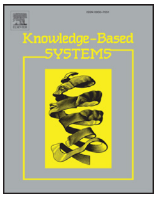
Knowledge-Based Systems 105 (2016) 225–235

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Overlapping community detection based on node location analysis

Wang Zhi-Xiao[a,b], Li Ze-chao[b,*], Ding Xiao-fang[a], Tang Jin-hui[b]

[a] *School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China*
[b] *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China*

ARTICLE INFO

Article history:
Received 9 July 2015
Revised 12 May 2016
Accepted 13 May 2016
Available online 14 May 2016

Keywords:
Overlapping community detection
Topology potential
Peak–valley structure
Node mass calculation
Node location analysis

ABSTRACT

As a novel overlapping community detection theory, topology potential has inspired many methods. However, these methods ignore the mass difference between nodes, leading to inaccurate topological potential values of nodes. Moreover, additional strategies are needed to determine the community affiliation of nodes, further complicating the process of community detection. In this paper, we propose a new overlapping community detection method based on node location analysis. In the proposed method, the PageRank algorithm is used to evaluate the node mass, and the community affiliation of nodes is determined based on their positions in the inherent peak-valley structure of the topology potential field. Experimental results show that the proposed method exhibits excellent performance on artificial and real-world networks and outperforms other topology-potential-based and most non-topology-potential-based methods.

1. Introduction

Many real networks exist in the form of complex networks [1], such as a personal relationship network in a social system and a food chain network in a biological system. These complex networks present a “community structure”, i.e. groups of vertices that have a higher density of edges within them and a lower density of edges between them [2]. Community identification is beneficial for understanding the structural properties of networks and forecasting the behavior of networks [3]. For example, in social networks, community detection can be used to forecast the information propagation among users [4] or predict the missing links among them [5], and in the field of bioengineering, it can be applied to the recognition of the functions of proteins.

Currently, some works concentrate on community detection in static networks [6], while others focus on the scenario of dynamic networks [7]. No matter for static networks or dynamic ones, previous works mainly address the problem by identifying a disjoint community structure [8,9], where one node belongs to only one community. However, it is very common that a node belongs to more than one community in the real networks. Take social networks as an example. A person is usually involved in several social groups such as family, friends, and colleagues [10]. This also happens in biological networks, where a single protein may participate in several function modules because of its functional diversity. Therefore, Kelley et al. [11] pointed out that **overlapping** is indeed a significant feature of many real-world networks.

There are three major types of overlapping community detection methods [12]: local expansion [13–15], fuzzy detection [16,17], and agent-based algorithms [18–20]. Among these three types, local expansion is a type of agglomerate and seed-centric approaches that locally expands or merges the selected seeds. In these methods, a seed may be a single node [21], a sub-graph [14] or a community [13,22]. The topology potential method [23] is a seed-centric approach which depicts the interaction and the association among nodes for overlapping community detection. Because of its inherent advantages in terms of time complexity and performance, it has attracted considerable attention [24].

However, there exist some defects in the current topology-potential-based community detection methods. Firstly, almost all these methods ignore the difference between nodes and assume that all nodes have the same mass. In fact, the mass of a node is supposed to reveal its inherent properties such as importance and influence. Obviously, different nodes have different inherent properties. For example, in a social network, **public figures** apparently are more influential than general people [2]. Thus, this hypothesis will decrease the accuracy of the node topology potential calculation. Secondly, additional strategies are needed to determine the community affiliation of nodes, such as the benefit function in [23] and the adjustable parameter in [24]. These additional strategies complicate the community detection process. For example, in [24], the community affiliation of nodes and the community scale are totally dependent on the adjustable parameter. However, in real-world networks, it is almost impossible to estimate

[*] Corresponding author.
E-mail address: zechao.li@njust.edu.cn (L. Ze-chao).

http://dx.doi.org/10.1016/j.knosys.2016.05.024
0950-7051/© 2016 Elsevier B.V. All rights reserved.

the community scale in advance, so it is very difficult to set a reasonable value.

In order to solve the above-mentioned problems, in this paper, we propose a new overlapping community detection method based on node location analysis. The main contributions of the paper are summarized as follows:

(1) The proposed method leverages the node mass evaluation to ensure the accuracy of topology potential calculation and seeds selection for community detection. In order to evaluate the node mass, the proposed method associates it with node's importance and influence in the network, and uses the PageRank algorithm to evaluate the node mass. The more important a node is, the greater its mass will be.

(2) The proposed method determines the community affiliation of nodes based on node location analysis in the topology potential field, rather than other additional strategies. Topology potential field presents a natural peak-valley structure, thus three types of node positions are defined in the topology potential filed: peak, valley, and slope. Peak nodes are the representative nodes of each community, slope nodes are the internal nodes of each community, and valley nodes are overlapping nodes among communities. By considering this analysis, our community detection process is simple but effective, which has been validated experimentally.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the topology potential field. Section 4 presents the node mass evaluation method. Section 5 analyses the node location in the peak-valley structure of the topology potential field. Section 6 introduces the proposed overlapping community detection algorithm. Section 7 discusses the experiments and results. Finally, Section 8 presents the conclusion of this paper.

2. Related work

Previous overlapping community detection methods can be roughly divided into three types [12]: local expansion, fuzzy detection and agent-based algorithms, among which the first type is the most related to our work. Local expansion methods are among the most successful strategies for overlapping community detection. Most local expansion methods rely on a local benefit function that characterizes the quality of a densely connected group of nodes. In general, local expansion methods consist of two steps. The first step is to identify seed nodes, and the second step is to locally expand or merge these seeds until a local density function cannot be improved any further.

Through adopting different techniques at these two steps, various local expansion methods have been developed. Seen from the first step, i.e. seed selection, local expansion methods can be divided into three groups based on the idea of taking a single node [15,21], a sub-graph [14,25,26], or a community [13,22] as a seed. The first group of methods takes a single node as a seed. Wang et al. [21] assumed that a node with a considerably large degree is likely the core of a community, thus selected these nodes as the core-vertices of communities. Yakun Li et al. [27] claimed that each community has multiple cores, not only one. These centers are the seeds of each community. Each community structure can be obtained by local expansion from these cores. Gan et al. [23] took the maximum-potential nodes in a topology potential field as seeds. Similar methods were also described in [24] and [28]. In the second group, a seed may be a sub-graph. Li et al. [14] took maximal cliques as seeds. These maximal cliques are identified by deep and broad searching and then merged into a larger sub-graph on the basis of given rules. Cui et al. [25] regarded maximal sub-graphs extracted from the original networks as seeds and then merged them by considering the clustering coefficient of two neighboring maximal sub-graphs. Other similar methods were described in [26], [29] and [30]. The third group of methods treats a community as a seed. In [13], all the seed communities are first extracted, and then more community members are absorbed by seed communities using the absorbing degree function. Another similar method was described in [22] and [31]. Zhan Bu et al. [32] regarded each node as a separate seed community in a network. These seed communities with the global maximum ΔQ will be merged into one community. Among the three groups of local expansion methods, the first group is relatively simple, but the accuracy of seed selection is very crucial because all local expansions are based on these single nodes. The rest two types are more complex in seed selection. Moreover, they need to determine the proper scale of each sub-graph or seed community. Even with the same type of seeds, different approaches leverage different functions or parameters in the second step to carry out local expansion. For example, both [29] and [30] regard sub-graphs as seeds, but they adopt different local expansion strategies. In [29], nodes are absorbed by seeds based on the dependence degree of a node on its neighbors and on a cluster. Eustace et al. [30] utilized the overlapping neighborhood ratio to assign nodes to their corresponding communities. Traditional local expansion strategies are usually computationally expensive. A simple but effective local expansion strategy is the key for a local expansion method.

The topology-potential-based method is a typical and important local expansion approach which has attracted considerable attention. This method takes a single node as a seed. Gan et al. [23] used a topology potential field to describe the interaction and the association among network nodes. Each community is treated as a local high-potential area, and the maximum-potential nodes in each local high-potential area are considered as seeds. The community structure can be uncovered by detecting all local high-potential areas surrounded at the margins by low-potential nodes. Based on topology potential, Zhang et al. [24] proposed a variable scale network for overlapping community identification, in which an adjustable parameter is defined to determine the community affiliation of nodes. Han et al. [28] put forward another topology-potential-based overlapping community detection algorithm, which divides networks into separate communities by "spreading outward from each local maximum potential node". In their method, a benefit function is defined to guide the local expansion process.

Despite their successful application in overlapping community detection, current topology-potential-based methods ignore the mass difference between nodes, thus suffering from inaccurate topology potential calculation and seed selection. Besides, they need additional strategies, e.g. adjustable parameter [24] and benefit function [28], to determine the community affiliation of nodes, which complicates the community detection process. Aiming at these problems, in this paper, we propose an improved topology-potential-based overlapping community detection method based on node location analysis. Compared with other overlapping community detection methods, the proposed method is characterized by two advantages:

(1) The proposed method leverages the node mass evaluation to ensure the accuracy of topology potential calculation and seeds selection for community detection. We associate node mass with its importance and influence in the network, and use the PageRank algorithm to evaluate the node mass.

(2) The proposed method determines the community affiliation of nodes based on node location analysis in the natural peak-valley structure of topology potential field, instead of other additional strategies. Thus our process is very simple but effective.