Knowledge-Based Systems 52 (2013) 268–278

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Efficient community detection with additive constrains on large networks

Yakun Li, Hongzhi Wang *, Jianzhong Li, Hong Gao

Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

ARTICLE INFO

Article history:
Received 13 December 2012
Received in revised form 29 July 2013
Accepted 2 August 2013
Available online 13 August 2013

Keywords:
Community detection
Feedback control
Additive constrains
Large networks
Scale-free network
Graph algorithm

ABSTRACT

The community structure is one of the most important patterns in network. Since finding the communities in the network can significantly improve our understanding of the complex relations, lots of work has been done in recent years. Yet it still lies vacant on the exact definition and practical algorithms for community detection. This paper proposes a novel definition for community which overcomes the drawbacks of existing methods. With the new definition, efficient community detection algorithms are developed, which take advantage of additive topological and other constrains to discover communities in arbitrary shape based on the feedback. The algorithm has a linear run time with the size of graph. Experimental results demonstrate that the community definition in this paper is effective and the algorithm is scalable for large graphs.

1. Introduction

Complex networks have been widely used in many applications. Representative complex networks include Internet [10], email communication networks [9], social networks [43], mobile call networks [28], instant-messaging networks [23], citation networks [8], and biological networks [13]. Because of its importance, the mining of complex network attracts the attentions of researchers in the literature. Many data mining problems over complex network have been studied. One of them is community detection.

The goal of community detection is to cluster the similar vertices into one community and separate to the others. Since the vertices in the same community share similar properties, the communities in the network make users understand the complex relations deeply. As an example, the Zachary's karate club network is shown in Fig. 1. The friendship relationships among 34 members of karate club at a US University in the 1970s [45] are modeled as a network. Because of the conflict of opinion between the administrator and the instructor, the club can be split into two communities. If the community detection algorithms are applied to the network, before the artificial classification is performed, these two communities can be known. The problem is that when there are millions of persons in one organization, artificial classification becomes infeasible. In such cases, community detection techniques are in demand.

The Zachary's karate club network is a typical social network. Since it is modeled based on real social relationships and the community structure is obvious after the break up, it has been analyzed by almost all community detection algorithms. In this paper, we choose it as the motivate example in this paper.

Community detection has many applications in systems related to complex networks. The examples include classification in social dimensions [40], finding influential bloggers [1] and recommendation system [44]. The social dimensions describe the affiliation of an actor. And the affiliations can be learned automatically in presence of community labels. Influential bloggers impact the followers in various ways, and they are often the most representative actors in a community. Community detection could help to find such bloggers.

Even though many community detection methods have been proposed due to its importance, the requirements of the applications have not been satisfied. Current methods have the following shortcomings.

First of all, community is not defined perfectly. Intuitively, a community is a group of vertices in the network, within which the connections are dense, but between which the connections are sparse. Although many researchers have tried to give an exact definition of the community [30,27,39,33,16], none of them have been generally accepted. The major reason is that existing definitions of community are based on the models far from the real-world situations. The method in [27,39] requires a fixed threshold.

* Corresponding author. Address: P.O. Box 750, Harbin Institute of Technology, 150001 Harbin, China. Tel.: +86 451 86403492 810; fax: +86 451 86415827.
E-mail addresses: liyakun.hit@gmail.com (Y. Li), wangzh@hit.edu.cn (H. Wang), lijzh@hit.edu.cn (J. Li), honggao@hit.edu.cn (H. Gao).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.08.003

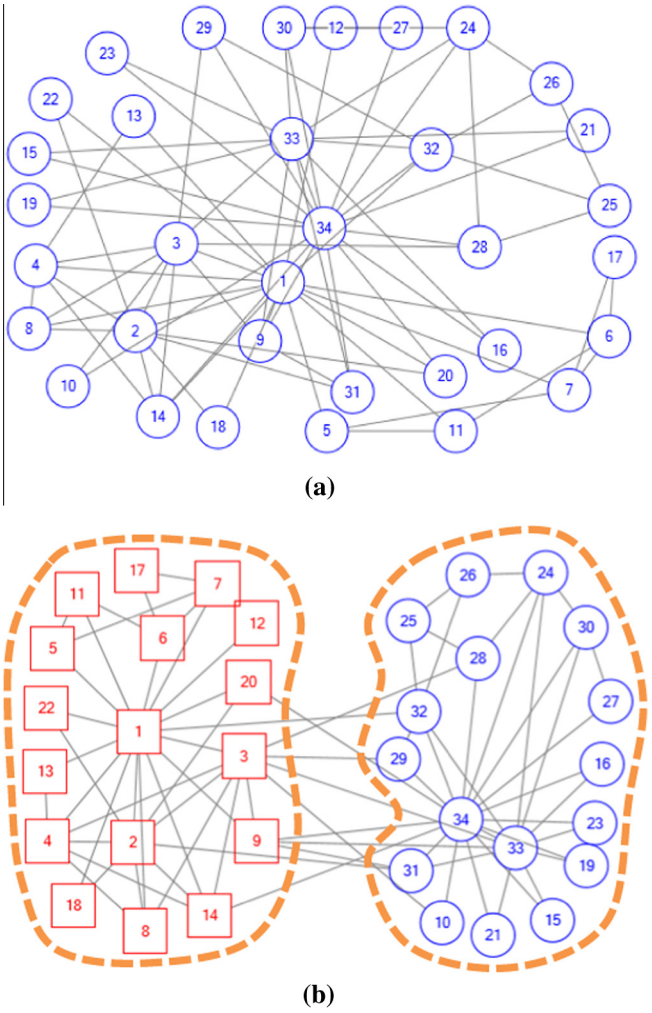

Fig. 1. The network of Zachary's karate club. Each vertex represents an individual in the club. The administrator and the instructor are represented by node 1 and node 33, respectively. Each edge represents the interaction between two members. All the relations among individuals are represented in (a), and the two clubs after the splitting are represented in (b).

This makes them unsuitable for the communities in various forms. The definition in [30] assumes that in a stochastic network, each vertex connects to other vertices in the same probability. Clearly, for a complex network, this assumption does not always hold. The strong and weak definitions in [33] both have their drawbacks, since the strong definition is so strict that will miss some vertices like V_8 in Fig. 2. In the other extreme, the weak definition is too loose and more like just a feature of community than an exact definition. That is, the whole network will meet the weak definition to be a community, which is incomprehensible.

The second is that current methods are not suitable for massive data. The time complexity of most classic clustering algorithms is more than $O(n^2 \log n)$, where n is the number of vertices. The naïve divisive [13] algorithms have time complexity $O(m^2 n)$ where m is the number of edges. Though many optimization techniques [30,41,31,35,14] are proposed in recent years, none of them is suitable for massive data due to the nature spirit of division that has to find the weakest edges or vertices and the time complexity is not improved. Ref. [42] introduced Markov Cluster Process for the partition, while its time complexity is $O(n^3)$. Some heuristic methods [3,21] have been proposed, but none of them have time complexity assurance. All the modularity maximization algorithms are approximate methods. The time complexity of [29] is $O((m+n)n)$, the time complexity of [32] is $O(mn^2)$, and that of [7] is $O(md\log n)$, where d is the depth of the "dendrogram". Ref. [38] has used the method in [7] as a step, which makes its time complexity no smaller than $O(md\log n)$. However, many applications such as social network involve huge graphs and require methods with linear or sub-linear time complexity.

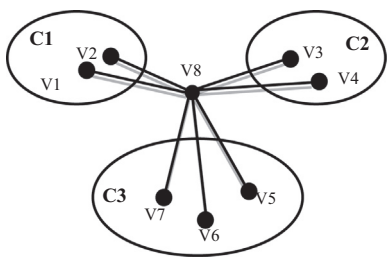

Fig. 2. The unreasonable situation for the strong community.

The third is that they do not take advantages of feedback information and structural features of graphs sufficiently. Different users may have different requirement for communities. For example, the number of communities divided by an election should be exact the number of candidates. Feedback can help the system to discover the intensions of users. The knowledge of network structural features can also improve both efficiency and effectiveness of community detection. For example, if we know that the network is modeled as a forest, then the community should be easily found from the root vertex. Unfortunately, all existing algorithms are closed systems. It means that if the communities are generated, they are fixed without modifications. Feedback and structural information are never considered.

To overcome the shortcomings of current approaches, we propose a novel definition of community as well as two community detection algorithms. Our definition considers scale-free pattern for complex network, such that the definition is more convincing and closer to the real-world situations. The community detection algorithms proposed in this paper are based on the new community definition. The linear time complexity makes them suitable for web scale networks. Additionally, they can make sufficient usage of feedback information and structural information to improve the efficiency and effectiveness.

The contributions can be summarized as follows:

(1) The first contribution is the Peaks model and a novel definition for community. This definition is based on the scale-free pattern and the community structure pattern. Based on the Peaks model, a community definition matching the real-world situations is proposed. In this definition, each community has some centers, called "Core". A community should satisfy that the distance from each vertex to the cores of the community should not be larger than those to any other community. And when it comes to an equal distance, a vertex should belong to the community with the most connections.
(2) The second contribution is the practical criterion to quantify the community structure. The most impartial evaluation is when there is an exact division in real-world data. And it is convincing.
(3) Two efficient community detection algorithms are proposed for different applications. These two algorithms can perform community detection according to our model in time complexity linear to the network size and are easy to make use of feedback information and structural features.
(4) Extensive experiments demonstrate that our method outperforms existing methods and scales well for large network.

Organization: The remaining part the paper is organized as follow. Section 2 introduces the related work on community