Journal of Computational Science 16 (2016) 51–58

Contents lists available at ScienceDirect

Journal of Computational Science

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

Detecting the missing links in social networks based on utility analysis

Luo Peng[a,*], Li Yongli[b,*], Wu Chong[a], Chen Kun[c]

[a] Harbin Institute of Technology, Harbin 150001, China
[b] Northeastern University, Shenyang 110819, China, China
[c] South University of Science and Technology of China, Shenzhen 518055, China

ARTICLE INFO

Article history:
Received 19 September 2015
Received in revised form 31 March 2016
Accepted 28 April 2016
Available online 29 April 2016

Keywords:
Social network
Link detection
Logistic regression
Network analysis
Utility function

ABSTRACT

This paper proposes a new model for detecting missing links in social networks. A utility function is introduced that considers the node attributes as well as the network structure for the individuals to decide whether to form a link. At the same time, logistic regression is also adopted to estimate the parameters of the algorithm based on the observed network. Furthermore, this paper validates this new missing link detection method in online social networks that were established from Facebook via comparison analysis. The results demonstrate that our method outperforms other algorithms in detecting the existent links in the original network. We also perform scalability analysis with respect to our method, analyze the complexity of method and attempt to reduce our method's complexity by deleting some of the parameters. Moreover, this study also applies our method to network evolution analysis, and it enables us to uncover the factors that promote network evolution.

1. Introduction

Network modeling has become a widely used tool for analyzing complex systems, and the study of networks has impacted many different scientific fields [1]. By the aid of network analysis, many social, information, and economic systems can be described as network models, in which nodes denote individuals, computers or users, and connections represent the relationships or interactions among them. Considering the large amount of effort that has been applied to studying networks, such as in community identification [2,3], network evolution [4,5], network sampling [6,7], and other problems, link detection is one of the most fundamental problems; in link detection, we attempt to detect those links that are missing in observed networks [8].

Link detection is applicable to a wide variety of application. In biological networks, it can be used to detect protein–protein interactions and to predict the possible missing relationships between proteins, which could reduce the high cost of laboratory experiments [9]. In online social networks, link detection helps to detect highly likely but non-existent links, which can be used as friend recommendations and to enhance the user's loyalty to the web site [10,11]. In E-commerce, one of its applications is to build commodity recommendation systems that enable companies to increase their profits and the satisfaction of the consumers [12]. In a collaboration network, link detection can investigate promising co-authors for the purpose of completing great scientific research [13].

In social networks, the network participants are people, and thus, the characteristics of people should be considered when making models of the link detections or the network's evolution. Accordingly, we define a utility function that considers a node's characteristics as well as the influence of its friends, for the network nodes to decide whether to form a link. In addition, logistic regression is applied to estimate the parameters of the algorithm because the structure of link detection in this paper can be handled via traditional logistic regression. Then, this paper further compares the proposed link detection method with the existing methods as applied to five social networks, and the results show that our method can be much more effective in detecting the missing links that are originally existent and not existent in the sampled network. Finally, we apply our proposed link detection method in the analysis of network evolution.

This paper is organized as follows. Section 2 reviews the related work in brief. Section 3 presents our research methodology including the utility function, the research model and the parameter estimation method. In Section 4, numerical experiments are designed to uncover our method's properties via comparison analysis, and an application is presented to explain network evolution. Section 5 provides conclusions and discusses possible future work.

* Corresponding authors.
E-mail addresses: luopeng_hit@126.com (P. Luo), ylli@mail.neu.edu.cn (Y. Li).

http://dx.doi.org/10.1016/j.jocs.2016.04.016
1877-7503/© 2016 Elsevier B.V. All rights reserved.

2. Related work

Considering an undirected network $G(V, E)$, V denotes the set of nodes, and E denotes the set of links, where self-links and multiple links are not allowed [14]. There are $|V|(|V|-1)/2-|E|$ links that do exist but are not currently detected or are nonexistent, where $|V|$ and $|E|$ are the number of nodes and links, respectively. The goal of our study is to determine the undetected links.

The widely used link detection methods are similarity-based algorithms. As introduced in Lu and Zhou [1], the algorithms give each pair of nodes that do not have a link between them, with a score defined as the similarity measure between them. Those pairs of nodes that have higher scores are expected to have higher likelihoods of connecting to each other. First, we introduce the Common Neighbors (CN) method. This method considers that two nodes (x, y) are more likely to form a link if they have more common neighbors. The algorithm is written as

$$s_{xy}^{CN} = \left|\Gamma(x) \cap \Gamma(y)\right|, \tag{1}$$

where $\Gamma(x)$ $(\Gamma(y))$ are the set of neighbors of node x (y). Moreover, based on the CN similarity-based algorithm, many other methods are proposed, such as the Salton Index [15], Jaccard Index [16], Hub Promoted Index [17], Resource Allocation Index [18] and others. Among these indexes, the Jaccard Index (JI) is defined as

$$s_{xy}^{JI} = \frac{\left|\Gamma(x) \cap \Gamma(y)\right|}{\left|\Gamma(x) \cup \Gamma(y)\right|}, \tag{2}$$

Unlike the above algorithms, which are based on the network's local structure, some other similarity-based algorithms are defined based on the global network structures. For example, the Katz Index (KI) [19] considers the ensemble of all paths; specifically, it sums over all of the paths and provides a free parameter β to control the path weights:

$$s_{xy}^{KI} = \sum_{l=1}^{\infty} \beta^l \left|\mathrm{path}_{xy}^{\{l\}}\right| = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \cdots, \tag{3}$$

where $path_{xy}^{\{l\}}$ denotes the set of paths that have length l between x and y.

Similarly, the SimRank (SR) algorithm was first researched by Jeh and Widom [20], and it assumed that two nodes were similar if they were connected to similar nodes:

$$s_{xy}^{SR} = c \times \frac{\sum_{z \in \Gamma(x)} \sum_{z' \in \Gamma(y)} s_{zz'}^{SR}}{k_x k_y} \tag{4}$$

where $s_{xx} = 1$, k_x or k_y is the degree of node x or y and $c \in [0, 1]$ is the decay factor. Except for these two measures, the Leicht- Holme-Newman Index [21], Average Commute Time [22], Cosine based on L^+ [23], and Random Walk with Restart [24] have been proposed, with the idea that these indexes ask for all of the topological information rather than only local similarity algorithms.

Apart from the above methods, the probabilistic model is another stream of development for link detection methods. For example, the hierarchical structure model was proposed by Clauset et al. [25] to predict the missing links via maximum likelihood estimation. As studied in White et al. [26] and Holland et al. [27], the stochastic block model was also discovered, where the nodes are grouped into different categories, and the probability of forming a link depends on the categories that they belong to. The other examples include the Probabilistic Relational Model [28], Probabilistic Entity Relationship Model [29], and Stochastic Relational Model [30]. The common aspect of these mentioned models is that they all predict the missing links based on machine learning techniques.

Furthermore, some methods have been recently proposed that aim at social network link detection. For example, Liu et al. [31] studied a hidden link detecting method (SLD1, for short hereafter) by considering social characteristics, which is similar to our method. Bai et al. [32] developed a similarity index by combining the resource allocation index and local path index, and the method performed well in social network link detecting (SLD2). Some of the above mentioned methods will be benchmarks for a comparison with the proposed method in this paper.

3. Methodology

3.1. Utility function

In many Econometric papers in the literature [33–35], the network participants are treated as individuals who have limited rationality. Thus, the links between them (nodes) are formed if both individuals in a pair view their links as beneficial. This approach is based on the strategic network formation model [36,37], which focuses on network formation based on individual choices that are motivated by utility maximization. For example, Christakis et al. [38] studied an empirical model for strategic network formation that is also based on the assumption that the participants make link formations according to their utility. Jackson and Watts [39] also examined the dynamic formation and stochastic evolution of networks, which also held that the formation of connections was based on the payoff of individuals. The foundation of strategic network formation theory is that the participants will receive feedback from the link formations. Following these approaches in the literature, we can design a utility function by considering the current state of the network (denoted by G_t) and the characteristics of all of the individuals (denoted by a matrix C). Each row in the matrix C, such as C_i, denotes the characteristics of an individual i. G_t includes the set of network nodes (V_t) and the set of network links (E_t) in the time period t, and it can be denoted by $G_t = (V_t, E_t)$. We also use M to denote the adjacency matrix of the network. The utility function of individual i can be written as

$$U_i(M, C). \tag{5}$$

Then, when there is no link between individual i and j in the state of the network G_t, the net utility of individual i forming a link with individual j at time t is

$$\Delta U_{i \to j}(M, C) = U_i(M + \{m_{ij} = 1\}, C) - U_i(M, C), \tag{6}$$

where $m_{ij} = 1$ means that the two individuals (i, j) form a link at time t. Likewise, the net utility of individual j $(\Delta U_{j \to i}(M, C, t))$ can also be obtained. If the individuals only consider their own net utility, then they will establish a link if and only if

$$\Delta U_{i \to j}(M, C, t) > 0\text{and}\Delta U_{j \to i}(\mathrm{M, C, t}) > 0; \tag{7}$$

Otherwise, they will not form the link.

3.2. Research model

In social networks, as Christakis and Fowler [40] proposed, the Rule of "Three Degrees of Influence" means that everything we do or say tends to ripple through our social network, impacting our friends (one degree), our friends' friends (two degrees), and even our friends' friends' friends (three degrees). Based on the people's behavior in social networks, many researchers have explored the factors that influence friendship formation and the participants' behaviors. For example, Katona et al. [41] modeled the adoption decision as a binary choice that is influenced by neighbors'