Iraqi Journal of Science

ISSN: 0067-2904
GIF: 0.851

A Heuristic Multi-objective Community Detection Algorithm for Complex Social Networks

Bara'a Ali Attea*, Wisam A. Hariz
Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

Abstract

Nowadays the characteristic of many systems can be captured and investigated as networks of connected communities. Recently, large research interests are devoted towards unraveling natural divisions in such complex networks. Due to problem complexity, the field of multi-objective evolutionary algorithms (MOEAs) reveals outperformed results, however, they lack the introduction of some problem-specific heuristic operators that realize their principles from the natural structure of communities. The main contribution of this paper is to introduce a heuristic perturbation operator that can as a local search operator. Thewell known multi-objective evolutionary algorithm with decompositions (MOEA/D) is adopted with the proposed perturbation operator to identify the overlapped community sets in complex networks. The performance of the proposed MOEA/D is evaluated under a set of experiments on real-world social networks of different complexities. The results prove the positive impact of the proposed heuristic operator to harness the strength of MOO model in both terms of convergence velocity and convergence reliability.

Keywords: Community-less nodes, graph co-clustering, MOEA/D, MOO, non-dominated solution, NP-hard, Pareto Front, social networks.

خوارزمية أرشادية متعددة الأهداف لكشف الجاليات في الشبكات الأجتماعية المعقدة

براء علي عطية*، وسام اركان حريز
قسم الحاسبات ، كلية العلوم ، جامعة بغداد ، بغداد ، العراق

الخلاصة

في وقتنا الحاضر من الممكن تمثيل العديد من على هيئة شبكات متداخلة من مجموعة جاليات مترابطة بكثافة داخليا وبشكل متطفل مع بعضها البعض. عدة بحوث حديثا أهتمت بتطوير خوارزميات للكشف عن هذه الجاليات المتداخلة. ولكن على الرغم من وجود جهود ملحوظة في تصميم خوارزميات متعددة الأهداف لحل هذه المشكلة، تفتقر هذه الخوارزميات الى تبني عامل أرشادي لتركيز عملية البحث حول العلاقات الداخلية والعلاقات الدفينة بين الأفراد كمحاولة لمساعدة الخوارزمية للوصول الى الحل الصحيح بشكل أسرع أو لزيادة موثوقية الحل. يساهم هذا البحث،لحل مشكلة الكشف عن الجاليات في الشبكات الأجتماعية، بأقتراح عامل أرشادي جديد يستمد لتطوير أداء واحد من أكثر الخوارزميات التطورية متعددة الأهداف المعروفة (والذي يسمى MOEA/D).لتقييم أداء الخوارزميةالمقترحة، تم أجراء المحاكاة على أربعة شبكات أجتماعية قياسية معروفة. النتائج المدونة توضح التأثير الأيجابي للعامل الأرشادي المقترح في تطوير أداء الخوارزمية التطورية متعددة الأهداف.

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

*Email: baraaali@yahoo.com

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 [1]. 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.

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 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, bottom-up (agglomerative hierarchical) methods, initialize each node as one community. It then iteratively merges similar communities according to some quality measures [9 – 13].

In EA-based literature, community detection model, $\Phi(\mathcal{C})$,often exploit information gathered from the *density of links* within and among communities of a given partition. One of recent and effective $\Phi(\mathcal{C})$ examples provided in the literature is C. Pizzuti's model [14]. However, few of such attempts proposed heuristic evolutionary operators that can deduce their mechanisms from the definition of community structure. To this end, the contribution of this paper is two-fold. First, three new definitions to qualify the neighborhood relation of a given node in the network are introduced. Second, based on the qualitative definition of the node, a heuristic perturbation operator is proposed to harness the performance of the multi-objective community detection model, $\Phi(\mathcal{C})$. Moreover, the prominent multi-objective evolutionary algorithm with decomposition (MOEA/D) [15] and the multi-objective $\Phi(\mathcal{C})$ model of C. Pizzuti's [14] are adopted in this paper to evaluate our contribution.

The remainder of this paper is organized as follows. Section 2presents basic concepts relating to the community detection problem. Section 3 reviews existing state-of-the-art multi-objective community detection algorithms. Section 4 introduces the formulation and the algorithmic steps for the proposed heuristic multi-objective community detection algorithm. Section 5 reports experimental results and, finally, section 6 presents our conclusions and suggestion of further research directions.

2. Clustering vs. Bi-clustering

In contrast to data clustering, community sets detection is defined to be a bi-clustering (i.e., co-clustering) problem. Consider an $n \times m$ data set matrix A consisting of n objects, each being characterized by m features, i.e. $A = [a_{ij}]$, $i = 1, \dots, n$and $j = 1, \dots, m$. 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., $A = [a_{ij}]$, $i, j = 1, \dots, n$). 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$C_{k1} \cap C_{k2} = \emptyset$. 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$C_{k1} \bigcap C_{k2} = \emptyset$ 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 $V(G) = \{v_1, \dots, v_n\}$ is the set of vertices (or nodes) and $E(G) = \{e_1, \dots, e_m\}$ 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 $n(\cdot)$ is used to represent cardinality concept, while $m(\cdot)$ is used to represent volume concept.

Now, consider partitioning V of G into a co-clustering solution $\mathcal{C} = \{C_1, \dots, C_K\}$ 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: $E(C_i, C_j), 1 \le i, j \le K\ and\ i \ne j$ 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 $\{C_1, \dots, C_K\}$. On the other hand, the right operand in Eq. 1 recommends that $\mathcal{C}$ with few edges fall at random without regarding the structure of $\{C_1, \dots, C_K\}$ modules.

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

3. Literature Review

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. 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 $v \in C_i$:

- $m(v, C_i) = |\{(v, w) \in E\ \wedge w \in C_i| = \sum_{w \in C_i} A(v, w)$ is the number of intra-edges of v, and
- $\overline{m}(v, C_i) = |\{(v, w) \in E\ \wedge w \notin C_i| = \sum_{w \notin C_i} A(v, w)$is the number of inter- edges of v.

To this end, we can generalize the language of intra- and inter- connections to a single community C_i and to the whole partition $\mathcal{C}$ as:

- $m(C_i) = |E(C_i, C_i)| = \sum_{v \in C_i} m(v, C_i)$is the number of *intra-cluster* connections of C_i.
- $\overline{m}(C_i) = \left|E(C_i, C_j)\right|_{\forall j \ne i} = \sum_{v \in C_i} \overline{m}(v, C_i)$is the number of *inter-cluster* connections ofC_i.
- $m(\mathcal{C}) = |E(\mathcal{C})| = \left|\{E(C_i, C_i)\}_{i=1}^{K}\right|$ is the number of *intra-partition* connections of $\mathcal{C}$, and
- $\overline{m}(\mathcal{C}) = |E(G)|\ /\ |E(\mathcal{C})|$ is the number of *inter-partition* connections of $\mathcal{C}$, and

Note that we usually refer to $m(v)$ as the degree of vertex v, while for a cluster or group of vertices C, $m(C)$ is said to be the *volume* of C. For example, in [14] Pizzuti refers to $m(C)$ as the volume of community C, while the number of nodes in C, i.e. $|C|$ is referred to as its *cardinality*. According to the volume of a community C, Radicchi *et al.* [8] semantically define C as either:*weak community*: if $m(C) > \overline{m}(C)$, or *strong community*: if $\forall v \in C \Rightarrow m(v) > \overline{m}(v)$.

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's nodes according to these subgroups into sub-networks [16] – [18]. 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], [19–24] and [14],[25–28] with paramount performance for the multi-objective evolutionary algorithms (MOEAs).

Shi *et al.* [26] define the community detection problem as a multi-objective minimization problem. These objective functions are the two terms of the modularity function Q in Eq. 1. Noting that Q is defined as a maximization function, and using a minimization formula:
$\Phi_{Shi\ et.al.}(\mathcal{C}) = Minimize\ \{\Phi_1(\mathcal{C}), \Phi_2(\mathcal{C})\}$, MOO can be expressed as:

$$\Phi_1(\mathcal{C}) = 1 - \sum_{i=1}^{K} \frac{m(C_i)}{m(\mathcal{C})} \tag{2}$$

$$\Phi_2(\mathcal{C}) = \sum_{i=1}^{K} \left(\frac{\sum_{v \in C_i} m(v)}{2m(\mathcal{C})}\right)^2 \tag{3}$$

Pizzuti [14] formulated a multi-objective maximization model:
$\Phi_{Pizzuti}(\mathcal{C}) = Maximize\ \{\Phi_1(\mathcal{C}), \Phi_2(\mathcal{C})\}$. For a partition $\mathcal{C}$, the first objective is to maximize *community score* [23] while the second objective is to maximize the *community fitness* proposed by Lancichinetti*et al.* [18]. Formally speaking:

$$\Phi_1(\mathcal{C}) = \sum_{i=1}^{K} \frac{\sum_{v \in C_i} \left(\frac{m(v)}{n(C_i)}\right)^r}{n(C_i)} * m(C_i) \tag{4}$$

Where $r > 0$ controls the size of community C_i found. For a given communityC_i, its fitness $f(C_i)$ is maximized by maximizing the fitness of its nodes, i.e.:

$$f(C_i) = \frac{m(C_i)}{(m(C_i) + \bar{m}(C_i))^{\alpha}} \tag{5}$$

Also, here $\alpha > 0$ control the size of community C_i. Then Pizzuti [14] defines $\Phi_2(\mathcal{C})$ as:

$$\Phi_2(\mathcal{C}) = \sum_{i=1}^{K} f(C_i) \tag{6}$$

After evolving a set of solutions, Pizzuti suggested to select the partition with the maximum modularity value $Q(\mathcal{C})$, however, she concluded that $Q(\mathcal{C})$ may fail to represent the true partition. As can be seen from both formulations of Shi *et al.* [26] and Pizzuti [14], the emphasize goes to $m(C_i)$ while the impact of inter-cluster connections $\bar{m}(C_i)$ is either indirectly or implicitly optimized (see Eq. 2 and Eq. 5).

If assuming and exploiting connections among nodes of a social network is essential in community detection problem, then it will be wise to isolate those connections that connect nodes within one community and those connect nodes within different communities. Recently, Gong *et al.* [28] formulate a very effective MOO model that explicitly emphasizes the impact of $m(C_i)$and $\bar{m}(C_i)$. The first objective concerns with maximizing the density of intra-community links, while the second objective concerns with minimizing the density of inter-community links. In the language of minimization, Gong *et al.*'smodel is $\Phi_{Gong\ et.al.}(\mathcal{C}) = Minimize\ \{\Phi_1(\mathcal{C}), \Phi_2(\mathcal{C})\}$. $\Phi_1(\mathcal{C})$is the kernel $K - means$ (KKM) while ratio cut (RC) is used to denote $\Phi_2(\mathcal{C})$, as expressed next.

$$\Phi_1(\mathcal{C}) = 2(n - K) - \sum_{i=1}^{K} \frac{m(C_i)}{n(C_i)} \tag{7}$$

$$\Phi_2(\mathcal{C}) = \sum_{i=1}^{K} \frac{\bar{m}(C_i)}{n(C_i)} \tag{8}$$

4. Multi-objective Evolutionary Algorithm with Decomposition: General Review

By nature, many real life problems have contradictory objectives to be fulfilled simultaneously. Due to its success, the field of multi-objective optimization (MOO) has, recently, attracted several researchers in formulating and solving multi-objective optimization problems (MOPs). Instead of single optimal or near-optimal solution, a set of non-dominated solutions can simultaneously be obtained, by a MOO model, providing the decision maker with an optimal tradeoff between the conflicting objectives. Generally, MOP is formulated according to [15], [29],[30] as a vector function $\mathbb{F}(\mathbb{X}) = [\mathbb{f}_1(\mathbb{X}), \mathbb{f}_2(\mathbb{X}), \dots, \mathbb{f}_k(\mathbb{X})]^T$ where $\mathbb{X} = [x_1, x_2, \dots, x_n]^T$ is the vector of decision variables. $\mathbb{F}(\mathbb{X})$ is optimized (in terms of *domination*) to find a non-dominated vector $\mathbb{X}^* = [x_1^*, x_2^*, \dots, x_n^*]^T$. Let us consider two solution vectors $\mathbb{U}$ and $\mathbb{V}$ from the solution space $\Omega(\mathbb{X})$. Then, solution $\mathbb{U}$ is said to dominate $\mathbb{V}$ if and only if the following two conditions hold:

1. The solution $\mathbb{U}$ is no worse than $\mathbb{V}$ in all objectives, or formally, $\mathbb{f}_i(\mathbb{U}) \ntriangleright \mathbb{f}_i(\mathbb{V})$ for all$i = 1,2, \dots, k$.Forexample in maximization, the word "no worse" means$\mathbb{f}_i(\mathbb{U}) \nless \mathbb{f}_i(\mathbb{V})$.
2. The solution $\mathbb{U}$ is strictly better than $\mathbb{V}$ in at least one objective, or formally, $\mathbb{f}_i(\mathbb{U}) \lhd \mathbb{f}_i(\mathbb{V})$for at least one $i \in 1,2, \dots, k$.Forexample in maximization, the word "strictly better" means $\mathbb{f}_i(\mathbb{U}) > \mathbb{f}_i(\mathbb{V})$.

The notation $(i \lhd j)$ is used to denote that solution i is better than solution j regardless of the type of the optimization problem at hand (maximization or minimization). Also, the notation $(i \rhd j)$is used in the same way to express that solution j is better than solution i. Hence, a dominated set can be defined as: among a set of solutions $\Omega(\mathbb{X})$, the non-dominated solutions set $\overline{\Omega}(\mathbb{X}) \subset \Omega(\mathbb{X})$are those that are not dominated by any member of the set $\Omega(\mathbb{X})$.

Among the famous multi-objective evolutionary algorithms being successfully applied to many real-world problems is the multi-objective evolutionary algorithm with decomposition (MOEA/D) put forward by Zhang and Li [15]. Consider a formulated MOP with k objective functions:

$$Maximize\ \mathbb{F}(\mathbb{X}) = [f_1(\mathbb{X}), f_2(\mathbb{X}), \ldots, f_k(\mathbb{X})]^T \quad (9)$$

Also, consider a reference point $z^* = (z_1^*, \ldots, z_k^*)$ to hold the best value obtained so far by MOEA/D for each objective function. The basic idea behind MOEA/D is to decompose (using Tchebycheff approach) the MOP into N scalar optimization sub-problems and treat each sub-problem as a complete individual solution. Each individual i is associated with one weight vector $\lambda^i = (\lambda_1^i, \lambda_2^i, \ldots, \lambda_k^i), s.t. \sum_{j=1}^{k} \lambda_j^i = 1$ from a set of N even spread weight vectors $\lambda^1, \lambda^2, \ldots, \lambda^N$. Each individual i is evolved using information gathered from its T neighbor solutions. Neighbor solutions to i, denoted by $B(i)$, are those with the closest (using Euclidean distance) weight vectors to λ^i. Thus, $B(i) = \{i_1, i_2, \ldots, i_T\}$with $B_\lambda(i) = \{\lambda^{i_1}, \lambda^{i_2}, \ldots, \lambda^{i_T}\}$.

The problem of approximating the Pareto Front (PF) of the MOP defined in Eq. (13) can be decomposed into N scalar optimization sub-problems, each with its objective function:

$$\forall i, j\ s.t., 1 \leq i \leq N\ and\ 1 \leq j \leq k$$

$$g_i(\mathbb{X}|\lambda^i, z^*) = max\{\lambda_j^i |f_j(\mathbb{X}) - z_j^*|\} \quad (10)$$

MOEA/D minimizes all these g_i objective functions simultaneously in a single run. MOEA/D with the Tchebycheff approach evolves a population of Nsolutions $\mathbb{X}^1, \mathbb{X}^2, \ldots, \mathbb{X}^N \in \Omega(\mathbb{X})$, where $\mathbb{X}^i$ is the current solution to the i^{th} sub-problem with $\mathbb{F}(\mathbb{X}^i) = [f_1(\mathbb{X}^i), f_2(\mathbb{X}^i), \ldots, f_k(\mathbb{X}^i)]^T$. Also, MOEA/D maintains an external population EP, for archiving the non-dominated solutions found during the search. At each generation t, MOEA/D performs four main operations while generating N new solutions $\mathbb{Y}^1, \mathbb{Y}^2, \ldots, \mathbb{Y}^N$. Firstly, $\forall i, 1 \leq i \leq N$, it produces a new offspring $\mathbb{Y}^i$, using problem-specific genetic operators (e.g., crossover and mutation), from $\mathbb{X}^i$'s neighbors $B(i)$. Secondly, it updates the reference points. $\forall j, 1 \leq j \leq k$, if $z_j^* < f_j(\mathbb{Y}^i)$, then it sets $z_j^* = f_j(\mathbb{Y}^i)$. Thirdly, it updates the neighbors $\mathbb{X}^{i_l} \in B(i)$): $\forall l, 1 \leq l \leq T$, if $g_i(\mathbb{Y}^i|\lambda^{i_l}, z^*) \leq g_i(\mathbb{X}^{i_l}|\lambda^{i_l}, z^*)$, then it sets $\mathbb{X}^{i_l} = \mathbb{Y}^i$ and $\mathbb{F}(\mathbb{X}^{i_l}) = \mathbb{F}(\mathbb{Y}^i)$. Finally, it updates EP by removing from it all solutions $\mathbb{Y}$ where $\mathbb{F}(\mathbb{Y}^i) \lhd \mathbb{F}(\mathbb{Y})$ and insert $\mathbb{Y}^i$ into EPif $\nexists \mathbb{Y} \in EP \rightarrow \mathbb{F}(\mathbb{Y}) \lhd \mathbb{F}(\mathbb{Y}^i)$.

4. Multi-objective Community Detection Algorithm

In what follow, we introduce three different definitions to qualify the *neighborhood relation* of node v in terms of its connections and community belongingness (the fourth relation will be defined later in section 5). Recall that a node v is neighbor to node w if and only if $\exists E(v, w) \in E(G)$. This will implies $A(v, w) = 1$and $A(w, v) = 1$. Also, recall that $m(v, C_i)$ is defined as the number of intra-edges of v within community C_i.

4.1Node Neighborhood Relation: Definition and Formulation

Now, the number of *intra-neighbors* of community C_i can be defined as:

$$Neighbor(C_i) = \sum_{\forall v, w \in C_i} A(v, w) = \frac{m(C_i)}{2} \quad (11)$$

Definition A (Strongly-Neighborhood Node) Given a partition $\mathcal{C}$and a set of communities $\mathrm{C_i} \in \mathcal{C}, \mathrm{i} = 1, \ldots, \mathrm{K}$. A node v is said to be a strongly-neighborhood node to the nodes of community $\mathrm{C_i} \in \mathcal{C}$ if and only if there exists no other community $\mathrm{C_j} \in \mathcal{C}$ such that $\mathrm{m(v, C_i)} \leq \mathrm{m(v, C_j)}$.

Definition B (Neutrally-Neighborhood Node) Given a partition $\mathcal{C}$and a set of communities $\mathrm{C_i} \in \mathcal{C}, \mathrm{i} = 1, \ldots, \mathrm{K}$. A node v is said to be a neutral-neighborhood node to the nodes of community $\mathrm{C_i} \in \mathcal{C}$ if and only if there exists at least another community $\mathrm{C_j} \in \mathcal{C}$ such that $\mathrm{m(v, C_i)} = \mathrm{m(v, C_j)}$.

Definition C (Weakly-Neighborhood Node) Given a partition $\mathcal{C}$and a set of communities $\mathrm{C_i} \in \mathcal{C}, \mathrm{i} = 1, \ldots, \mathrm{K}$. A node v is said to be a weakly-neighborhood node to the nodes of community $\mathrm{C_i} \in \mathcal{C}$ if and only if there exists at least another community $\mathrm{C_j} \in \mathcal{C}$such that $\mathrm{m(v, C_i)} < \mathrm{m(v, C_j)}$.

4.2 Individual Representation and Genetic Operators

The choice for a good genotype encoding (i.e. individual representation) is an essential issue for the applicability and effectiveness of any evolutionary algorithm. It is highly problem-related decision step. In all related works [14], [25 – 28] the adopted representation is the locus-based adjacency representation being proposed by Park and Song [31]. In locus-based representation, each individual I is represented as a fixed-length vector of n genes where n is the total number of nodes in the network. The allele value of each gene can be varied from 1to n. Thus, $I = (I_1, I_2, \dots, I_n), s.t. I_{i,1 \le i \le n} \in \{1,2,\dots,n\}$.

The decoding function δ of individual I will outline the structure of the communities of the network, i.e. $\delta(I): \mathcal{C} = \{C\}_{i=1}^{K}$. By its nature, the locus-based representation can automatically determine the number of communities, K, being encoded in each individual I. Consider gene i is assigned with value j. This means that nodes i and j will be in the same community C. However, this decoding function may hold in some cases infeasible solutions if node j has no connection with all nodes (including i) of community C (i.e. $\forall i \in C, A(i,j) = 0$).

Given that MOEA/D is population-based optimization algorithm, then a population ρ of N solutions can be formally represented as:

$$\rho = \{I^1, I^2, \dots, I^N\} \tag{12}$$

Now, the adopted MOEA/D can be described as an iterative evolution function $\Psi: \{\rho, EP\} \to \{\rho', EP'\}$with $\Psi(\rho_t) = \rho_{t+1}$, where ρ_t and ρ_{t+1} are the population at generation t and $t+1$, respectively. EPandEP' are the non-dominated set of solutions at generation t and $t+1$, respectively. The population starts with an initial random population ρ_0 and continues until a maximum number of iterations max_t has been reached.

Uniform crossover and mutation operators are used with probability $p_c$and p_m, respectively. Consider two individuals I^1 and I^2 to be the two participating parents in the crossover. A child I' can be formally generated by:

$$\forall i, 1 \le i \le n$$

$$I'_i = \begin{cases} I_i^1 \; if \; r \le 0.5 \\ I_i^2 \; otherwise \end{cases} \tag{13}$$

where$r \sim [0,1]$ is a uniform random number. For the mutation operator, the allele of the mutated gene I_i can be altered to any value j such that $A(I_i, j) = 1$.

4.3The Proposed Heuristic Migration Operators

Almost all related works [14], [25] – [28]adopt similar genetic operators. However, to improve the performance of any evolutionary algorithm, one should design some problem-specific operators. This motivates us to propose a heuristic mutation operator coined as *heuristic migration* operator to be applied with probability p_{hm}. This operator is proposed to act as a heuristic partition generator that can exploit information from the neighborhood relations between nodes of the network.

For an individual I and under the control of p_{hm}, the proposed heuristic migration operator will change the community belongingness of node I_i if it appears to be either weakly- or neutrally-neighborhood node within other nodes of its community. If I_i is a weakly-neighborhood node in community C, then the migration operator will migrate it to another community that would satisfy with its nodes the highest strongly-neighborhood relation. Otherwise if I_i is a neutrally-neighborhood node in community C, then the migration operator will either leave the node inside its community or migrate it to another community that would also satisfy inside it an equal neighborhood relation. Algorithm 1 recapitulates the main steps of the proposed heuristic migration operator.

```
Algorithm 1: Heuristic Migration ($I$, $n$, $A$,$p_{hm}$)
set $\mathcal{C} \leftarrow \delta(I)$ // decode $I$
for $i = 1$ to $n$ do
    // migrate node $I_i$ with control
    if( $rand \leq p_{hm}$)
        set $C_i \leftarrow Community\_ID(I_i)$
        set $K \leftarrow max(Community\_ID(I))$
        set $k\_I_i\_in \leftarrow m(I_i, C_i)$
        set $k\_I_i\_out \leftarrow \overline{m}(I_i, C_i)$
        if ($k\_I_i\_in < k\_I_i\_out$) //weakly neighborhood node
            set $C \leftarrow argmax_{C_j \in \mathcal{C}}(m(I_i, C_j))$
            set $Community\_ID(I_i) \leftarrow C$
        elseif($k\_I_i\_in = k\_I_i\_out$) //neutrally neighborhood node
            // migrate or leave node $I_i$ with equal probability
            if( $rand$ ≤0.5)
                set $C \leftarrow argmax_{C_j \in \mathcal{C}, C_j \neq C_i}(m(I_i, C_j))$
                set $Community\_ID(I_i) \leftarrow C$
            end if
        end if
    end if
end for
```

5. Simulation Results

In this section, we will test the performance of the proposed heuristic MOO model under one, more or less, commonly used setting found in the literature for an evolutionary algorithm. Population size $N = 300$, neighborhood size $T = 5$, maximum number of generation $max_t = 100$, and $p_c = 0.8$. Also, the results report the impact of the proposed heuristic migration operator on the final performance of the competent MOO models. Either heuristic migration operator with $p_{hm} \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$or mutation operator with $p_m = 0.2$ (i.e. $p_{hm} = 0.0$) is used.

First, a test-bed of four commonly used real life networks is explained in the light of their bounded difficulties. Then, the performance of the algorithm is evaluated in terms of *convergence reliability* and *convergence velocity*. Convergence reliability is expressed by evaluating the average of normalized mutual information (NMI) over ten different runs for each network. Normalized mutual information between two partitions $\mathcal{A}$ and $\mathcal{B}$ of a network $\mathcal{N}$ of n nodes, is the normalization of the mutual information (MI) score between $\mathcal{A}$ and $\mathcal{B}$ being scaled between 0 (no mutual information) and 1.0 (perfect correlation) [32]. Consider the confusion matrix $c = [c_{ij}]$, $i = 1, \dots, K_{\mathcal{A}}$and $j = 1, \dots, K_{\mathcal{B}}$, where c_{ij} be the number of nodes of community i of $\mathcal{A}$ that are also in community j of $\mathcal{B}$. Then,

$$NMI(\mathcal{A},\mathcal{B}) = \frac{-2\sum_{i=1}^{K_{\mathcal{A}}}\sum_{j=1}^{K_{\mathcal{B}}} c_{ij} log(c_{ij} * n / c_i c_j)}{\sum_{i=1}^{K_{\mathcal{A}}} c_i log(c_i/n) + \sum_{j=1}^{K_{\mathcal{B}}} c_j log(c_j/n)} \quad (19)$$

wherec_i and c_j are the sum of elements of community i in $\mathcal{A}$ and community j in $\mathcal{B}$, respectively. On the other hand, convergence velocity is evaluated by the maximum number of generations required to get the optimal NMI^* in all ten runs, if such case appears.

5.1 Data Sets

The first famous network used by all community detection algorithms is *Zachary's network* drawn from his 2-years "karate club" study [33]. The friendship relations among 34 club's members were considered in his study to construct networks of ties. However, due to a dispute between the club's administrator (denoted in Zachary's study by node number 1) and one of the club's instructors (node 33), the club's members were split into two smaller club communities. The communities are: C_1 of $|C_1| = 16$ members centered around administrator 1 and C_2 of $|C_2| = 18$ members centered around instructor 33. The total number of friendship relations in the network is $m(\mathcal{N}) = |E| = 78$.

The second well-known network is *Bottlenose Dolphins network* [34]. A New Zealand's population of 62 bottlenose dolphins living off Doubtful Sound was compiled by Lusseau's study to draw a seven year complex couple relations. A total of $m(\mathcal{N}) = |E| = 159$ relations is explored in this network with two large groups.

The third network is *American football* game of Division I-A colleges being compiled by Girvan and Newman [5]. The network consists of 115 teams playing championship games against each other during the season of fall 2000. The teams are divided into 12 conferences coming from roughly 12 different geographic grounds. With a total of 613 games, each conference plays the majority of games within its own teams. However, teams from two different conferences can play games against each other. One can formally state this network as $n(\mathcal{N}) = 115$, $m(\mathcal{N}) = 613$ and $|\mathcal{C}^*| = 12$.

Finally, the fourth network is the *Krebs' books* on American politics being compiled by Krebs [7]. This network consists of 105 US politics books sold by the online bookseller Amazon.com. A total of 440 co-purchasing by the same buyers is found in this network being divided into two political alignment groups and one small un-aligned group of 13 books.

5.2 Results and Discussions

Table-1 and Table-2 report, respectively,averageand best results over ten runs of MOEA/D (without and with heuristic, denoted respectively by, $MOEA/D$ and$hMOEA/D$) on the four networks. For both Zachary's karate club network and Bottlenose Dolphins network, the performance should also be justified according to the convergence velocity while approachingoptimal $NMI^* = 1$. Moreover, Figure 1 and 2 depict detection results on Zachary's karate club network and Dolphin network, respectively. For Zachary's karate club network (see Figure 1-a), the results clarify that when no heuristic is used, $MOEA/D$fails to detect the community belongingness of the neutrally-neighborhood node number 10 in all ten runs, resulting in local optima at $NMI = 0.8372$ (see square-node connected with circle-nodes community in Figure 1-b). Note that member 10 in this club network has only two neighborhood relations, each being belong to a distinct club community. Moreover, $MOEA/D$ also fails (in two runs) to detect the correct community of the second neutrally-neighborhood node (i.e. member 3) which has five neighborhood relations inside its community and five other relations with other nodes of the opposite community. This results in approaching another local optima at local optima at $NMI = 0.8365$ (average of 10 runs, then, will yield $NMI = 0.8370$ as reported in Table 1). However, with the help of the proposed heuristic migration operator, $hMOEA/D$succeed in detecting the correct community of these neutrally-neighborhood nodes (see Figure 1-c). Moreover, the results demonstrate that increasing impact of the proposed heuristic operatoraccelerates $hMOEA/D$ to approach the correct detection solution in all runs.

Also, in Bottlenose Dolphin network (see Figure 2-a),$MOEA/D$ and $hMOEA/D$resemble its corresponding behavior in Zachary's network. For $MOEA/D$, we see that it does not reach the optimal division, but it converges to different local optima with maximum reliability at $NMI = 0.9022$ (see square-nodes 8, 20, and 28 being isolated from their correct square-nodes community in Figure 2-b). The correct division $\mathcal{C}^*$ (see Figure 2-c) is obtained in all runs of $hMOEA/D$ (with $p_{hm} \geq 0.6$).

For the remaining two networks, again $hMOEA/D$ performs better, on average, than $MOEA/D$. However, both $MOEA/D$ and $hMOEA/D$ do not reach the global optimum solution, and this is due to the increased complexities and nodes overlapping exist in these networks. This suggests further modification to the current work.

Table 1-Convergence reliability (and convergence velocity written between parentheses if $NMI^* = 1$ is approached) on four real-life networks.Results reported as average NMI over ten runs.

Network	p_{hm}					
	0.0	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
Zachary's club	0.8370	0.8556	1(23)	1(8)	1(5)	1(2)
Bottlenose Dolphins	0.8073	0.9504	0.9669	1(13)	1(18)	1(4)
Football 2000	0.6194	0.7653	0.8076	0.8379	0.8721	0.8768
Kreb's books	0.5691	0.5825	0.5921	0.5950	0.5925	0.5906

Table 2-Convergence reliability (and convergence velocity written between parentheses if $NMI^* = 1$ is approached) on four real-life networks. Results reported as best NMI over ten runs.

Network	$\boldsymbol{p_{hm}}$					
	0.0	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
Zachary's club	0.8372	1	1	1	1	1
Bottlenose Dolphins	0.9022	1	1	1	1	1
Football 2000	0.6886	0.8032	0.8466	0.8662	0.9011	0.9083
Kreb's books	0.6850	0.6588	0.6456	0.6446	0.6310	0.6138

6. Conclusion

Three types of node-neighborhood relations are introduced in this paper. Moreover, a problem specific heuristic operator is proposed, based also on the defined neighborhood relations, to improve the convergence velocity and convergence reliability of the adopted multi-objective optimization model. The basic idea of this heuristic operator is to allow nodes to migrate between different communities. The performance of the proposed MOEA is evaluated under four common real-life social networks. The results demonstrate the positive impact of injecting the roles of the defined neighborhood relations into the multi-objective community detection model. Moreover, remarkable improvement comes after introducing the heuristic operator, allowing the multi-objective community detection model to transcend its limits. Further research direction can be followed after the current work. One suggestion, which is our current interest, is to redirect the design of $\Phi(\mathcal{C})$ according to the *neighborhood relations* of intra-community and inter-community nodes and thus 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.

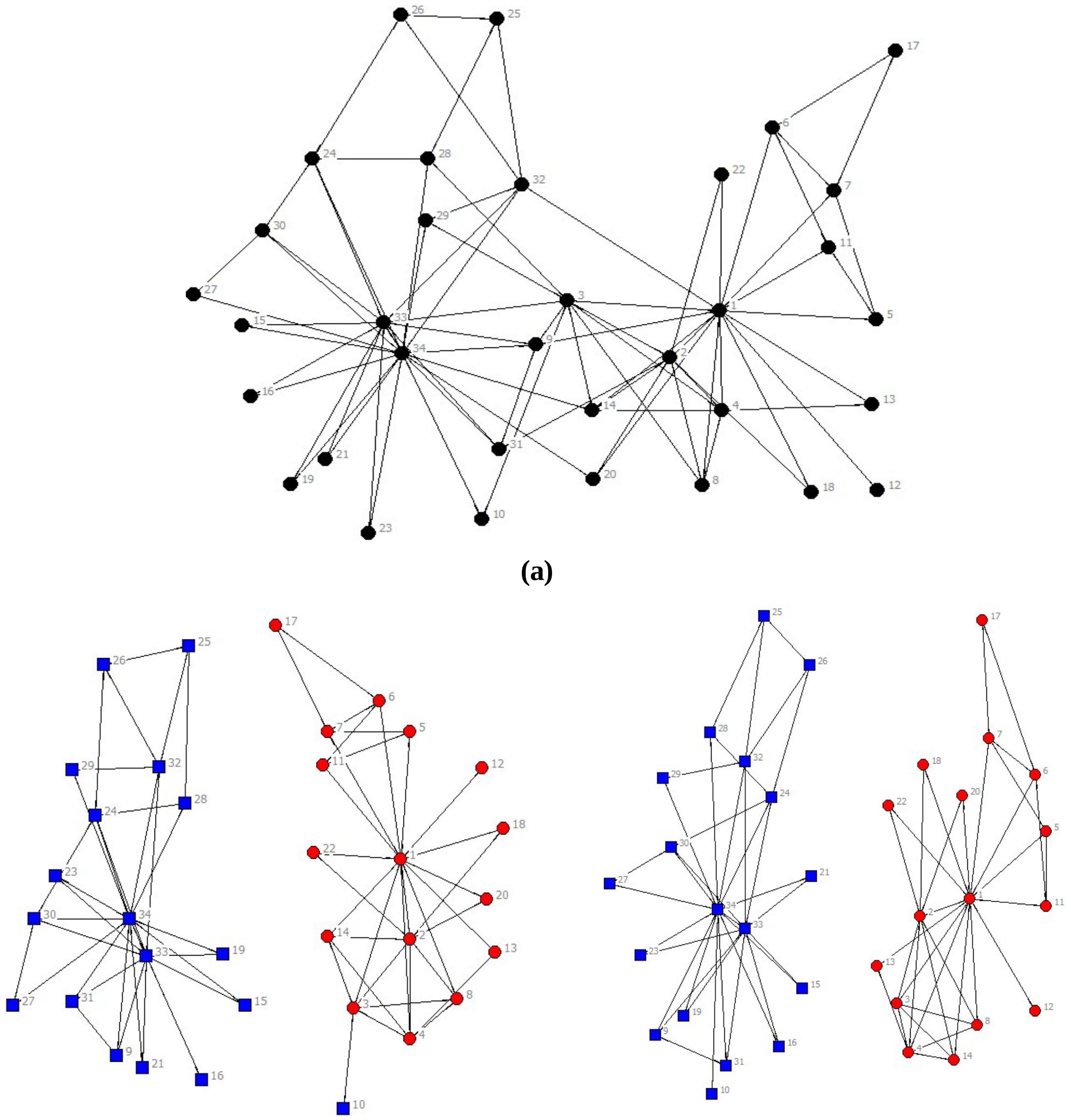
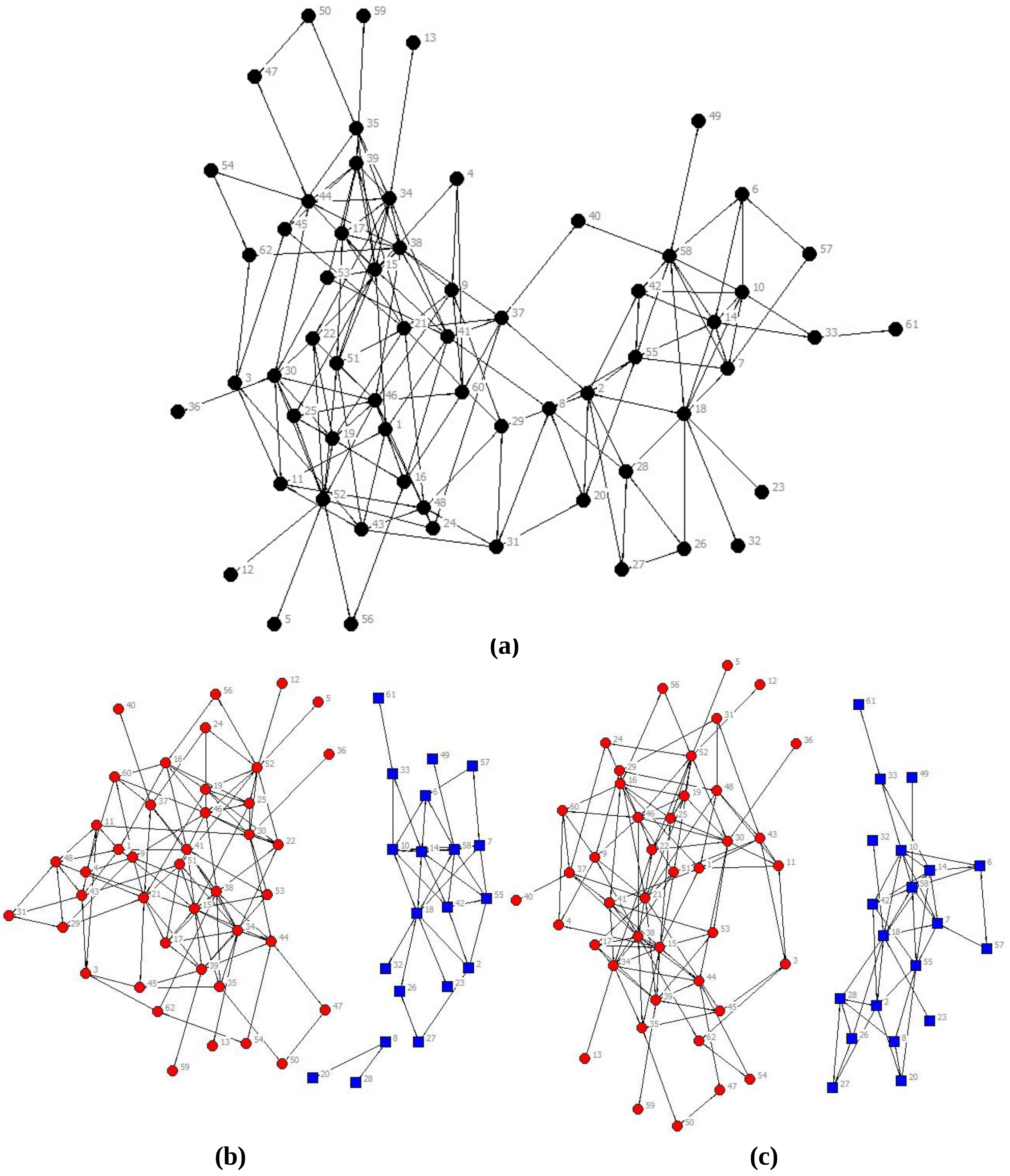


Figure 1- (a) Zachary's Karate club network. (b) Local optimum community structure obtained by $MOEA/D$(at $NMI = 0.8372$). (c)Correct community structure obtained by $hMOEA/D$(at $NMI = 1.0$).



(a)

(b) **(c)**

Figure 2- (a) Dolphin network. (b) Local optimum community structure obtained by $MOEA/D$(at $NMI = 0.9022$). (c) Correct community structure obtained by $hMOEA/D$(at $NMI = 1.0$).

References

1. Newman, M. E., and Girvan, M. **2004**. Finding and evaluating community structure in networks. *Physical review E*, 69(2), 026113.
2. Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. **2008.** On modularity clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 20(2), 172-188.
3. Schaeffer, S. E. **2007.** Graph clustering. *Computer Science Review*, *1*(1), 27-64.
4. Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. **2008.** Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*,P10008.
5. Girvan, M., and Newman, M. E. **2002**. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), 7821-7826.

6. Lozano, S., Duch, J., and Arenas, A. **2007.** Analysis of large social datasets by community detection. *The European Physical Journal Special Topics*, 143(1), 257-259.

7. Newman, M. E. **2006.** Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577-8582.

8. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. **2004.** Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9), 2658-2663.

9. Clauset, A., Newman, M. E., and Moore, C. **2004.** Finding community structure in very large networks. *Physical review E*, 70(6), 066111.

10. Lipczak, M., and Milios, E. **2009.** Agglomerative genetic algorithm for clustering in social networks. In Proceedings of the 11th Annual conference on Genetic and evolutionary computation (pp. 1243-1250). *ACM*.

11. Newman, M. E. **2004.** Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6), 066133.

12. Pons, P., and Latapy, M. **2006.** Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2), 191-218.

13. Pujol, J. M., Béjar, J., and Delgado, J**. 2006.** Clustering algorithm for determining community structure in large networks. *Physical Review E*, 74(1), 016107.

14. Pizzuti, C. **2012**. A multiobjective genetic algorithm to find communities in complex networks. *Evolutionary Computation, IEEE Transactions on*, 16(3), 418-430.

15. Zhang, Q., and Li, H. **2007.** MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *Evolutionary Computation, IEEE Transactions on*, 11(6), 712-731.

16. Arenas, A., and Diaz-Guilera, A. **2007.** Synchronization and modularity in complex networks. *The European Physical Journal Special Topics*, 143(1), 19-25.

17. Schuetz, P., and Caflisch, A. **2008.** Multistep greedy algorithm identifies community structure in real-world and computer-generated networks. *Physical Review E*, 78(2), 026112.

18. Lancichinetti, A., Fortunato, S., and Kertész, J. **2009.** Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3), 033015.

19. Feng, Z., Xu, X., Yuruk, N., and Schweiger, T. A. **2007.** A novel similarity-based modularity function for graph partitioning. In Data Warehousing and Knowledge Discovery Springer Berlin Heidelberg, pp:385-396.

20. Firat, A., Chatterjee, S., and Yilmaz, M. **2007.** Genetic clustering of social networks using random walks. *Computational Statistics and Data Analysis*, 51(12), 6285-6294.

21. He, D., Wang, Z., Yang, B., and Zhou, C. **2009.** Genetic algorithm with ensemble learning for detecting community structure in complex networks. In Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. Fourth International Conference on *IEEE*. pp:702-707.

22. Shi, C., Wang, Y., Wu, B., and Zhong, C. **2009.** A new genetic algorithm for community detection. In Complex Sciences *Springer Berlin Heidelberg*, pp:1298-1309.

23. Pizzuti, C. **2008.** Ga-net: A genetic algorithm for community detection in social networks. In Parallel Problem Solving from Nature–PPSN X *Springer Berlin Heidelberg*, pp:1081-1090.

24. Chira. C, Gog. A. **2011.** Collaborative Community Detection in Complex Networks, *Hybrid Artificial Intelligent Systems,Lecture Notes In Computer Science*, 6678, pp:380-387.

25. Agrawal, R. **2011.** Bi-objective community detection (bocd) in networks using genetic algorithm. In Contemporary Computing, *Springer Berlin Heidelberg*, pp:5-15).

26. Shi, C., Yan, Z., Cai, Y., and Wu, B. **2012.** Multi-objective community detection in complex networks. *Applied Soft Computing*, 12(2), 850-859.

27. Hafez, A. I., Al-Shammari, E. T., ella Hassanien, A., and Fahmy, A. A. **2014.** Genetic algorithms for multi-objective community detection in complex networks. In Social Networks: A Framework of Computational Intelligence *Springer International Publishing*, pp:145-171.

28. Gong, M., Cai, Q., Chen, X., and Ma, L. **2014**. Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition. *Evolutionary Computation, IEEE Transactions on*, 18(1), 82-97.

29. Coello, C. A. C., Van Veldhuizen, D. A., and Lamont, G. B**. 2002**. *Evolutionary algorithms for solving multi-objective problems*, 242, New York: Kluwer Academic.

30. Srinivas, N., and Deb, K**. 1994.** Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary computation*, 2(3), 221-248.

31. Park. Y. J. and Song. M. S. **1989.** A genetic algorithm for clustering problems, *Proc. 3rd Annu. Conf. Genet. Algorithms*, pp. 2-9.

32. MacKay D. J. C, **2002.** Information theory. *Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge University Press.

33. Zachary, W. W. **1977.** An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 452-473.

34. Lusseau, D. **2003**. The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(Suppl 2), S186-S188.