Information Sciences 355–356 (2016) 37–57

Contents lists available at ScienceDirect

Information Sciences

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

A weighted local view method based on observation over ground truth for community detection

Yanmei Hu[a], Bo Yang[a,*], Hau-San Wong[b]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan, China
[b] Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Hong Kong, China

ARTICLE INFO

Article history:
Received 30 March 2015
Revised 6 January 2016
Accepted 16 March 2016
Available online 24 March 2016

Keywords:
Community detection
Social networks
Personalized PageRank vector
Ground-truth communities
Community structure

ABSTRACT

Community detection is a fundamental problem for many networks, and there have been a lot of methods proposed to discover communities. However, due to the rapid increase of the scale and diversity of networks, the modular organization at the global level in many large networks is often extremely difficult to recognize. In these cases, many existing methods fail to discover the latent community structure, because they follow a paradigm of discovering communities from a global view of networks. In this paper, we propose a weighted local view method based on an interesting observation on ground-truth communities, with the aim of revealing community structure in large real networks. This is achieved by the following steps: 1) a set of nodes which can well represent their neighboring nodes are chosen by local seeding strategies; 2) each chosen node explores the community in its local view to the whole network, using an improved approximate personalized PageRank-based community finder which is based on an interesting observation on large real networks with ground-truth communities; 3) all explored local communities are merged to form the global community structure. We evaluate the weighted local view method against the state-of-the-art community detection methods on large real networks with ground-truth communities. Experiments show that the proposed method can not only improve the detected communities, but can also scale to very large networks with good computational efficiency compared with other methods, which indicates that the weighted local view method has great potential for overlapping community detection in large networks.

1. Introduction

A community (also referred to as a cluster) usually consists of a set of nodes which are densely connected with each other, while sparsely connected to the rest of the network. Community structure is a fundamental characteristic of many networks. Exploring communities may help identify functional or organizational subunits of the networks, to reveal mesoscopic structure of networks, and to uncover relationships among nodes that are not apparent in the absence of detailed information. Thus the community detection problem has been attracting a lot of attention [15,27].

However, with the rapid increase of the scale and diversity of networks, many community detection methods fail to uncover the community structure. In networks with small scale, the denser areas are easily identifiable, while for large-scale

* Corresponding author. Tel.: +86 28 61831656.
E-mail addresses: huym260@126.com (Y. Hu), yangbo@uestc.edu.cn (B. Yang), cshswong@cityu.edu.hk (H.-S. Wong).

http://dx.doi.org/10.1016/j.ins.2016.03.028
0020-0255/

Algorithm 1 ApproximatePageRank.

Input $G = (V, E)$: the network structure; s: the seed node; α, ε: the control parameter, $\alpha \in (0, 1]$, $\varepsilon \in (0, 1)$
Output $\mathbf{p}$: the approximate personalized PageRank vector

1. Let $\mathbf{p} = \vec{0}$, and let the element of the residual vector $\mathbf{r}$ corresponding to the seed node s be 1 and other elements be 0.
2. While $\max_{u \in V} \mathbf{r}(u) \geq \varepsilon$:
   (1) Choose any vertex u that satisfies $\mathbf{r}(u) \geq \varepsilon$.
   (2) Let $\mathbf{p}(u) = \mathbf{p}(u) + (1 - \alpha) \cdot \mathbf{r}(u)$.
   (3) For each v such that $(u, v) \in E$ do:
   $\mathbf{r}(v) = \mathbf{r}(v) + \alpha \cdot \mathbf{r}(u)/(2d(u))$.
   (4) Let $\mathbf{r}(u) = \alpha \cdot \mathbf{r}(u)/2$.

networks the problem of finding denser areas becomes much harder. Not only is it very computationally intensive to find denser parts from the global structure of large-scale networks, the modular structure of many large networks is also hard to distinguish at the global level, since at such scales the organization of the system becomes too complex [11]. Thus, many community detection methods usually fail to handle large networks, because they extract communities by optimizing a particular score function (e.g., maximizing the modularity, or maximizing the likelihood of the underlying network conditioned by the latent community structure) from a global view of networks [9,35,44].

On the other hand, each node often has a clear perspective of the part around it [11]. The perspective of each node may be incomplete relative to the entire network, but it will be easier to understand and analyze. Solving the problem of community detection from the local view thus becomes a good alternative, since it only needs local information which is generally a small and also easily understandable part of the entire network, and thus is much easier to handle [11]. In addition, in the case where global information is not accessible, a local method is also a good alternative for discovering communities. In this paper, we thus explore a local view method to detect global communities in large real networks. Note that the local view methods for community detection in this paper refer to the ones that find global communities in the network by checking the nodes' local views, and are different from the local community detection methods which only detect a single local community to which the starting node belongs.

There have been a few studies on figuring out the global community structure from the local views of nodes [11,41]. Coscia et al. propose a democratic bottom-up mining approach (DEMON) to identify communities in large networks. In this approach, each node votes for the communities which are present in its egonet, and then all the votes are combined together to form the estimate of real communities. Since DEMON needs to consider each node's votes for communities, the corresponding computational time and storage requirements will be significant for very large networks. On the other hand, Whang et al. first chose a number of nodes as seeds, and then apply an approximate personalized PageRank-based community finder to explore the local views of those seeds. Specifically, an approximate personalized PageRank vector from the seed is first computed (see Algorithm 1 in Section 4.1), and then a sweep over the PageRank vector using conductance is performed to generate the community around the seed (see Steps 2 and 3 of Algorithm 3 in Section 4.2). The identification of the underlying community structure is achieved through merging the local communities expanded from the chosen seed set. In general, this class of techniques is effective for exploring communities around seeds [2,30,41,43].

When computing the approximate personalized PageRank vector, each node equally distributes its probability mass to its neighbors (see Step 2.3 of Algorithm 1 in Section 4.1). In other words, each neighbor is equally important to a node. However, by exploring several large real networks with ground-truth communities, we observe that in those community structures, many of the edges are contributed by connecting nodes with similar degrees, and the difference of degrees (abbreviated as *dod*) of connected nodes follows a heavy-tailed distribution with highly right-skewed tail, which implies that in a community structure it is more probable for nodes with similar degrees to be connected than those with dissimilar degrees. This observation implies that nodes should treat neighbors differently. As a result, when computing the approximate personalized PageRank vector to explore the local community around the starting node, some neighbors should receive less probability mass, while some should receive more.

In view of the above observation, we propose a weighted local view method based on an improved approximate personalized PageRank-based community finder (WAPR for short), with the aim of discovering communities similar to the ground-truth communities in large real networks. In particular, according to the observation on large real networks, we develop an improved algorithm for computing the approximate personalized PageRank vector from a given node, i.e., for each node encountered in the process of computing the approximate personalized PageRank vector, we first weight its neighbors based on the observation, and then spread probability mass to each neighbor according to the weight. The amount of probability mass spread to each neighbor is proportional to the weight of this neighbor, so we name the improved algorithm as WApproximatePageRank. The community from the local view of the given seed is then generated by performing a sweep over the computed weighted personalized PageRank vector by conductance. To detect communities at global level in large real networks, we explore good seeds which are selected based on local information only, and then apply the improved approximate personalized PageRank-based community finder to explore the local communities around those seeds. The set of local communities from the local views of seeds are finally merged to form the community structure at the global level.

Due to the local nature, our weighted local view method can avoid the problem of failing to detect the modular structure of large networks at global level, and has the capability to scale to large networks as well. The proposed method also has