Neurocomputing 207 (2016) 488–500

Contents lists available at ScienceDirect

Neurocomputing

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

LED: A fast overlapping communities detection algorithm based on structural clustering

Tinghuai Ma [a,b,*], Yao Wang [b], Meili Tang [c], Jie Cao [d], Yuan Tian [e], Abdullah Al-Dhelaan [e], Mznah Al-Rodhaan [e]

[a] CICAEET, Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210-044, China
[b] School of Computer & Software, Nanjing University of Information Science & Technology, Jiangsu, Nanjing 210-044, China
[c] School of Public Administration, Nanjing University of Information Science & Technology, Nanjing 210-044, China
[d] School of Economics & Management, Nanjing University of Information Science & Technology, Nanjing 210-044, China
[e] Computer Science Department, College of Computer and Information Science, King Saud University, Riyadh 11362, Saudi Arabia

ARTICLE INFO

Article history:
Received 29 May 2015
Received in revised form
3 February 2016
Accepted 8 May 2016
Communicated by Y. Chang
Available online 25 May 2016

Keywords:
Overlapping communities
Structural similarity
C-DBLP

ABSTRACT

Community detection in social networks is a fundamental task of complex network analysis. Community is usually regarded as a functional unit. Networks in real world more or less have overlapping community structure while traditional community detection algorithms assume that one vertex can only belong to one community. This paper proposes an efficient overlapping community detection algorithm named LED (*Loop Edges Delete*). LED algorithm is based on Structural Clustering, which converts structural similarity between vertices to weights of network. The evaluations of the LED algorithm are conducted both from classical networks from literature and C-DBLP, which is a huge and real-life co-author social network in China. The results show that LED is superior to other methods in accuracy, efficiency, comparing with FastModurity and GN algorithm.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Many dataset can be represented as networks. Networks are usually modeled as a graph $G(V, E)$, where V is a set of vertices and E is a set of edges. For example, an online social network can be modeled as a graph, where users are represented as vertices that are connected by an edge when one user follows another. Community structure [1] is a common property of many networks. In terms of our experience on online social networks, it is a common sense that people with same interests trending to get together as communities: subsets of vertices within which vertex to vertex connections are dense, but between which connections are sparse [2,3,5,6]. Also, some active members of social network may take part in many communities simultaneously, which are regarded as overlapping vertices in the network.

Network clustering (or *graph partitioning*) aims at detecting community structure. Community detection is a useful tool to mine the hidden information from networks. Xu et al. [7] proposed a community forest model which assumes a community as a tree. A tree consists of trunks and leaves, while a community consists of vertices and relations. Strong relations are like a tree trunk, some core vertices are like tree roots and some border vertices are like leaves. Community detection is just like finding trees in a forest. Overlapping is a common phenomenon in networks; overlapping community detection takes this factor into count detecting overlapping or hierarchical communities.

Nowadays we are in the era of information explosion, networks are getting larger and larger. We need much more efficient communities detection algorithms to analyze network with millions of vertices. In this paper, we propose a new method for very large networks which have linear time complexity. The time complexity of our algorithm is linear, only a few state of the art algorithms [9,10] can do this. The goal of our algorithm is to find communities and overlapping vertices in very large networks. To achieve this goal, our algorithm must be fast and precise. Our algorithm is inspired by density-based algorithm DBSCAN [4]. We transfer the structural relationship between vertices to their weight by computing their structural similarity. Structural similarity represents connection strength of two vertices, because it counts all common neighbors the two vertices share. It is a common sense that two people who share many friends should be clustered in the same community. We delete the edges of small weights to disconnect vertices which should not be clustered in the same community. From the deleted edges, the initial communities are formed. Our algorithm only needs to traverse all the edges of the network for once, calculating the structure similarity of the edge's two vertices

* Corresponding author.
E-mail address: thma@nuist.edu.cn (T. Ma).

http://dx.doi.org/10.1016/j.neucom.2016.05.020
0925-2312/© 2016 Elsevier B.V. All rights reserved.

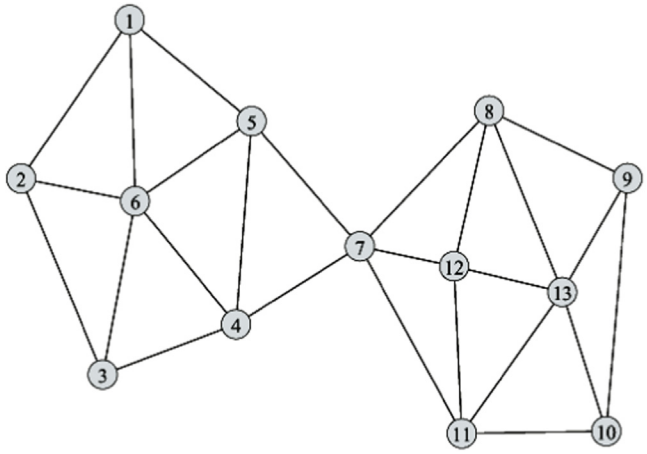

Fig. 1. A network with two communities and one overlapping vertex.

and deleting the edges with structure similarity below the threshold α (manually specified).

In many real world networks, the overlapping communities are universal [8,25]. For example, the network in Fig. 1, it is a simple network, so we can artificially determine that it consists of two communities and one overlapping *vertex*7. It is divided into two communities by GN algorithm [11]. Vertices from *vertex*1 to *vertex* 6 are in one community, the rest of vertices are within another community. Infomap algorithm [12] gives exactly the same partition. Apparently, the *vertex*7 connects to the two communities with almost the same strength. *Vertex*7 should be regard as member of two communities at the same time. Algorithms we mentioned before do not have the ability to identify overlapping vertices while we extend the ability to LED. The overlapping vertex often plays an important role in the network. As we all know, community is regarded as functional unit in network. The overlapping vertices promote the interaction between functional units. In real networks, overlapping vertex is useful in web search engine optimization [13,16], viral marketing [14], epidemiology [15] and so on.

The paper is organized as follows. In Section 2, we introduce the related work of overlapping community detection. Section 3 describes the algorithm we propose, including Structural Similarity, Loop Edges Delete Process, Overlapping Vertices Detection and LED Algorithm. In Section 4 we analysis the time complexity of our algorithm in theory and the selection of parameter. In Section 5, we run our algorithm on C-DBLP and other real world social networks. Finally, we present our conclusions and future work in Section 6.

2. Related work

In this section we survey related works firstly and then give a toy motivating example.

2.1. Related work

Community detection is a fundamental task of complex networks analysis. It has great contribution to other studies such as recommender system [17] and privacy protection [18]. A great deal of work has been done to detect communities in complex networks; they can be categorized into two classes according to whether allow overlapping: disjoint community detection and overlapping community detection.

In early studies of community detection, researchers rarely took the overlapping community into their definition of community. An important kind of disjoint community detection algorithm is poposing a measure of quality of community and then using a greed algorithm to maximize the measure. The GN algorithm [11] which is probably the best-known algorithm for finding community structure is based on the betweenness centrality measure [26]. Structure entropy [24] is an information theoretical measure of complexity of networks. Eustace et al. [8,25] proposed local neighborhood ratio to define community structure. They all have similar greed algorithms but with different measures. CONGA [23], proposed by Steve Gregory, is based on Girvan and Newman's algorithm [11]. On GN algorithm's basis, CONGA extends a novel method of splitting vertices for allowing overlapping communities. It performs well on randomly generated networks and has interesting results on a range of real-world networks. It is based on an old algorithm, so it is not efficient enough for large networks. One year later, Steve Gregory proposed a new algorithm named CONGO [27] (*CONGA Optimized*). GN algorithm and CONGA both relay on betweenness which is a global centrality measure. In CONGO, Steve Gregory refined the global betweenness to a local betweenness which greatly reduced the time complexity.

In recent years, overlapping community detection has attracted more and more attention. The area of overlapping community detection is the hot area currently. Clique percolation method (*CPM*) [19] may be the most famous and widely used overlapping community detection algorithm which is based on graph theory [20]. CPM claims "The basic observation on which our community definition relies is that a typical community consists of several complete (*fully connected*) subgraphs that tend to share many of their nodes." CPM finds *k*-cliques (*k fully connected subgraphs*) and the *k*-cliques's series of adjacent *k*-cliques (*where adjacency means two k-cliques sharing* $k-1$ nodes) to detect overlapping community. However, CPM has many problems. It has an inappropriate community definition to detect overlapping community for different kinds of networks. Although CPM can adapt to different networks by selecting a different *k*, it has only limited effects. When the network is too dense, CPM detects giant communities. When the network is too sparse, it finds no communities at all. Also it is a very time-consuming algorithm, in theory, its time complexity is $O(n^3)$.

Yong-Yeol et al. [21] considered community to be a set of closely connected links instead of vertices with many links between them. The link communities incorporate overlapping vertices as vertices can belong to many links of different communities. In general, the number of links is much more than the number of vertices, so it is a more time-consuming algorithm. Brian et al. [22] proposed a probabilistic model of link communities to detect communities, either overlapping or not, and used a fast, closed-form expectation–maximization algorithm to analyze networks of millions of vertices in reasonable running time.

Many attempts were made to detect overlapping communities. Du et al. [28] use the maximal cliques for community detection. The algorithm finds all the maximal cliques as the initial communities, then assigns the rest vertices to the closest initial community. EAGLE [29] and GCE [30] also work this way. Our algorithm's main process is similar to theirs, but we do not use cliques to get initial communities. BELP, proposed by Chen et al. [33], is based on SLPA [34] (*Speaker–listener Label Propagation Algorithm*) which is an unstable algorithm. For edges connecting two communities, it uses average degree to make decision for which community the vertices should belong to. This strategy inspires us. We use a similar strategy to identify overlapping vertices. Embedding is another popular way to perform community detection. SSDE-Cluster [31] first embeds the graph and then performs a metric clustering using a Gaussian Mixture Model (*GMM*). Creating an approximate embedding is useful for clustering in metric spaces and GMM can be adapted to get overlapping

Download English Version:

https://daneshyari.com/en/article/494478

Download Persian Version:

https://daneshyari.com/article/494478

Daneshyari.com