Iraqi Journal of Science

ISSN: 0067-2904

An Evolutionary Bi-clustering Algorithm for Community Mining in Complex Networks

Saja Salah Abdul Emeer*, Saja Hattem Kareem, Baraa Ali Atea

Department of Computer Science, College of Science, Baghdad University, Baghdad, Iraq

Abstract

A network (or formally a graph) can be described by a set of nodes and a set of edges connecting these nodes. Networks model many real-world phenomena in various research domains, such as biology, engineering and sociology. Community mining is discovering the groups in a network where individuals group of membership are not explicitly given. 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 evolutionary algorithms (EAs) takes a remarkable interest. To this end, the aim of this paper is to present the general statement of community detection problem in social networks. Then, it visits the problem as an optimization problem where a *modularity-based* (Q) and *normalized mutual information* (NMI) metrics are formulated to describe the problem. An evolutionary algorithm is then expressed in the light of its characteristic components to tackle the problem. The presentation will highlight the possible alternative that can be adopted in this study for individual representation, fitness evaluations, and crossover and mutation operators. The results point out that adopting NMI as a fitness function carries out more correct solutions than adopting the modularity function$Q$. Moreover, the strength of mutation has a background role. When coupled with non elite selection, increasing mutation probability could results in better solutions. However, when elitism is used, increasing mutation probability could bewilder the behavior of EA.

Keywords: complex networks, graph co-clustering, EA, NP-hard.

خوارزمية تطورية ذات تصنيف ثنائي الأبعاد لكشف الجاليات في الشبكات المعقدة

سجى صلاح عبد الأمير*، سجى حاتم كريم، براء علي عطية

قسم الحاسبات ، كلية العلوم ، جامعة بغداد ، بغداد ، العراق

الخلاصة

يمكن وصف الشبكة كمخطط من خلال مجموعة من العقد ومجموعة من الروابط التي تربط هذه العقد. تعتبر الشبكات نموذج لعديد من الظواهر في العالم الحقيقي و في المجالات البحثية المختلفة ، مثل علم الأحياء والهندسة وعلم الاجتماع. الشبكات الاجتماعية والشبكات البايولوجية، الشبكة العالمية، والإنترنت، وشبكات التعاون وشبكات الطاقة، والفيسبوك والبيئية والاتصالات وشبكات النقل ماهي الا أمثلة .و دراسة هذه الشبكات المعقدة تشمل باحثين من تخصصات مختلفة كثيرة، على سبيل المثال ،علوم الكمبيوتر، والهندسة، وعلم الأحياء، والرياضيات، والفيزياء، وعلم الاجتماع، مما يؤدي إلى تشكيل العديد من المجالات المتعددة التخصصات. اكتشاف المجتمع هو اكتشاف المجموعات المرتبطة بالشبكة من حيث انها عضو صريح في الشبكة او لا. ومن بين الطرق المقترحة في هذا المجال الخوارزميات التطورية (EAs) والتي تأخذ اهتماما ملفتا للنظر في الفترة الاخيرة . فالهدف من هذا البحث هو تقديم بيان عام للمشكلة وكشف المجتمعات في الشبكات الاجتماعية و تطمح هذه الرسالة الى النظر للمشكلة كونها مشكلة امثلية مستندة الى مقياسي (Q)

*Email:Saja.salah46@yahoo.com

و (NMI) واللتان تعتبران مقياسان لوصف المشكلة . ثم يتم التعبير عن الخوارزمية التطورية في ضوء العناصر المميزة لها لمعالجة هذه المشكلة. و سوف يتم تسليط الضوء على البدائل الممكنة التي يمكن اعتمادها في هذه الدراسة لتمثيل الأفراد. وتشير النتائج إلى أن اعتماد NMI بوصفها مقياس لكفاءة الفرد تنفذ حلول أكثر دقة من اعتماد مقياس Q. اضافة لذلك، احتمالية قوة الطفرة (pm) لها دوركبير. عندما يقترن الافراد مع عدم اختيار النخبة منهم، وزيادة احتمال الطفرة يمكن أن يؤدي إلى حلول أفضل. ومع ذلك، عند استخدام النخبة، وزيادة احتمال الطفرة قد يربك سلوك الـ EA.

1 Introduction

Complex networks constitute an efficacious formalism to represent the relationships among the objects composing many real world systems. Collaboration networks, the Internet, the world-wide-web, biological networks, communication and transport networks, social networks are just some examples. 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 [1].

In the last few years many different approaches have been proposed to uncover community structure (i.e. to detect communities) in networks. In general, these techniques can be categorized into three main approaches: 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 [2-4]. 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 [5, 6].

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 [7, 8]. The main aim of this paper is to revisit and elaborate both modularity (Q) and normalized mutual information (NMI) metrics as an optimization models that can cast on the properties of community structure. Then, based on the model definition, an evolutionary algorithm (EA) is proposed to tackle the problem. The remainder of this paper is organized as follows. Section 2 presents basic concepts relating to the community detection problem. Section 3 presents related works while section 4 introduces our formulation for the evolutionary community detection problem. Results on two commonly used social networks are reported in section 5. Finally, conclusions and future work are pointed out in section 6.

2 Problem Statement

A complex network is a representation of a complex system from real life in terms of nodes and edges, where a node is an individual member in the system and an edge is a link between nodes according to a relation in the system [9]. As an example, in a social network, a node represents a person and an edge represents social interaction between two people. One of the main problems in the study of the complex networks is the detection of community structure. There are two main challenges in discovering communities. The first is that it is not known a priori the number of groups present in a given network. The second is that the communities may overlap, i.e. some nodes can belong to more than one cluster. The membership of an entity to many groups is very common in real world networks. For example, in a social network, a person may participate to many interest groups. Also in real world, objects often have multiple roles. Example, a professor collaborates with researchers in different fields; a person has his family group, as well as, friends group at the same time etc.

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 according to the correlation among n objects. Thus, if C_{k1} and C_{k2} 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 (see Figure-1) [10]. In Figure 1, one can see that the traditional clustering (left matrix) searches a partition of all objects into k disjoint groups (here $k = 4$). The right matrix, however, depicts bi-clustering where a set of blocks containing a *consistent local pattern* is to be found (here $k = 3$). Note that it is not generally possible to display several bi-clusters at the same time as contiguous blocks.

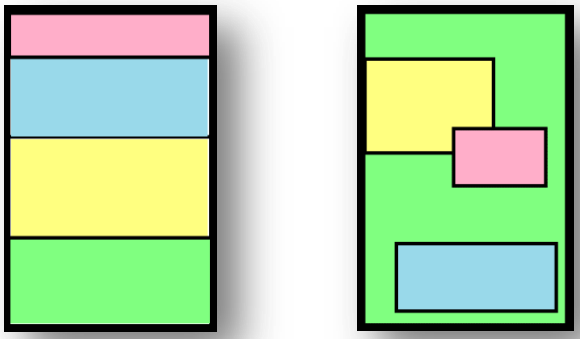

Figure 1-Comparison between clustering (left) and bi-clustering (right)

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 [11]. 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.

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

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
- $\bar{m}(v, C_i) = |\{(v, w) \in E \wedge w \notin C_i| = \sum_{w \notin C_i} A(v, w)$ 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.
- $\bar{m}(C_i) = \left|E\left(C_i, C_j\right)\right|_{\forall j \neq i} = \sum_{v \in C_i} \bar{m}(v, C_i)$ is the number of *inter-cluster* connections of C_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
- $\bar{m}(\mathcal{C}) = |E(G) / E(\mathcal{C})|$ 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 [12] 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.* [4] semantically define C as a *weak* community if $m(C) > \bar{m}(C)$, or as *strong* community if $\forall v \in C \Rightarrow m(v) > \bar{m}(v)$. Formally speaking,

Definition 1: Strong Community

The sub graph C is a community in a strong sense if

$$\boldsymbol{k_i^{in}(C) > k_i^{out}(C), \forall_i \in C .} \quad \textbf{(2)}$$

In a strong community each node has more connections within the community than with the rest of the graph.

Definition 2: Weak Community

The sub graph C is a community in a weak sense if

$$\boldsymbol{\sum_{i \in v} k_i^{in}(C) > \sum_{i \in v} k_i^{out}(C)} \quad \textbf{(3)}$$

In a weak community the sum of all degrees within C is larger than the sum of all degrees toward the rest of the network. Clearly a community in a strong sense is also a community in a weak sense, while the converse is not true.

3 Related work

During the past decade, the research on analyzing the community structure in complex networks has drawn a great deal of attention. Dominated ones are:

M. Girvan and M. E. J. Newman (2002) *proposed* a divisive hierarchical clustering method to identify communities. The algorithm looks for the edges in the network that are most "between" other vertices, meaning that the edge is, in some sense, responsible for connecting many pairs of others, or in other words, looks edges that lie between communities. Such edges need not be weak at all in the similarity sense. They tested this method on computer generated and real-world graphs whose community structure is already known and they detect community with high degree of success [13].

Radicchi et al. (2004) proposed a divisive hierarchical algorithm to identify communities based on the concept of edge-clustering coefficient, defined in analogy with the node clustering coefficient. The edge-clustering coefficient is the number of triangles an edge participates, divided by the number of triangles it might belong to, given the degree of the adjacent nodes. Their algorithm works like that of Newman and Girvan, but it is faster. The main difference is that instead of choosing to remove the edge with the highest edge "betweenness", the removed edges are those having the smallest value of edge-clustering coefficient [4].

Pons and Latapy (2006) introduced an agglomerative hierarchical algorithm to compute the community structure of a network. The algorithm starts from a partition of the graph in which each node is a community, and then merges the two adjacent communities (i.e. having at least a common edge) that minimize the mean of the square distances between each vertex and its community. The distances between communities are recomputed and the previous step is repeated until all the nodes belong to the same community. In order to decide the best partitioning to choose, the modularity criterion of Girvan and Newmann is adopted [14].

Clauset et al. (2004) proposed agglomerative hierarchical clustering method to find communities. They used it to analyze a network of items for sale on the web-site of a large online retailer (amazon .com), The network has more than 400 000 vertices and 2 million edges. They could extract meaningful communities from this network [5].

Recently, the relaxed nature of meta-heuristic evolutionary algorithm based optimization methods (i.e. EA-based community detection algorithms), makes them very suitable to reduce the complexity of the problem and to approach adequate and more reliable solutions than the existing state-of-the-art community detection algorithms. The prominent EA-based community detection algorithms that successfully beat existing ones are: Pizzuti's MOO model (2012), Shi, Yan, Cai, & Wu's MOO model [15], and Gong, Cai, Chen & Ma's MOO model [16].

Shi, Yan, Cai, & Wu [15] they defined 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.

Pizzuti [12] formulated a multi-objective maximization model for a partition $\mathcal{C}$, the first objective is to maximize *community score* [17] while the second objective is to maximize the *community fitness* proposed by Lancichinetti, Fortunato & Kertész [18]. Formally speaking:

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

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

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

Also, here $\alpha > 0$ control the size of community C_i. Then [12]) defines $\Phi_2(\mathcal{C})$ as:

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

After evolving a set of solutions, Pizzuti suggested selecting the partition with the maximum modularity value $Q(\mathcal{C})$.

4 The proposed community mining algorithm

An evolutionary algorithm (EA) evolves a constant-size population of elements (called chromosomes) by using the genetic operator of *reproduction*, *crossover* and *mutation*. Each chromosome represents a candidate solution to a given problem and it is associated with a *fitness value* that reflects how good it is, with respect to the other solutions in the population. Generally, a chromosome is encoded as a string of bits from a binary alphabet. The reproduction operator copies elements of the current population into the next generation with a probability proportional to their fitness (this strategy is also called roulette wheel selection scheme). The crossover operator generates two new chromosomes by crossing two elements of the population selected proportional to their fitness. The mutation operator randomly alters the bits of the strings.

4.1 Genetic Representation and Initialization

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 [19], the adopted representation is the locus-based adjacency representation being proposed by Park and Song . 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 (see Figure 2). The allele value of each gene can be varied from 1 to 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 $\mathcal{C}$ (i.e. $\forall i \in C, A(i,j) = 0$).

4.2 Fitness Function

The quality of each individual can be evaluated in terms 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) [20]. 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,

$$\boldsymbol{NMI(\mathcal{A},\mathcal{B})} = \frac{-2 \sum_{i=1}^{K_{\mathcal{A}}} \sum_{j=1}^{K_{\mathcal{B}}} \boldsymbol{c_{ij} log(c_{ij} * n / c_i c_j)}}{\sum_{i=1}^{K_{\mathcal{A}}} \boldsymbol{c_i log(c_i/n)} + \sum_{j=1}^{K_{\mathcal{B}}} \boldsymbol{c_j log(c_j/n)}} \quad \textbf{(7)}$$

where c_i and c_j are the sum of elements of community i in $\mathcal{A}$ and community j in $\mathcal{B}$, respectively. Consider a correct partition of social network $\mathcal{A}$ has $K_{\mathcal{A}}$ communities and also consider a candidate partition $\mathcal{B}$ of a chromosome has $K_{\mathcal{B}}$ communities. Then, for the above formulation, one can see that the confusion matrix c will have $K_{\mathcal{A}}$ rows and $K_{\mathcal{B}}$ columns, where each entery c_{ij} in c features the nodes being belong to the correct community i in partition $\mathcal{A}$ and the candidate community j in partition $\mathcal{B}$.

Another alternative to quantify EA individuals is modularity defined as:

$$\boldsymbol{Q(\mathcal{C}) = \sum_{i=1}^{K} \left[\frac{|E(C_i, C_i)|}{m(\mathcal{C})} - \left(\frac{\sum_{v \in C_i} m(v)}{2m(\mathcal{C})} \right)^2 \right]} \quad \textbf{(8)}$$

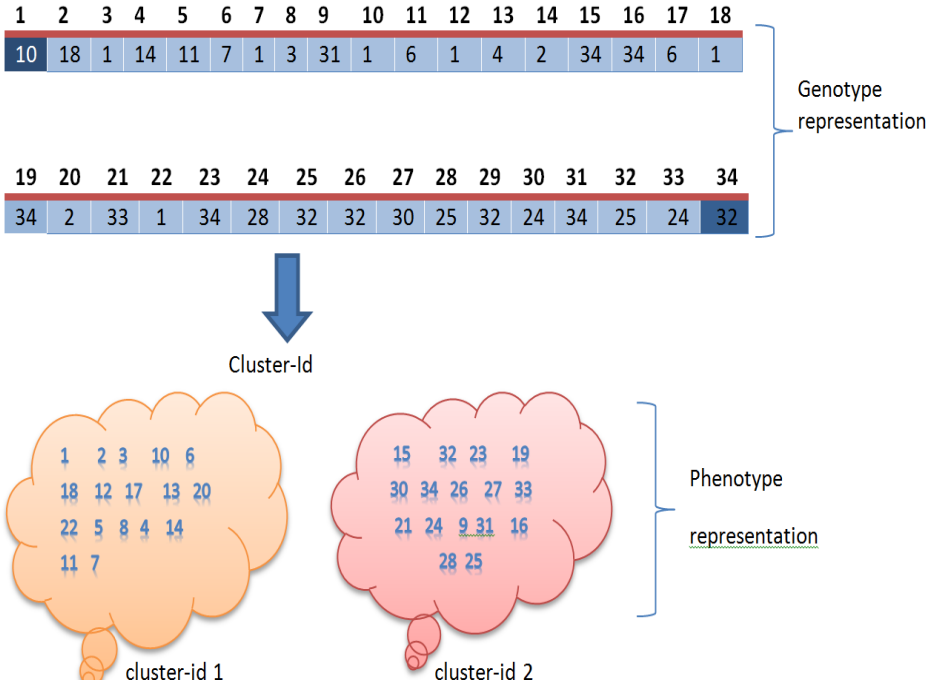

Figure 2-Individual representation.

4.3 Evolutionary Operators

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:

$\forall i, 1 \leq i \leq n$

$$\boldsymbol{I'_i = \begin{cases} I_i^1 \; if \; r \leq 0.5 \\ I_i^2 \; otherwise \end{cases}} \quad \textbf{(9)}$$

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 $A(I_i, j) = 1$.

5 Results and Discussions

In this section, we test the performance of the proposed model. The characteristic components of the EA are quantified to their, more or less, commonly used setting found in the literature. Population size $N = 25$, maximum number of generation $max_t = 25$, and $p_c = 0.8$. The results report the impact of the proposed fitness models, selection with elitism, mutation strength. Two real life networks with known community structures (i.e. correct partition) are experimented with. The performance of the algorithm is evaluated (over five different runs for each network) in terms of confusion matrix, modularity, and NMI.

The well-known "karate club" study of Zachary [21] forms a social network consisting of 34 members of a karate club over a period of two years. During the course of the study, a disagreement developed instructor, which ultimately resulted in the instructor's leaving and starting a new club, taking about a half of the original club's members with him. Zachary constructed a network of friendships between members of the club, using a variety of measures to estimate the strength of ties

between individuals. Figure-3 shows the network, with the instructor and the administrator represented by nodes 1 and 34, respectively. Figure-4 shows the correct division of this network into two communities (depicted in two different colors) of roughly equal size.

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. In this network, dolphins represented as vertices have a link with each other if they are observed together more often than expected by chance over a period of seven years from 1994 to 2001. A total of $m(\mathcal{N}) = |E| = 159$ relations is explored in this network with two large groups. Figure 3 and 4 depict the original network with its correct partition, respectively.

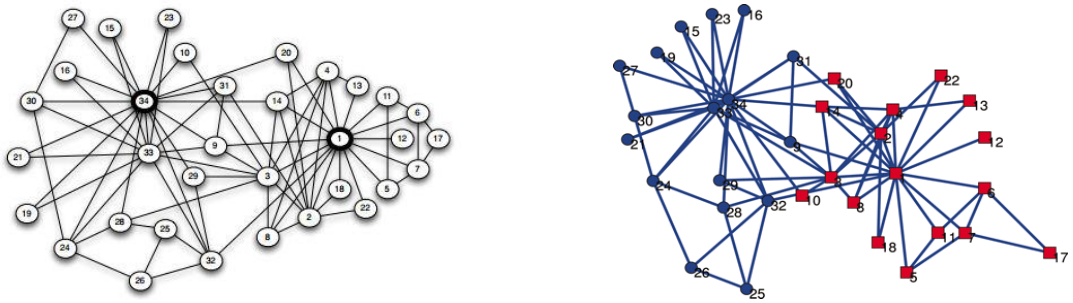

Figure 3- Left: the friendship relations of Zachary's karate club network. Rigth**:** correct partition of Zachary's karate network into two communities.

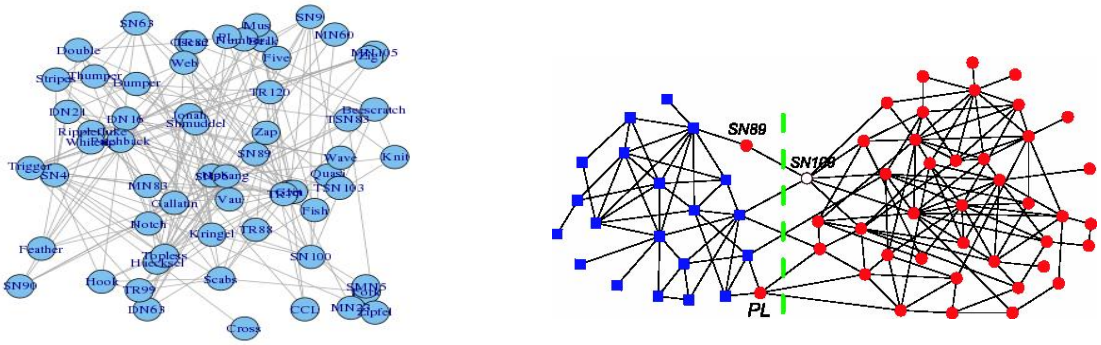

Figure 4- Left: the friendship relations of Dolphin network. Rigth**:** correct partition of Dolphin network into two communities.

Table 1- Quantitative average results for 5 different runs on Zacharys Karate Club where NMI.

Generations	Without Elitism		With Elitism	
	Pm=0.1	Pm=0.7	Pm=01	Pm=0.7
1	0.68852	0.69254	0.6953	0.70282
5	0.72790	0.70644	0.78622	0.72628
10	0.70400	0.6061	0.80694	0.78064
15	0.65098	0.59642	0.81766	0.81768
20	0.62610	0.56448	0.81794	0.8372
25	0.64970	0.56086	0.81794	0.8372

Table 2- Quantitative average results for 5 different runs on Zacharys Karate Club.

Generations	Without Elitism				With Elitism			
	Pm=0.1		Pm=0.7		Pm=01		Pm=0.7	
	Q	NMI	Q	NMI	Q	NMI	Q	NMI
1	0.33	0.71	0.24	0.71	0.33	0.74	0.16	0.75
5	0.04	0.73	0.45	0.73	0.16	0.74	0.16	0.75
10	0.26	0.73	0.43	0.72	0.16	0.74	0	0.75
15	0.29	0.70	0.42	0.70	0.16	0.74	0	0.75
20	0.16	0.70	0.21	0.71	0	0.75	0	0.75
25	0.21	0.70	0.24	0.67	0	0.75	0	0.75

Table 3- Quantitative average results for 5 different runs on Bottlenose Dolphins where NMI

Generations	Without Elitism		With Elitism	
	Pm=0.1	Pm=0.7	Pm=01	Pm=0.7
1	0.6596	0.67602	0.6596	0.6596
5	0.72282	0.6563	0.72282	0.72282
10	0.80044	0.5971	0.80044	0.80044
15	0.80868	0.65462	0.80868	0.80868
20	0.85666	0.58298	0.85666	0.85666
25	0.87926	0.59034	0.87926	0.87926

Table 4- Quantitative average results for 5 different runs on Bottlenose Dolphins.

Generations	Without Elitism				With Elitism			
	Pm=0.1		Pm=0.7		Pm=01		Pm=0.7	
	Q	NMI	Q	NMI	Q	NMI	Q	NMI
1	0.69	0.744	0.596	0.69	0.76	0.75	0.612	0.724
5	0.616	0.694	0.554	0.666	0.60	0.77	0.742	0.774
10	0.628	0.692	0.522	0.67	0.71	0.77	0.734	0.786
15	0.48	0.66	0.57	0.664	0.81	0.78	0.734	0.786
20	0.4	0.666	0.528	0.702	0.83	0.78	0.734	0.786
25	0.508	0.676	0.4	0.67	0.85	0.78	0.84	0.796

The results reported in Tables 1 – 4 reveal the ability of the tested EA on social networks and the impact of different characteristic components of the proposed EA, point out that adopting NMI as a fitness function carries out more correct solutions than adopting the modularity function$Q$. Moreover, the strength of mutation has a background role. When coupled with non elite selection, increasing mutation probability could results in better solutions. However, when elitism is used, increasing mutation probability could bewilder the behavior of EA.

The structure of Bottlensoe Dolphin network seems to be easier tackling by the proposed EA than the structure of Zachary's karate club network. Some EA's runs reach the correct (i.e. optimal solution at $NMI = 1.0$) for Bottlensoe Dolphin network. However in Zachary's karate club network, the proposed EA mis-detect at least one node (either node 3 or node 10).Node's characteristics (represented by number of its connections) have also an impact to the quality of the final solution provided by the proposed EA. For example, in Zachary's karate club network, the mis-detection of only one, but different, node may not imply similar quality of the final result. Mis-detection of node 3 results in final solution quality measured at $NMI = 0.8365$. However, mis-detection of node 10 has a solution with quality $NMI = 0.8372$. This is expected as node 10 has less neighborhoods (only two connections) than node 3 (has ten connections). As compared with non elitist, using elitism with selection can preserve the best qualified solution to further generations which in turns causes EA to reach more correct solutions.

The strength of mutation (measured by its probability of occurance) has a secondary role. When coupled with non elite selection, increasing mutation probability could results in better solutions. However, when elitism is used, increasing mutation probability could bewilder the behavior of EA, resulting in improper improvement in the solution's quality.

6 Conclusions

The results reported in this paper reveal the ability of EA to handle community detection problem in social networks. Also, the results reports the impact of different characteristic components of the proposed EA.

The work drafted in this paper can be extended to include further investigations including, but not limted to another chromosome representation can be explored. For example, one can consider that one complete cluster, rather than group of clusters, can be encoded in the chromosome. The EA, here, has to deliver at the end of generations, one community with its intra-connected nodes. Remaining un-detected nodes and un-detected communities are subsequently determined by successive application of EA and Further community-structure wise fitness functions can be proposed, designed, and investigated. Mutation operator can be re-designed to approach a more effective one. The mutation operator can work as a local search operator to modify the community belongingness of those nodes satisfying a pre-determined condition. For example, a mutation operator can move a node from the community that forms with its nodes less-intra connections to a community that would form with its nodes more-intra connections.

References

1. Newman, M. E. **2004.** Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6), p:066133.
2. Newman, M. E., and Girvan, M. **2004.** Finding and evaluating community structure in networks. *Physical review E*, 69(2), p:026113.
3. 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*, 2008(10), p:10008.
4. 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), pp:2658-2663.
5. Clauset, A., Newman, M. E., and Moore, C.**2004.** Finding community structure in very large networks. *Physical review E*, 70(6), p:066111.
6. Pujol, J. M., Béjar, J., and Delgado, J. **2006.** Clustering algorithm for determining community structure in large networks. *Physical Review E*, 74(1), p:016107.
7. Arenas, A., and Diaz-Guilera, A. **2007.** Synchronization and modularity in complex networks. *The European Physical Journal Special Topics*, 143(1), pp:19-25.
8. Lancichinetti, A., Fortunato, S., and Kertész, J. **2009.** Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3), p:033015.
9. Dorogovtsev, S. N., and Mendes, J. F. F. **2001.** Effect of the accelerating growth of communications networks on their structure. *Physical Review E*, 63(2), p:025101.
10. Tanay, A., Sharan, R., and Shamir, R. **2005.** Bi-clustering algorithms: A survey. *Handbook of computational molecular biology*, 9(1-20), pp:122-124.
11. Newman, M. E. **2006.** Modularity and community structure in networks.*Proceedings of the national academy of sciences*, 103(23), pp:8577-8582.
12. Pizzuti, C. **2012.** A multiobjective genetic algorithm to find communities in complex networks. *Evolutionary Computation, IEEE Transactions on*, 16(3), pp:418-430.
13. Girvan, M., and Newman, M. E. **2002.** Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12), pp:7821-7826.
14. Pons, P., and Latapy, M. **2006.** Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005*, pp. 284-293. Springer Berlin Heidelberg.
15. Shi, C., Yan, Z., CAI, Y., and Wu, B. **2012.** Multi-objective community detection in complex networks. *Applied Soft Computing*, 12(2), pp:850-859.
16. 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), pp:82-97.
17. 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.
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), p:033015.

19. Agrawal, R. **2011.** Bi-objective community detection (bocd) in networks using genetic algorithm. In *Contemporary Computing* (pp:5-15). Springer Berlin Heidelberg.
20. MacKay, D. J. **2003.** *Information theory, inference and learning algorithms*. Cambridge University Press.
21. Zachary, W. W. **1977**. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pp:452-473.