Pattern Recognition Letters 70 (2016) 66–72

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Fuzzy community detection via modularity guided membership-degree propagation☆

Hengyuan Zhang, Xiaowu Chen*, Jia Li, Bin Zhou

State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, 37 Xueyuan Road, Beijing 100191, China

ARTICLE INFO

Article history:
Received 2 February 2015
Available online 2 December 2015

Keywords:
Fuzzy community detection
Membership-degree propagation
Seeds selection

ABSTRACT

In complex network analysis, fuzzy community detection is a challenging task that aims to reveal the network structure by assigning each vertex quantitative membership-degrees to various communities. In this paper, we propose a fuzzy community detection method that iteratively propagates membership-degrees of all vertices. In each iteration, a candidate seed vertex of a potential community is first selected according to the topological characteristics. After that, the membership-degrees are propagated among adjacent vertices so that a number of communities can be obtained with respect to all selected seeds. To ensure that the modularity keeps improving, in each iteration we discard the selected seeds that decreases the modularity of the community decomposition. In this manner, the topological information about the network can be fully utilized, and communities gradually emerge along with the acceptance of new seeds. Experimental results on real-world and synthetic networks demonstrate that our approach has impressive performance and is robust on both disjoint and fuzzy community detections. Moreover, the proposed approach exhibits a high flexibility between computational complexity and overall performance.

1. Introduction

In network analysis, complex networks are usually decomposed into communities that can be defined as groups of vertices. Usually, vertices in the same group have dense or tight edges, while such edges become sparse or lose if vertices are in different groups. By detecting such communities, various applications such as discovering functional modules in biological networks and identifying organizations in social networks [16] becomes feasible and practical.

In the past decades, a great deal of methods have been proposed for community detection [7]. Among them, label propagation and modularity optimization are two of the most popular solutions. The former is an intuitive idea with logical and physical significance, which deems that the community membership of a vertex is decided by that of its neighbors [21,26]. The latter converts the task of community detection into an optimization problem. The objective function is often defined on community modularity, which is a widely accepted quality measurement of community decomposition introduced in Newman and Girvan [19]. By optimizing the modularity, approaches such as [1,2,22] demonstrated impressive robustness and effectiveness in detecting disjoint communities. However, these approaches only output the qualitative memberships of vertices to disjoint communities, while an important fact that each vertex may appear in multiple communities with different belonging coefficients, is ignored [9]. For instance, a person can associate with several social communities with different levels of commitment regarding occupation and interests.

To address this problem, fuzzy community detection is proposed to measure the membership of vertices in communities with quantitative indicators, or namely, membership-degrees. Compared with disjoint community detection, fuzzy community detection can reveal not only the affiliations of vertices but also the network structure. It can provide more information for applications such as discovering overlapping communities for visualization [25]. Some fuzzy community detection methods have been proposed in recent years. However, many of them require the priori knowledge about the community structure (e.g. number of communities) [29,31,32] or need to fine-tune intricate parameters (e.g. probability threshold) [6,20], which may degrade the performance when processing complex networks.

In this paper, we propose a fuzzy community detection approach called *Membership-Degree Propagation* (*MDP*), which is motivated by the ideas of both label propagation and modularity optimization. The method iteratively propagates membership-degrees of all vertices. In each iteration, a candidate seed vertex of a potential community is first selected according to the topological characteristics. After that, the membership-degrees of selected seeds are propagated

☆ This paper has been recommended for acceptance by Andrea Torsello.
* Corresponding author: Tel.: +86 10 82317610.
E-mail address: chen@buaa.edu.cn, zhang-hy@buaa.edu.cn (X. Chen).

http://dx.doi.org/10.1016/j.patrec.2015.11.008
0167-8655/

to non-seed vertices, so that a number of communities can be obtained. To ensure that the modularity keeps improving, whenever a new candidate seed is selected, we discard the seeds which decreases the modularity of the community decomposition. With such a strategy, communities gradually emerge along with the acceptance of new seeds, and the method becomes flexible between computational complexity and overall performance. Compared to several state-of-the-art algorithms, experiments show that our method has competitive scores in various performance indices.

To sum up, this paper offers two main contributions:

1. We specify the selection order of community seeds according to the topological characteristics of vertices. The seeds selection is accompanied by seeds updating, which is under the guidance of modularity. From the seeds, the community structure can be gradually discovered with high flexibility between computational complexity and overall performance.
2. The membership-degrees of seeds are propagated to non-seed vertices by fully utilizing the topological information about the network. Experimental results show that the proposed approach has impressive performance in detecting fuzzy communities.

The rest of this paper is organized as follows. Section 2 describes the technical details of our method. Experiments are given in Section 3, and the paper is concluded in Section 4.

2. Methods

2.1. Problem formulation

In this study, a complex network is defined as an undirected weighted or un-weighted network $G = \langle V, E\rangle$. Here, $V = \{v_i | i = 1, 2, \dots, N\}$ is the set of vertices, and $E = \{e_{ij} | i \neq j\}$ is the set of edges (M edges in total). Thus the objective of fuzzy community detection is 1) to determine the community number C, and 2) to calculate the membership-degrees of each vertex to all the C communities.

Toward this end, we define the membership-degrees of v_i as a column vector $\mathbf{u}_i$ with C components. The cth component $\mathbf{u}_i[c]$ is a real-valued number that reflects the degree of v_i belonging to the cth community. $\mathbf{u}_i$ satisfies below constraints:

$$0 \preceq \mathbf{u}_i \preceq 1 \text{ and } \|\mathbf{u}_i\|_1 = 1. \tag{1}$$

In this work, we suppose each community in G contains a single seed vertex. We consider the memberships of seeds are crisp. That is, if v_i is the seed of the cth community, $\mathbf{u}_i[c] = 1$.

From the seeds in G, we can thus generate the membership-degrees of all non-seed vertices and decompose G into disjoint communities accordingly. Here we define the membership-label l_i of v_i as

$$l_i = \arg\max_k \{\mathbf{u}_i[k]\}. \tag{2}$$

Given the membership-labels, the quality of disjoint community decomposition can be measured through Newman's modularity [19]:

$$Q = \frac{1}{\sum_{i,j} w_{ij}} \sum_{i,j} \left(w_{ij} - \frac{\mathcal{D}(v_i) \cdot \mathcal{D}(v_j)}{\sum_{i,j} w_{ij}} \right) \cdot \delta(l_i, l_j), \tag{3}$$

where w_{ij} is the weight of the edge e_{ij}, and $\delta(l_i, l_j)$ is a binary indicator which equals to 1 if $l_i = l_j$ and 0 otherwise. $\mathcal{D}(v_i) = \sum_k w_{ik}$ is the vertex degree of v_i. The modularity $Q \in [0, 1]$ becomes higher for better result of community decomposition.

Based on the above descriptions, we formulate the task of fuzzy community detection as three problems:

1. How to select seeds for potential communities?
2. How to estimate the membership-degrees of non-seed vertices based on the selected seeds?
3. How to update the seeds so that the community modularity can be optimized?

2.2. Membership-degree propagation

Seeds selection. Many studies approve that there exist seed vertices in communities [5,21,26,27]. It is often acknowledged that a community seed has two significant topological characteristics [3,14,30]. First, it usually has high vertex degree. Second, it has significant local centrality in network (*i.e.*, a vertex degree of local maximum). Suppose the set of neighbors of v_i is $\mathbb{N}_i$, we define $\mathcal{D}_{\mathbb{N}}(v_i) = \sum_{v_j \in \mathbb{N}_i} \mathcal{D}(v_j)$. The smaller $\mathcal{D}_{\mathbb{N}}(v_i)$ is, the higher local centrality of v_i.

In order to effectively select the community seeds, we sort all the vertices into a candidate queue of seeds by referring to their topological characteristics. The queue is denoted as $S = \{s_i | s_i \in V\}$, in which

$$\mathcal{D}(s_i) \geq \mathcal{D}(s_{i+1}) \text{ and } \mathcal{D}_{\mathbb{N}}(s_i) \leq \mathcal{D}_{\mathbb{N}}(s_{i+1}). \tag{4}$$

When our approach tries to detect a new community, s_0 (the most likely community seed currently) is popped out from S and selected to be a candidate seed. However, every previously selected seed may be cancelled during the following seeds updating process until the algorithm is completed.

The computation complexity of our approach increases as S expands. Since vertices with low degree are unlikely to be seeds, we define a threshold T_s to filter out the vertices in S with $\mathcal{D}(s) < T_s$. By tuning T_s, our approach could make trade-off between computational complexity and overall performance.

Membership propagation. From the selected seeds, we propagate their membership-degrees to the non-seeds vertices following the principle of label propagation. Specifically, the membership-degrees of a non-seed vertex is regarded as the weighted average of those of its neighbors. That is, for a non-seed vertex v_i, we have

$$\mathbf{u}_i = \left(\sum_{v_j \in \mathbb{N}_i} w_{ij} \cdot \mathbf{u}_j \right) \Big/ \sum_{v_j \in \mathbb{N}_i} w_{ij}. \tag{5}$$

The propagation is an iterative process, in which $\mathbf{u}$ of seeds are fixed, and that of the non-seed vertices are updated according to (5). In each iteration, we first update the neighboring vertices of the seeds, the next neighbors of these vertices, and so forth. As the updates are repeated, the influences of seeds are extended to the whole network according to the topology. In this manner, the topological information loss of label propagation can be eliminated [17].

When the membership-degrees of each non-seed vertex converges, the propagation is completed. That is,

$$\Delta\mathbf{u} = \max\{\|\mathbf{u}_i - \mathbf{u}_i'\|_2, \forall v_i \in V\} < \epsilon_s, \tag{6}$$

where $\|\mathbf{u}_i - \mathbf{u}_i'\|_2$ is the variation of the membership-degree of v_i (measured by the $\ell 2$ distance). ϵ_s is a predefined small threshold. To speed up the propagation, we initialize $\mathbf{u}_i[c]$ to be inversely proportional to the distance from v_i to the seed of the cth community.

We demonstrate the process of propagation with Karate network [18]. Suppose the red and green vertices are the selected seeds (Fig. 1(a)), whose membership-degrees are {1, 0} and {0, 1}. In order to visualize the process, we initialize the membership-degrees of all the non-seed vertices as {0.5, 0.5} in this example, and color them by blending the seed-colors according to $\mathbf{u}$. Fig. 1(b) and (c) are the states after the 1st and 5th iteration of the propagation, respectively. After 12 iterations, $\Delta\mathbf{u} < \epsilon_s$ (we set $\epsilon_s = 10^{-4}$ in this example), and the propagation is terminated. Fig. 1(d) shows the final result. Blue dashed line indicates the disjoint community decomposition obtained according to the definition of membership-label. The decomposition exactly matches the real community structure of the Karate club. The vertex in blue square is a typical common one shared between the two communities because of the tiny difference between the two components of $\mathbf{u}$ ({0.508, 0.492}).

*Seeds updating.*It is obvious that the performance of propagation is determined by the selected seeds. According to (2), the propagating