

Evolutionary Detection of Community Structures in Complex Networks: a New Fitness Function

Camelia Chira
Department of Computer Science
Babes-Bolyai University
Kogalniceanu 1
Cluj-Napoca 400084, Romania
Email: cchira@cs.ubbcluj.ro

Anca Gog
Department of Computer Science
Babes-Bolyai University
Kogalniceanu 1
Cluj-Napoca 400084, Romania
Email: anca@cs.ubbcluj.ro

David Iclănzan
Department of Electrical Engineering
Sapientia Hungarian
University of Transylvania
Tg-Mureş 540485, CP 4, OP 9
Romania

Abstract—The discovery and analysis of communities in networks is a topic of high interest in sociology, biology and computer science. Complex networks in nature and society range from the immune system and the brain to social, communication and transport networks. The key issue in the development of algorithms able to automatically detect communities in complex networks refers to a meaningful quality evaluation of a community structure. Given a certain grouping of nodes into communities, a good measure is needed to evaluate the quality of the community structure based on the definition that a strong community has dense intra-connections and sparse outside-community links. We propose a new fitness function for the assessment of community structures quality which is based on the number of nodes and their links inside a community versus the community size further reported to the size of the network. A novel aspect of the proposed fitness function refers to considering the way nodes connect to other nodes inside the same community making this second level of links contribute to the strength of the community. The introduced fitness function is tested inside a collaborative evolutionary algorithm specifically designed for the problem of community detection in complex networks. Computational experiments are performed for several real-world complex networks which have a known real community structure. This allows the direct verification of the quality of evolved communities via the proposed fitness function emphasizing extremely promising numerical results.

Index Terms—Community Detection, Complex Networks, Evolutionary Algorithms, Partition Fitness

I. INTRODUCTION

Networks are a central model for the description of many complex phenomena. Typical examples of complex networks in nature and society include metabolic networks, the immune system, the brain, the human social networks, communication and transport networks, the Internet and the World Wide Web. The basic unit of the system is reduced to simple nodes (or vertices) connected by edges (or links) depicting their pairwise relationships. The study of real-world networks revealed features such as degree distribution, average distance between vertices, network transitivity and community structure [1], [7], [12], [15], [21].

In a graph representation of a complex system as a network, nodes with similar properties (or function) have a higher chance to be linked to each other compared to random pairs of nodes. Such nodes tend to form a consistent subgraph

(called community) highlighted by the dense interconnections. A community in a network can be defined as a group of nodes densely connected with each other but sparsely connected with nodes belonging to other communities [15], [16]. An efficient community structure detection can facilitate the identification of functional subunits of the system providing at the same time a powerful tool for the visualization and representation of the network structure. Examples of communities include groups of mutual acquaintances in social networks, web pages grouped on the same subject and functional modules in protein interaction networks [12]. The detection of communities in complex networks has been intensively investigated over the last few years resulting in the development of many varied techniques and algorithms [6], [7], [12], [15].

Given the rather vague existing definitions of communities, one of the major challenges in the development of algorithms able to detect community structures refers to what measure should be used to meaningfully assess the quality of a certain grouping of nodes into communities. An important contribution in this direction comes from Newman and Girvan [7], [15], [16] who introduced a measure of the quality for a partitioning called *modularity*. Given a division of a network into communities, the modularity is based on the difference between the proportion of edges that connect vertices in a community and the proportion of edges with at least one node in the community (computed at the level of each community). Modularity quantifies the deviation of number of interconnections inside a community from the expected density of the same group of nodes in random graphs (with the same expected degree sequence). Higher values of the modularity indicate stronger community structures. A popular approach to detect communities in complex networks consists in the optimization of the modularity as a quality function [4], [5], [9], [15], [20]. However, recent studies indicate that this approach does not necessarily lead to good community structures of networks [11], [12]. The main drawback is the resolution limit of modularity maximization (even with the introduction of a resolution parameter [13]). The maximum value of modularity is essentially unreachable although finding a good approximation of the modularity maximum can be relatively easily achieved by many algorithms. In [11], a comparative analysis

of community detection algorithms emphasizes weak results of the modularity optimization approach for benchmark graphs with built-in community structure. Recently, a study of various fitness functions for community structures in the context of evolutionary algorithms revealed a weak correlation between the modularity and the Normalized Mutual Information (NMI) between the detected community structure and the known real partition [2]. This implies that maximizing the modularity as the fitness of evolved partition solutions does not imply maximizing the similarity (computed by NMI) of potential solutions with the known real partition.

In this paper, a new fitness function for the assessment of community structures quality is proposed. The quality of a network partition in communities is evaluated by looking at the contribution of each node and its neighbors to the strength of its community. The number of nodes and their links inside a community versus the community size further reported to the size of the network also contribute to the proposed partition fitness. The way nodes connect to other nodes inside the same community forms a second level of links which contribute to the strength of the community. Both the internal and external degrees of a node reported to the community size are used to quantify the contribution of each node to the quality of its community. The proposed fitness function is engaged in the framework of an evolutionary algorithm designed to detect network partitions. Individuals represent network partitions and are evolved towards better fitted solutions triggered by the proposed fitness. A collaborative evolutionary algorithm [3] is used for this purpose: genetic material saved from the best and worst individuals as well as the best of a line of related individuals informs the selection and recombination process in a standard evolutionary model (all the features of the proposed approach are detailed in section three). The collaborative evolutionary algorithm based on the introduced fitness function is tested in a set of computational experiments for several real-world complex networks which have a known real community structure. This allows the direct verification (using NMI) of the quality of evolved communities via the proposed fitness function emphasizing extremely promising numerical results.

The rest of the paper is structured as follows: section two presents a brief review of related work in the area of community detection algorithms with a particular focus on evolutionary models and existing fitness functions to assess the quality of a network partitioning, sections three and four describe the proposed fitness function and the evolutionary model used in community detection, section five presents the computational experiments and results and, finally, section six contains the conclusions and directions for future research.

II. RELATED WORK

The detection of communities in complex networks is a challenging problem intensively studied in recent years. Many techniques have been proposed for finding community structures (see [6] for a recent comprehensive survey). Community detection methods range from hierarchical clustering

[19] (using similarity metrics for the strength of connection between vertices) and divisive algorithms [7], [18] (using the edge betweenness as a weight measure) to random search methods such as evolutionary algorithms [3], [17].

A problem of great importance is defining a good quality measure for the distribution of nodes into communities. One of the most well known quality measures is the modularity proposed by Newman and Girvan [15]:

$$Q = \sum_{i=1}^k (e_{ii} - a_i^2) \quad (1)$$

where k is the number of communities, e is a symmetric matrix of size $k \times k$, each element e_{ij} represents the fraction of edges that connect nodes from community i to nodes in community j (i.e. therefore, e_{ii} is the proportion of edges that connect vertices inside community i) and $a_i = \sum_j e_{ij}$ (i.e. the proportion of edges with at least one node in the community i).

Higher values of the modularity indicate stronger community structures. However, as already indicated in the previous section, modularity optimization - although popular in the literature [4], [5], [9], [15], [20] - is not able to efficiently provide good solutions to the community structure problem in complex networks [2], [11]–[13].

A simple community quality measure has recently been described in [10]. The fitness of a community G is defined by:

$$f_G = \frac{k_{in}^G}{(k_{in}^G + k_{out}^G)^\alpha} \quad (2)$$

where k_{in}^G represents the total internal degree of the nodes in community G , k_{out}^G represents the total external degree and α is a positive real-valued parameter controlling the size of the communities. The fitness of a division P of nodes into communities is the average value of the communities fitness:

$$f_P = \frac{1}{n_c} \sum_{i=1}^{n_c} f_{G_i} \quad (3)$$

where n_c is the number of communities.

Some evolutionary approaches to the problem of detecting community structures in complex networks have been proposed in the literature [3], [17], [20]. Generally, evolutionary computation provides promising algorithms for addressing various NP-hard problems. In [20], the authors propose a genetic algorithm based on the network modularity [15] as fitness function. Individuals are represented as arrays of integers, where each position corresponds to a node in the network and each value in the array represents the id of the corresponding community. In [17], a genetic algorithm called *GA-Net* is proposed to discover communities in networks. Individuals are represented as array of integers of size N (number of nodes in the network) where each position $i, i = 1 \dots N$ has a value $j, j = 1 \dots N$ with the meaning that nodes i and j will be placed in the same cluster. The concept of community score - a quality

measure of a partitioning favoring highly intra-connected and sparsely inter-connected communities - is engaged as fitness function. The community score of a clustering S_1, \dots, S_k of a network is defined as:

$$CS = \sum_{i=1}^k score(S_i) \quad (4)$$

The *score of a community* S is defined as:

$$score(S) = M(S) * v_s \quad (5)$$

where $M(S)$, the *power mean of S of order r* , is:

$$M(S) = \frac{1}{|S|} \sum_{i \in S} \left(\frac{k_i^{in}(S)}{|S|} \right)^r \quad (6)$$

and v_s is the volume of a community S defined as the number of edges connecting vertices inside S .

In [3], a collaborative evolutionary algorithm is proposed for the community detection problem. In this evolutionary approach, collaborative selection and recombination sustain a balanced search process. Each individual has information about its best ancestor and the global optimal and worst solutions already detected. Selection of parents considers individuals which are not genetically related while the recombination operator takes into account the intermediary global best and worst solutions. The fitness function used in [3] is the community score proposed by Pizzuti [17]. The results obtained for several real-world networks are shown to be superior to those reported by a standard evolutionary approach as that reported in [17] using the same representation and fitness function.

The capability of the above described fitness functions (the modularity [15], [16], the fitness described in [10] and the community score [17]) to lead to good partitioning solutions (via evolutionary search) is limited as the correlation between the fitness and NMI of the same evolved community structure is rather weak [2].

III. PROPOSED FITNESS FUNCTION FOR THE EVALUATION OF COMMUNITY STRUCTURES

The fitness function evaluating the quality of the overall community structure is based on the score (community fitness) computed for each community. At the community level, each node contributes to the community fitness based on the internal and external degrees reported to the community size. Furthermore, the neighbors of each node form a second level which have a weighted contribution (computed in the same way as for first-level nodes) to the community fitness.

Let $\mathcal{C}_1, \dots, \mathcal{C}_{nc}$ be a division of a network into nc communities. Each community \mathcal{C}_i contains a certain number of nodes each having links with other nodes inside the same community (representing the internal degree k_{in} of the node) and/or nodes outside the community (representing the external degree k_{out} of the node). Obviously, the total degree of a node satisfies $k = k_{in} + k_{out}$.

Given a node x belonging to a community \mathcal{C} , a node score evaluating the contribution of the node to the strength of its community is computed using the following measure:

$$SNode(x) = \frac{k_{in}^x - k_{out}^x}{|\mathcal{C}|} \quad (7)$$

where $|\mathcal{C}|$ represents the size of community \mathcal{C} (i.e. the number of nodes in \mathcal{C}) to which node x belongs to.

A high positive value of $SNode(x)$ indicates an important contribution of node to the strength of its community. In this case, the number of links inside the community is higher compared to the external links and the node degree difference is significant enough reported to the community size. It should be noted that a negative $SNode(x)$ score is an indication of more connections of the node outside the community compared to the number of intra-connections.

A community \mathcal{C} is evaluated by considering the contribution $SNode(x)$ of each node $x \in \mathcal{C}$ (first level nodes) and furthermore the contribution of all nodes $y \in \mathcal{C}$ (second level nodes) which are linked with x . The measure used to compute the fitness of each community is defined as follows:

$$SComm(\mathcal{C}) = \sum_{x \in \mathcal{C}} \left[SNode(x) + \frac{1}{2} \sum_{\substack{y \in \mathcal{C} \\ (x,y)=1}} SNode(y) \right] \quad (8)$$

The measure $SComm(\mathcal{C})$ takes into account the number of links of each node inside the community versus the community size and also the connections of each neighboring node inside and outside community. This second contribution is weighted in $SComm(\mathcal{C})$ (currently by a fixed value of 0.5) as nodes are naturally considered several times as contributing neighbors (in addition to the one full contribution as the main node). The main idea is to reward nodes with many interconnections which further connect with nodes densely connected with nodes in the same community and sparsely connected with other nodes from the network.

The overall fitness of a certain network partitioning \mathcal{P} into nc communities $\mathcal{C}_1, \dots, \mathcal{C}_{nc}$ is defined as follows:

$$SPart(\mathcal{P}) = \frac{1}{nc} \sum_{i=1}^{nc} SComm(\mathcal{C}_i) \cdot \frac{v(\mathcal{C}_i)}{|\mathcal{C}_i|} \quad (9)$$

where $v(\mathcal{C}_i)$ is computed as

$$v(\mathcal{C}_i) = \frac{edges(\mathcal{C}_i)}{edges(\mathcal{N})} \quad (10)$$

i.e. the number of edges connecting nodes inside community \mathcal{C}_i reported to the total number of edges in the network.

The proposed fitness takes into account the score $SComm(\mathcal{C})$ of each community weighted by the volume of the community reported to the total number of edges in the network and further to the number of nodes in the community. An interesting aspect highlighted by extensive preliminary numerical experiments refers to the radical change in behavior (directly affecting search capabilities) when apparently small

changes are brought to this fitness function. For instance, if the number of edges is not further reported to the number of total edges in the network as well as the number of nodes in the community (in *SPart*) or if the node degree difference is reported to the total degree of the node (similar to the function in Eq. 2 [10]) instead of community size (in *SNode*), the performance of the evolutionary search guided by the modified fitness function drops dramatically (in some cases, from obtaining maximum NMI values of 1 to the minimum NMI of 0 corresponding to completely different detected partitions compared to real known partition).

IV. A COLLABORATIVE EVOLUTIONARY ALGORITHM FOR COMMUNITY DETECTION IN COMPLEX NETWORKS

An evolutionary algorithm (EA) based on collaborative operators (called collaborative evolutionary algorithm [3]) is developed for the problem of community detection in complex networks. The fitness function *SPart* introduced in the previous section (given by Eq. 9) is engaged to evolve individuals towards good-quality partitioning solutions. Every individual is represented as an integer array of size N , where N is the number of nodes in the network. Each position i in the array assumes a value j (where j can be an integer number from 1 to N) which is translated to a partitioning in which nodes i and j belong to the same cluster. The number of network communities emerges after the chromosome is decoded to a partitioning. The initial value of each chromosome is randomly generated with the restriction that each value j assigned to a position i means that edge (j, i) exists in the given network.

Features related to selection and application of search operators are inherited from collaborative evolutionary algorithms [3], [8]. Individuals have knowledge about the best potential solution (*GlobalBest*), the lowest fitness solution (*GlobalWorst*) already obtained in the search process and the individual's best ancestor (*LineBest*). The ancestors represent all individuals that have existed in one of the previous generations and have contributed to the creation of the current individual. If within a single ancestral line there are multiple ancestors with the same best fitness values, the closest ancestor is chosen. If within the two ancestral lines of an individual the best individuals have identical fitness, one of them is randomly chosen. In the initial population, the *LineBest* of each individual is the individual itself. *GlobalBest*, *GlobalWorst* and *LineBest* guide the search process in the form of passing *relevant* genetic material to the individuals.

The main framework of the collaborative EA proposed for community detection in complex networks is depicted in algorithm 1.

Each individual carries information about its *LineBest* and has access to the current *GlobalBest* and *GlobalWorst* values. An elitist strategy is used by which the best individuals from the current population are automatically selected for next population. The rest of the individuals are selected according to a roulette scheme. This intermediary formed population represents the pool from which parents are selected for recombination according to a collaborative selection scheme.

Algorithm 1 Collaborative EA

```

 $t \leftarrow 1$ 
Initialize  $P^t$ 
Evaluate and sort  $P^t$  using the SPart fitness (Eq. 9)
while  $t \leq \text{MaxNumberOfGenerations}$  do
  for each individual in  $\text{Elite}(P^t)$  do
    Copy individual in  $P^{t+1}$ 
  end for
  for  $i = 0$  to  $\text{PopulationSize} - \text{EliteSize}$  do
     $X = \text{RouletteSelection}(P^t)$ 
    Copy  $X$  in  $\text{SelPool}$ 
  end for
   $P^{t+1} = P^{t+1} \cup \text{SelPool}$ 
  Set GlobalBest and GlobalWorst individuals
  Update Groups
  for  $i = \text{EliteSize} + 1$  to  $\text{PopulationSize}$  do
    if XOR probability met then
       $\text{Parent1} \leftarrow P^{t+1}[i]$ 
       $\text{RandGroup} \leftarrow \text{RandomSelection}(\text{Groups} - \text{Group}(\text{Parent1}))$ 
       $\text{Parent2} = \text{BinaryTournament}(\text{RandGroup})$ 
       $X = \text{BW}X(\text{Parent1}, \text{Parent2})$ 
      Set LineBest( $X$ )
      if  $\text{Fitness}(X) > \text{Fitness}(\text{Parent1})$  then
         $P^{t+1}[i] \leftarrow X$ 
      end if
    end if
    if Mutation probability met then
       $\text{Parent} \leftarrow P^{t+1}[i]$ 
       $Y = \text{Mutation}(\text{Parent})$ 
      Set LineBest( $Y$ )
      if  $\text{Fitness}(Y) > \text{Fitness}(\text{Parent})$  then
         $P^{t+1}[i] \leftarrow Y$ 
      end if
    end if
  end for
  Update GlobalBest and GlobalWorst
   $t \leftarrow t + 1$ 
end while

```

Every generation t , the individuals within the population $P(t)$ are grouped by their *LineBest*. These groups are formed so that they represent a partition of $P(t)$ and all the individuals that belong to one cluster have the same *LineBest*. The selection of a mate for recombination purposes is performed such that an individual with a different *LineBest* than the first parent (i.e. belonging to a different group than the group of the first parent) contributes to the generation of the offspring. The mate is selected from a particular group according to a tournament scheme. A recombination operator called *Best-Worst Recombination* (BW X) is applied to such a pair of parents with a certain probability. BW X follows a particular uniform crossover scheme in which the *GlobalBest* and *GlobalWorst* information is used in addition to that from the two parents as follows:

For each chromosome position i ,

- If the value at i in the first parent is identical with $GlobalBest[i]$ then it is inherited in the offspring;
- If the value at i in the first parent is identical with $GlobalWorst[i]$ then another random neighbor of node i is set for position i in the offspring;
- Otherwise, the value at position i from the second parent is inherited.

The resulted offspring replaces the first parent if it has a higher fitness. Mutation is applied to each offspring with a certain probability and brings random re-initializations at the level of each node (selected from the neighboring nodes).

V. COMPUTATIONAL EXPERIMENTS AND RESULTS

This section presents the computational experiments and the analysis of results for the proposed evolutionary approach to community detection. The networks tested are described, the parameter setting and evaluation tools are explained and the results are discussed.

A. Benchmark Networks

Experiments focus on three real-world networks extensively used in testing community detection algorithms: Zachary's karate club network [22], Bottlenose Dolphins [14] and Krebs network of political books [16].

The first network analyzed is the Zachary's karate club network [22] which describes the friendship of 34 members of a karate club over two years. The network has 34 nodes (representing the members) and 78 edges (representing the social interactions between members). Because of disagreements, two groups of almost same size have been naturally formed (one around the administrator of the club and a second group around the coach).

The social network of bottlenose dolphins was generated by Lusseau [14] who observed the behavior of 62 dolphins over a period of 7 years. The number of edges is 159, each emphasizing a statistically significant frequent association. The dolphins network is naturally divided into two large groups.

The Krebs network of political books [16] contains 105 nodes representing books on American politics bought from Amazon.com. Edges connect two books frequently purchased by the same person. The books were divided by Newman [16] in three groups according to their political alignment: liberal, conservative and a small group of books with other or no clear affiliation.

B. Parameter Setting and Evaluation Tools

The proposed algorithm is applied for each network with the following parameters: population size is 500, number of generations is 100, crossover probability is 0.8, mutation probability is 0.5 and elite size is 10% from the population. These parameter values have been chosen based on the best results observed in an extensive phase of preliminary experiments. Binary tournament is used for selecting an individual (to act as a parent) from a cluster.

For each partitioning solution, the NMI - Normalized Mutual Information, as given in [11], [12], is computed to evaluate the results. NMI represents a similarity measure between two partitions and is based on evaluating the Shannon information content of partitions. Let x and y be the cluster labels of a node in two different partitions \mathcal{X} and \mathcal{Y} . We assume that cluster labels x and y are the values of two random variable X and Y with joint distribution:

$$P(x, y) = P(X = x, Y = y) = n_{xy}/n \quad (11)$$

where n_{xy} is the number of overlapping nodes between the two clusters labeled by x and y .

The marginal probability distribution of X , respectively Y , is defined as $P(x) = P(X = x) = n_x/n$, respectively $P(y) = P(Y = y) = n_y/n$, where n_x and n_y represent the cluster sizes for labels x and y .

The mutual information $I(X, Y)$ of two random variables X and Y is defined as:

$$I(X, Y) = \sum_x \sum_y P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \quad (12)$$

or

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \quad (13)$$

where $H(X)$ represents the entropy of random variable X associated with a partition and $H(X, Y)$ is the joint entropy.

The mutual information $I(X, Y)$ measures the information that X and Y share (or how much we learn about X if we know Y). For comparing network partitions, $I(X, Y)$ has an important limitation: given a partition \mathcal{X} , all partitions derived from \mathcal{X} by further partitioning some of its clusters would have the same mutual information with \mathcal{X} . To avoid this problem, the normalized mutual information has been proposed [4] and is currently extensively used in testing community detection algorithms.

The normalized mutual information is defined as follows [4], [11], [12]:

$$NMI(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)} \quad (14)$$

NMI is expressed as a real number between 0 and 1 with the following significance: a value of 1 for NMI means the two partitions compared are identical whereas a NMI value of 0 suggests that two completely different (independent) partitions are compared.

Using NMI, we compare the detected partitions for the considered networks to the known real partitions. For computing the NMI in our experiments we have used the source code made available by Lancichinetti et al [11] which can be freely downloaded from [23].

TABLE I
RESULTS OBTAINED BY THE PROPOSED ALGORITHM OVER 10
INDEPENDENT RUNS (BEST SOLUTION) FOR THE THREE CONSIDERED
COMPLEX NETWORKS.

Network	Fitness	NMI
Zachary's karate club network [22]	0.594	1
Dolphins Network [14]	0.368	1
Krebs network of political books [16]	0.339	0.663

C. Experimental Results

The best solution from 10 runs detected by our algorithm for each network is reported in Table I.

For the Zachary's karate club network (see Fig. 1) and the dolphins network (see Fig. 2), the proposed fitness is able to guide the evolutionary search process towards network partitions having maximum NMI value of 1 (meaning that the exact real-world known partition has been detected) while a NMI value of 0.663 is obtained for the Krebs network of political books. These NMI values have been computed for the best fitted solutions evolved by the proposed approach. Table I presents both the fitness and the corresponding NMI of the best individual in the final generation of the algorithm. The correlation between the fitness of an individual and the computed NMI generally follows the expected pattern of increasing NMI with increasing fitness values. However, some exceptions have been observed and this behavior should be further analysed and confirmed by computational experiments.

Table II presents the results obtained compared to the collaborative EA results [3] where the community score of a clustering proposed in [17] is used as a fitness function. Besides the NMI, the modularity Q [15] for each reported solution is calculated to facilitate the analysis of results. The modularity quantifies the strength of the community structure in the following way: $Q = 0$ means no quality at all (similar to placing nodes in communities completely at random) whereas $Q = 1$ is the maximum possible value indicating a strong community structure. Newman and Girvan [15] emphasize that Q values for strong community structures partitionings are in practice within the interval $[0.3, 0.7]$.

The results (in terms of modularity Q and NMI) obtained by the collaborative EA based on the $SPart$ fitness function are given in the last two columns of Table II. These results outperform the previous results reported in [3] where the same collaborative EA has been used based on a different fitness function. In [3], it has been shown that the performance of collaborative EAs is better than that of the GA-Net method [17] (both evolutionary algorithms using the same representation and fitness function with the main difference being the evolutionary search framework - a standard one in [17] and a collaborative one in [3]). Furthermore, the results in Table II clearly emphasize the better comparative performance of the proposed fitness function guiding the collaborative evolutionary search as opposed to the fitness used in [3].

Table II also presents the modularity Q of all solutions reported. Firstly, it should be noted that all partitionings have a Q value inside the interval $[0.3, 0.7]$ emphasizing indeed a

TABLE II
COMPARATIVE NUMERICAL RESULTS: MODULARITY (Q) AND
NORMALIZED MUTUAL INFORMATION (NMI) ARE PRESENTED FOR THE
SAME EA USING THE COMMUNITY SCORE FITNESS [17] AS IMPLEMENTED
IN [3] AND THE PROPOSED FITNESS FUNCTION. RESULTS ARE THE BEST
OBTAINED OVER 10 INDEPENDENT RUNS.

Network	EA from [3]		EA with $SPart$ Fitness	
	Q	NMI	Q	NMI
Zachary's karate club network [22]	0.399	0.825	0.371	1
Dolphins Network [14]	0.458	0.568	0.383	1
Krebs network of political books [16]	0.501	0.604	0.443	0.663

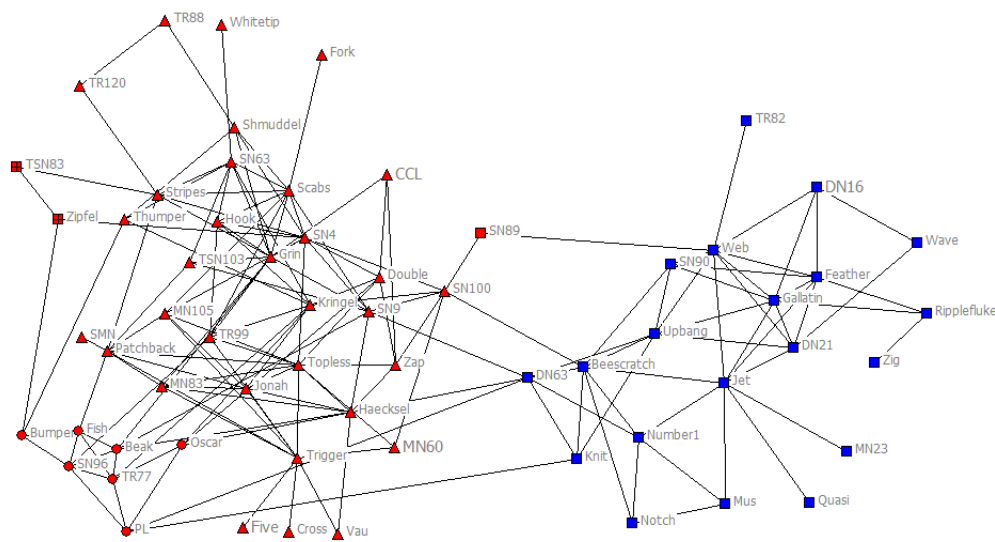
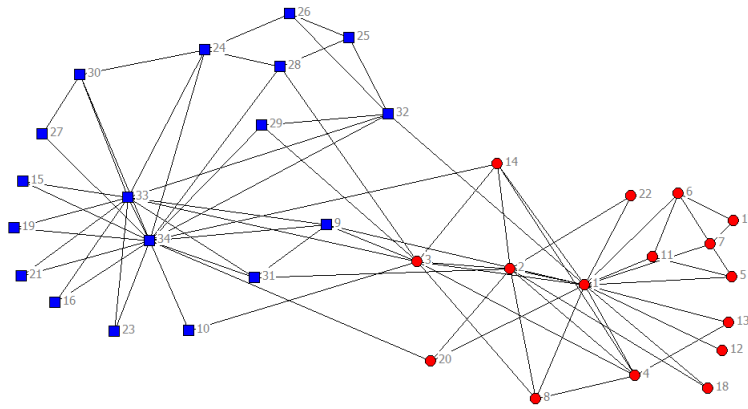
strong community structure (for both the EA results reported in [3] and the new obtained results). Secondly, an interesting aspect related to the Q results is that the modularity of the new solutions detected having higher NMI values actually have a lower corresponding Q value than those initially obtained in [3]. For instance, the modularity of a partitioning solution having NMI value 1 evolved by the proposed fitness function is 0.383 while the previously obtained result with a significantly lower NMI of 0.568 actually triggers a higher modularity of 0.458. While both solutions still have $Q \in [0.3, 0.7]$, the fact that a lower modularity corresponds to a higher NMI confirms the limits of modularity optimization to produce good close-to-real partitioning solutions.

VI. CONCLUSIONS AND FUTURE WORK

A new fitness function has been presented to evaluate the quality of community structures in complex networks. In the proposed approach, the strength of each community is measured by looking at the difference between the internal and external degree reported to the community size and further added by the weighted contribution (computed in the same way) of each neighboring node. A collaborative evolutionary algorithm relying on the proposed fitness has been implemented to address the problem of community structure detection.

The experiments performed for three real-world complex networks indicate the ability of the proposed fitness function to guide the evolutionary search process towards good partitioning solutions, many of them being the exact known real-world community structures.

Future work focuses on investigating the effect of individual representation choice and variation operators on the effectiveness of the search process. Also, an extensive analysis of the behavior of the proposed fitness function will be carried out to check and explain the correlation between the changes in fitness and the trend of the corresponding NMI line. Furthermore, we plan to extend the computational experiments to other network data including overlapping community structure detection, online social networks and dynamic complex networks.



- [8] Gog, A., Dumitrescu, D., Hirsbrunner, B.: Community Detection in Complex Networks Using Collaborative Evolutionary Algorithms. ECAL 2007, pp. 886-894 (2007)
- [9] Guimera, R., Amaral, L. A. N., Functional cartography of complex metabolic networks, *Nature* 433, 895900 (2005)
- [10] Lancichinetti, A., Fortunato, S., Kertesz, J., Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics* 11 (2009) 033015
- [11] Lancichinetti, A., Fortunato, S., Community detection algorithms: A comparative analysis, *Phys. Rev. E* 80, 056117 (2009).
- [12] Lancichinetti, A., Radicchi, F., Ramasco, J.J., Fortunato, S., Finding statistically significant communities in networks, *PloS One* 6, e18961 (2011)
- [13] Lancichinetti, A., Fortunato, S., Limits of modularity maximization in community detection, *Phys. Rev. E* 84, 066122 (2011)
- [14] Lusseau, D., The emergent properties of dolphin social network. *Biology Letters, Proc. R. Soc. London B* (2003)
- [15] Newman, M. E. J., Girvan, M., Finding and Evaluating Community Structure in Networks. *Physical Review E*, 69 (2004) 026113-1
- [16] Newman, M. E. J., Modularity and community structure in networks, *Proc. Natl. Acad. Sci. USA* 103, 8577 (2006).
- [17] Pizzuti, C., GA-NET: a genetic algorithm for community detection in social networks. In *Proceedings of the 10th International Conference on Parallel Problem Solving from Nature* (2008) 1081-1090

- [18] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D., Defining and Identifying Communities in Networks. Proceedings of National Academy of Science in USA, vol. 101 (2004) 2658 - 2663
- [19] Scott, J., Social Network Analysis, A Handbook, Sage Publication, London (2000)
- [20] Tasgin, M., Bingol, H., Community Detection in Complex Networks using Genetic Algorithm. cond-mat/0604419 (2006)
- [21] Watts, D.J., Six degrees: The Science of a Connected Age. Gardner's Books, New York (2003)
- [22] Zachary, W.W., An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33 (1977) 452-473
- [23] <http://sites.google.com/site/andrealancichinetti/mutual>