Expert Systems with Applications 41 (2014) 4309–4321

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An efficient algorithm for community mining with overlap in social networks

Delel Rhouma [a,*], Lotfi Ben Romdhane [b]

[a] MARS (Modeling of Automated Reasoning Systems) Research Group, FSM/University of Monastir, Tunisia
[b] Insitute of Computer Science and Telecom (ISITCom), University of Sousse, Tunisia

ARTICLE INFO

Keywords:
Social networks
Communities
Overlap
Objective function
Fuzzy membership degree

ABSTRACT

Detecting communities in social networks represents a significant task in understanding the structures and functions of networks. Several methods are developed to detect disjoint partitions. However, in real graphs vertices are often shared between communities, hence the notion of overlap. The study of this case has attracted, recently, an increasing attention and many algorithms have been designed to solve it. In this paper, we propose an overlapping communities detecting algorithm called DOCNet (Detecting overlapping communities in Networks). The main strategy of this algorithm is to find an initial core and add suitable nodes to expand it until a stopping criterion is met. Experimental results on real-world social networks and computer-generated artificial graphs demonstrate that DOCNet is efficient and highly reliable for detecting overlapping groups, compared with four newly known proposals.

1. Introduction

Graphs play a central role in the field of complex systems. Indeed they are the preferred tool for mathematical modeling. We find it naturally in the study of several areas: sociology, biology, linguistics, physics, computer science ... (Pons, 2004). These graphs can reach large sizes and in more than hundred nodes, it becomes difficult to understand their structures and to view it legibly (Pons, 2004). The search for strongly linked groups of vertices can provide a simplified representation of the structure of large graphs: this fact is important for the end user because it allows to understand very intuitively the modeled social network. Thus, it brings the members of network by affinity or common characteristics, it is enrolled in the context of community detection. This represents one of the key problems in social network analysis and it has been extensively studied (Pons, 2004). These studies are divided into two families: finding homogenous communities (Fortunato, 2010; Zardi & Ben Romdhane, 2013; Cheong, Huynh, Lo, & Goh, 2013) or extracting a set of pairs of communities that behave in opposite ways with one another (exhibiting antagonistic behaviors) (Zhang, Lo, Lim, & Prasetyo, 2013; Lo, Surian, Prasetyo, Zhang, & Lim, 2013).

However, we should notice that in most of existing approaches the computed partitions are disjoint; i.e., each vertex is assigned to a single community. However, it is well understood that people in a social network are naturally characterized by multiple community memberships, hence the notion of overlap between communities. For example, a person usually has connections to several social groups like family, friends and colleagues. He can be an active member simultaneously in the fields of mathematics, biology, science, etc. Another typical example is in the PPI networks (protein–protein interaction) (Fortunato, 2010) in which we seek to identify functional classes. Indeed, many proteins have multiple functions depending on different tissues. They may belong to more than one functional unit and sometimes they act as a bridge that allows the transfer of information. So the assignment of this gene to a single class is not justifiable. For this reason, overlapping community detection algorithms have been investigated (Xie, Kelley, & Szymanski, 2013).

In this paper, we propose an efficient algorithm to identify overlapping nodes. It is based on the local optimization of a fitness function and a fuzzy belonging degree of different nodes. This membership is not only based on the number of link which connects the node to the community, but also on the size of the community and the shortest path from the node to all its members. We propose an objective function to qualify the overall quality of a partition; and present DOCNet, an algorithm for its optimization. The rest of this paper is organized as follows. In Section 2, we review related work; while in Section 3 we introduce preliminary material. Section 4 outlines the fundamentals of our model. In Section 5, we report experimental results and the final section offers concluding remarks and sheds light on future research directions.

2. Related works

Detecting overlapping communities is a task of a great importance in our world. Indeed, it is treated by several approaches which are reviewed and categorized into five classes that reflect how communities are identified (Xie et al., 2013).

* Corresponding author. Tel.: +216 95574879.
E-mail addresses: rh.delel@gmail.com (D. Rhouma), lotfi.ben.romdhane@usherbrooke.ca (L.B. Romdhane).

0957-4174/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2014.01.002

2.1. Graph theory partition

The first family is based on graph theory and the most popular technique in this approach is the Clique Percolation Method (CPM) proposed by Palla, Derenyi, Farkas, and Vicse (2005). It is based on the concept that the internal links in a community are likely to form cliques due to their high density. The main idea of this method is to move a clique on a graph, in some way, so it would probably be trapped inside its original community because it could not cross the bottleneck formed by the inter-community links. CPM is suitable for networks with dense connected parts. However K-clique cannot reach vertices with degree one (“leaves”) (Fortunato, 2010). In addition, it is very costly (Palla et al., 2005).

2.2. Link partition

Another line of research is link partition (Evans & Lambiotte, 2009; Ahn, Bagrow, & Lehmann, 2009; Wu, Lin, Wan, & Tian, 2010; Kim & Jeong, 2011). Indeed, it may happen that communities are joined to each other through their overlapped nodes without an inter-cluster edge. So it has been recently suggested to define community as sets of edges (Xie et al., 2013). The basic idea of Evans's method (Evans & Lambiotte, 2009) is to transform the original graph to a line graph i.e., each vertex in the line graph corresponds to an original edge and a link in the line graph represents the adjacency between two edges in the original graph. Nevertheless, this algorithm is memory inefficient (Tang, Wang, & Liu, 2012), so it cannot be applied to large social networks. Ahn and Al (Ahn et al., 2009) suggested an hierarchical clustering of links and computed the similarity between two links using Jaccard Index. The time complexity of this algorithm is $O\left(nk_{max}^2\right)$, where k_{max} is the maximum degree of node and $n$is the number of vertices in the network.

2.3. Local expansion and optimization partition

The idea of growing a partial community has also been explored (Xie et al., 2013). It relies on a fitness function characterizing the local quality of dense groups of nodes. Different overlapped groups can be locally optimal, so the vertices can be shared between communities. Baumes, Goldberg, Krishnamoorthy, Magdon-Ismail, and Preston (2005) proposed the iterative scan algorithm (IS) which starts with a candidate and adds or removes vertex as long as the function related to the density of link strictly improves. LFM (Lancichinetti, Fortunato, & KertTsz, 2009) develops a community from a random starting node until the objective function is not maximized. This method depends on a parameter that controls the size of formed groups. EAGLE (Shen, Cheng, Cai, & Hu, 2008) uses the agglomerative framework to produce a dendrogram. First, all maximal cliques are found and considered as first communities. Then the pair of communities with maximum similarity are merged. The optimal cut in the dendrogram is determined by the modularity. EAGLE is computationally expensive with complexity $O(n^2 + (h+n)s)$ (Xie et al., 2013), where s is the number of maximal cliques and h is the number of pairs of maximal cliques which are neighbors. GCE (Lee, Reid, McDaid, & Hurley, 2010) identifies cliques as seeds and expands them in greedy way. GCE also removes the communities that are similar using a function which computes the distance between communities. OSLOM (Lancichinetti, 2011) which is a multi-purpose technique, tests the statistical significance of a cluster with respect to a global null model during community expansion. Its main idea is to progressively add and remove vertices within the community so that to improve its fitness function. This process is repeated several times starting from different nodes in order to explore different regions of the graph. Its time complexity is $O(n^2)$. This family neglects communities of small sizes. An improvement of GCE and OSLOM is given by its conjunction, with WERW-Kpath (Fiumara, De Meo, Ferrara, & Provetti, 2013) algorithm which is a preprocessing step in which edges are weighted according to their centrality. This algorithm enhances the modularity and the quality of the community structure of these methods.

2.4. Fuzzy partition

The fourth approach is based on fuzzy clustering (Xie et al., 2013). It quantifies the strength of association between all nodes and communities and determines its adhesion to a group or not according to this degree. The most famous algorithm in this class is FCM which minimizes the intra-cluster variance by reducing its objective function (Bezdek, Ehrlich, & Full, 1984). FCM loses the graph structure because it takes into account only the distances between nodes. Nepusz, Petrczi, Ngyessy, and Bazs (2008) modeled the overlapping community detection as a nonlinear constrained optimization problem which can be solved by simulated annealing methods. NMF (Psorakis, Roberts, & Ebden, 2011) is a model based on Bayesian nonnegative matrix factorization. We may cite also OSBM (Latouche, Birmele, & Ambroise, 2011; Gregory, 2010), etc. Yet, these fuzzy approaches compute communities with spherical shapes mainly due to the constraints imposed on the membership degrees (Bezdek et al., 1984; Latouche et al., 2011; Psorakis et al., 2011). This is a major shortcoming since in real-networks, communities are of arbitrary shapes.

2.5. Agent-based partition

Finally, the Agent-based (Xie et al., 2013) approach uses labels to identify the membership of vertices and propagate it between neighbors, a node can have more than one label. In COPRA (Gregory, 2010), nodes update their belonging coefficients by averaging the coefficients from all its neighbors in a synchronous way. Its time complexity is $O(\log(\frac{vm}{n}))$ by iteration, where n is the number of vertices and m is the number of links and v is a parameter. SLPA (Xie, Szymanski, & Liu, 2011) spreads labels between nodes according to pairwise interaction rules and provides each node with a memory to store received information. Multi-state spin models (Reichardt & Born holdt, 2004) aim to minimize the equation of Hamiltonian. Despite, the high speed of this type of methods, they produce only small communities in some networks.

Despite the attempt of various methods to overcome the detection of overlapping communities, this problem still remains. Since it is an NP-hard problem (Fortunato, 2010) and some unstable nodes lying at the border between communities are often hard to classify into one community. Inspired by the above approaches, in this paper, a new Local Expansional algorithm called DOCNet, based on node fuzzy membership degree, proposed to detect the overlapping community structures. But, before going into its details, we need to introduce the following preliminary concepts.

3. Preliminaries

3.1. Problem formulation

We consider an undirected graph $G = (V, E)$, with $n = \mid V \mid$ nodes and $m = \mid E \mid$ edges. The purpose of the detection of overlapping communities in G is to determine a partition $P = \{C_1, \ldots, C_x\}$ of all the nodes of G where communities may be joined to each others (overlapped) ($\exists C_i \cap C_j \neq \emptyset,\ i \neq j$). A community may generally be described as group of nodes that probably share common properties and/or play similar roles within a network (Fortunato, 2010). It is a tight group with a high density of inter-community connections and a low density of intra-community connections (two communities can overlap since a node can belong to more than one).