Knowledge-Based Systems 87 (2015) 3–15

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Detecting hierarchical structure of community members in social networks

Fengjiao Chen, Kan Li *

Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, School of Computer, Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, Beijing, China

ARTICLE INFO

Article history:
Received 26 October 2014
Received in revised form 19 May 2015
Accepted 28 May 2015
Available online 3 June 2015

Keywords:
Social network
Community detection
Hierarchical structure
Random walk
Linear regression

ABSTRACT

Current methods often predefine fixed roles of members and only detect fixed hierarchy structures that are not consistent with real-world communities; methods with hand-crafted thresholds bring difficulties in real applications, while choosing the community corresponding to the maximal belonging coefficient for each node results in a single boundary and neglects the multi-resolution of communities. In order to solve the limitations above, we propose a novel structure to dig finer information by partitioning the members into several levels according to their belonging coefficients. We call this novel structure Hierarchical Structure of Members (HSM) and discuss its properties in continuity, comparability, consistency and stability which reveal the multi-resolution of community as well as the intra-relations among members. We propose a two-phrase method, Random Walk and Linear Regression (RWLR), to detect HSM. The method measures the belonging coefficients of members by random walk and then divides the members into multiple segments by linear regression. Experiments show that members in the same level hold the same properties and HSM reveals multi-resolution of community. Besides, the comparison in benchmarks shows the efficiency in community detection. Finally, we apply HSM to analyze social networks, including visualization of community structures in large social networks and interactive recommendations in Amazon network.

1. Introduction

Many social networks display communities, groups of vertices with a higher-than-average density of edges connecting them. Community structure is fundamental for uncovering the links between structure and function in complex networks and for practical applications in many disciplines [1,17]. Members in a community play different roles, such as cores, bridges and fringes, which demonstrates the inner hierarchy of a community. Core members have large influence on other members, while bridge members help to keep communication with outside. Since the boundary of a community in social network is often fuzzy [8], a community can have different sizes from different views. Modeling a community by multiple boundaries is more suitable for real social networks, and we call such model as multi-resolution of a community.

Existing community detection methods often predefine fixed roles of members, resulting in a fixed hierarchical structure of community members [16,22,26]. However, communities in the real world usually have a variety of structures rather than a specific one. Take the relation of employees inside a company as an example, the hierarchy structure reflects the leadership of the company. Some companies have chairmen, department managers and their staffs. While other companies have chairmen, department managers, team leaders and team members. The hierarchy structures are different. In addition, current work of detecting multi-resolution of a community mainly follows two trends. One is to set hand-crafted thresholds to filter out the multi-resolution of community [18]. The choice of thresholds is not determined by algorithms, which brings difficulties in real applications. The other one is to choose the community corresponding to the maximal belonging coefficient for each node, which is easy to result in a single boundary. In this case, we need to propose a new method to detect the roles of community members as well as the multi-resolution of community.

To address these problems, we propose a new concept, called hierarchical structure of members (HSM), which describes a community by a ‘level’ structure according to the belonging coefficients. The belonging coefficients reflect the strength of relation between a member and its community. Members with similar belonging coefficients are in the same level. In the first level,

* Corresponding author.
E-mail addresses: cfjmonkey@hotmail.com (F. Chen), likan@bit.edu.cn (K. Li).

http://dx.doi.org/10.1016/j.knosys.2015.05.026
0950-7051/

members have the highest belonging coefficients, while those in the last level have the lowest belonging coefficients. The former ones are regarded as core members and the latter ones are treated as marginal members. From the first level to the last level, one can form the multi-resolution of community from the seed community to the whole network.

HSM is different from the hierarchical structure of communities (HSC) which gains attentions in recent years [4,10,11,14,30,31]. HSC describes nested community structures and shows the relations of communities, while HSM describes the levels of members and shows the relations of nodes.

An ideal HSM detection approach should maintain continuity, comparability, consistency and stability at the same time (see Section 3), which becomes the most challenging part of this problem. Besides, the automatic determination of the number of levels requires consideration as well. In this paper, we propose the RWLR (Random Walk and Linear Regression) method to solve these problems. It measures the belonging coefficients by random walk from a seed community and divides the members into levels by linear regression on the sorted sequence of belonging coefficients. Our method achieves a good performance on the benchmark datasets as well as on the real-world networks, which demonstrates the usefulness of HSM.

The rest of the paper is organized as follows. Section 2 reviews some related work on belonging coefficients measurement and hierarchical community detection. Section 3 defines the problem in a more formal way and then illustrates the RWLR method with a community detection framework. Section 4 shows the experimental results and the statistical analysis. Section 5 discusses the advantages and the limitation of RWLR method. Finally, we draw the conclusions in Section 6.

2. Related work

According to the belonging coefficients, nodes are assigned into different levels in HSM. Nodes play different roles in the community. There are some works that divide members into predefined roles. Nepusz et al. [22] detected the fuzzy communities and recognized three kinds of nodes ('outlier', 'bridge' and 'regular') according to the belonging coefficients. Huang et al. [11] expanded communities locally from all the nodes to get the overlapping hierarchical structure of communities (HSC) and then separated the homeless nodes as 'hub' or 'outlier'. Stanoev [26] used dynamic process to reveal the fuzzy communities and assigned nodes to one of the three roles, i.e. 'leader', 'follower' and 'proxy'. Leskovec [16] and Guimera [9] focused on two kinds of members, core members and peripheral members. However, roles in these methods are predefined, which limits the generalization ability of HSM.

The most relevant to ours is the work by Havemann [10] which aims to detect HSC. They calculated all the proper values of parameter alpha which represents stable community structures to improve the Fitness function proposed by Lancichinetti et al. [14]. Although the multi-resolution of communities can be produced, their method cannot guarantee the consistency of belonging coefficients in each level.

Measurement of belonging coefficients has been studied for years, and their corresponding methods are often called fuzzy or overlapping methods. Liu [18] extended the modularity to fuzzy modularity based on a random walk process. Psorakis et al. [23] utilized a Bayesian nonnegative matrix factorization (NMF) model to assign the participation scores of nodes. Nepusz et al. [22] considered the fuzzy community detection as a constrained optimization problem which minimized the difference between adjacency matrix. The similarity matrix is generated by belonging coefficients of nodes. Zhang et al. [34] mapped network nodes to Euclidean space based on a generalized modularity and applied fuzzy c-means to obtain a soft assignment. Steve et al. [7] extended the label and propagation dynamic process to fuzzy community detection. Similarly, based on information dynamic process, Xie et al. [29] defined the membership strength as the probability of observing a label in a node's memory.

Although there are many measurements, they are oriented to nodes, i.e. the sum of belonging coefficients of a node to all communities is equal to 1. Since the normalization is independently for each node, these measurements are not suitable for comparison among nodes. Unlike the methods above, another measurements are oriented to the community, where the sum of all nodes' belonging coefficients to a community is equal to 1 [25]. But, the belonging coefficients are dependent on the resolution of community, which is complex to get multi-resolution of community. Other generative models have the same constraints [20].

In addition, current work of detecting multi-resolution of a community mainly follows two trends. One is to set hand-crafted thresholds to filter out the multi-resolution of community, which brings difficulties in real applications. The other one is to choose the community corresponding to the maximal belonging coefficient for each node, which is easy to result in a single boundary. In this case, we need to propose a new method to detect the multi-resolution of community.

In this paper, we propose a method called RWLR to detect HSM. First, we measure the belonging coefficients oriented to communities by random walk and sort them in descending order. Second, we divide the order of belonging coefficients into multiple segments by linear regression and then get the HSM, where belonging coefficients are consistent in each level and the community structure is stable.

3. Method

3.1. Problem formulation

To simplify the problem, we mainly focus on the undirected, unweighted and simple network. The hierarchical structure of members for community C is defined as

$$Hier(C) = \{level_i\}, \quad i = 1, 2, \ldots, K, \tag{1}$$

where K is the number of levels. Each $level_i$ is a subset of members in community C and satisfies constraints below

$$\bigcup_i level_i = S_C, \tag{2}$$

$$level_i \cap level_j = \varnothing, \quad \forall i \neq j, \tag{3}$$

$$Max\{BC(\alpha), \alpha \in level_i\} < Min\{BC(\beta), \beta \in level_j\}, \quad \forall i > j, \tag{4}$$

where S_C is the set of members in the community C and $BC(\alpha)$ indicates the belonging coefficient of node α. $level_1$ has the highest belonging coefficient where the members are called core members; $level_K$ has the lowest belonging coefficient where the members are called marginal members.

Besides the definition of basic structure, HSM should hold several properties.

- Continuity: Nodes in any top levels will construct a connected component. Formally, for any $level_k$, each pair of node $\alpha, \beta \in \bigcup_{i=1}^{k} level_i$, exists a path $\alpha, p_1, \ldots, p_r, \beta$, where $p_i \in \bigcup_{i=1}^{k} level_i$.
- Comparability: The belonging coefficients of nodes to a community should be comparable. In other words, the belonging coefficients of a node to all communities should not be normalized independently.