Discrete Applied Mathematics 176 (2014) 89–99

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

Community detection in networks via a spectral heuristic based on the clustering coefficient

Mariá C.V. Nascimento *

Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, Rua Talim, 330 - Vila Nair - São José dos Campos/SP CEP: 12231-280, Brazil

ARTICLE INFO

Article history:
Received 28 September 2012
Received in revised form 26 August 2013
Accepted 27 September 2013
Available online 21 October 2013

Keywords:
Community detection in networks
Spectral heuristic
Clustering coefficient
Unweighted graphs
Graph clustering

ABSTRACT

The community detection problem in networks consists of determining a clustering of "related" vertices in a graph or network. Nowadays, studies involving this problem are primarily composed of modularity maximization based heuristics. In this paper, the author proposes a spectral heuristic based on a measure known as clustering coefficient to detect communities in networks. This measure favors clusterings with a strong neighborhood structure inside clusters, apparently, overcoming the scale deficiency of the modularity maximization problem. The computational experiments indicate a very successful performance by the proposed heuristic in comparison with other community detection heuristics in the literature.

1. Introduction

The community detection problem in networks aims at finding partitions of a graph with clusters with "related" vertices. Its nomenclature derives from the sociological problem of detecting communities in social networks in order to explain similar behaviors among groups of individuals. As a consequence, the representation of social networks as graphs (or networks) enables their analysis by using graph clustering methods. Apart from sociology, there are numerous other areas where the community detection problem might be suitable, for example, biology [14].

Closely related to the community detection problem is the graph partitioning problem, whose goal is to find either balanced or equal-sized clusters [17,2]. The main difference between the community detection problem and the graph partitioning problem is that the former aims at clustering a graph according to the connections between the vertices of its clusters, where the vertices are not necessarily equally distributed among the clusters. For this reason, these two problems have been analyzed using different trends and applications. In certain applications like in the VLSI domain [17], the constraint of the equal-sized cluster is required.

There are a few measures that evaluate the clustering tendency of the vertices inside the clusters of a graph partition, such as modularity [24]. This measure, proposed by [14], evaluates the clustering tendency of graph partitions by using statistical network concepts. Apart from the edge betweenness algorithm [14], the label propagation algorithm [28] and a few other community detection algorithms, heuristics based on the modularity maximization problem are, nowadays, frequently used to detect clusters in networks.

Although very effective, the modularity measure has some limitations, as pointed out by [13,27]. The resolution limit that will be addressed in this paper in the next sections is possibly the worst case scenario for modularity maximization based algorithms. Summarizing, for small sized clusters, modularity maximization based algorithms seem to provide a less strong

* Tel.: +55 12 33099595; fax: +55 12 3309 9500.
E-mail address: mcv.nascimento@unifesp.br.

0166-218X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dam.2013.09.017

cluster than expected. For this reason, new studies and alternatives for the community detection problem in networks are desirable and relevant for this research topic.

A measure specially designed to assess the clustering tendency of the vertices in a graph is known as clustering coefficient. It was originally proposed by [35] for unweighted graphs in order to verify whether a given vertex in a graph had high cluster tendency in comparison with its neighboring vertices. This measure is often applied to graphs whose experiments require the analysis of their vertices' main characteristics. For this purpose, the clustering coefficient has been widely used, particularly, for making inferences about the clustering tendency of the vertices from some complex networks. In this paper, the clustering coefficient will be used as the basis to derive a novel evaluation measure for graph clustering partitions. Based on this new measure, a spectral heuristic for finding clusterings by maximizing the measure, here named SPECtral Clustering Coefficient (SPEC^3), is proposed here.

This paper is organized as follows. Section 2 presents a brief literature review on the topic of community detection in networks with problems related to the adopted point of view. Section 3 presents a concise description of the clustering coefficient measures found in the literature. Section 4 introduces the proposal of the assessment measure for communities in networks based on the clustering coefficient measure. Section 5 describes the proposed spectral heuristic for finding partitions with the maximum clustering coefficient. Section 6 presents the computational experiments which compare the proposed spectral heuristic with other heuristics in literature. To sum up, Section 7 concludes the paper with some final remarks.

2. Related work

Before presenting the studies related to the performed investigation found in the literature, some graph notations that will be used throughout the paper are presented. Let $G = (V, E)$ be a graph with n vertices, represented by the set $V = \{1, 2, \ldots, n\}$ and m edges, represented by tuples $(i, j) \in E$ where $i, j \in V$. Let G be an undirected graph and consider a complete set as a pairwise adjacent vertex set. A clique is defined as an inclusion-wise maximal complete set [6]. Furthermore, a *maximal clique* is defined as a clique that cannot be enlarged by including one more adjacent vertex, which implies that it is not a subgraph of a larger clique.

Graphs are combinatorial structures that may represent a wide number of real systems [15]. Regarding these systems, Watts and Strogatz [35] started to investigate the topology of these graphs, sometimes called complex networks, in order to extract useful information from them. In this sense, the community detection problem appears as an appealing issue since finding clusters of nodes that are highly related may provide interesting properties with respect to the "particles" of the systems. Additionally, it provides an easier view of the network, since studying individual vertices (microscopic structures) may not be as informative as analyzing a group of individuals (mesoscopic structure).

The community detection problem has been the focus of a large amount of research in the last decade due to advances on this topic by [14]. In their study, a measure called modularity was derived in order to evaluate the quality of the communities (partitions or clusterings) found through graph clustering algorithms. Given a partition with κ groups $\mathcal{C} = \{C_1, C_2, \ldots, C_\kappa\}$, its modularity can be defined as:

$$Q(\mathcal{C}) := \frac{1}{2m} \sum_{C \in \mathcal{C}} \sum_{i,j \in C} \left[a(i,j) - \frac{d(i)d(j)}{2m} \right], \tag{1}$$

where $a(i, j)$ indicates the number of edges between vertices i and j and $d(i)$ refers to the degree of vertex i.

It can be noticed that $\sum_{i,j \in C} \left[a(i,j) - \frac{d(i)d(j)}{2m} \right]$, for each $C \in \mathcal{C}$, represents the difference between the number of edges inside cluster C and the expected number of edges inside this cluster in a random graph with the same degree sequence as G (null model). Another formulation for the modularity measure can be found in [22].

In order to find communities in networks, integer programs that aim at finding the partitions with the maximum modularity have been successfully used [8,1,7]. Since the decision version of the modularity maximization problem is NP-complete [7] and graph clustering instances are usually large, exact algorithms have not been thoroughly explored for this problem. As a result, an overwhelming amount of research towards heuristics can be observed for this study topic, like, for example, spectral clustering algorithms [23]. Spectral clustering algorithms are attractive solution methods due to their speed and good quality of the results they produce [23].

Although very effective for finding clusterings, one may find some drawbacks regarding the modularity maximization based algorithms. One of the most serious drawbacks is possibly the fact that clusters smaller than a scale are not identified by these algorithms. In this line of study, [13] performed a detailed investigation after which they concluded that some modules, which clearly represent clusters, could not be identified by modularity based frameworks. A standard example which precisely presents this fact is a graph with κ cliques, where each one is connected to two other cliques by a bridge. In this case, one would have a ring graph, as exemplified in Fig. 1, where it is possible to observe a 30-clique ring graph with five vertices each.

Considering a ring with κ cliques, [13] derived that the partition with the maximum modularity must have no more than $\sqrt{m}$ clusters. For a ring graph like the one previously mentioned, one may observe that the number of edges inside each cluster would be dependent on the number of vertices in each cluster, n_c. In this case, the number of edges inside each cluster would be $\frac{n_c(n_c-1)}{2}$, and the whole graph would have $\kappa \frac{n_c(n_c-1)}{2} + \kappa - 1$ edges. Therefore, for the ring graph from Fig. 1, the value for m is 329. Although this graph possesses 30 natural clusters, the partition with maximum modularity would have