



A robust ant colony optimization-based algorithm for community mining in large scale oriented social graphs [☆]



L. Ben Romdhane ^a, Y. Chaabani ^{b,*}, H. Zardi ^b, MARS Research Group ¹

^a ISIT'COM, University of Sousse, Tunisia

^b Faculty of Sciences, University of Monastir, Tunisia

ARTICLE INFO

Keywords:

Ant colony optimization
Community detection
Social networks
NP-complete

ABSTRACT

Community detection plays a key role in such important fields as biology, sociology and computer science. For example, detecting the communities in protein–protein interactions networks helps in understanding their functionalities. Most existing approaches were devoted to community mining in undirected social networks (either weighted or not). In fact, despite their ubiquity, few proposals were interested in community detection in oriented social networks. For example, in a friendship network, the influence between individuals could be asymmetric; in a networked environment, the flow of information could be unidirectional. In this paper, we propose an algorithm, called *ACODIG*, for community detection in oriented social networks. *ACODIG* uses an objective function based on measures of density and purity and incorporates the information about edge orientations in the social graph. *ACODIG* uses ant colony for its optimization. Simulation results on real-world as well as power law artificial benchmark networks reveal a good robustness of *ACODIG* and an efficiency in computing the real structure of the network.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction and related work

Many complex systems, including physical, biological and social systems as well as many man-made technical systems can be modeled by networks (Albert & Barabási, 2002). Networks can be represented by a graph $G = (V, E)$, where V is the set of vertices (or nodes) and E is the set of edges (or links) representing system units and relations between these units, respectively. When the edges have a direction, the network is called directed and; otherwise, it is called undirected.

Many networked systems are found to divide naturally into modules or communities; i.e., groups of vertices with relatively dense connections within groups but sparser connections between them. Detecting such communities could have a wide range of potential applications in real-world as detecting genes with the same functionality in a biological network, detecting the actors of influence in a political network, etc. From a theoretical perspective, community detection in a social network can be modeled as graph partitioning problem which is known to be NP-complete (Fortuna-

to, 2009). To overcome this inherit difficulty, researchers proposed several methods to obtain the best partitioning of the given network (Fortunato, 2009; Leicht & Newman, 2007; Zardi & Ben Romdhane, 2013; Zhang, Zhu, Wang, & Zhao, 2013). At this stage, it is important to notice that most existing methods can only be applied to undirected networks. However, many complex networks in the real-world are directed. Among these networks, let us mention PPI (protein–protein interaction) networks in biology; the World Wide Web; citation networks in the research community; telecommunication/phone call networks; and email networks to highlight just a few. In this kind of networks, the direction of a link contains important information such as asymmetric influence or information flow. A link between a pair of nodes may represent fundamentally different dynamics when its direction is reversed. Therefore, any kind of approach that disregards the direction of links may fail to understand the dynamics and the function of these directed networks. Also, any kind of community detection approach may fail to detect the communities correctly if the direction of the link is not considered properly. In this regard, several recent proposals (Chen & Saad, 2010; Kim, Son, & Jeong, 2010; Lai, Lu, & Nardini, 2010; Leicht & Newman, 2007) have tried to resolve this problem. A common background between all of these methods is that each of them outlines its own definition of a community in a directed network. Actually, no definition is universally accepted. Hereafter, we will outline the most known methods for community detection in oriented social graphs. For a substantial review of the existing approaches as well as their basics, we refer the interested

[☆] This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-No Derivative Works License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

* Corresponding author. Tel.: +216 20360782.

E-mail addresses: lotfi.ben.romdhane@usherbrooke.ca (L. Ben Romdhane), chaabanijamin@gmail.com (Y. Chaabani), zardidarine@gmail.com (H. Zardi).

¹ Modeling of Automated Reasoning Systems.

reader to the specialized literature – see for example (Fortunato, 2009).

Random walks is a widely used technique for community detection in directed (as well as undirected graphs). The basic idea is to use a random walker from one vertex to another; and edges (resp. nodes) “central” to a community will form a trap in a random walk journey. Lai et al. (2010) uses PageRank random walk induced network embedding to transform a directed network into an undirected one, where the information on edge directions is effectively incorporated into the edge weights. The purpose of network embedding is to represent each vertex of a network as a low dimensional vector that preserves these similarities between the vertex pairs, usually measured by the edge weights. Starting from this new undirected weighted network, previously developed methods for undirected social graphs can be used without any modification. In Kim et al. (2010), propose a similar model but based on the importance of links rather than nodes. In fact, the authors proposed a new generalization of modularity based on LinkRank, which is a quantity that indicates the importance of links in a directed social graph. The proposed generalized modularity is the fraction of time spent by a random walker moving within communities minus the expected value of this fraction. In a second phase, any existing model for community detection in undirected networks can be used to optimize this generalized modularity. Although, transforming a directed social network into a weighted undirected gives us the valuable advantage of directly using previously developed methods for undirected networks to find communities in directed ones; there is no real guarantee that this transformation is done without “information loss”; i.e., without altering the hidden real community structure of the original oriented network. Stated otherwise, ignoring edge orientations may discard potentially useful information. In addition, these “random walk”-based methods suffer from the high spatial complexity (Kim et al., 2010).

Graph mining is also another used technique for community detection in oriented social networks. In Chen and Saad (2010), the authors define a community as being a “dense area” in the original graph. Thereby, the problem of community detection is reduced to the problem of extracting the set of meaningful dense subgraphs from a given sparse graph. The proposed idea in the algorithm bears some similarities with the problem of reordering/blocking matrices in sparse matrix techniques and which utilizes the cosine similarity of matrix columns. Doing it this way, a partial clustering of the vertices in a graph is computed, where each cluster represents a dense subgraph. The proposed algorithm is parametric and requires a density threshold above which the output subgraphs are considered to be dense. Unfortunately, the proposed method is unable to detect communities with unbalanced sizes (Chen & Saad, 2010). Stated otherwise, in the presence of large dense communities; smaller sparse ones will be ignored.

Hierarchical clustering is a widely used technique, among others, which puts together similar vertices into larger communities. Hierarchical clustering algorithms build the communities gradually in a hierarchical manner. Sometimes we use some terminating conditions to select the partition or the group of partitions that satisfy a given criteria such as the number of communities desired, the minimum (or maximum) number of objects in each community, the optimization of an objective function, etc. (Fortunato, 2009). In Leicht and Newman (2007), the well-known modularity function is generalized in a principled fashion to incorporate the information contained in edge directions. Then the community structure of the networks is computed by maximizing this generalized modularity function using an hierarchical clustering approach; i.e., over several possible divisions of the network. Unfortunately, as any modularity-based model, this approach will ignore small communities which will be merged with bigger ones.

This is known in the literature as the “resolution limit” problem and is discussed in details in Fortunato and Barthélemy (2007).

In this paper, we propose an algorithm, called ACODIG, for community detection in directed social networks. In ACODIG, we define an objective function that takes into account edge orientation, and we use ant colony for its optimization. A strong feature of our model is its robustness in detecting communities of unbalanced sizes; and thereby avoids the “resolution limit” problem. The rest of this paper is structured as follows. In Section 2, we outline preliminary material. Section 3 details our proposals; illustrates it on a sample network; and analyzes its time and space complexity. In Section 4, we conduct an experimental analysis and compare our model to other recent proposals using large scale real-world and synthetic social networks; while the last section offers concluding remarks.

2. Basic concepts

2.1. Problem formulation

A social network can be modeled by a graph $G = (V, E)$ where V denotes the set of vertices and $E \subseteq (V, V)$ denotes the set of edges. In a directed network an edge is an ordered vertex pair (i, j) , while in an undirected one both the vertex pair (i, j) and (j, i) represent the same edge. A network can be represented by an adjacency matrix A whose elements are non negative; i.e., A_{ij} is positive if there is an edge between vertex i and vertex j , and 0 otherwise. We should note that A is asymmetric for directed graphs. The problem of community detection could be defined as follows.

Definition 1 (Community detection). Let $G = (V, E)$ be a directed graph modeling the directed social network at hand. Let f be an objective function measuring the quality of a partitioning of G . The problem of community detection consists in finding a partitioning $P = \{C_1, \dots, C_k\}$ of G such that: (i) $C_i \subset V$; (ii) $C_i \cap C_j = \emptyset \forall C_i, C_j \in P$; (iii) $\bigcup_{i=1}^k C_i = V$; and (iv) $f(P)$ is optimal.

From Definition 1, we can state that a good partitioning is that optimizing a given objective function measuring the overall quality of the detected communities. In a computed partitioning, we assume that communities do not overlap. We should notice that we do not know a priori neither the size nor the number of communities.

2.2. Notations and basic definitions

The main goal of this section is to introduce preliminary concepts and the basic definitions fundamental to our model.

Notation 1. Given a directed graph $G = (V, E)$, we denote by $|V|$ its number of vertices and $|E|$ its number of edges.

Definition 2 (Degree of vertex). Given a graph $G = (V, E)$, then the degree of a vertex s is the number of incoming and outgoing edges from s given by:

$$\text{Degree}(s) = \sum_{i \in G} E(s_i, s) + E(s, s_i) \quad (1)$$

Definition 3 (Importance of a vertex). Given a graph $G = (V, E)$, we define the importance of a vertex s as the fraction of the sum of the outgoing edges from s divided by the sum of the maximum number of incoming edges and the maximum outgoing edges existing in G . It is given by:

$$\text{Importance}(s) = \frac{D_{out}(s)}{D_{max_in} + D_{max_out}} \quad (2)$$

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات