Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Iterated Greedy algorithm for performing community detection in social networks

Jesús Sánchez-Oro *, Abraham Duarte

Dept. Computer Sciences, Universidad Rey Juan Carlos, Tulipán s/n, 28933 Móstoles, Madrid, Spain

HIGHLIGHTS

- We propose a new algorithm for detecting communities in social networks.
- The algorithm optimizes the modularity of the communities detected.
- The proposed method favorably compares with the best previous method.
- The results highlight the relevance of using modularity for detecting communities.

ARTICLE INFO

Article history:
Received 20 October 2017
Received in revised form 28 May 2018
Accepted 7 June 2018
Available online xxxx

Keywords:
Social networks
Community detection
Iterated greedy
Metaheuristics

ABSTRACT

The spreading of social networks in our society has aroused the interest of the scientific community in hard optimization problems related to them. Community detection is becoming one of the most challenging problems in social network analysis. The continuous growth of these networks makes exact methods for detecting communities not suitable for being used, since they require large computing times. In this paper, we propose a metaheuristic approach based on the Iterated Greedy methodology for detecting communities in large social networks. The computational results presented in this work show the relevance of the proposal when compared with traditional community detection algorithms in terms of both quality and computing time.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Social networks have become one of the main media all over the world in the last years, as the number of users is in continuous growth [1,2]. The rationale behind this exponential expansion might be related to the immediacy of the information. Nowadays, any new information is firstly published in social networks and, after that, in traditional media. Furthermore, users are getting used to obtain information from social networks instead of considering traditional media [3].

The transmission of information through social networks has created new lines of research, like viral information detection [4], analysis of the relevance of social network users [5], and community detection [6], among others (see [7,8]). In this work, we focus on the detection of communities in social networks, which is a relevant problem not only in social network analysis, but also in areas like natural disaster management [9], biology [10], semantic web [11], or cybersecurity issues [12].

The community detection problem consists in dividing a network of users into an unknown number of groups, with the objective of optimizing the value of a function that determines the quality of the division. Although this problem has been widely studied from both, exact and heuristic perspectives [13–15], the best objective function used to find the best partition of a network in groups is still under discussion [16].

The quality of a community detection over a social graph has been widely studied from both exact and heuristic perspectives. The Louvain algorithm [13] is focused on maximizing the modularity. It is a heuristic algorithm that follows a greedy criterion to insert a node in a community. Specifically, a node will be added to a community if and only if it leads to an increment in the modularity value, stopping when no improvement is found. The Infomap algorithm [17] focuses on finding the minimum information description of a random walk, using the Minimum Description Length objective function. Finally, the Label Propagation algorithm [14] tries to find the best communities by iteratively assigning to each node the community where most of its adjacent nodes belong to, trying to maximize the modularity metric. These algorithms are based on the structure of the network in order to improve the community detection. However, some works include additional

* Corresponding author.
E-mail address: jesus.sanchezoro@urjc.es (J. Sánchez-Oro).

https://doi.org/10.1016/j.future.2018.06.010
0167-739X/© 2018 Elsevier B.V. All rights reserved.

information of the network in the community detection, like the traffic between two nodes [18] or in wireless sensor networks [19].

As far as we know, the best heuristic method for finding high quality communities in graphs derived from social network is a bioinspired algorithm based on Ant Colony Optimization [6]. This algorithm is focused on detecting communities in Ego Networks, where a user (node) is selected as the center of the graph (Ego) and then all the connected users (nodes) are added, together with the relations (edges) between each pair of users. The main objective in community detection over Ego Networks is to find the groups connected to a certain user in a social network [20].

In this paper we propose a new Iterated Greedy algorithm [21] for detecting communities in Ego Networks. This algorithm starts from an initial solution, constructed by a heuristic procedure. Then, it iteratively improves it by performing two well differenced phases: destruction and reconstruction.

The remaining of the paper is structured as follows: Section 2 formally describes the problem under consideration, Section 3 presents the algorithmic description of the Iterated Greedy method proposed, Section 4 describes the computational experiments performed for analyzing the quality of the proposal, and, finally, Section 5 summarizes the conclusions derived from this research.

2. Problem definition

Before presenting the problem under consideration, it is necessary to provide a formal definition of a network of users. Specifically, given a set of users connected in a social network, we define the graph $G = (V, E)$ where V is the set of n nodes (each user is represented by a unique node) and E is the set of m edges. An edge $(u, v) \in E$, with $u, v \in V$ represents that there is a connection between users u and v. Notice that the meaning of the connection (both users are friends, work in the same company, etc.) is totally dependent of the nature of the social network.

It is important to remark that we are facing an unsupervised clustering problem, since the optimal clustering is not usually known in advance. Therefore, we need to focus on metrics that are able to evaluate the quality of a partition without knowing the optimal one. The most relevant metrics are based on maximizing the density of edges that connect nodes in the same cluster (intra-cluster edges) and, at the same time, minimizing the number of edges that connect nodes located in different clusters (inter-cluster edges).

In this context, there are three main metrics for evaluating the quality of a given partition [22]: modularity, conductance, and coverage. The three considered metrics are normalized in the range 0–1, where 1 is the optimal score for coverage and modularity, an 0 for conductance. Notice that not all networks can reach the optimal score due to their internal structure.

Before formally defining each metric, it is necessary to introduce the solution structure for the Community Detection Problem (CDP). A solution (or partition) for the CDP is represented as the set of clusters $\mathcal{K}$, where each node $v \in V$ is assigned to a different cluster K_i, with $\bigcup_{1 \leq i \leq |\mathcal{K}|} K_i = V$ and $K_i \cap K_j = \emptyset$, with $1 \leq i, j \leq |\mathcal{K}|$. Additionally, φ is a function defined as $\varphi : V \rightarrow \{1, 2, \ldots |\mathcal{K}|$ that represents the cluster to which it belongs a particular node. For example, for a given node $v \in V$, $\varphi(v) = 2$ would indicate that v is located at cluster K_2.

The most simple metric is the coverage [23], which analyzes the number of intra-cluster edges in a given solution with respect to the total number of edges in the network. More formally,

$$Cv(G, \varphi) = \frac{|(u, v) \in E : \varphi(u) = \varphi(v)|}{|E|}$$

Notice that the optimization of this metric can eventually lead to the trivial clustering where all the nodes are in the same cluster.

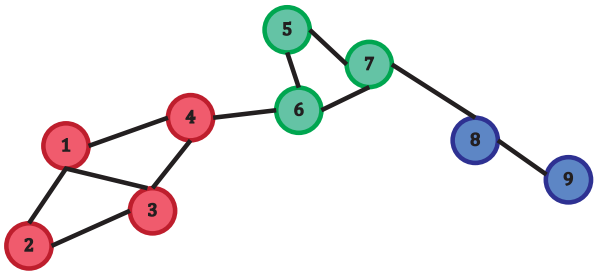

Fig. 1. Example graph with a possible community detection (each community corresponds to a different color). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The conductance of a cluster compares how many inter-cluster edges are in a particular cluster with respect to the total number of edges with an endpoint in that cluster or those with no endpoint in the cluster. In mathematical terms,

$$Cn_k(G, \varphi) = \frac{|(u, v) \in E : \varphi(u) \neq \varphi(v)|}{\min\left(E_k, \overline{E_k}\right)}$$

where $E_k = |(u, v) \in E : \varphi(u) = k \vee \varphi(v) = k|$ is the set of edges with an endpoint in cluster k and $\overline{E_k} = |(u, v) \in E : \varphi(u) \neq k \wedge \varphi(v) \neq k|$ is the set of edges with no endpoint in cluster k. Then, the conductance of a solution φ for graph G is evaluated as the average conductance among all clusters in the solution. More formally,

$$Cn(G, \varphi) = \frac{1}{|\mathcal{K}|} \sum_{k=1}^{|\mathcal{K}|} Cn_k(G, \varphi)$$

where $|\varphi|$ is computed as the number of clusters in solution φ.

The conductance can also be computed by considering intra-cluster edges. The aforementioned definition based on inter-cluster edges is focused on the inter-cluster sparsity, while the one based on intra-cluster edges emphasizes intra-cluster density [24].

The last metric considered is the modularity of a solution, which compares the actual intra-cluster edges with the probability of finding that edge in a random graph [25,26]. This metric has been widely used by the most relevant clustering algorithms in the literature [27,28], although it presents some limitations when considering large scale networks [29]. Modularity of a solution φ for a graph G is formally defined as:

$$Md(G, \varphi) = \sum_{k=1}^{|\mathcal{K}|} \left(e_{kk} - a_k^2\right)$$

where

$$e_{kk} = |(u, v) \in E : \varphi(u) = \varphi(v)| / |E|$$

represents the probability of intra-cluster edges in cluster k, while

$$a_k = |(u, v) \in E : \varphi(u) = k \vee \varphi(v) = k| / |E|$$

represents the probability of an edge with at least one endpoint in cluster k.

Fig. 1 shows an example graph where three communities have been detected, each one highlighted with a different color, together with the value of each one of the considered metrics. Specifically, the red community contains nodes 1, 2, 3, and 4; the green one contains nodes 5, 6, and 7; and the last one (blue) contains nodes 8 and 9. The values for the aforementioned metrics are $Cv(G, \varphi) = 0.82$, $Cn(G, \varphi) = 0.36$, and $Md(G, \varphi) = 0.42$.

This work is focused on optimizing the modularity of the community detection, since it is considered the most robust metric to evaluate the quality of the partition for several community detection algorithms [30].

Please cite this article in press as: J. Sánchez-Oro, A. Duarte, Iterated Greedy algorithm for performing community detection in social networks, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.06.010.

Download English Version:

https://daneshyari.com/en/article/11002439

Download Persian Version:

https://daneshyari.com/article/11002439

Daneshyari.com