Information Processing and Management 000 (2016) 1–13

Contents lists available at ScienceDirect

Information Processing and Management

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

A balanced modularity maximization link prediction model in social networks

Jiehua Wu[a,b,*], Guoji Zhang[a], Yazhou Ren[c]

[a] School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
[b] Department of Computing, Guangdong Polytechnic of Industry and Commerce, Guangzhou, China
[c] School of Computer Science and Engineering, Big Data Research Center, University of Electronic Science and Technology of China, Chengdu, China

ARTICLE INFO

Article history:
Received 24 June 2015
Revised 30 September 2016
Accepted 5 October 2016
Available online xxx

Keywords:
Link prediction
Social network
Community detection
Modularity

ABSTRACT

Link prediction has been becoming an important research topic due to the rapid growth of social networks. Community-based link prediction methods are proposed to incorporate community information in order to achieve accurate prediction. However, the performance of such methods is sensitive to the selection of community detection algorithms, and they also fail to capture the correlation between link formulation and community evolution. In this paper we introduce a balanced Modularity-Maximization Link Prediction (MMLP) model to address this issue. The idea of MMLP is to integrate the formulation of two types of links into a partitioned network generative model. We proposed a probabilistic algorithm to emphasize the role of innerLinks, which correspondingly maximizes the network modularity. Then, a trade-off technique is designed to maintain the network in a stable state of equilibrium. We also present an effective feature aggregation method by exploring two variations of network features. Our proposed method can overcome the limit of several community-based methods and the extensive experimental results on both synthetic and real-world benchmark data demonstrate its effectiveness and robustness.

1. Introduction

Link prediction is a fundamental data mining and machine learning technique for studying and understanding social networks. It aims to predict the presence and absence of links between disconnected nodes, which benefits many online applications such as friend recommendation (Hopcroft, Lou, & Tang, 2011) and collaboration discovery (Tang, Wu, Sun, & Su, 2012). In recent years, many link prediction algorithms have been proposed. The simplest framework is to assign a similarity score for each pair of nodes based on network structure (e.g. feature). Such idea leads to many kinds of similarity-based methods such as node-based link prediction (e.g., Common Neighbors (CN), Adamic Adar(AA) (Liben-Nowell & Kleinberg, 2007)), path-based link prediction (e.g. Katz (Liu & Lü, 2010), Local Path (Lü & Zhou, 2011)), and random walk-based link prediction (e.g., Superposed Random Walk, Local Random Walk (Liu & Lü, 2010)).

However, the mathematical definition of most existing similarity-based methods is simple. It is not sufficient enough to represent the deep knowledge of social network. Therefore, some novel methods (Cannistraci, Alanis-Lobato, & Ravasi, 2012; Soundarajan & Hopcroft, 2012; Yan & Gregory, 2012) tried to deal with this problem by providing community informa-

* Corresponding author.
E-mail address: wu.jiehua@mail.scut.edu.cn (J. Wu).

http://dx.doi.org/10.1016/j.ipm.2016.10.001
0306-4573/

tion. For example, Yan and Gregory (2012) have introduced an Enhanced Community-based Link Prediction (ECLP) method to supply node-based local predictors (e.g., CN, AA etc.) with community membership information. Valverde-Rebaza and de Andrade Lopes (2013) defined two new community-based features: the overlapping groups degree and the overlapping groups clustering coefficient, and presented a probabilistic naive Bayes link prediction model. Although the benefit of utilizing community information in addition to similarity-based predictors is obvious, community-based link prediction methods have the drawback of being sensitive to the choice of community detection algorithms, and more important, the key idea of those methods is to simply merge community property with the classic predictors, rather than to incorporate such property with the link prediction problem to build a generative framework. There is also an attempt to use generative relational model from a block or a membership view for the predictive task. These works Kemp, Tenenbaum, Griffiths, Yamada, and Ueda (2006) and Schölkopf, Platt, and Hofmann (2006), however, also fail to capture the correlation between link formulation and community detection to reflect the network dynamic in partitioned network. Specifically, nodes in a community are more densely connected than average, which means links tend to occur inside communities to formulate a more dense structure (Liu, He, Kapoor, & Srivastava, 2013; Yan & Gregory, 2012). For example, a movie social network is partitioned to several interest groups (communities). In each group, users maintain close relationships and share the same interest with each other. It is evidenced that missing relations or potential friends connections are more likely to be found within groups than elsewhere. This motivates us to ask a research question that whether such community detection theory can be adapted to link prediction scenario to reflect the network evolution.

To address this question, it is noteworthy that community (Newman, 2012) is a partitioned group (cluster, block) of nodes formulating a sub network within which links are dense, but between which links are sparse. Thus links in such partitioned network can be classified into two kinds (Granovetter, 1973): intra-community links (intraLinks) and inter-community links (interLinks). IntraLinks refers to connections occur inside a community (group) to form a closer relations while interLinks on the other hand mean a bridge between communities. In this way, there are two ways a link formulated: inside community or across communities. Then link prediction can be naturally transformed to the task of link formation inference problem. In order to infer link formulation, there are two fundamental challenges to be addressed. The first challenge is what we can learn from the community detection process. The second challenge lies in how to define the correlation between link formulation and community detection. The idea here is to incorporate two kinds of link formulation into a community network generative process to maximize modularity. Modularity is the number of links falling within communities minus the expected number in an equivalent network with links placed at random, which can be naturally incorporated in our link prediction framework. Particularly as mentioned above, if we formulate new (predicted) links to a given partitioned network, we should put a high premium on intraLinks, which correspondingly improves the whole network modularity. However, it is inadvisable to overemphasis interLinks, as interLinks also play an important role in information cascading and diffusion (Zhang, Wang, Yu, & Wang, 2013). To clearly address such question, we refer to a toy example illustrated in Fig. 1. The input of our study is a partitioned social network composed of users and user relations (links). The potential links (dash lines) lie either inside community or between communities. The predictive task not only tries to reveal whether links exist, but also attempt to figure out which kind of links really lie in the expected output. One natural method is to formulate more intraLinks (red lines). However, lack of interLinks (green lines) tend to cause broken communication among communities, which is not conform to the reality. And we need to make a trade-off between modularity maximization and network balance. To the best of our knowledge, no previous works have extensively studied such problem.

In this article, we propose a balanced Modularity-Maximization Link Prediction (MMLP) model to address the aforementioned limitations of community-based algorithms. MMLP borrows the idea of block model (Liu et al., 2013) and social circle model (McAuley & Leskovec, 2012, 2014), which are used to detect community. The mechanism of our framework is as follows: First, we cluster all nodes into communities. Second, we built a generative probabilistic model to put more emphasize on the formulation of intraLinks into the partitioned network. Next, we make the link probability as a function which associated with aggregative features and learn the feature weight in different communities. Then we predict the link formulation based on the learned parameter. The main contributions of this paper are summarized as follows:

- We formally formulate the problem of link prediction based on community modularity, and introduce a balanced Modularity-Maximization Link Prediction model (MMLP) to reveal the correlation between link prediction and community detection.
- We present a new perspective to investigate the diverse features role in different communities.
- Extensive experiments demonstrate the effectiveness and robustness of our proposed approach.

The outline of this papers is as follows: In Section 2 we review related works on link prediction and community detection. Section 3 presents our balanced modularity maximization generative model and the motivation for its parameterizations. Then we introduce the feature construction and inference procedure. After that we conduct extensive experiments on synthetic and real social networks in Section 4 followed with the conclusion in Section 5.

2. Related works

Link prediction is a social network analysis technique that has been widely used in various online services, such as user recommendation (Valverde-Rebaza & de Andrade Lopes, 2013) and information cascade inference (Zhang et al., 2013). Liben-Nowell and Kleinberg (2007) first addressed the link prediction problem with a ground breaking survey. They pointed out

Please cite this article as: J. Wu et al., A balanced modularity maximization link prediction model in social networks, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.10.001