Physica A 428 (2015) 470–480

Contents lists available at ScienceDirect

Physica A

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

Predicting the structural evolution of networks by applying multivariate time series

Qiangjuan Huang [a], Chengli Zhao [a], Xiaojie Wang [a], Xue Zhang [a], Dongyun Yi [b,a,*]

[a] School of Science, National University of Defense Technology, Changsha, Hunan, 410073, China

[b] State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha, Hunan 410073, China

HIGHLIGHTS

- We propose a method for link prediction in time-varying networks.
- We combined time series methods with traditional indexes to improve prediction accuracy.
- We discuss the importance of temporal information and topological information for dynamic network analysis.

ARTICLE INFO

Article history:
Received 11 September 2014
Received in revised form 4 December 2014
Available online 17 February 2015

Keywords:
Time-varying network
Link prediction
Topological structure
Time-series analysis
Multivariate analysis

ABSTRACT

In practice, complex systems often change over time, and the temporal characteristics of a complex network make their behavior difficult to predict. Traditional link prediction methods based on structural similarity are good for mining underlying information from static networks, but do not always capture the temporal relevance of dynamic networks. However, time series analysis is an effective tool for examining dynamic evolution. In this paper, we combine link prediction with multivariate time series analysis to describe the structural evolution of dynamic networks using both temporal information and structure information. An empirical analysis demonstrates the effectiveness of our method in predicting undiscovered linkages in two classic networks.

1. Introduction

Link prediction is an important bridge between complex networks and information science, addressing basic issues such as restoring and predicting missing information. In complex networks, link prediction explores the existence of linkages using the attribute information of nodes and the topological structure of the network. The diversity of applications means that link prediction has attracted considerable attention from different areas. In biological research, predicting the interaction of proteins can improve an experiment's efficiency and reduce its cost. In social networks, the problem can be used to establish a relationship between strangers, such as the likelihood of two scientists collaborating on a paper, while in e-commerce, predicting links between product recommendations can help a website to attract more customers.

Traditional link prediction algorithms can be divided into two categories. The first type uses attribute information from each node. These include probabilistic relational models [1,2], Markov relation networks [3], structural logistic

* Corresponding author at: State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha, Hunan 410073, China.

E-mail address: 350149321@qq.com (D. Yi).

http://dx.doi.org/10.1016/j.physa.2015.02.019

regression models [4], and random relational models [5]. With sufficient information, these methods achieve precise predictions. However, in real life, it is often difficult to attain the necessary information. Therefore, techniques based on the topological network structure are often used. For example, Libel-Nowell and Kleinberg proposed a method based on similarity information about the nodes, paths, clusters, and so on. Compared to node attributes, structure information is more accessible.

The link prediction algorithms mentioned above were established on static networks, which ignore the time-varying properties of complex systems. To study link prediction in dynamic networks, they mostly just consider changes in one attribute over time, such as the network's density or diameter [6]. In 2009, Huang and Lin proposed a hybrid method that combined link prediction based on structural similarity with time series analysis [7]. This used both the time-varying information and structure information to improve prediction accuracy. Soares and Prudencio proposed a time series analysis approach based on link prediction indexes related to the efficiency of prediction [8]. However, these methods cannot reflect changes in the relationship between the nodes. These linkages have correlative relationships, so they cannot be considered individually. In this paper, we consider both the horizontal structure information and the vertical time-varying information, which means that changes in one edge are affected by its past evolution as well as the relative linkage evolution. Thus, we employ the theory of multivariate time series analysis to the study of temporal and structural correlation.

In the next section, we review several common link prediction models based on structure similarity in a static network. We then apply time series analysis to time-varying networks in Section 3, and propose a multivariate time series analysis method to study time-varying characteristics and structural correlation in Section 4. In Section 5, we examine two classic networks, the Enron email network and the High Energy Particle Physics co-authorship network, to verify the effectiveness and precision of our method. Finally, in Sections 6 and 7, we summarize the results of this paper and identify the details needed for further discussions.

2. Link prediction based on structure similarity

An important premise of link prediction based on structure similarity is that a linkage between two nodes becomes more likely to exist as the similarity of these nodes increases [9]. In this paper, we will introduce five types of link prediction index, which can be divided into two classes [10]. The first, based on local structure information, includes the Common Neighbor Index (CN), Adamic-Adar Index (AA), Resource Allocation Index (RA), and LHN-I index (LH). The second class utilizes global information, and includes the Katz Index.

We must first define an undirected network $G = (V, E)$, where V is the set of nodes and E is the set of edges. In a time-varying network, we have a sequence of network structures from time $t = 1 \dots T$, i.e., $(G_1, G_2, \dots, G_T)$. Accordingly, we obtain a sequence of network adjacency matrices $(M_1, M_2, \dots, M_T)$. We denote the number of linkages between node v_i and node v_j at time t as $M_t(i, j)$.

As we are concerned with whether a linkage will occur at the next point in time, rather than the linkage weights, we perform some pretreatment on the data. We first reduce the graph series $(M_1, M_2, \dots, M_T)$ to a single, weighted graph $M_{1\sim T}$, where $M_{1\sim T}(i, j) = \sum_{t=1}^{T} M_t$. Then, we simplify the adjacency matrix to a simple matrix $S_{1\sim T}$, where $S_{1\sim T}(i, j) = 1$ if $M_{1\sim T}(i, j) > 1$, and $S_{1\sim T}(i, j) = 0$ otherwise. We are then able to predict whether a linkage will exist at time $T + 1$ using the indexes mentioned above.

2.1. Similarity indexes that use local information

(1) CN Index [11]: this takes the hypothesis that two nodes that have many common neighbors are more likely to be similar. For example, in a social network, if two strangers have many friends in common, there is a greater chance that they will be friends [12]. The index is defined as:

$$S(i, j) = \|\Gamma(i) \cap \Gamma(j)\|, \tag{1}$$

where $\Gamma(i)$ and $\Gamma(j)$ represent the set of common neighbor nodes of nodes v_i and v_j, respectively.

(2) LHN-I Index [13]: this index is based on a similar principle to CN, and is defined as:

$$S(i, j) = \frac{\|\Gamma(i) \cap \Gamma(j)\|}{d(i) \cdot d(j)}, \tag{2}$$

where $d(i)$ and $d(j)$ indicate the degree of nodes v_i and v_j, respectively.

(3) AA Index [14]: the idea of this index is that nodes of small degree make a greater contribution than those of large degree. For example, in terms of micro-blogging, the people who receive the most attention tend to be experts or celebrities in a certain field, but this does not mean that their followers have similar interests. Thus, AA assigns a weight to every common neighbor node that is directly proportional to the logarithm of the node's degree. The definition is:

$$S(i, j) = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log d(z)}. \tag{3}$$