Information Sciences 374 (2016) 1–14

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Sampling-based algorithm for link prediction in temporal networks

Nahla Mohamed Ahmed[a,b], Ling Chen[a,c,*], Yulong Wang[d], Bin Li[a,c], Yun Li[a], Wei Liu[a]

[a] College of Information Engineering, Yangzhou University, Yangzhou China, 225009
[b] College of Mathematical Sciences, Khartoum University, Khartoum, Sudan
[c] State Key Lab of Novel Software Tech, Nanjing University, Nanjing China, 210093
[d] College of Agriculture, Yangzhou University, Yangzhou China, 225009

ARTICLE INFO

Article history:
Received 18 February 2016
Revised 8 September 2016
Accepted 11 September 2016
Available online 12 September 2016

Keywords:
Link prediction
Random walk
Sampling
Temporal network

ABSTRACT

The problem of link prediction in temporal networks has attracted considerable recent attention from various domains, such as sociology, anthropology, information science, and computer science. In this paper, we propose a fast similarity-based method to predict the potential links in temporal networks. In this method, we first combine the snapshots of the temporal network into a weighted graph. A proper damping factor is used to assign greater importance to more recent snapshots. Then, we construct a sub-graph centered at each node in the weighted graph by a random walk from the node. The sub-graph constructed consists of a set of paths starting from the given node. Because the similarity score is computed within such small sub-graphs centered at each node, the algorithm can greatly reduce the computation time. By choosing a proper number of sampled paths, we can restrict the error of the estimated similarities within a given threshold. While other random walk-based algorithms require $O(n^3)$ time for a network with n nodes, the computation time of our algorithm is $O(n^2)$, which is the lowest time complexity of a similarity-based link prediction algorithm. Moreover, because the proposed method integrates temporal and global topological information in the network, it can yield more accurate results. The experimental results on real networks show that our algorithm demonstrates the best or comparable quality results in less time than other methods.

1. Introduction

Networks can naturally describe various social structures. In such networks, vertices denote individuals, while the edges represent relations among the individuals, such as corporations or companionship. Social network analysis has drawn increasing attention in the fields of sociology, computer science, and physics. It analyzes and explores the potential relations between social objects. Recently, complex network analysis has also drawn much attention in many commerce fields, such as e-business analysis and market modeling.

One of the most important research areas in network analysis is link prediction [46]. The objective of link prediction is to forecast prospective links from existing topological information of the network or identify unobserved links from the

* Corresponding author. Fax: +8651487887937.
E-mail addresses: nahlaibrahim@uofk.edu (N.M. Ahmed), yzulchen@163.com (L. Chen).

http://dx.doi.org/10.1016/j.ins.2016.09.029
0020-0255/

existing network structure. Link prediction is exploited to identify and categorize the human behavior and activity [1] in social networks. Link prediction can be applied to detect criminals and terrorists via their secret contacts [29] in social security networks. Link prediction is also employed to analyze the trend of changes in sensor networks [48], to perform web searches in the World Wide Web [17], to obtain the best possible routing [18], and to guarantee the confidentiality of information transmission [27]. In recent years, bipartite link prediction has been widely applied in areas such as recommendation [32,42,47], scientist-paper cooperation analysis [24], scientific paper impact prediction [22], medical parameter network analysis [19,20], and protein interaction prediction [13].

As relations between individuals in networks vary dynamically, links in social networks are continuously changing. Old links will possibly disappear from the network, while new ones may emerge constantly. For example, email communications between friends, transactions between businesses, and partnerships between scientific researchers change over time. Thus, the link prediction methods must be able to detect the changes of relationships among individuals in a dynamic network. In recent years, many methods for identifying latent or prospective links in dynamic networks have been proposed.

Similarity-based methods are the most common approach used for link prediction. In such methods, each pair of node is associated with an index to indicate the similarity between the corresponding nodes. This similarity quantifies the likelihood of link existence in the graph. Some essential attributes of the nodes can be used to define their similarity, such as common features or topological structures between the nodes [44]. Many studies in social networks have shown that a higher similarity may exist between individuals who are close to each other [2,11]. Structural similarity indexes are often used in popular similarity-based methods. There are three categories of similarity indexes: local indexes, quasi-local indexes, and global indexes. To calculate local indexes, only the neighbor information of each node is required. Such local indexes include Common Neighbors, Jaccard, Salton, Sorensen, Preferential Attachment, Hub Depressed, Hub Promoted, Adamic–Adar, Resource Allocation, and Leicht–Holme–Newman (LHN1) indexes [29]. Quasi-local indexes require more structural information than local indexes and less information than global indexes. Quasi-local indexes include Local Random Walk, Superposed Random Walk [26], and Local Path Index [28,50]. Global indexes require comprehensive information for link prediction tasks. They use global topological information of networks, such as the Katz, Matrix Forest Index (MFI), and Leicht–Holme–Newman (LHN2) indexes [29]. In general, the use of global indexes can yield higher-quality prediction results than quasi-local and local indexes. However, local indexes require less information than global ones. Another class of similarity-based methods is random walk methods. Those methods include SimRank [3], Random Walk with Restart, Cos+, and Average Commute Time [29]. Based on random walk, B. Chen et al. [9] presented an algorithm for predicting links to nodes of interest. The method first constructs a subgraph centered at the node the user is interested in. Then, it computes the similarity score in the subgraph.

Some methods predict potential links by exploring the structural characteristics of the network. Purnamrita et al. [36] presented a nonparametric method for link prediction in temporal networks. This method partitions the time domain into subsequences represented by graph snapshots. Their method predicts connections between nodes based on their topological features and local neighbors. Kim et al. [21] proposed an approach to identify potential links in networks. Their approach is based on node centrality, which can predict the future importance of nodes. Murata et al. [30] investigated the relationship between graph structure and link occurrence. They advanced a weighted proximity-based method for link prediction in social networks.

For temporal network link prediction, some methods employ machine-learning techniques. Vu et al. [45] proposed a continuous-time regression model that can integrate time-varying regression coefficients and time-dependent network statistics. Pujari et al. [35] used a supervised rank aggregation approach to predict potential links in temporal networks. Zeng et al. [49] presented a method using semi-supervised learning. To predict potential links in a network, their method exploits the latent information of the node pairs that are not currently linked. He et al. [15] advanced a link prediction ensemble algorithm using an ordered weighted averaging operator. The algorithm assigns weights for nine local information-based link prediction algorithms and then aggregates their results to obtain the final prediction scores. Bao et al. [4] advanced a network link predictor using principal component analysis to identify features that are important to link prediction. Madadhain et al. [33] proposed an event-based link prediction approach on temporal networks. By applying machine-learning and data-mining techniques, their approach is able to forecast potential cooperation between individuals in social events. Using data-mining techniques such as frequent-pattern and association-rule mining, Bringmann et al. [7] advanced a method for link prediction in temporal networks. To avoid the high computational cost of optimization in machine-learning methods, some heuristic methods are employed in link prediction. Catherine et al. [6] presented a method for predicting future links by applying the covariance matrix adaptation evolution strategy. Based on ant colony optimization, Sherkat et al. [43] introduced an unsupervised link prediction algorithm.

Probabilistic model-based methods are also used in link prediction on complex networks. Hu et al. [16] presented a probabilistic model to discover individual actions in social networks. They also proposed an approach employing a genetic algorithm to optimize the model. Barbieri et al. [5] proposed a stochastic link prediction model on directed graphs with node attribute features. In addition to predicting links, the model also provides explanations for the links detected. To estimate the probability of a link appearance, Gao et al. [12] presented a model that exploits various types of information in the temporal network. For link prediction in a user-object network, Ji Liu et al. [25] proposed an approach that takes both time attenuation and diversion delay into consideration. By extending the exponential random graph model, Hanneke et al. [14] advanced a set of statistical models for dynamic network link prediction. Because the probabilistic model proposed needs to know the distribution of link occurrence, it is impractical for link prediction in a real-world network.