evolving. Finally, while static similarity predictors can successfully predict future links with new nodes in SCHOLAT network (table 2), temporal similarity methods utilizing nodal attribute can further improve the predictive performance. Taken together, these results indicate that temporal similarity methods can better deal with future links with new nodes. In addition, these results also raise an important question: prediction accuracy in temporal similarity methods is enhanced by which network structural property or nodal attribute?

To answer this question, we further deeply analyze how the network structure and node attributes affect the prediction performance by setting different the parameter $\beta$. As shown in figure 4, the most obvious finding is that the parameter $\beta$ has a major impact on the prediction accuracy of ASSPL in short-path network (SCHOLAT) and long-path networks (funding networks). The performance of ASSPL increases monotonously and eventually reaches a stable level with the increase of the $\beta$. This means that nodal attributes and network
Figure 4. The performance of temporal similarity predictors measured by AUC under the variation of the parameter $\beta$. (a) Funding0809, (b) Funding1314, (c) Funding1415, and (d) SCHOLAT. (a)–(c) belong to long-path networks and (d) is short-path network. All the results in this figure are based on the average over 1000 independent runs of simulation.

structure can work together well for predicting future links with new nodes in time-evolving networks. However, $\beta$ almost has the same impact on the prediction accuracy of TCN, TAA, TRA in short-path network and long-path networks when $\beta < 1.5$. The prediction accuracy of TCN, TAA, TRA in short-path network has a downward trend when $\beta > 1.5$. This indicates that the parameter $\beta$ may have an inhibitory effect on the prediction accuracy of TCN, TAA, TRA after exceeding the threshold.

5. Conclusion and discussion

To summarize, we extend the structural equivalence and shortest path length hypotheses to predict future links between new nodes and old nodes in long-path networks with time-evolving. To solve the temporal link prediction problems with new nodes, in this study, we take into account nodal attribute similarity and shortest path length, ASSPL, to predict future links with new nodes. The results tested on scholar social network and academic funding networks show that it is highly effective and applicable for ASSPL in funding networks with time-evolving. In addition, to study how network structure and nodal attribute affect the prediction accuracy in temporal similarity methods, we further deeply analyze the prediction performance by setting different parameter $\beta$. We find that ASSPL can obtain the best predictive performance when taking advantage of nodal attribute similarity and network structure in time-evolving networks. Meanwhile, our finding shows that the more similar funds are, the more likely for them to jointly fund the same research. More importantly, this finding will bring new insights into the standardized funding of the fund in a work.

In this work, we propose a new similarity-based framework that integrates the network structure and nodal attribute for the temporal link prediction with new nodes. This may have a problem that the nodal attribute in most real networks cannot utilize. In the future, for the link prediction task of whether there are new nodes, we will take advantage of the representation learning techniques to further study. In addition, we will take into account the time series utilizing only network structure to predict future links with new nodes.
Acknowledgments

We thank Leilin Feng for her help in funding datasets collection. This work is supported by National Natural Science Foundation of China (61803047, 61603309), Industry-University-Research Innovation Fund for Chinese Universities (No. 2021ALA03016), Major Project of the National Social Science Foundation of China (19ZDA149, 19ZDA324), Fundamental Research Funds for the Central Universities (14370119, 14390110). Yijun Ran is supported by China Scholarship Council (CSC No. 202006990042).

Data availability statement

The SCHOLAT dataset used in this study is available from SCHOLAT website. The funding datasets used in this study are available from the corresponding author upon reasonable request.

ORCID iDs

Yijun Ran https://orcid.org/0000-0002-7047-3343
Ke-Ke Shang https://orcid.org/0000-0002-7454-4276
Tao Jia https://orcid.org/0000-0002-2337-2857

References

[1] Zeng G et al 2020 Multiple metastable network states in urban traffic Proc. Natl. Acad. Sci. USA 117 17528–34
[2] Zeng G, Li D, Guo S, Gao L, Gao Z, Stanley H E and Havlin S 2019 Switch between critical percolation modes in city traffic dynamics Proc. Natl. Acad. Sci. USA 116 23–8
[3] Liu Y-Y, Slotine J-J and Barabási A-L 2011 Controllability of complex networks Nature 473 167–73
[4] Li A, Cornelius S P, Liu Y-Y, Wang L and Barabási A-L 2017 The fundamental advantages of temporal networks Science 358 1042–6
[5] Zhang J et al 2016 Neural, electrophysiological and anatomical basis of brain-network variability and its characteristic changes in mental disorders Brain 139 2307–21
[6] Girardi-Schappo M 2021 Brain criticality beyond avalanches: open problems and how to approach them J. Phys. Complex 2 031003
[7] Dubois M, Gascuel D, Goll M and Claudet J 2019 Recovery debts can be revealed by ecosystem network-based approaches Ecosystems 22 658–76
[8] Chakraborti A, Hrishidev K, Sharma K and Pharasi H K 2020 Phase separation and scaling in correlation structures of financial markets J. Phys. Complex 2 015002
[9] Zhou T 2021 Representative methods of computational socioeconomics J. Phys. Complex 2 031002
[10] Jia T, Wang D and BoleslawSzymanski K 2017 Quantifying patterns of research-interest evolution Nat. Human Behav. 1 78
[11] Ran Y, Deng X, Wang X and Jia T 2020 A generalized linear threshold model for an improved description of the spreading dynamics Chaos 30 083127
[12] Barabási A-L. 2012 The network takeover Nat. Phys. 8 14–6
[13] Fortunato S et al 2018 Science of science Science 359 eaao0185
[14] Zhou T 2021 Progresses and challenges in link prediction Sci.ence 24 103217
[15] Watts D J and Strogatz S H 1998 Collective dynamics of small-world networks Nature 393 440–2
[16] Barabási A-L. and Albert R 1999 Emergence of scaling in random networks Science 286 509–12
[17] Linhuan L, Pan L, Zhou T, Zhang Y-C and Stanley H E 2015 Toward link predictability of complex networks Proc. Natl. Acad. Sci. USA 112 2325–30
[18] Linhuan L, Chen D, Ren X-L, Zhang Q-M, Zhang Y-C and Zhou T 2016 Vital nodes identification in complex networks Phys. Rep. 650 1–63
[19] Sun J, Feng L, Xie J, Xiao M, Wang D and Hu Y 2020 Revealing the predictability of intrinsic structure in complex networks Nat. Commun. 11 1–10
[20] Liben-Nowell D and Kleinberg J 2007 The link prediction problem for social networks J. Am. Soc. Inf. Sci. 58 1019–31
[21] Barzel B and Barabási A-L 2013 Network link prediction by global silencing of indirect correlations Nat. Biotechnol. 31 720–5
[22] Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, Peng J, Chen L and Zeng J 2017 A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information Nat. Commun. 8 1–13
[23] Kovács I A et al 2019 Network-based prediction of protein interactions Nat. Commun. 10 1240
[24] Mariani M S and Lü L 2020 Network-based ranking in social systems: three challenges J. Phys. Complex 1 011001
[25] Tang D, Du W, Shkfehman L, Wang Y, Havlin S, Cao X and Yan G 2020 Predictability of real temporal networks Nat. Sci. Rev. 7 929–37
[26] Shang K-K, Yang B, Moore J M, Ji Q and Small M 2020 Growing networks with communities: a distributive link model Chaos 30 041101
[27] Xian X, Wu Y, Gao L, Wang X-Z, Wang W and Liu Y 2020 Netsre: link predictability measuring and regulating Nonlinear Syst. 196 105800
[28] Cui M, Mariani M S and Medo M 2022 Algorithmic bias amplification via temporal effects: the case of pagerank in evolving networks Commun. Nonlinear Sci. Numer. Simul. 104 106029
[29] Shang K-K, Yan W-S and Small M 2016 Evolving networks—using past structure to predict the future Physica A 455 120–35
[30] İsmail G, Gündüz-Öğüdicü Ş and Çataltepe Z 2016 Link prediction using time series of neighborhood-based node similarity scores Data Mining and Knowledge Discovery 30 147–80
[31] Wang T, He X-S, Zhou M-Y and Fu Z-Q 2017 Link prediction in evolving networks based on popularity of nodes Sci. Rep. 7 1–10
[32] Chi K, Yin G, Dong Y and Dong H 2019 Link prediction in dynamic networks based on the attraction force between nodes Knowl.-Based Syst. 181 104792
[33] Bürün E and Kaya M 2019 Predicting citation count of scientists as a link prediction problem IEEE Trans. Cybern. 50 4518–29
[34] Linyuan L and Zhou T 2011 Link prediction in complex networks: a survey Physica A 390 1150–70
[35] Zhou T, Lee Y-L and Wang G 2021 Experimental analyses on two-hop-based and three-hop-based link prediction algorithms Physica A 564 125532
[36] Ran Y, Liu T, Jia T and Xu X-K 2021 A novel similarity measure for mining missing links in long-path networks Chin. Phys. B (https://doi.org/10.1088/1674-1056/ac4483)
[37] Xu R-Q, Zhou M-Y and Liao H 2021 Par: how to optimally combine different link prediction approaches? Inf. Sci. 584 342–59
[38] Clauset A, Newman M E J and Moore C 2004 Finding community structure in very large networks Phys. Rev. E 70 066111
[39] Bieł D M, Ng A Y and Jordan M I 2003 Latent Dirichlet allocation J. Mach. Learn. Res. 3 993–1022
[40] Wang D, Pedreschi D, Song C, Giannotti F and Barabási Á-L 2011 Human mobility, social ties, and link prediction Proc. 17th ACM SIGKDD Int. Conf. Knowledge discovery and Data Mining pp 1100–8
[41] Sun J 2012 Jieba Chinese word segmentation tool (https://github.com/fxsjy/jieba)
[42] Rehurek R and Sojka P 2010 Software framework for topic modelling with large corpora Proc. LREC 2010 Workshop on New Challenges for NLP Frameworks (ELRA) pp 45–50
[43] Cao R-M, Liu S-Y and Xu X-K 2019 Network embedding for link prediction: the pitfall and improvement Chaos 29 103102
[44] Zhou T, Lü L and Zhang Y-C 2009 Predicting missing links via local information Europhys. Lett. 89 18001
[45] Newman M E J 2001 Clustering and preferential attachment in growing networks Phys. Rev. E 64 025102
[46] Adamic L A and Adar E 2003 Friends and neighbors on the web Soc. Netw. 25 211–30
[47] Ghasebian A, Hosseinmardi H, Galstyan A, Airoldi E M and Clauset A 2020 Stacking models for nearly optimal link prediction in complex networks Proc. Natl Acad. Sci. USA 117 23393–400