Physica A 393 (2014) 553–559

Contents lists available at ScienceDirect

Physica A

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

Degree-corrected stochastic block models and reliability in networks

Xue Zhang, Xiaojie Wang, Chengli Zhao, Dongyun Yi, Zheng Xie *

Department of Mathematics and Systems Science, College of Science, National University of Defense Technology, Changsha, 410073, China

HIGHLIGHTS

- A link prediction method based on a degree-corrected stochastic block model is proposed.
- The method could be used in networks containing multi-links and self-links.
- It outperforms the method based on a stochastic block model in predicting missing links.

ARTICLE INFO

Article history:
Received 25 April 2012
Received in revised form 15 March 2013
Available online 4 September 2013

Keywords:
Stochastic block models
Complex networks
Link reliability
Bayesian estimation

ABSTRACT

Plenty of algorithms for link prediction have been proposed to extract missing information, identify spurious interactions, reconstruct networks, and so on. Stochastic block models are one of the most accurate methods among all of them. However, this algorithm is designed only for simple graphs and ignores the variation in node degree which is typically displayed in real-world networks. In this paper, we propose a corresponding reliable approach based on degree-corrected stochastic block models, which could be applied in networks containing both multi-edges and self-edges. Empirical comparison on five disparate networks shows that the overall performance of our method is better than the original version in predicting missing links, especially for the interactions between high-degree nodes.

1. Introduction

Many complex systems in the real world such as social, biological, and information systems can be naturally described as networks, with nodes representing entities (individuals, proteins, web pages, and so on) and links denoting relations or interactions between them. The study of complex networks has therefore become a common focus of many branches of science. While making great efforts to understand the structural features and evolutionary mechanism of networks, scientists gradually realized that the inaccuracy and incompleteness of data sets is a significant obstacle to the research [1,2]. To address this issue, link prediction algorithms have been adopted to extract the missing information, identify spurious interactions, reconstruct networks, and so on [3].

The natural framework of link prediction methods is the similarity-based algorithm [4]. Simple similarity-based measures such as neighborhood-based measures, e.g., Common Neighbors (CN) [5], Jaccard index [6], Adamic–Adar index (AA) [7] and Resource Allocation index (RA) [8], only require consideration of the local structure of the networks. These local similarity indices are widely used in practice because of their low computation cost and relatively good accuracy. Other similarity-based measures such as path-based methods, e.g., Katz index [9] and Random Walk with Restart (RWR) [10], focusing on the globe structure of the networks, are more effective but have a high computational complexity. Recently, some variants

* Corresponding author. Tel.: +86 158 7495 0863.
E-mail address: leonzhengxie@gmail.com (Z. Xie).

0378-4371/$ – see front matter
http://dx.doi.org/10.1016/j.physa.2013.08.061

of local similarity indices are proposed. These methods argue that different common neighbors may play different roles and contribute differently. Thus, each common neighbor is treated differently according to its degree centrality, closeness centrality and betweenness centrality in these algorithms [11,12]. Meanwhile, the weights of links are also taken into account to estimate the likelihood of the existence of links in weighted networks, and it surprisingly turns out that the well-known Weak-Ties Theory also takes a role in the link prediction field [13].

An algorithm based on Bayesian estimation is another direction in the study of link prediction. Inspired by empirical studies that many real-world networks exhibit hierarchical organizations, Clauset, Moore and Newman [14] present a method inferring hierarchical structure from network data and use the knowledge of hierarchical structure to predict the missing connections in partly known networks. Guimerà and Sales-Pardo [15] suppose that the observed network is an implementation of the stochastic block model (SBM), in which nodes are partitioned into groups and the probability that two nodes are connected depends only on the groups to which they belong. Under this assumption, they build a mathematical and computational framework to reliably identify both missing and spurious interactions in noisy network observations. The SBM method outperforms other link prediction algorithms on a variety of networks, yet it ignores heterogeneity in the degree of nodes, which is typically displayed in real-world networks and may influence the prediction accuracy somehow. So far as we know when the stochastic block model is used as a tool for detecting community structure in networks, the degree-corrected version dramatically outperforms the uncorrected one in both real-world and synthetic networks [16]. However, whether it still has an advantage in the link prediction area is still unclear. In this paper, we make use of the degree-corrected stochastic block model (DCSBM) to assess the reliability of network links, conduct experiments on several real-world networks and compare it with other two algorithms—SBM, the original version, and CN which is regarded as a benchmark. The experimental results show that the proposed algorithm unsurprisingly overmatches CN and works even better than SBM in identifying the missing interactions, especially at finding the connections between high-degree nodes.

The rest of the paper is organized as follows. In Section 2, we review the theoretical framework of Bayesian estimation for link prediction, and propose our algorithm. Experimental results on real-world networks are presented in Section 3. Finally, we draw our conclusions in Section 4.

2. The method

2.1. General reliability formalism

Consider an observed undirected network, including multi-edges and self-edges, whose adjacency matrix A^O is conventionally defined as follows. A^O_{ij} is equal to the number of edges between node i and j when $i \neq j$, but the diagonal element A^O_{ii} is equal to twice the self-edges from i to itself.

We assume that this observed network is a realization of an underlying probabilistic model [17]. Let $\mathcal{M}$ be the set of generative models that could conceivably give rise to the observed network, and $p(M|A^o)$ the probability that $M \in \mathcal{M}$ is the model that gave rise to the observation A^O. The probability $p(X = x)$ and expected value $\omega(X)$ for an arbitrary network property X are

$$p(X = x|A^O) = \int_{\mathcal{M}} dM p(X = x|M)p(M|A^O) \tag{1}$$

$$\omega(X|A^O) = \int_{\mathcal{M}} dM \omega(X|M)p(M|A^O), \tag{2}$$

where $p(X = x|M)$ is the probability that $X = x$ in a network generated by model M, $\omega(X|M)$ is the corresponding expected value of property X. Using the Bayes theorem, Eqs. (1) and (2) can be rewritten as

$$p(X = x|A^O) = \frac{\int_{\mathcal{M}} dM p(X = x|M)p(A^O|M)p(M)}{\int_{\mathcal{M}} dM' p(A^O|M')p(M')} \tag{3}$$

$$\omega(X|A^O) = \frac{\int_{\mathcal{M}} dM \omega(X|M)p(A^O|M)p(M)}{\int_{\mathcal{M}} dM' p(A^O|M')p(M')}, \tag{4}$$

where $p(A^O|M)$ is the probability that model M gives rise to A^O among all possible adjacency matrices, and $p(M)$ is the *priori* probability that model M is the correct one. $p(X = x|A^O)$ is called the reliability of the $X = x$ measurement and $\omega(X|A^O) = \int_x x p(X = x|A^O)\mathrm{d}x$ is the expected value of X.

2.2. Degree-corrected stochastic block model

From an information-theoretic viewpoint, an edge between two high-degree vertices is less surprising than an edge between two low-degree vertices. The BA model [18] which is well-known for mimicking the evolution of real networked systems is also biased to the nodes with more connections. Such empirical knowledge has been incorporated into the degree-corrected stochastic block model [16].