Journal of Theoretical and Applied Information Technology
31st October 2019. Vol.97. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ENHANCED EVOLUTIONARY ALGORITHM WITH LOCAL HEURISTIC APPROACH FOR DETECTING COMMUNITY IN COMPLEX NETWORKS

[1] DHUHA ABDULHADI ABDULJABBAR, [2] SITI ZAITON MOHD HASHIM, [3] ROSELINA SALLEHUDDIN

[1, 2, 3] School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia

[1] Department of Computer Science, Baghdad University, Baghdad, Iraq

E-mail: [1] dhuhaftsm@gmail.com, [2] sitizaiton@utm.my, [3] roselina@utm.my

ABSTRACT

These days, the properties of numerous systems in biology, engineering and sociology paradigms can be captured and analysed as networks of connected communities. The increasing emergence of these networked systems has fuelled the desire to study and analyse them into several sub-networks called communities. Community detection in a complex network is an ill-defined problem. Evolutionary Algorithms (EAs) have shown promising performance in community detection, but it's difficult to identify natural divisions included in such complex networks accurately and effectively without designing a problem-specific operator that exploits domain knowledge and guides the search process. Moreover, most of the contemporary studies only employ EA-based models to detect communities, which may not be adequate to represent the real community structure of networks due to the limitation in their topological properties. Thus, to enhance the predictive power of the state-of-the-art EA-based models, the main contribution of this research work is to put forward a framework that integrates evolutionary algorithm (EA) with a local heuristic approach. In the experiments, we select and optimise four well-known community detection models within the evolutionary algorithm framework, i.e. expansion model, scaled cost function model, conductance model, and internal density model. Then, the proposed heuristic approach is employed to locally aid along with the optimisation model, in which the nodes having dense intra-connections with nodes of other communities are moved to neighbouring communities. In the experiments, the performance of the optimisation models has been examined on both synthetic and real-world networks that are publicly available. The results show that the put forward local heuristic approach has a positive effect that significantly enhanced the existing optimisation models' detection ability.

Keywords: *Community Detection, Complex Network, Evolutionary Algorithm, Optimisation Model, Local heuristic Approach*

1. INTRODUCTION

In the last decades, amongst the fields of the science of complex systems, the science of networks is experienced the highest level of activity. Many real complex systems have been represented as complex networks, in which the nodes represent the system's entities, while the links signify the interactions between the entities. Complex network analysis has been an interesting subject in the field of data mining and has attracted many researchers from different communities such as sociology [1], epidemiology [2], recommendation systems [3], email communication [4], and so on [5]. Few examples pertaining to the complex networks include communication networks, transport networks, collaboration networks, biological networks, the Internet, the world-wide-web, paper citation networks, neural networks, metabolic and protein-protein interaction networks (PPI) [6-8]. The main feature of complex networks is that the entities tend to group in a bid to form communities. The analysis of communities may help find out the structural features of the networks and simplify applications as targeted marketing and advertising, and discovering influential individuals [5]. Because of the relevance of communities' structure and their prominent role in a large number of applications, partitioning complex networks into communities or clusters is a widely studied issue (called community detection issue) and represented in the literature as graph clustering issues. In complex network analysis, community detection (CD) is a typical problem which has been proven as an NP-hard problem [9-11], and the greatest challenge in CD is that there is no universal definition describing community structure [12]. Due to the inherent complexity of the problem, often cannot be well solved by traditional optimisation methods. For this reason, evolutionary algorithms have been adopted as a major tool for dealing with CD problems. The relaxed nature of evolutionary algorithms is proved to be essential for competitively transcending the limits of other approaches in solving CD problem [6]. The literature studies mentioned a series of different models that were proposed to determine the quality pertaining to a set of predicted communities in a graph. Most of these works have been employed for detecting community in social networks as well as complex detection in PPI networks. Some of the popular detection models in the literature include expansion, internal density [15], conductance [13, 14], and scaled cost function [16] that used the evolutionary algorithms (EAs) as a tool to resolve the issue. It is difficult to identify natural divisions included in the complex networks accurately and effectively without designing a problem-specific approach that exploits domain knowledge, and also, reduces the random impact of the evolutionary operators that often generate solutions involving nodes belonging to the wrong community. Thus, to guide the search process, and to enhance the community division's quality, the main contribution of this work is to put forward a framework that integrates evolutionary algorithm (EA) with a local heuristic approach, where it is performed after the end of every generation, serves as a locally common assisted operator to enhance the predictive power of the state-of-the-art EA-based models.

The rest of the paper is structured as follows: Section 2 introduces some of the prominent studies in the literature based on EAs for detecting community. This is followed by presenting the details of the proposed EA framework including the fitness functions, and the adopted evolutionary operators which would be utilised, and discusses the recommended local heuristic approach. The experimental framework (encompassing test data sets and assessment metrics) is presented in Section 4; experimental outcomes which back the favourable effect of the recommended local heuristic approach to further rectify the community structures of the well-known community detection models are presented in Section 5; Section 6 presents the conclusion.

2. COMMUNITY DETECTION PROBLEM IN THE LITERATURE

In literature, the CD problem has been addressed with 3 different approaches (top-down clustering, bottom-up clustering, and optimisation methods) as a graph mining problem.

In the top-down method or divisive hierarchical method, the whole network is initiated as a one community before the detection and removal of the weak edges iteratively that link different communities [15, 17, 18]. Contrarily, each node in the network is initialised as one community in the bottom-up or agglomerative hierarchical methods before iteratively merging the similar communities based on some quality metrics [19-21].

However, CD problems have been defined and formulated as optimisation problems by adopting several algorithms. The optimisation methods have a similar objective which is optimised one or more objective functions by investigating the relationship between the featured subgroups and partition the nodes of the network into sub-networks based on subgroups' relations [22].

The relaxed nature of the meta-heuristic-based optimisation methods has made them suitable for the reduction of the problem complexity and to achieve adequate solutions. For this reason, methods based on evolutionary computation have attracted the interest of many researchers in different fields to solve the community detection problem. For example, Pizzuti (2008) [23] proposed a single objective genetic algorithm (GA-net) for CD in social networks. The researcher introduced the community score (CS) concept to measure the quality of a partitioning of a network in communities and tried to optimise this quality measure thru a genetic algorithm. Only one objective function (CS) was optimised so that only a certain solution was obtained in one run. Unlike most existing methods, the algorithm did not require the number of communities in advance. This number was automatically determined by the optimal value of the community score.

Gong et al. 2011 [24] suggested a memetic algorithm for optimising the modularity density (MemeNet), to detect the community structure of a network. Meme-Net also revealed its ability to explore the network at different resolutions and showed the hierarchical structure of the network.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Liu and Li (2017) [25] suggested a Modified Genetic Algorithm (MGA) using alleles encoding and half uniform crossover to detect community structure. In MGA, each allele of the chromosome denoted a community index of the corresponding node. If two nodes had the same community index, they would have the same allele values. At the same time, the half uniform crossover could better prevent elite individuals from destroying. And the researchers chose the modularity function as a fitness function. Experimental results on complex networks showed that MGA algorithm had an effective performance and got comparable results.

Rao et al. (2018) [9] presented a novel framework called Efficient Reduced-Bias GA (ERBGA) that addressed the issues of redundant representation and linearity assumption associated with the previous approaches. The method was flexible as it was easily adapted to any given mathematical objective, but the performance of the method still needs to be improved by considering vertices information to break up large dense networks into distinct clusters.

A new framework called Clustering Coefficient-based Genetic Algorithm (CC-GA) was proposed by Said et al. (2018) [26] for the detection of social and complex networks communities through the optimisation of the community structures' modularity. This work contributed to the use of CC for generating the initial population of the genetic algorithm, and a new mutation method which is based on the discovered community structure for reconnecting the existing connections. Experiments' results showed that CC-GA could compete with other algorithms on various types and sizes of real-world networks.

Although CD problem in a complex network has been tackled in the literature approximately since the last decade using the computational methodology based on graph clustering concept, discovering communities continues being an active area in terms of research. Moreover, most of the contemporary studies only employ EA-based models to detect communities, which may not be adequate to represent the real community structure of networks due to the limitation in their topological properties, and in this study, we have developed a framework that integrates evolutionary algorithm (EA) with a local heuristic approach to enhance the predictive power of the state-of-the-art EA-based models.

3. MATERIAL AND METHODS

Modelling of a complex network $\mathcal{N}$ is done as an undirected graph G, in which the pair (V, E) signifies the whole pairwise connections between various objects in the network. In $\mathcal{N}$, the set of v objects is regarded as a set of nodes of graph, i.e. $G: V(G) = \{v_1, v_2, \dots v_n\}$ while representation of the mutual connection between any pair of nodes in $\mathcal{N}$ is done as the edges (v_i, v_j) in E, viz., $E(G) \subseteq V(G) \times V(G) = \{(v_i, v_j) | 1 \le i, j \le n \wedge i \ne j\}$.The problem of clustering complex networks can be presented as detecting dense regions, which means partitioning a complex network in the form of a graph into sets or clusters of nodes that possess dense intra-connections as well as sparse inter-connections. Consequently, a CD problem has been modelled as an optimisation issue, wherein developing various optimisation functions was done by employing different kinds of quality metrics. Here, we put forward utilising four well-known quality metrics. Furthermore, a new guided local heuristic approach has been designed with an aim to enhance the identification of communities in complex networks.

3.1 Fitness Functions

Suppose $\{C_1, \dots, C_K\}$ represents a partition of complex network $\mathcal{N}$, that having n nodes and m links, in k clusters (or communities), and let us assume a community C_i of $\mathcal{N}$ that possesses $n(C_i)$ nodes and $m(C_i)$ edges. For any of the node $v \in C_i$, the number of connections incident to node v as a degree of v can be defined, which can be formally noted as $m(v)$, and can be divided into two concepts: inter-edges pertaining to node v signified as $\bar{m}(v, C_i) = |\{(v, w) \in E \wedge w \notin C_i\}|$ and intra-edges pertaining to node v denoted as $m(v, C_i) = |\{(v, w) \in E \wedge w \in C_i\}|$. Subsequently, the concept pertaining to intra- and inter-connections could be popularized to a single community C_i, in which $m(C_i) = \sum_{v \in C_i} m(v, C_i)$ characterises the number of internal-connections of the nodes of C_i, and $\bar{m}(C_i) = \sum_{v \in C_i} \bar{m}(v, C_i)$ denotes the number of outside-connections of the nodes of C_i [22, 17]. The following metrics, derived from [27, 28] and designed to catch the concept of clustering quality, can be defined as:

Conductance (CO): measures the fraction of edges that point outwards of the clustering.

$$min\ CO\ (C) = \sum_{i=1}^{k} \frac{\sum_{v \in C_i} \bar{m}(v, C_i)}{2m(C_i) + \sum_{v \in C_i} \bar{m}(v, C_i)} \quad (1)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Expansion (EX): measures the number of edges per nodes which pointing outwards of the clustering.

$$min\ EX(\mathcal{C}) = \sum_{i=1}^{K} \frac{\sum_{v \in C_i} \bar{m}(v, C_i)}{n(C_i)} \quad (2)$$

Internal Density (ID): signifies a partitioning solution based on the internal edge density.

$$min\ ID(\mathcal{C}) = \sum_{i=1}^{K} 1 - \frac{m(C_i)}{n(C_i)(n(C_i)-1)/2} \quad (3)$$

Scaled Cost Function (SCF): measures the total number of bad connection incidents along with node v, i.e. connections between v and a node that do not fall under the same cluster of v or connections not existing between v and any other node pertaining to the same cluster of v, in terms of the area size impacted by v in the clustering.

$$minSCF = \frac{n-1}{3} \sum_{v \in V} \frac{(\bar{m}(v, C_i) + \bar{l}_{c_i}(v))}{|nn(v) \cup \{u \in C_i\}|} \quad (4)$$

Here, $\bar{l}_{c_i}(v) = |\{u \in C_i | (v, u) \notin E\}|$ denotes the number of nodes present in C_i and are not linked to v, while $nn(v)$ can be defined as the set of neighbour nodes of v.

3.2 Method Representation and Evolutionary Operators

While employing an evolutionary algorithm, choosing a good encoding scheme representation is a crucial issue. Park and Song [29] put forward the locus-based adjacency representation, which was adopted in this study. In this representation, each individual I includes g genes $\{I_1, I_2, \dots, I_g\}$, in which g denotes the total number of nodes in the network. Every gene is linked to only one of its neighbouring nodes, in order to achieve good divisions to the network by containing only connected nodes. The decoding step is crucial for each individual I to identify the structure of the communities of a network. Next, the initialisation process will take into account only the effective connections, i.e. one neighbouring node is chosen for each gene. The type of adopted crossover operator is a uniform crossover, which is applied with a particular probability of p_c. By considering two parents, generating of a random binary vector is done. Then, based on the uniform crossover, the genes are chosen from the first parent when the vector value equals to 0; while selecting genes from the second parent when the vector value equals to 1, then integrating genes for producing the child. For the mutation operator, under the control of the probability p_m, a random change to the allele value of I_i, the mutated gene, is done to one of its neighbouring nodes [30].

3.3 The Proposed Heuristic Approach

The main contribution of our work involves combining EA with a local heuristic approach to enhance the obtained solution, post each generation. Formally speaking, consider a node I_i falling under a community C of the clustering $\mathcal{C} = \{C_1, \dots, C_K\}$, generated by such an individual, accordingly we can define *Intra-Connections Score (ICS)* pertaining to node I_i in its community C as:

$$I_i\_ICS(C) = \frac{m(I_i, C)_{I_i \in C}}{n(C)} * 2m(C) \quad (5)$$

From C community $C \in \mathcal{C}$ and based on ICS concept, a node I_i will get deleted from its home community, and allocated to another community, only if the ICS pertaining to node I_i in its community C is evaluated smaller than its ICS in another community $C' \in \mathcal{C}$ after I_i moves from C to C' ($i.e.\ I_i\_ICS(C) < I_i\_ICS(C' \cup I_i)$). In other words, a change in node I_i regarding its community belongingness occurs via the proposed heuristic approach to a different neighbouring community, aiming to achieve the highest ICS value of the node I_i within the new community. Algorithm 1 highlights the major steps pertaining to this local heuristic approach.

Algorithm 1: Local Heuristic Approach

Input : I: individual, n: number of nodes in the network
Output: I': individual after applying the proposed heuristic approach

set $\mathcal{C} = \{C_1, C_2, \dots, C_k\} \leftarrow \delta(I)$ // decode I into its community structure
for $i \leftarrow 1$ **to** n **do**
  set $C_i \leftarrow Community\_ID(I_i)$
  set $K \leftarrow max(Community\_ID(I))$
  set $k\_I_i\_ICS(C_i) \leftarrow \frac{m(I_i, C_i)_{I_i \in C_i}}{n(c_i)} * 2m(c_i)$
  set $k\_I_i\_ICS(C'_j \cup I_i) \leftarrow \frac{m(I_i, C'_j)_{C'_j \in \mathcal{C} \wedge i \neq j}}{n(C'_j)} * 2m(C'_j)$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

if ($k\_I_i\_ICS(C_i) < k\_I_i\_ICS(C'_j \cup I_i)$)
  set $C' \leftarrow argmax_{C'_j \in C, j \neq i}(ICS\left(C'_j \cup I_i\right))$
  set $Community\_ID(I_i) \leftarrow C'$
end if
end for
Return (I')

As per the general layout pertaining to EA as represented in [31], algorithm 2 shows the key steps that combine EA with the proposed local heuristic approach.

Algorithm 2: An Enhanced EA (E_EA)

Input: $\mathcal{N}$: network, n: a number of nodes in the network, pop: Population size, p_c: Crossover rate, p_m: Mutation rate, S: Selection operator, ι: $Stopping\ criterion\ \{true, false\}$.
Output: I^*:Best solution

$t \leftarrow 0$;
$Initialize\ P(t) \leftarrow \{I_1, I_2, I_3, \ldots, I_{pop}\}$;
while ($\iota(P(t)) \neq true$) **do**
  for $i \leftarrow 1$ ***to*** pop ***do***
    evaluate $F(I_i(t))$;
  end for
  $P(t) \leftarrow S(P(t))$;
  $P(t) \leftarrow$ Perform uniform Crossover $(P(t), p_c)$;
  $P(t) \leftarrow$ Perform neighbour Mutation $(P(t), p_m)$;
  for $i \leftarrow 1$ ***to*** pop ***do***
    $I_i'(t) \leftarrow$ Perform Local Heuristic Approach $(I_i(t), n)$
  end for
  $t \leftarrow t + 1$;
end while
$I^*(t) \leftarrow$ Community structure of individual having best fitness value
Return $I^*(t)$

4 EXPERIMENTAL SETTING

4.1 Dataset

This study will avail from a recently developed context to test CD algorithms with synthetic and real-world dataset, in which the correct partition is available as the ground fact. The put forward approach will evaluate based on four real-world networks employed by various societies, and, moreover, evaluation of six random computer-generated networks pertaining to Girvan and Newman [18] $\mathcal{N}et\_GN_\gamma$ is done. A concise portrayal of these networks is mentioned below.

Zachary's network: It is one of the well-known real social networks, compiled by Zachary [34], and widely used by researchers to evaluate community detection algorithms. Zachary observed 34 members of a karate club over a period of two years. Because of fierce dispute developed between the manager and the coach of the club, eventually, the club is separated into two groups. The network includes 34 nodes and 78 ties or edges in total, which are segmented into 2 groups.

Dolphin network: It denotes a network of 62 bottlenose dolphins that live in Doubtful Sound, New Zealand, which was described in [32]. Depending on the statistically frequent association, a tie was established between two dolphins. Partition of the network naturally into two groups was done (i.e. male and female). The network includes 62 nodes and 159 ties or edges in total, which are segmented into 2 groups.

Krebs' Books: This can be defined as the social network derived from US politics books as compiled by Krebs [33]. Each node in this network represents a book about US politics, and each edge connects two books which are frequently co-purchased by the same person. Classification of the books was done into three: conservative, liberal and neutral. The network includes 105 nodes and 440 edges in total, which are segmented into 3 groups.

Football network: The American College Football network was derived by Girvan and Newman [18]. It was created using college football matches in USA. The nodes of this network represent football teams, while the edges represent the matches played between the teams. The network consists of 115 nodes and 613 edges grouped in 12 teams.

$\boldsymbol{\mathcal{N}et\_GN_\gamma}$: Six random networks pertaining to Girvan and Newman [18] are generated and studied with the help of corresponding mixing parameter $\gamma \in \{0.05, 0.3\}$. Wherein the mixing parameter is increased with a step-size equal to (0.5) causing the network structure to become more ambiguous. In GN benchmark, every network $\mathcal{N}$ includes 128 nodes and 1024 edges, wherein each network was segmented into four different communities, each including 32 nodes.

4.2 Evaluation Measures

For the purpose of quality assessment of models' outcomes, the validity indices are required

Journal of Theoretical and Applied Information Technology
31st October 2019. Vol.97. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

to be defined. Both of the modularity (Q) [17] and the Normalized Mutual Information (NMI) [35] are used to evaluate the quality of detected community structures. NMI is a standard external measure, used to assess the identified communities' quality on the complex networks that have a known community structure. NMI, between a predicated partitioning result $\mathcal{C}$ and the proper partition S of a network $\mathcal{N}$ having n nodes, is the MI (mutual information) normalisation score between $\mathcal{C}$ and S being measured from 0 when no MI exists among the 2 solutions and until 1.0 when there is an absolute correlation between both solutions [35]. Consider the confusion matrix $c = [c_{ij}]$, $i = 1, \dots, K_{\mathcal{C}}$ and $j = 1, \dots, K_S$, where c_{ij} represents the number of nodes in the community i of $\mathcal{C}$ which are in community j of S as well. Then,

$$NMI(C,S) = \frac{-2\sum_{i=1}^{K_{\mathcal{C}}}\sum_{j=1}^{K_S} c_{ij} log(c_{ij} * n / c_i c_j)}{\sum_{i=1}^{K_{\mathcal{C}}} c_i log(c_i/n) + \sum_{j=1}^{K_S} c_j log(c_j/n)} \quad (6)$$

where c_i and c_j denote the sum of the elements of the community i in $\mathcal{C}$ as well as community j in S, respectively.

On the other hand, the modularity measure (noted as Q), the most common internal quality measure, presented by Newman and Girvan [17] to evaluate the quality of community detection depending on the separations and connections between the communities without using any additional knowledge of the domain (Equation (7)). The value of Q ranges from 0 to 1; networks that has a strong community structure presents a higher Q value.

$$Q(\mathcal{C}) = \sum_{i=1}^{K}\left[\frac{2m(C_i)}{m(\mathcal{C})} - \left(\frac{\sum_{v \in C_i} m(v)}{2m(\mathcal{C})}\right)^2\right] \quad (7)$$

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

The functioning of 4 state-of-the-art evolutionary-based community detection models (which are, $EA\_CO, EA\_EX, EA\_ID, and\ EA\_SCF$) is studied and examined. For the sake of fairness, while examining the computational abilities of the present models, the distinctive elements of EA technique are quantified to their used general settings present in the literature. Population size $pop = 100$, maximum number of iteration $max_t = 100$, probability of mutation $p_m = 0.2$, and probability of crossover $p_c = 0.8$. Moreover, the outcomes report the effect of the recommended local heuristic approach on the EA models' ultimate performance. Each of the EA models will be tested using two versions: first, when the models function with Traditional EA support (mentioned henceforth as T_EA); second, when the models function with Enhanced EA support that integrates EA with the recommended local heuristic approach (mentioned henceforth as E_EA).

Tables (1-4), show the outcomes of the EA models testing on real networks (Zachary's network, Dolphin network, Krebs' Books network and Football network), and artificial networks ($\mathcal{N}et\_GN_{\gamma \in \{0.05,0.1,0.15,0.2,0.25,0.3\}}$). The quality of community detection in a complex network for all EA models is presented in terms of average best NMI value (NMI_{avg}) over the 10 distinct runs in the two versions given above. The best outcome found for each model is shown in bold in the tables.

Table 1: Quality of community detection in terms of NMI_{avg}(T_EA: Traditional EA framework, E_EA: Enhanced EA framework) for all EA models in real- world networks: Zachary and Dolphin.

EA Model	NMI_{avg}			
	Zachary's Network		Dolphin Network	
	T_EA	E_EA	T_EA	E_EA
EA_CO	0	**0.226**	0.2	**0.7540**
EA_EX	0	**0.1356**	0.6391	**0.7754**
EA_ID	0.2238	**0.2675**	0.2181	**0.6428**
EA_SCF	0.4389	**0.4544**	0.5231	**0.6150**

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TABLE 2: Quality of community detection in terms of NMI_{avg}(T_EA: Traditional EA framework, E_EA: Enhanced EA framework) for all EA models in real- world networks: Krebs' Books and Football.

EA Model	NMI_{avg}			
	Krebs' Books Network		Football Network	
	T_EA	E_EA	T_EA	E_EA
EA_CO	0.2964	**0.5649**	0.1874	**0.7283**
EA_EX	0.2863	**0.6003**	0.0643	**0.7745**
EA_ID	0.1756	**0.3819**	0.2077	**0.8041**
EA_SCF	0.469	**0.5426**	0.3192	**0.8822**

Table 3: Quality of community detection in terms of NMI_{avg}(T_EA: Traditional EA framework, E_EA: Enhanced EA framework) for all EA models in synthetic networks:$\mathcal{N}et\_GN_{0.05}$, $\mathcal{N}et\_GN_{0.1}$, and $\mathcal{N}et\_GN_{0.15}$.

EA Model	NMI_{avg}					
	$\mathcal{N}et\_GN_{0.05}$		$\mathcal{N}et\_GN_{0.1}$		$\mathcal{N}et\_GN_{0.15}$	
	T_EA	E_EA	T_EA	E_EA	T_EA	E_EA
EA_CO	0.7017	**1**	0.3431	**1**	0	**0.9302**
EA_EX	0.7210	**0.9857**	0.5616	**0.9782**	0.1125	**0.8687**
EA_ID	0.2293	**0.5629**	0.111	**0.3415**	0.1161	**0.1706**
EA_SCF	0.9344	**1**	0.7778	**1**	0.4168	**1**

Table 4: Quality of community detection in terms of NMI_{avg}(T_EA: Traditional EA framework, E_EA: Enhanced EA framework) for all EA models in synthetic networks:$\mathcal{N}et\_GN_{0.2}$, $\mathcal{N}et\_GN_{0.25}$, and $\mathcal{N}et\_GN_{0.3}$.

EA Model	NMI_{avg}					
	$\mathcal{N}et\_GN_{0.2}$		$\mathcal{N}et\_GN_{0.25}$		$\mathcal{N}et\_GN_{0.3}$	
	T_EA	E_EA	T_EA	E_EA	T_EA	E_EA
EA_CO	0	**0.7153**	0	**0.3955**	0	**0.1636**
EA_EX	0	**0.7857**	0	**0.3066**	0	**0.2588**
EA_ID	0.0989	**0.1423**	0.1234	**0.1274**	0.0897	**0.1297**
EA_SCF	0.1946	**1**	0.1561	**0.8632**	0.1229	**0.7572**

The findings confirm clearly the positive effect of the recommended local heuristic approach on all the EA models' performance. The association of the EA models and recommended local heuristic approach caused the better quality of EA models' detection of communities and enabled them to approach from the optimum solution (that is, $NMI = 1$). In Tables (1 and 2), the most efficient performance noted when applying the proposed local heuristic approach with the EA_SCF model on the tested real networks: Zachary and Football, while EA_EX model showed the most reliable performance (at E_EA version) on the tested real networks: Dolphin and Krebs' Books.

From the other hand, the EA_SCF model had superior performance compared to the other models in Tables (3 and 4). The competitive performance was noted between the models EA_CO and EA_EX. While the EA_ID model showed a poor performance in most networks.

A close examination of the individual runs, in case of the real-world networks, reveals that all tested model at E_EA version stopped at different local optima. In Zachary's network, we noted at the single run level, the highest performance in terms of NMI_{max} was reported by EA_SCF and EA_ID models which were, respectively, 0.5543 and 0.5491. Moreover, the performance of EA_CO and EA_EX models improved with the heuristic approach and attained their highest NMI value equal to 0.2260.

Journal of Theoretical and Applied Information Technology
31st October 2019. Vol.97. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In the Dolphin network, we noted at the single run level, the most excellent performance was $NMI_{max} = 0.9022$ attained by the EA_ID model, whereas the other models improved and reported their best performance at $NMI = 0.8888$ with EA_CO and EA_EX models, and $NMI = 0.7534$ with EA_SCF model. As for Krebs' Books network, the recommended approach facilitated EA_SCF model to attain highest local optima equalled to $NMI_{max} = 0.6414$, whereas the other models improved and reported their best performance at $NMI = 0.6326$ with EA_CO and EA_EX models, and $NMI = 0.5236$ with EA_ID model. A significance improvement was observed when applying the proposed local approach along with the EA models on the real Football network, which presented, at the single run level, the best performance in terms of NMI_{max} equal to 0.892 with EA_SCF model, while the other models (EA_CO, EA_EX, and EA_ID) nearly showed a comparable performance in terms of NMI (i.e. $0.8433, 0.8639$, and 0.8358, respectively).

For the synthetic benchmark networks, the recommended heuristic approach enabled the EA_SCF model to achieve the correct divisions in all the runs at the networks $\mathcal{N}et\_GN_{0.05-0.2}$, and in some operations with the rest other networks. The EA_CO model approached the optimal solution in all the runs at the networks and $\mathcal{N}et\_GN_{0.05,0.1}$, and in two runs at the network $\mathcal{N}et\_GN_{0.15}$, while EA_EX model attained the optimum solution in (9, 7, 2, 1) runs at the networks $\mathcal{N}et\_GN_{0.05}$, $\mathcal{N}et\_GN_{0.1}$, $\mathcal{N}et\_GN_{0.15}$, $\mathcal{N}et\_GN_{0.2}$ respectively.

Finally, the performance of the EA_ID model is improved but failed to approach the optimal solution in all GN networks, and it stagnated at various local optima even when applying the local heuristic approach. This can be turned back that the cooperation between the proposed approach and optimisation function of the EA_ID model resulted in generating solutions included many communities (or clusters) with small sizes comparing to the other models.

However, the outcomes shown in Tables (1-4) substantiate the positive effect of the recommended approach to enhance the reliability of detection of all the models towards near optimal, or even, the optimal partitioning solutions.

Next, the performance of the EA models with the proposed heuristic approach was compared with five algorithms, in terms of Q for the four real networks, including the GN algorithm [18]; the Fast Newman (FN) algorithm proposed by Newman [19]; CCGA (Clustering Combination Based Genetic Algorithm) proposed by He et al. [36]; LGA (Genetic Algorithm with local search) presented by Jin et al. [37], and MGA (Modified Genetic Algorithm) proposed in 2017 by Liu and Li [25]. The results of the $GN, FN, CCGA, LGA$, and MGA in Table 5 are taken from those of Ref. [25]. It can be observed from Table 5 that EA_CO, EA_EX, EA_ID, and EA_SCF models showed more reliable performance than the other algorithms in terms of the average values of Modularity (Q_{avg}) on the real-world networks.

Table 5: Comparison performance in terms of Q_{avg} for the four real-world networks

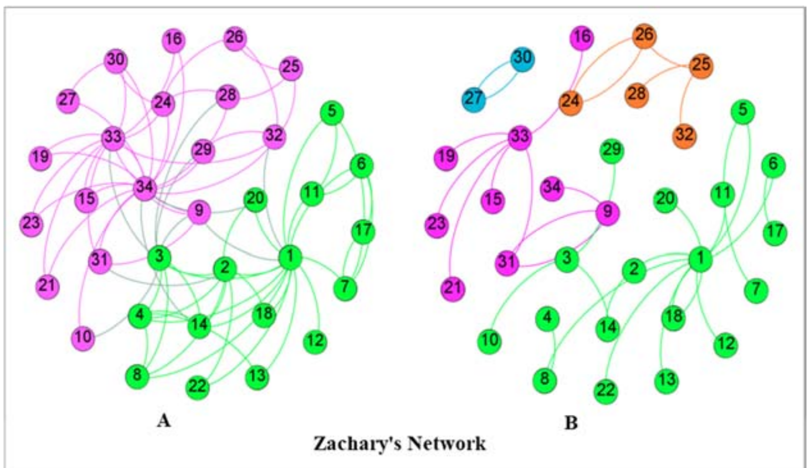
Algorithm \ Network name	Zachary	Dolphin	Krebs' Books	Football
GN	0.4013	0.4706	0.5168	0.5996
FN	0.3807	0.4955	0.4993	0.5647
CCGA	0.4198	0.5273	0.4445	0.6054
LGA	0.4198	0.5280	0.5272	0.6046
MGA	0.4198	0.5280	0.5260	0.6054
EA_CO	**0.7447**	**0.7445**	**0.7882**	**0.6842**
EA_EX	**0.7468**	**0.7512**	**0.8081**	**0.6757**
EA_ID	**0.6279**	**0.6957**	**0.6083**	**0.6761**
EA_SCF	**0.6095**	**0.7275**	**0.7429**	**0.6811**

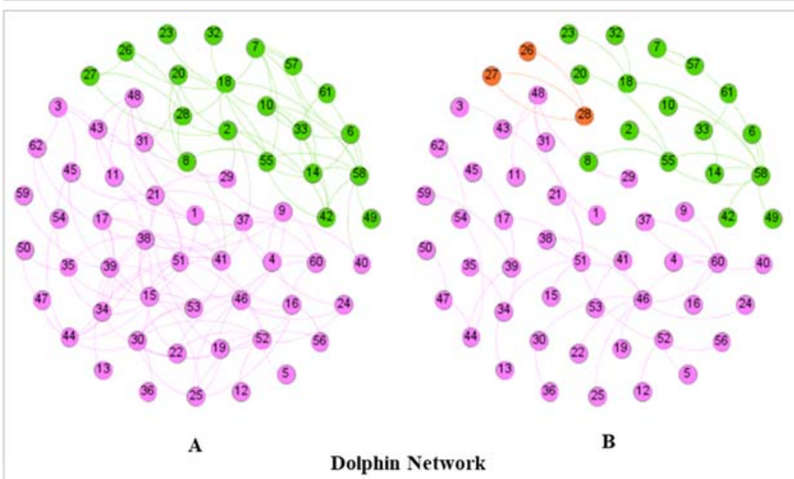
Finally, Figures (1-5) display the configuration of the network. Part A denotes the original configuration of each tested network, whereas part B denotes the best community structure of the related network that was acquired by the EA model in the version of E_EA. In Figure 1, Zachary's network (part B) demonstrates the highest near-optimal solution $NMI_{max} = 0.5543$ acquired using the EA_SCF model, and moreover, the Dolphin network (part: B) demonstrates the highest near-optimal solution $NMI_{max} = 0.9022$ acquired using the EA_ID model.

In Figure (2), Krebs' Books network and Football network (part B) demonstrate the highest local optima $NMI = 0.6414$ and $NMI = 0.8920$ respectively, acquired by the EA_SCF model. In Figure 3, the $\mathcal{N}et\_GN_{0.05}$ and $\mathcal{N}et\_GN_{0.1}$ (part: B) demonstrate the optimum solution $NMI = 1$ obtained by applying the EA_CO model. In Figures (4 and 5), the networks $\mathcal{N}et\_GN_{0.15}$, $\mathcal{N}et\_GN_{0.2}$, $\mathcal{N}et\_GN_{0.25}$, $\mathcal{N}et\_GN_{0.3}$ (part: B) demonstrate the optimum solution $NMI = 1$ obtained by applying the EA_SCF model.

6. CONCLUSION

In this study, we have recommended a heuristic approach based on the evolutionary algorithm for effective detection of community in large complex networks. The recommended approach in this study is a promising idea to get good community division in the static and unsigned networks thru guiding the search process by moving nodes having dense intra-connections with nodes of other communities to the neighbour communities. From the results of the experiments on the real and synthetic complex networks, we can conclude that integrating EA models with the proposed local heuristic approach yielded results more accurate than the corresponding EA models (that did not use the local heuristic approach) and more reliable than some of the related algorithms. In the forthcoming days, we wish to integrate this approach with multi-objective EA models. Also, it is desirable to additionally examine the local heuristic approach in other types of complicated networks, like dynamic and signed networks.





ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 1: Community structures of Zachary's network and Dolphin network (A: original structure of the network, B: the best community structure (in terms of NMI_{max}) obtained by the EA model in E_EA version). In each network, each community distinguished with a different colour.

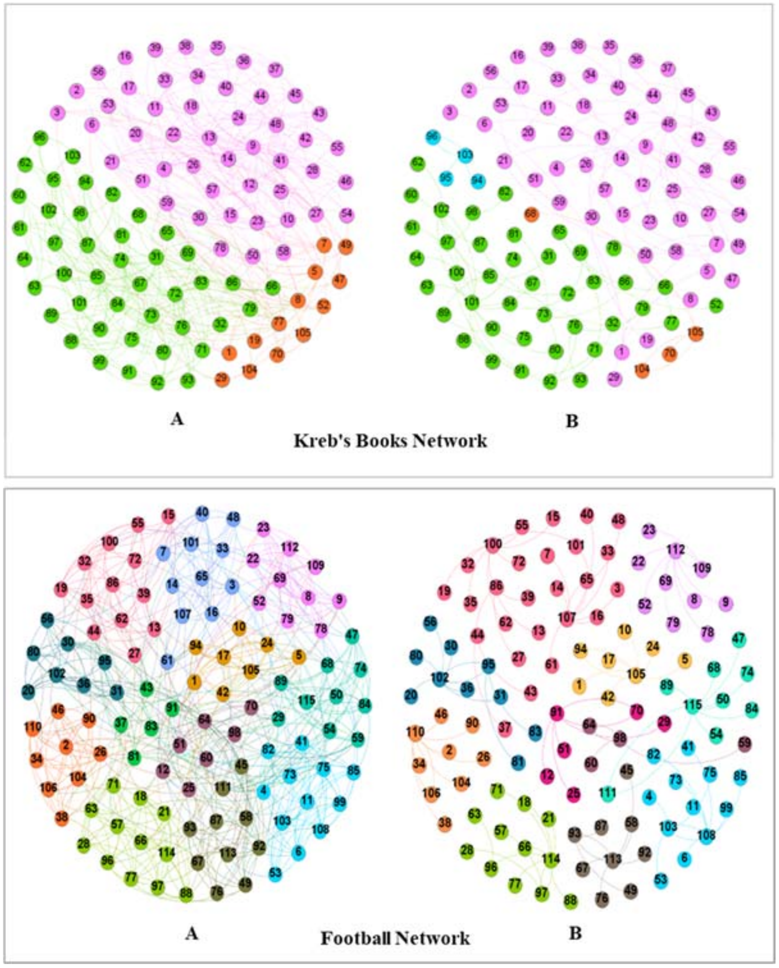


Figure 2: Community structures of Krebs' Books network and Football network (A: original structure of the network, B: the best community structure (in terms of NMI_{max}) obtained by the EA model in E_EA version). In each network, each community distinguished with a different colour.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

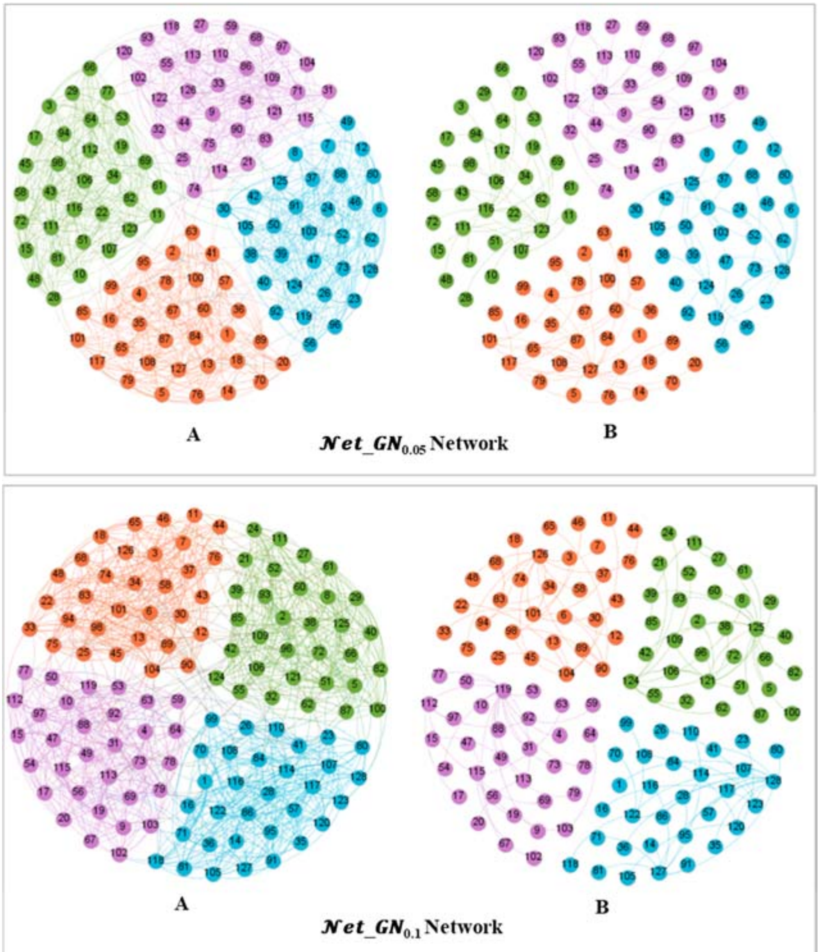


Figure 3: Community structures of $\mathcal{N}et\_GN_{0.05}$ network and $\mathcal{N}et\_GN_{0.1}$ network (A: original structure of the network, B: the best community structure (in terms of NMI_{max}) obtained by the EA model in E_EA version). In each network, each community distinguished with a different colour.

Journal of Theoretical and Applied Information Technology
31st October 2019. Vol.97. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

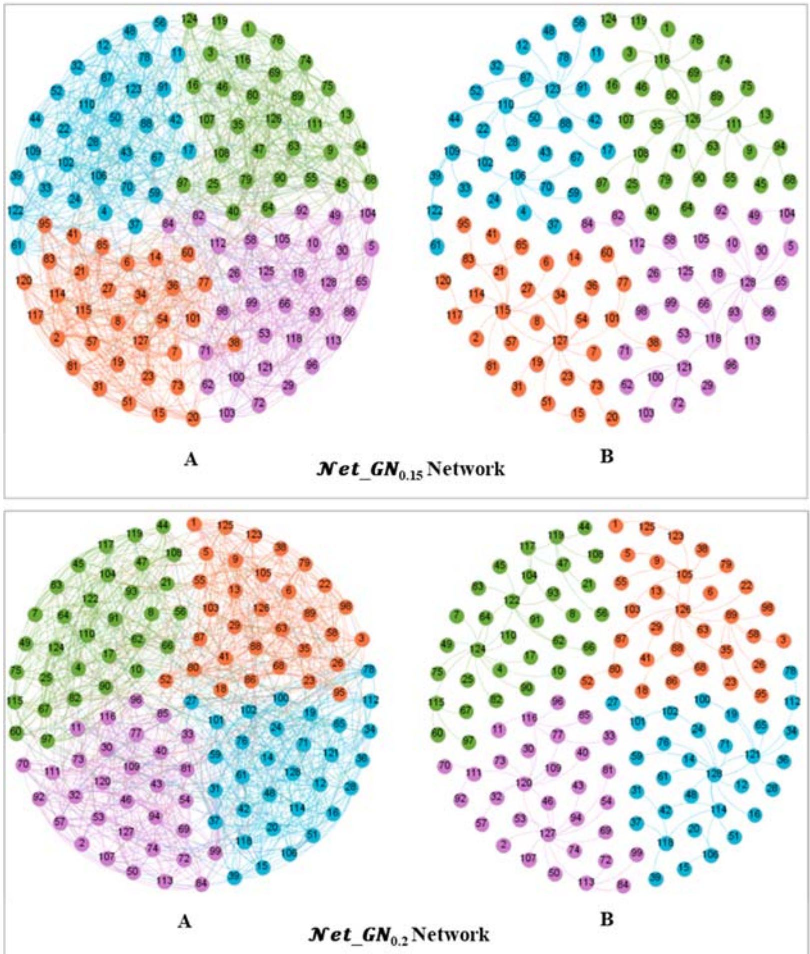


Figure 4: Community structures of $\mathcal{N}et\_GN_{0.15}$ network and $\mathcal{N}et\_GN_{0.2}$ network (A: original structure of the network, B: the best community structure (in terms of NMI_{max}) obtained by the EA model in E_EA version). In each network, each community distinguished with a different colour.

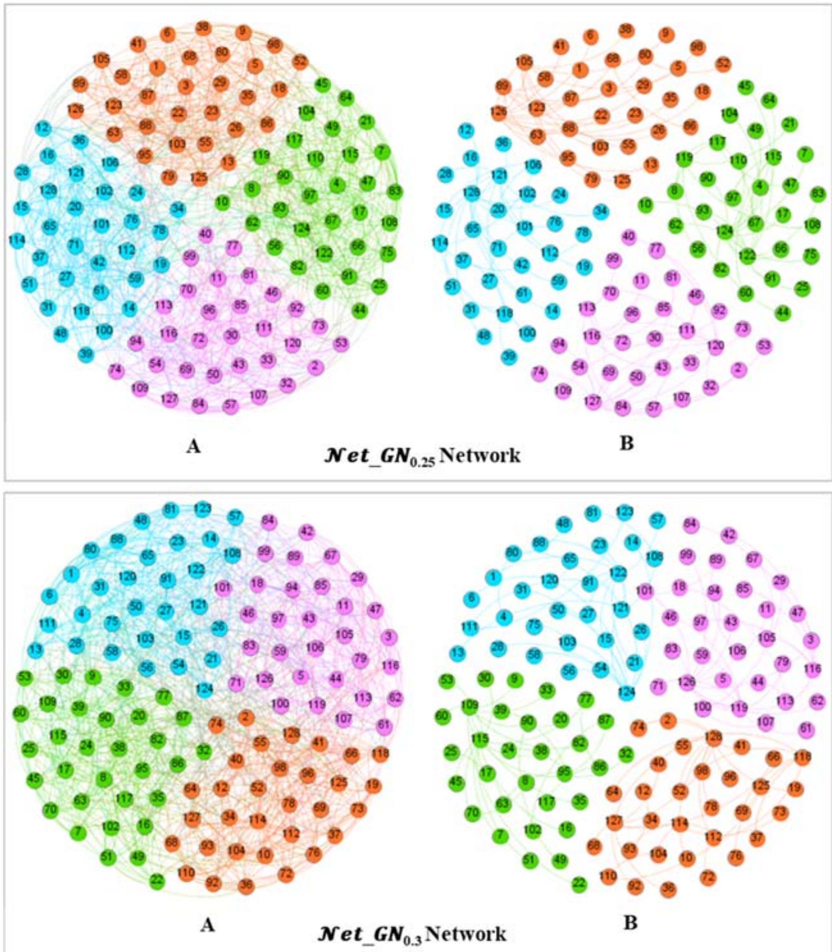


Figure 5: Community structures of $\mathcal{N}et\_GN_{0.25}$ network and $\mathcal{N}et\_GN_{0.3}$ network (A: original structure of the network, B: the best community structure (in terms of NMI_{max}) obtained by the EA model in E_EA version). In each network, each community distinguished with a different colour.

REFERENCES:

[1] Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications (Vol. 8). Cambridge university press.

[2] Meyers, L. A., Newman, M. E. J., & Pourbohloul, B. (2006). Predicting epidemics on directed contact networks. Journal of theoretical biology, 240(3), 400-418.

[3] Palau, J., Montaner, M., López, B., & De La Rosa, J. L. (2004, September). Collaboration analysis in recommender systems using social networks. In International Workshop on Cooperative Information Agents (pp. 137-151). Springer, Berlin, Heidelberg.

[4] Tyler, J. R., Wilkinson, D. M., & Huberman, B. A. (2005). E-mail as spectroscopy: Automated discovery of community structure within organizations. The Information Society, 21(2), 143-153.

[5] Boujlaleb, L., Idarrou, A., & Mammass, D. (2017). Tracking community evolution in social networks. Journal of Theoretical & Applied Information Technology, 95(22).

[6] Pizzuti, C. (2018). Evolutionary computation for community detection in networks: a review. IEEE Transactions on Evolutionary Computation, 22(3), 464-483.

[7] Cheng, F., Cui, T., Su, Y., Niu, Y., and Zhang, X. (2018). A local information based multi-objective evolutionary algorithm for community detection in complex networks. Applied Soft Computing, 69, 357-367.

[8] Pourkazemi, M., and Keyvanpour, M. R. (2017). Community detection in social network by using a multi-objective evolutionary algorithm. Intelligent Data Analysis, 21(2), 385-409

[9] Rao, A. K. G., Janikow, C., Bhatia, S., and Climer, S. (2018). Efficient Reduced-Bias Genetic Algorithm (ERBGA) for Generic Community Detection Objectives.

[10] Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2008). On modularity clustering. IEEE transactions on knowledge and data engineering, 20(2), 172-188

[11] Schaeffer, S. E. (2007). Graph clustering. Computer science review, 1(1), 27-64

[12] Fortunato, S., and Hric, D. (2016). Community detection in networks: A user guide. Physics Reports, 659, 1-44.

[13] Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. Departmental Papers (CIS), 107.

[14] Sinclair, A. and Jerrum, M., 1989. Approximate counting, uniform generation and rapidly mixing Markov chains. Information and Computation, 82(1), pp.93-133.

[15] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. Proceedings of the National Academy of Sciences, 101(9), 2658-2663.

[16] King, A. D., Pržulj, N., & Jurisica, I. (2004). Protein complex prediction via cost-based clustering. Bioinformatics, 20(17), 3013-3020.

[17] Newman, M. E., and Girvan, M. (2004). Finding and evaluating community structure in networks. Physical review E, 69(2), 026113.

[18] 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.

[19] Newman, M. E. (2004). Detecting community structure in networks. *The European Physical Journal B*, *38*(2), 321-330.

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

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

[22] 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.

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

[24] Gong, M., Fu, B., Jiao, L., & Du, H. (2011). Memetic algorithm for community detection in networks. Physical Review E, 84(5), 056101.

[25] Liu, S., & Li, Z. (2017, February). A modified genetic algorithm for community detection in complex networks. In Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), 2017 International Conference on (pp. 1-3). IEEE.

[26] Said, A., Abbasi, R. A., Maqbool, O., Daud, A., and Aljohani, N. R. (2018). CC-GA: A clustering coefficient based genetic algorithm for detecting communities in social networks. Applied Soft Computing, 63, 59-70.

[27] Leskovec, J., Lang, K. J., and Mahoney, M. (2010, April). Empirical comparison of algorithms for network community detection. In Proceedings of the 19th international conference on World wide web (pp. 631-640). ACM

[28] Pizzuti, C., and Rombo, S. E. (2014). Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. Bioinformatics, 30(10), 1343-1352.

[29] Park, Y., and Song, M. (1998, July). A genetic algorithm for clustering problems. In Proceedings of the Third Annual Conference on Genetic Programming (pp. 568-575).

[30] Pizzuti, C., and Rombo, S. (2012). Experimental evaluation of topological-based fitness functions to detect complexes in PPI networks. In Proceedings of the 14th annual conference on Genetic and evolutionary computation (pp. 193-200). ACM.

[31] Coello, C. A. C., Lamont, G. B., & Van Veldhuizen, D. A. (2007). Evolutionary algorithms for solving multi-objective problems (Vol. 5, pp. 79-104). New York: Springer

[32] Lusseau, D. (2003). The emergent properties of a dolphin social network.Proceedings of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Royal Society of London B: Biological Sciences, 270(Suppl 2), S186-S188.

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

[34] Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. Journal of anthropological research, 452-473.

[35] MacKay, D. J. (2003). Information theory, inference, and learning algorithms (Vol. 7). Cambridge: Cambridge university press.

[36] He, D. X., Zhou, X., Wang, Z., Zhou, C. G., Wang, Z., & Jin, D. (2010). Community mining in complex networks clustering combination based genetic algorithm. *Acta Automatica Sinica*, *36*(8), 1160-1170.

[37] Jin, D., Liu, J., Yang, B., He, D. X., & Liu, D. Y. (2011). Genetic algorithm with local search for community detection in large-scale complex networks. *Acta Automatica Sinica*, *37*(7), 873-882.