Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 90 – 99

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Community detection by consensus genetic-based algorithm for directed networks

Stefano B. B. R. P. Mathias[a], Valério Rosset[a], Mariá C. V. Nascimento[a,*]

[a]*Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo (UNIFESP)*
Av. Cesare M. G. Lattes, 1201, Eugênio de Mello, São José dos Campos-SP, CEP: 12247-014, Brasil

Abstract

Finding communities in networks is a commonly used form of network analysis. There is a myriad of community detection algorithms in the literature to perform this task. In spite of that, the number of community detection algorithms in directed networks is much lower than in undirected networks. However, evaluation measures to estimate the quality of communities in undirected networks nowadays have its adaptation to directed networks as, for example, the well-known modularity measure. This paper introduces a genetic-based consensus clustering to detect communities in directed networks with the directed modularity as the fitness function. Consensus strategies involve combining computational models to improve the quality of solutions generated by a single model. The reason behind the development of a consensus strategy relies on the fact that recent studies indicate that the modularity may fail in detecting expected clusterings. Computational experiments with artificial LFR networks show that the proposed method was very competitive in comparison to existing strategies in the literature.

Peer-review under responsibility of KES International

Keywords: genetic algorithms; consensus clustering; modularity; directed networks.

1. Introduction

Many elements in our daily life such as the internet, the transportation systems, the city mapping, among others, can be represented as relational structures. Graphs or networks are among the many forms to represents such data. Extracting relevant information from networks is of utmost importance and, for this, there are specific tools. The identification of patterns in the data may enable the adequate analysis of these graphs. In this sense, if the graph topology characterizes a clustering tendency, there exists a number of works that attempt to identify groups of highly related vertices (communities). Finding communities (also known as clusters) in networks allows a different type of inference with regard to the information about the elements of the groups instead of the individual elements. Examples of graphs with these characteristics are those represented by a social network data, such as the widely known online social network, the *facebook*.

* Corresponding author.
E-mail address: mcv.nascimento@unifesp.br.

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.103

Finding groups of vertices in a graph is known in the literature as the graph clustering problem or community detection in networks. This problem has been investigated for decades and currently there are several approaches to solving it. Among them, we can highlight those that optimize assessment measures which classify the quality of the solutions found in [1,2,3]. Most of the problems of finding the communities that optimize quality measures is $\mathcal{NP}$-complete [4]. Consequently, the heuristics are the most explored methods for this problem, due to the sizes of the networks that represent most of the applications. In particular, to address this problem in directed networks remains a challenge due to the few existing algorithms that take the arcs directions into account. It is considerably difficult to define a consistent measure for evaluating the quality of the communities in directed networks. However, some assessment measures, as the modularity in directed networks [5] and the *map equation* proposed in [2], attempt to evaluate the quality of communities considering the asymmetric relations in the networks.

This paper presents a novel strategy for detecting communities in directed networks through a consensus genetic-based algorithm. As the modularity poses as a good measure despite the resolution limit drawback, the consensus strategy has this measure as objective function (or fitness function). The resolution limit in the modularity refers to the size of the clusters and the number of arcs in the networks. For networks with small-sized communities, the modularity maximization-based algorithms may not correctly identify the groups [6]. An advantage of developing a consensus-based strategy is that it enables the search for partitions with different traits than those found in algorithms that specifically maximize this measure.

Computational experiments carried out using artificial and real networks indicate a very good potential of the proposed strategy. The combination of a consensus strategy considering a fine tuning of the resolution parameter and the evolutionary traits found in genetic algorithms contributed in the robustness of the proposed strategy. The results achieved by the consensus genetic-based algorithms outperformed heuristics widely employed found in the literature.

2. Community detection in directed networks

A simple digraph, oriented graph, can be defined by a tuple $D = (V(D), A(D))$, in which $V(D)$ represents its set of vertices and $A(D)$, its set of arcs. In this study, the notations of the sets of vertices and arcs will be simplified to V and A, respectively. The set V has its elements represented by the numbers belonging to the set of integers $\{1, 2, 3, ..., |V|\}$. The numbers of vertices and arcs in a digraph are represented by, respectively, n and m. Moreover, each element of the set of arcs A is a tuple (i, j), being i the end corresponding to the source of the arc and j, the terminal vertex of the arc. In this case, we say that the vertices i and j are adjacent and that arc (i, j) is different from arc (j, i). The number of times a vertex i is a source vertex defines the out-degree of a node (d_i^+) whereas the number of times it is a terminal indicates the in-degree of a node i (d_i^-). A vertex i is said a neighbor of j if they are adjacent.

Digraphs are structures extensively employed to represent relational data as, e.g., the web links of the internet. In this example, each vertex of the corresponding network refers to the site and the links between a pair of sites can be represented by arcs. The representation of a social network, as, e.g., *twitter* social network, can be performed by associating each user with a vertex of the graph whereas the arcs may indicate the dominance between the pair of vertices: if there exists an arc (i, j), it means that the individual i follows the individual j inside the social network. Another example is the data that represent a transportation network that can also be represented by a directed graph. For this, simply consider each vertex as a city and the arcs as the streets and roads that connect these cities [7]. Detecting communities in this type of networks is one of the most difficult problems in clustering. Since the arcs do not necessarily possess symmetric relations, the communities may have different interpretations, that may not take into account the reciprocity of the arcs. Consequently, many consolidated formulations and algorithms for detecting communities in undirected networks may not be an option for considering the directed networks. In [3], the authors thoroughly discuss the existing methods for finding clusters in directed networks and mention as a gap in the literature methods specially designed for this type of networks.

In [3], the authors propose a categorization of these clustering types into distinct groups. These groups are the naive transformation approach, the transformations maintaining directionality, the extending objective functions and methodologies for directed networks and the other alternative approaches. For being beyond the scope of this paper to go into detail about all these approaches, we draw the attention to the strategies based on the paradigm of extending methodologies for tackling directed networks, within which the proposed strategy fits.

In line with our proposal, a way to define clustering in directed networks is by assessing the network topology, i.e. the amount of arcs inside cluster must be greater than the number of existing links between the clusters. This