Physica A 407 (2014) 380–393

Contents lists available at ScienceDirect

Physica A

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

A local multiresolution algorithm for detecting communities of unbalanced structures

Krista Rizman Žalik *, Borut Žalik

University of Maribor, Faculty of Electrical Engineering and Computer Science, Slovenia

HIGHLIGHTS

- We propose a fast local expansion algorithm for community detection named LSE.
- It uses quality function that enables detecting communities independently of their sizes that can be of different densities.
- The proposed algorithm can detect multiresolution community from a source vertex.
- The experimental results verify that LSE can uncover rich information on networks.

ARTICLE INFO

Article history:
Received 26 March 2013
Received in revised form 5 February 2014
Available online 27 March 2014

Keywords:
Modularity
Objective function
Community detection
Dense subgraphs
Networks

ABSTRACT

In complex networks such as computer and information networks, social networks or biological networks a community structure is a common and important property. Community detection in complex networks has attracted a lot of attention in recent years. Community detection is the problem of finding closely related groups within a network. Modularity optimisation is a widely accepted method for community detection. It has been shown that the modularity optimisation has a resolution limit because it is unable to detect communities with sizes smaller than a certain number of vertices defined with network size. In this paper we propose a metric for describing community structures that enables community detection better than other metrics. We present a fast local expansion algorithm for community detection. The proposed algorithm provides online multiresolution community detection from a source vertex. Experimental results show that the proposed algorithm is efficient in both real-world and synthetic networks.

1. Introduction

Complex networks commonly hide a community structure, while groups of vertices of networks are more densely connected to each other than to other vertices [1]. A precise common definition of communities is still lacking. Many methods have been proposed using different definitions. The problem is made more difficult because in reality these communities often overlap, such that each vertex may occur in more than one community and communities can have different hierarchical orders.

Different community detection methods have been proposed that seek to identify natural groups of related vertices within networks (for review see Ref. [2]). Two distinct types of identified communities are possible. In crisp communities each vertex belongs fully to one community of which it is a member and no overlapping is identified. In overlapped communities each vertex belongs to more communities to a different extent.

* Corresponding author.
E-mail address: krista.zalik@um.si (K. Rizman Žalik).

http://dx.doi.org/10.1016/j.physa.2014.03.059

Community detection is the process of finding dense groups on graphs such as web-pages having the same topics, substances within the metabolic, chemical or biological networks or groups of friends within social networks. Community members have more common properties amongst themselves than with non-members of the community and the identification of community structure can help when analysing the functionalities of networks [3]. Each dense group within a network is expected to match-up to the individual functions performed by that network, such that functions can be inferred from topology.

Many complex systems in the real-world may be represented as networks. A network can be modelled using a graph $G(V, E)$ with a set of vertices $V = (v_1, v_2, \ldots, v_n)$ that models objects and a set of edges E that models relationships between objects. Edges connect pairs of vertices that interact within the complex metabolic, biological or social systems or they are just similar enough, or connected as in citation networks. Individual vertices are connected with edges which describe a certain degree of interaction between vertices. Communities are defined as natural groups of densely interconnected vertices that are only sparsely connected to the rest of the network. Community detection methods find the partition $P = C_1 \bigcup C_2 \bigcup \cdots \bigcup C_m$ where $C_1, C_2, \ldots, C_m$ are sets of vertices called clusters or communities that can be disjointed or can overlap.

A brief overview of efforts towards efficient community detection follows. Different measures have been proposed to measure the quality of division of a network into communities. Global criterion requires knowledge of the entire network.

Since global methods are inappropriate when working with large and dynamic networks or when real-time results are expected, several techniques have been proposed to identify local community structure. The most popular methods based on the optimisation of quality function named modularity Q_N was introduced by Newman and Girvan [4]. Newman reformulated modularity in terms of eigenvectors of a new characteristic matrix for the network and obtained a time complexity $O(n^2 \log n)$ for sparse graphs. (See Ref. [5] for details.)

The idea of modularity Q is that vertices linked in a random way should not form communities. Modularity Q_N measures the fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random:

$$Q_N = \sum_{i=1}^{m} \left(\frac{l_i}{L} - \left(\frac{d_i}{2L} \right)^2 \right) \tag{1}$$

where l_i is the total number of internal edges for subgraph C_i and the total number of edges in the subgraph C_i is d_i: $d_i = \sum_{j \in C_i} k_j$; k_j is the total number of edges connected with a vertex j, and L is the total number of edges in graph $G : L = \frac{1}{2} \sum_{j \in G} k_j$. The bigger the Q_N the better the partition. Q_N quantifies the difference in the density of the internal edges from the expected edge density in an equivalent network with a similar number of vertices, that are connected randomly. In a random graph each vertex can be connected to any other vertex of the network. However in very large networks the number of edges between two groups decreases and a single edge between the two groups can represent a strong correlation between two groups that can result in merging of these two groups. Optimising modularity Q_N within large networks fails to detect small communities. It has been shown that modularity has a resolution limit as it cannot detect communities with sizes smaller than a threshold, depending on the network size [6]. An algorithm for the optimisation of modularity measure Q_N needs to be a global method that requires complete knowledge of the entire network (total number of edges L). Many networks are large such as huge social networks or the Internet, and their sizes grow over time. It is impossible for such networks to know the entire network and its global structure. Therefore such global community detection methods are inappropriate for handling large real-world networks. Local community detection methods that can be used for only part of a network are necessary.

Several methods have been proposed recently using local edge structure. Clauset [7] has proposed a local modularity measure R for the sharpness of a community's border B. He has proposed an algorithm that infers a hierarchy of vertices that enclose a given vertex by exploring a graph one vertex at a time. The algorithm adds the neighbouring vertex that results in the largest increase or smallest decrease in R to current community. This process continues until it has agglomerated either a given number of vertices k or the increase of R is below some threshold.

$$R = B_{\text{in}}/(B_{\text{in}} + B_{\text{out}}) \tag{2}$$

where B_{in} is the number of edges that connect the boundary vertices with other vertices in a community C and B_{out} is the number of edges that connect the boundary vertices with those vertices from graph G that is not in community C. Communities are detected by local optimisation of a local modularity metric that only considers vertices on the boundary of a subgraph.

A divisive heuristic based on modularity maximisation that is locally optimal in the sense that each of the successive bipartition is done in a provably optimal way has already been proposed [8]. The proposed heuristics provides better results than the agglomerative heuristic of Clauset et al.

Bagrow et al. [9] proposed a method for detecting local communities. It spreads an l-shell outward from a starting vertex. In each step all neighbour vertices of vertices included in the shell are added. l is the distance from the starting vertex to all shell vertices. The process of spreading stops when the total emerging degree tends to increase more than some given threshold and a community is found. The emerging degree of a vertex is defined as the number of edges that connects that vertex to vertices the l shell has not already visited as it expanded from previous $l-1, l-2 \ldots$ shells. This approach works