Swarm and Evolutionary Computation 26 (2016) 137–156

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

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

Improving the performance of evolutionary multi-objective co-clustering models for community detection in complex social networks

Bara'a A. Attea [a],*, Wisam A. Hariz [a], Mayyadah F. Abdulhalim [b]

[a] Department of Computer Science, University of Baghdad, Iraq
[b] Department of Computer Science, University of Bahrain, Bahrain

ARTICLE INFO

Article history:
Received 6 May 2015
Received in revised form
27 August 2015
Accepted 16 September 2015
Available online 28 September 2015

Keywords:
Community-less nodes
Graph co-clustering
MOEA/D
Non-dominated solution
NP-hard
Social networks

ABSTRACT

Due to globalization, the characteristic of many systems in biology, engineering and sociology paradigms can nowadays be captured and investigated as networks of connected communities. Detecting natural divisions in such complex networks is proved to be extremely NP-hard problem that recently enjoyed a considerable interest. Among the proposed methods, the field of multi-objective evolutionary algorithms (MOEAs) reveals outperformed results. Despite the existing efforts on designing effective multi-objective optimization (MOO) models and investigating the performance of several MOEAs for detecting natural community structures, their techniques lack the introduction of some problem-specific heuristic operators that realize their principles from the natural structure of communities. Moreover, most of these MOEAs evaluate and compare their performance under different algorithmic settings that may hold unmerited conclusions. The main contribution of this paper is two-fold. Firstly, to reformulate the community detection problem as a MOO model that can simultaneously capture the intra- and inter-community structures. Secondly, to propose a heuristic perturbation operator that can emphasize the search for such intra- and inter-community connections in an attempt to offer a positive collaboration with the MOO model. One of the prominent multi-objective evolutionary algorithms (the so-called MOEA/D) is adopted with the proposed community detection model and the perturbation operator to identify the overlapped community sets in complex networks. Under the same MOEA/D characteristic settings, the performance of the proposed model and test results are evaluated against three state-of-the-art MOO models. The experiments on real-world and synthetic social networks of different complexities demonstrate the effectiveness of the proposed model to define community detection problem. Moreover, the results prove the positive impact of the proposed heuristic operator to harness the strength of all MOO models in both terms of convergence velocity and convergence reliability.

1. Introduction

Many complex real-world systems in almost every discipline of biology, sociology, and engineering can be represented as graphs, or networks. Social networks, protein networks, World Wide Web, the Internet, collaboration networks, power grids, communication and transport networks are just some examples. Natural divisions within such networks, follow a general heterogeneous connections rule, known as *modules* or *communities* where densely intra-connected groups of nodes are also sparsely inter-connected with other groups. In different context, other terms such as cluster, partition, group, and cohesive subgroup can be used to describe a community set. The growing demand for algorithms to detect such community structure in networks comes from its considerable extent of applications. For example, in social networks, individuals or organizations are tied through various social contacts, familiarities, or profiles. Social modularity means, then, a set of social individuals which satisfy dense convergence of contacts. In protein-protein interaction (PPI) networks, all cell activities can be understood by analyzing those proteins structured as interacting and separable modules. Thus, PPI modularity refers to a set of physically or functionally interacted proteins work together to accomplish particular functions. Another example is in recommendation systems where latent similarities between users (in terms of friendship, commenting, items, and etc.) can be used to help such system to work. With the growing demand for all these and other real-world applications, community structure aspires to capture the essential characteristics, topology, and functions of these networking systems.

In contrast to data clustering, community sets detection is defined to be a bi-clustering (i.e., co-clustering) problem. Consider

* Corresponding author. Tel.: +964 77 039 99959.
E-mail addresses: baraali@scbaghdad.edu.iq (B.A. Attea), wisan.arkan@yahoo.com (W.A. Hariz), mfaisal@uob.edu.bh (M.F. Abdulhalim).

http://dx.doi.org/10.1016/j.swevo.2015.09.003
2210-6502/© 2015 Elsevier B.V. All rights reserved.

an $n \times m$ data set matrix A consisting of n objects, each being characterized by m features, i.e. $A = [a_{ij}], i = 1, \ldots, n$ and $j = 1, \ldots, m$. Note that in community detection problem, both dimensions of A, called adjacency matrix, are identical, equal to the number of nodes n in the networks (i.e., $A = [a_{ij}], i, j = 1, \ldots, n$). Any clustering algorithm tries to partition the space of A into a set of K regions or clusters $\mathcal{C} = C_{k\,k=1}^{\,K}$ according to the correlation among n objects. Thus, if $C_{k1} = \{a_{ij}\}_{i=1,j=1}^{n1,m}$ and $C_{k2} = \{a_{ij}\}_{i=1,j=1}^{n2,m}$ are two clusters, then $C_{k1} \bigcap C_{k2} = \varnothing$. However, considering *both* correlation of features as well as objects in the light of clustering process, means to *simultaneously* select and group (i.e. *co-cluster*) both dimensions of A into sub-matrices, each of which consists of locally correlated objects under a subset of their features. Formally speaking, let $\mathcal{C} = C_{k\,k=1}^{\,K}$ be a set of K co-clusters and let $C_{k1} = \{a_{ij}\}_{i=1,j=l1}^{n1,u1}$ and $C_{k2} = \{a_{ij}\}_{i=1,j=l2}^{n2,u2}$ are two co-clusters belong to $\mathcal{C}$, then $C_{k1} \bigcap C_{k2} = \varnothing$ in *both* i and j dimensions.

Simultaneous matrix co-clustering needs a quality index that can capture the embedded sub-matrix structures. The *modularity* (noted as Q) index of Newman and Girvan, lays the foundation of many existing successful graph clustering algorithms [1]. The purpose of Q is to capture the hidden structure of community sets in complex networks by maximizing intra-cluster links while minimizing inter-cluster ones. Consider a network constituted by n nodes which can be formally described as a graph $G = (V, E)$, where $V(G) = v_1, \ldots, v_n$ is the set of vertices (or nodes) and $E(G) = e_1, \ldots, e_m$ is the set of edges (or connections) between nodes. Then, the *cardinality* of G, $n(G) = |V|$ and the *volume* of G, $m(G) = |E|$. The *degree* of any vertex, $m(v)$, is defined as the number of edges incident to v. Throughout this paper, the notation $n(\bullet)$ is used to represent cardinality concept, while $m(\bullet)$ is used to represent volume concept.

Now, consider partitioning V of G into a co-clustering solution $C = \{C_1, \ldots, C_K\}$ such that each vertex v_i, $1 \le i \le n$ is exactly assigned to one cluster C_j, $1 \le j \le K$. The impact of E in $\mathcal{C}$ can, now, be quantified in two distinct terms. The set of edges between vertices existing in two distinct clusters: $E(C_i, C_j)$, $1 \le i, j \le K \text{and} i \ne j$ and the set of edges found inside one cluster: $E(C_i, C_i)$, $1 \le i \le K$. Then, modularity in [1] will award $\mathcal{C}$ according to the fraction of connections inside its communities as formulated in Eq. (1), where two contradictory objectives are implicitly handled. The left operand in Eq. (1) biases towards a solution $\mathcal{C}$ that is covered with a densely intra-connected modules, i.e. many edges fall within $C_1, \ldots, C_K$. On the other hand, the right operand in Eq. (1) expresses that the expected value of the same edge density in $\mathcal{C}$ with the same community structure $C_1, \ldots, C_K$ but fall at random between the vertices should be small. Q will approach its minimum at 0 if the number of within-community edges is no better than random. On the other hand, values approaching $Q = 1$, which is the maximum, indicate strong community structure.

$$Q(\mathcal{C}) = \sum_{i=1}^{K} \left[\frac{|E(C_i, C_i)|}{m(\mathcal{C})} - \left(\frac{\sum_{v \in C_i} m(v)}{2m(\mathcal{C})} \right)^2 \right] \tag{1}$$

Community detection problem is proved to be an NP-hard problem [2,3] and can mainly be decomposed into two sub-problems. The first one considers the algorithmic aspect, trying to find an answer for how to partition a network (i.e., how to generate $\mathcal{C}$). The second problem is more semantically related with how to assess the quality of a given partitioning solution (i.e., how to define $\Phi(\mathcal{C})$ for some quality function Φ).

In literature, the detection of community structure has been addressed as a graph mining problem with three different methodologies. These are: top–down co-clustering methods, bottom–up co-clustering methods and optimization methods. The top–down (also called divisive hierarchical) methods initiate the whole network as one community and iteratively detect the weakest edges that connect different communities and remove them [1,4–8]. In contrary, a bottom–up (agglomerative hierarchical) method, initializes each node as one community. It then iteratively merges similar communities according to some quality measures [9–13].

Due to NP-completeness, many algorithms define and formulate the community detection problem as *modularity maximization* problem. These optimization methods share a common ground by trying to optimize one or two objective functions realizing correlation among featured subgroups and divide the network nodes according to these subgroups into sub-networks [14–16]. Recently, the relaxed nature of meta-heuristic based optimization methods makes them very suitable to reduce the complexity of the problem and to approach adequate solutions. The dominated optimization methods explored so far in this area of study is single- and multi-objective evolutionary algorithms (EAs) [10,17–22] and [23–27] with paramount performance for the multi-objective evolutionary algorithms (MOEAs).

In EA-based literature, community detection has been addressed in both algorithmic and semantic directions, with more focus on the first issue. For the quality measure $\Phi(\mathcal{C})$, however, there is a few of such attempts and, almost all the proposed $\Phi(\mathcal{C})$ metrics exploit information gathered from the *density of links* within and among communities of a given partition. Additionally, few attempts proposed heuristic evolutionary operators that can deduce their mechanisms from the definition of community structure. In this paper, we suggest to redirect the design of $\Phi(\mathcal{C})$ according to the *neighborhood relations* of intra-community and inter-community nodes. To this end, the contributions of this paper are:

- To revisit and elaborate modularity metric in a new multi-objective optimization model (MOO) that can rigorous cast on the two contradictory properties of community structure.
- Four new definitions have been introduced in this paper to qualify the neighborhood relation of a given node in the network.
- Based on the qualitative definition of a node and its neighborhood relations, a heuristic perturbation operator is proposed to control node right community-belongingness and thus to harness the performance of any model, including, the proposed MOO model.

The remainder of this paper is organized as follows. Section 2 presents basic concepts relating to the community detection problem. Section 3 introduces our formulation for the multi-objective community detection problem. Section 4 introduces the algorithmic steps used to solve the formulated problem. Section 5 reports experimental results and, finally, Section 6 presents our conclusions and suggestion of further research directions.

2. Preliminaries

The problem of community detection in social networks is modeled, in the literature, as graph partitioning or graph co-clustering problem. Finding a globally optimal solution to the graph co-clustering problem, however, is NP-hard. Informally, a community in a network is a sub-network having *dense* connections within its nodes and *loose* connections with other communities. Thereinafter, without loss of generality, the graphs considered to model social networks are undirected and un-weighted. Let $\mathbb{C}(G)$ be the space of all possible partitions $\mathcal{C}$ of a graph G. Also, let a cluster $C_i \in \mathcal{C}$ be a community belongs to a partition $\mathcal{C}$, and let $E(C_i, C_i)$ be the set of edges connecting vertices of C_i, i.e. $E(C_i, C_i) = (v, w) \in E \wedge v, w \in C_i$. Then, we can *quantitatively* and *semantically* formalize the following definitions. For vertex $v \in C_i$: