Expert Systems With Applications 54 (2016) 121–135

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A consensus graph clustering algorithm for directed networks

Camila Pereira Santos, Desiree Maldonado Carvalho, Mariá C.V. Nascimento*

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 12247-014, Brazil

ARTICLE INFO

Keywords:
Heuristic method
Community detection
Directed networks

ABSTRACT

Finding groups of highly related vertices in undirected graphs has been widely investigated. Nevertheless, a very few strategies are specially designed for dealing with directed networks. In particular, strategies based on the maximization of the modularity adjusted to overcome the resolution limit for directed networks have not been developed. The analysis of the characteristics of the clusters produced by these approaches is highly important since among the most used strategies for detecting communities in directed networks are the modularity maximization-based algorithms for undirected graphs. Towards these remarks, in this paper we propose a consensus-based strategy, named ConClus, for providing partitions for directed networks guided by the adjusted modularity measure. In the computational experiments, we compared ConClus with benchmark strategies, including Infomap and OSLOM, by using hundreds of LFR networks. ConClus outperformed Infomap and was competitive with OSLOM even for graphs with high mixture index and small-sized clusters, to which modularity-based algorithms have limitations. ConClus outperformed all algorithms when considering the networks with the highest average and maximum in-degrees among the networks used in the experiments.

1. Introduction

Detecting communities in networks, also known as graph clustering, plays an important role in pattern recognition research area. Roughly, it enables the identification of groups of highly related vertices in a graph, also known as clusters. It is a relevant issue, for example, to look into the communities that represent the functional activities of the brain, known as brain networks (Park & Friston, 2013). One reason is that, in some surgeries, this knowledge might enable a better assessment about the areas of the brain related to motor skills. Regardless the distance between two regions of the brain, they might be strongly related according to the functional activities.

In spite of most community detection strategies being designed for undirected networks, several applications to which community detection is highly relevant are better modeled in directed networks. We may cite, for example, social, informational, biological and neuroscience networks. For defining communities in these networks, the most employed approach consists in ignoring the arc directions of the networks to make use of strategies designed for undirected graphs.

However, Malliaros and Vazirgiannis (2013) point out that important characteristics of the network might be lost with this approach, the reason why arc directions should be considered. One reason is the non-existence of reciprocal relationship between vertices, created after ignoring the arc directions. For example, in citation networks, networks of scientific papers, the links are obviously directed and without symmetric arcs, since it is rare an article to cite and to be cited by the same paper. Consequently, to detect communities by ignoring the arcs directions could lead to communities different from the expected for a correct analysis.

Additionally, the uncertainty about the clustering structure of undirected networks has led the proposal of many new algorithms for detecting communities. Consequently, to determine which algorithm to adopt for general applications is hard, as pointed out in Lancichinetti and Fortunato (2009). Lancichinetti and Fortunato (2009) assess the quality of a number of community detection algorithms for undirected networks to attest which of them have a good performance. They performed the experiments using artificial graphs, known as LFR networks (Lancichinetti, Fortunato, & F, 2008), whose expected partitions are known. According to the experiments carried out by the authors, Infomap (Rosvall, Bergstrom, 2010) appears as the best algorithm since it outperformed all tested strategies, including the modularity maximization-based algorithms as, e.g., the Louvain method (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008).

* Corresponding author. Tel.:+55 12 33099595; fax.: +55 12 3309 9500.
E-mail addresses: camila.santos@unifesp.br (C.P. Santos), dmcarvalho@unifesp.br (D.M. Carvalho), mcv.nascimento@unifesp.br (M.C.V. Nascimento).

http://dx.doi.org/10.1016/j.eswa.2016.01.026
0957-4174/

It is worth to underline that modularity maximization-based algorithms, extensively adopted for the task of providing clusterings, are also known by their tendency to fail in detecting partitions with numerous small-sized clusters. As a consequence, these strategies tend to merge communities that represent individual groups (Fortunato & Barthélemy, 2007). A promising alternative for the modularity measure suggested in Reichardt and Bornholdt (2006) is the target of the study of this paper. Reichardt and Bornholdt (2006) proposed to fine-tune modularity through the inclusion of a parameter, here called resolution parameter. In Carvalho, Resende, and Nascimento (2014), the authors firmly establish the connection between some graph characteristics and the resolution parameter, for automatically adjusting it. For this, they proposed the use of a neural network, trained according to the topology of the graph. For each input graph, the output of the neural network is an interval of values, expected to be the most suitable options for defining the resolution parameter.

Among these algorithms for undirected networks, some already have its adaptation to tackle directed networks as, e.g., Infomap. More recently, Lancichinetti et al. (2011) proposed the Order Statistics Local Optimization Method (OSLOM), that outperformed Infomap in both undirected and directed LFR networks.

Bearing in mind the discussion outlined, this paper presents:

- A robust consensus clustering, named ConClus, based on arc contractions with a memory mechanism resulting in a strategy that unifies both diversification and intensification paradigms to detect communities in directed networks;
- A study about the performance of this modularity-based algorithm with the resolution parameter adjusted by a neural network trained according to the topology of a number of directed LFR networks;
- An experimental analysis of ConClus using 600 LFR networks with different sizes (from 1000 to 5000 nodes), mixture degrees (from 0.1 to 0.8), community sizes (small and large) and average/maximum in-degrees (20/50 and 40/100);
- A comparative analysis of the results achieved by ConClus with those obtained by the benchmark community detection algorithms: OSLOM, Infomap and the Label Propagation (LP);
- The competitive results of ConClus considering directed LFR networks with average/maximum in-degrees 20/50 in comparison to OSLOM and its better performance over Infomap and LP;
- The results indicating that ConClus outperformed all algorithms considering directed LFR networks with average/maximum in-degrees 40/100;
- A case study with real networks showing that ConClus achieved very accurate results for an undirected network and a directed network.

The remaining of this paper is organized as follows. Section 2 shows a brief review of related works about community detection algorithms in directed networks. Section 2.2.1 presents a comprehensive discussion about the modularity measure and the resolution limit focusing on directed networks. Section 3 presents the proposed strategy. Section 4 shows the computational experiments with real and artificial directed networks. To sum up, Section 5 presents the final remarks and directions of future works.

2. Related works

This section briefly reviews the main approaches for directed networks. The reader interested in a detailed survey in this topic, we indicate the reading of Malliaros and Vazirgiannis (2013). The most recent references are underlined in this paper. Before going into detail about the literature review, this section starts presenting some basic graph theory definitions to be used throughout the paper.

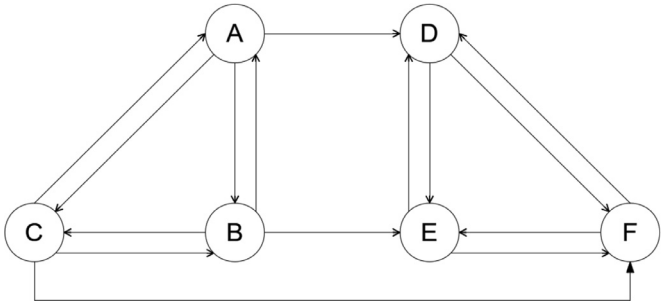

Fig. 1. A directed network composed by two natural communities.

2.1. Basic terminology and background

In this paper, a directed graph $G = (V(G), E(G))$ is represented by a set of vertices or nodes, $V(G)$, and a set of arcs, $E(G)$, where each arc $e := (v_i, v_j) \in E(G)$ is associated with an ordered pair of vertices of G. Additionally, a given arc $(v_i, v_j) \in E(G)$ has as ends the vertices v_i and v_j, where v_i is called the tail, v_j is called the head of the arc and $i, j \in \{1, 2, \ldots, |V(G)|\}$. The number of vertices and arcs of G are denoted in this paper by $n(G)$ and $m(G)$, respectively. The degree of a vertex v_i from G, $d_G(v_i)$, corresponds to the number of times v_i is an end vertex. The in-degree and out-degree of a vertex v_i from G, here called, respectively, $d_G^-(v_i)$ and $d_G^+(v_i)$, correspond to the number of times that a vertex v_i appears as an arc head and arc tail in G. A graph induced by a set of vertices $X \subseteq V$ is denoted by $G[X]$. $\mathcal{N}^-(v_i)$ and $\mathcal{N}^+(v_i)$ are the sets of vertices where v_i is an arc head and tail, respectively. Let $e' \cap e$, where e and $e' \in E(G)$, be the coincident end-vertices of the arcs e and e'.

The pattern recognition in graphs may be performed by identifying their groups of highly related vertices. For this, one way is to find communities through graph clustering algorithms. Among them, we underline those guided by evaluation measures that quantify the clustering quality. The definition of a clustering relies on the k-way partition of the vertex set. Let $\mathcal{C} = \{V_1, V_2, \ldots, V_k\}$, with $1 \leq k \leq n$, be a k-way partition of $V(G)$. The induced graph $G[\mathcal{C}] = (V(G), E(G[\mathcal{C}]))$, where $E(G[\mathcal{C}]) := \bigcup_{i=1}^{k} E(G[V_i])$ defines a graph clustering.

2.2. Community detection in directed networks

Malliaros and Vazirgiannis (2013) present in their survey a good overview of the existing approaches for detecting communities in directed networks. Although the relevance of the topic, they highlight the lack of a consensual general definition for this problem. In interpreting the problem as detecting a group of highly related vertices, what would be "highly related vertices"? What type of relations are expected inside the communities? To formally answer these questions is the first challenge the authors point up as suggestions for future works.

Consequently, it is common to approach the community detection in directed networks by simply ignoring arc directions. However, there is a strong evidence that, depending on the networks, this approach might fail in describing important characteristics of the reciprocity of the network links. Figs. 1 and 2 display an example of a directed network that, if having its arc directions ignored, algorithms may produce incorrect communities. Fig. 1 presents the directed network composed by two communities: {*A, B, C*} and {*D, E, F*}. However, in the network obtained by ignoring arc directions, presented in Fig. 2, it is not clear whether there is one cluster or the two original clusters, even being the expected communities maximal cliques.

As an attempt to overcome this misinterpretation of the arc directions, another approach for dealing with a directed network is