Information Sciences 462 (2018) 290–314

Contents lists available at ScienceDirect

Information Sciences

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

A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding

Gema Bello-Orgaz[a], Sancho Salcedo-Sanz[b], David Camacho[a,*]

[a] *Computer Science Department, Universidad Autónoma de Madrid, Madrid, Spain*
[b] *Department of Signal Theory and Communications, Universidad de Alcalá, Madrid, Spain*

ARTICLE INFO

Article history:
Received 27 December 2017
Revised 4 June 2018
Accepted 7 June 2018

Keywords:
Multi-Objective genetic algorithms
Overlapping community detection
Graph clustering
Network metrics
Edge-based encoding

ABSTRACT

The Community Detection Problem (CDP) in Social Networks has been widely studied from different areas such as Data Mining, Graph Theory Physics, or Social Network Analysis, among others. This problem tries to divide a graph into different groups of nodes (communities), according to the graph topology. A *partition* is a division of the graph where each node belongs to only one community. However, a common feature observed in real-world networks is the existence of *overlapping communities*, where a given node can belong to more than one community. This paper presents a new Multi-Objective Genetic Algorithm (MOGA-OCD) designed to detect overlapping communities, by using measures related to the network connectivity. For this purpose, the proposed algorithm uses a phenotype-type encoding based on the edge information, and a new fitness function focused on optimizing two classical objectives in CDP: the first one is used to maximize the internal connectivity of the communities, whereas the second one is used to minimize the external connections to the rest of the graph. To select the most appropriate metrics for these objectives, a comparative assessment of several connectivity metrics has been carried out using real-world networks. Finally, the algorithm has been evaluated against other well-known algorithms from the state of the art in CDP. The experimental results show that the proposed approach improves overall the accuracy and quality of alternative methods in CDP, showing its effectiveness as a new powerful algorithm for detecting structured overlapping communities.

1. Introduction

The Community Detection Problem (CDP) has been the subject of many recent studies in the field of Data Mining and Social Network Analysis [8]. It is an important problem with application in disciplines such as Sociology, Biology, Neuroscience, or Computer Science, whose information can be easily represented using connected networks or graphs. In Computer Science, the unsupervised process of identifying the underlying structure of the data, in terms of grouping the most similar elements, is called clustering. Therefore, graph or network clustering [15] can be understood as the process of grouping the vertices of the graph into clusters considering the graph structure. The goal of CDP is to find optimal groups of nodes, or *communities*, depending on the graph topology. Note that the CDP is very similar to a graph clustering problem in graph theory.

* Corresponding author.
E-mail addresses: gema.bello@uam.es (G. Bello-Orgaz), sancho.salcedo@uah.es (S. Salcedo-Sanz), david.camacho@uam.es (D. Camacho).

https://doi.org/10.1016/j.ins.2018.06.015
0020-0255/© 2018 Published by Elsevier Inc.

There is not an unique (fully accepted) definition of *community* in graph theory. There are, however, several variants used in the literature. Despite of this ambiguity, different graph clustering methodologies have been used to identify communities such as *Random Walks, Spectral Clustering, Modularity Maximization* or *Statistical Mechanics* among others [15]. Note that many of these algorithms are typically based on the topology information of the graph or network considered.

Regarding to the graph connectivity needed to define good quality solutions for the CDP, each cluster should be connected. This means that there always should exist paths connecting each pair of vertices within the cluster. It is generally accepted that a subset of vertices forms a good cluster if the induced sub-graph is *dense*, and there exist *connections* from the included vertices to the rest of the graph [26]. Considering both features (density and connectivity) a possible definition of a graph cluster could be a *connected component* or a *maximal clique*, i.e., a type of sub-graph in which no vertex can be added without losing the clique property.

Several graph clustering techniques are focused on finding disjoint communities. In this case, the network or graph is partitioned into dense regions, in which nodes have more connections to each other than to the rest of the network. However, it is not always clear that a vertex should be only assigned to a given cluster. On the contrary, in different real domains it is necessary that a vertex belongs to various clusters (this property is generally known as overlapping in CDP) [45]. For instance, people in a social network are usually members of multiple communities. Therefore, the overlap is a significant feature for many real-world networks. To solve this problem, fuzzy clustering algorithms have been applied to different graphs partition problems [40,49], and several overlapping approaches [9,35,45,46], have also been proposed. Note that all these families of algorithms show a high computational complexity in networks of very large size.

In order to obtain improved algorithms for the CDP (commonly known as Community Detection Algorithms (CDAs)), approaches based on evolutionary computation, such as Genetic Algorithms (GAs), have been designed. Several of these evolutionary graph clustering algorithms use a single optimization criteria as objective function, such as the modularity. However, there are also other GA-based approaches where the community detection is solved as a multi-objective optimization problem, generally using two criteria to be optimized [20,37].

Most of the GA-based approaches for CDP are based on the idea of node clustering. These algorithms commonly use an encoding scheme where the individuals represent the nodes belonging to the input graph [9,20]. However, recent studies have suggested that the use of edge information for partitioning the graph into clusters of edges is a good strategy. Therefore, new encodings based on a edge representation of the individuals have been proposed to identify overlapping communities using evolutionary approaches [36,42]. Particularly, these methods are quite suitable for real world networks, which tend to be sparse, and where the node-based methods often have difficulties to find large communities.

Another issue related to CDAs is derived from the fact that these algorithms are mainly based on the topology information of the graph to partitioning it. Due to this, it is intuitive that the performance of a CDA algorithm is very dependent on the graph structure. For example, the algorithms based on the greedy optimization of modularity measurement tends to form large communities rather than small ones, which often results in poor values of modularity. Therefore, these type of algorithms are not usually able to achieve good results for sparse graphs with small communities.

This paper presents a new Multi-Objective Genetic Algorithm (MOGA-OCD) for detecting overlapping communities, designed in such a way that it can solve some of the issues discussed before: The proposed algorithm uses a phenotype encoding, based on the edge information to represent the solutions. Also, it considers different connectivity measures from the network topology as fitness functions to guide the searching process, such as Density, Triangle Participation Ratio, Clique Number, Clustering Coefficient, Expansion, Separability and Cut Ratio. A comparative assessment of the previous measures (used as an optimization criteria) has been carried out in order to identify the most suitable metrics for partitioning the graph into communities. Regarding the experimental evaluation, the proposed algorithm has been tested using several real-world social networks, and its results have been compared against a subset of well-known algorithms from the state of the art in CDPs.

The rest of the paper has been structured as follows: Section 2 describes the previous work related to graph clustering, GAs, and community detection algorithms. Section 3 presents the proposed MOGA-OCD algorithm, the proposed encoding used, and the fitness functions implemented. Section 4 provides a description of the dataset used, the experimental setup of the algorithm, and a complete experimental evaluation against several CDP algorithms. In Section 5 some final remarks and future research lines are presented.

2. Background

Graph clustering [15] can be defined as the task of grouping the vertices of a graph into clusters, or communities, considering information from its internal structure. One of the first tasks in graph clustering is to look for a quantitative definition of community (or cluster), because that definition often depends on the specific application domain. However, and according to the nature of the considered problem, it is generally accepted that there should be more edges within each community than edges linking to the rest of the graph. Therefore, it is possible to find several definitions related to a *good community* within a graph in the literature. A large number of methodologies have been applied to solve this problem, most of them are typically based on the topology information of the graph, and in general terms, there are two main approaches to deal with CDPs [41]: ***partitional*** and ***overlapping*** (or non-exclusive) approaches. In this section we first present a general introduction to partitional techniques, which perform a disjoint division of the data, where each element belongs only to a single