Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 407–416

Procedia Computer Science

www.elsevier.com/locate/procedia

21st International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

An evidential influence-based label propagation algorithm for distributed community detection in social networks

Mehdi Azaouzi*, Lotfi Ben Romdhane

MARS☆ Research Laboratory LR17ES05, Higher Institute of Computer Science and Telecom (ISITCom), University of Sousse, Tunisia

Abstract

Community detection in social networks is a computationally challenging task that has attracted many researchers in the last decade. Most of approaches in the literature focus only on modeling structural properties, ignoring the social aspect in the relations between users. Additionally, they detect the communities, one after another, in a serial manner. However, the size of actual real-world social networks grows exponentially which makes such approaches inefficient. For this, several models tend to parallelize the community detection task. Unfortunately, social networks data often exhibits a high degree of dependency which renders the parallelization task more difficult. To overcome this difficulty, amongst the proposed distributed community detection methods, the label propagation algorithm (LPA) emerges as an effective detection method due to its time efficiency. Despite this advantage in computational time, the performance of LPA is affected by randomness in the algorithm. Indeed, LPA suffers from poor stability and occurrence of monster community. This paper introduces a new LPA algorithm for distributed community detection based on *evidence theory* which has shown a high efficiency in handling information. In our model, we will use the belief functions in the update of labels as well as in their propagation in order to improve the quality of the solutions computed by the standard LPA. The mass assignments and the plausibility, in our model, are computed based on *the social influence* for detecting the domain label of each node. Experimentation of our model on real-world and artificial LFR networks shows its efficiency compared to the state of the art algorithms.

Peer-review under responsibility of KES International

Keywords: Social networks; Distributed community detection; Theory of belief functions; Influence measure; Label propagation

1. Introduction and Related Work

With the overwhelming explosion of social networks, a large number of techniques to understand the structures of networks are developed. One of the most used is community detection that is based on the hypothesis that the set of individuals in a community should have strong interactions between them and little interactions with the outside [1].

☆ Modeling of Automated Reasoning Systems
* Corresponding author. Tel.: +216-98-100-843.
E-mail address: mehdi.azaouzi@gmail.com

1877-0509
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.045

Most of proposed methods in the literature detect communities in a serial manner as follows; i.e.; they detect only one community in each iteration following a given strategy such as optimization methods [1,2,4,5,6]; divisive clustering algorithms [3]; consensus strategies [7] and the relational concept analysis [8]. However, this *centralized* and global control is a significant bottleneck when used in very large scale networks because of high time and space complexities. In order to overcome this difficulty, many efforts have been taken to find various algorithms to efficiently and effectively address the community mining in large-scale networks. Recently, several methods based on *distributed* detection are proposed. These approaches could be roughly classified into two categories. The first and most studied one aims to distributed detection framework with shared-memory parallelism or "distributed computing" [9,10]. These parallelization strategies divide the social network into non-overlapping subnetworks that are treated in parallel. The main obstacle in this first category lies in the fact that the community mining algorithms often exhibit a high degree of data dependency. To address this issue, the second category – which is the focus of this paper – is presented. The rationale behind the models in this category is to detect several communities in a distributed manner using an *undistribtued* social graph. In this second category, a promising algorithm, called the *label propagation algorithm* (*LPA*), is proposed in [11]. The basic idea of LPA could summarized as follows. Initially, each node is assigned a unique label. Thereafter, labels are diffused through the network and each node decides of its label using a decision rule (generally the majority label). In addition to its simplicity and its distributed nature, LPA is able to detect communities in *linear time*. Thus, it can process extremely large networks. For this, several LPA-based algorithms were proposed in the literature– see for instance [12,13,14]. Unfortunately, the performance of LPA is affected by randomness in the algorithm. Indeed, LPA suffers from poor stability and occurrence of monster community. This instability is mainly caused by the order in which nodes are updated. In fact, in LPA, we have observed that the probabilistic system of updating label with the most frequent label among neighbors leads to the uncertain labeling. Explicitly, at step t, in the case of most frequent label among neighbors, one of them is chosen randomly. This probabilistic choice at t becomes certain in the next step $t + 1$; and thereby, in the rest of the diffusion process. Unfortunately, this leads to weak robustness.

In this paper, we propose a new LPA-based model for distributed community detection in mega-scale social networks that addresses each of the issues raised above. The central idea of our model, named Evidential Influence LPA (EILPA), is the use of influential nodes as "leaders" in community detection. In fact, the information is much easier to be propagated from influential to common nodes, and the influential ones can attract others to their communities. EILPA is run mainly in two phases. In a first phase, we extract a set of most influential nodes that could form the centers of communities. Then, each center will be assigned with unique label and the remained nodes are not labeled. In a second phase, we will apply LPA, in each part of the network, to propagate the label center based on the influence measure. EILPA adopts the framework of belief functions based on the K-nearest neighbor rule. Hence, in EILPA, an influence piece of evidence quantities the impact (influence) of each center on the current node; and is used to initialize nodes labels. Moreover, at stage t, EILPA avoids the randomness of LPA, by adopting a new measure of clustering coefficient based on the neighborhood labels at stage $(t - 1)$. The rest of this paper is structured as follows. Section 2 introduces background material about the standard LPA and the theory of belief functions. Section 3 introduces our model EILPA; while Section 4 outlines experimental results. The last section offers concluding remarks and suggests future research directions.

2. Background

2.1. The LPA algorithm

The label propagation algorithm (LPA) [11] identifies network communities by the following procedure. Initially, LPA assign a unique label to every node in the network and arrange them in a random order. Then, LPA updates the label of each node by choosing the most frequent label within its neighbors. If more than one label have the same frequency among neighbors, the algorithm chooses one of them randomly. This process is repeated until each node in the network gets the most frequent label from its neighbors. Consequently, densely connected groups reach a common label quickly. When many such dense (consensus) groups are created throughout the network, they continue to expand outwards until it is impossible to do so. We denote by $G(V, E)$ an undirected network graph modeling the