COMPARATIVE ANALYSIS OF QUALITY METRICS FOR COMMUNITY DETECTION IN SOCIAL NETWORKS USING GENETIC ALGORITHM

S. Kaur[*], *S. Singh*[*], *S. Kaushal*[*], *A.K. Sangaiah*[†]

Abstract: Web 2.0 has led to the expansion and evolution of web-based communities that enable people to share information and communicate on shared platforms. The inclination of individuals towards other individuals of similar choices, decisions and preferences to get related in a social network prompts the development of groups or communities. The identification of community structure is one of the most challenging task that has received a lot of attention from the researchers. Network community structure detection can be expressed as an optimisation problem. The objective function selected captures the instinct of a community as a group of nodes in which intra-group connections are much denser than inter-group connections. However, this problem often cannot be well solved by traditional optimisation methods due to the inherent complexity of network structure. Therefore, evolutionary algorithms have been embraced to deal with community detection problem. Many objective functions have been proposed to capture the notion of quality of a network community. In this paper, we assessed the performance of four important objective functions namely Modularity, Modularity Density, Community Score and Community Fitness on real-world benchmark networks, using Genetic Algorithm (GA). The performance measure taken to assess the quality of partitions is NMI (Normalized mutual information). From the experimental results, we found that the communities' identified by these objectives have different characteristics and modularity density outperformed the other three objective functions by uncovering the true community structure of the networks. The experimental results provide a direction to researchers on choosing an objective function to measure the quality of community structure in various domains like social networks, biological networks, information and technological networks.

Key words: *community detection, social network, optimization, objective function, Genetic Algorithm, Normalized mutual information*

Received: September 10, 2016 **DOI:** 10.14311/NNW.2016.26.036
Revised and accepted: October 11, 2016

[*]Simrat Kaur – Corresponding author, Sarbjeet Singh, Sakshi Kaushal, University Institute of Engineering and Technology, Panjab University, Chandigarh 160014, India, E-mail: simrat_bala@yahoo.co.in, sarbjeet@pu.ac.in, sakshi@pu.ac.in

[†]Arun Kumar Sangaiah, VIT University, Vellore-632014, Tamil Nadu, India, E-mail: arunkumarsangaiah@gmail.com

1. Introduction

Various real world complex systems of great significance can be modelled as networks [35]. Examples span various fields including Biological networks (food web, protein networks), Technological networks (power grid, road networks), Social networks (friendships, collaboration networks), Information networks (World Wide Web, co-citation), Distribution networks (blood vessels, postal delivery) and many more [28]. The network data can be modelled with graph theory. A graph $G = (V, E)$ comprises of a vertex set V, where V denotes number of nodes or vertices and an edge set E, where E represents interconnections among the nodes [37].

In the last decade, the exponential growth of social networks like Facebook, Twitter, YouTube and LinkedIn has attracted the research community to study and analyse their properties at a large scale [6, 24, 39, 40]. The increasing accessibility and availability of data from these networks has given a thrust to research in this area. In social networks, people are represented by nodes and edges between nodes represent different types of social interaction including friendship, collaboration or others. One of the imperative problem in studies of complex networks, especially in social networks, is finding out underlying sub-structures or community structures [25].

A community in a social network is defined as a set of nodes which has high edge density among themselves and a lower edge density with rest of the network [10]. The problem of identifying n communities in a network, where the number n is unknown, can be expressed as partitioning of the nodes in k subsets that have more connections within and sparser connections among groups.

Communities occur in many networked systems like computer science, politics, engineering, biology and economics. The Homophily principle – "the tendency of individuals to associate with those similar to themselves i.e. similarity breeds connection" has been discovered in many social networks [23]. So, people with similar characteristics (race, gender, age, family, co-workers etc.) and interests tend to form communities. Communities usually represent real social groupings that share some common properties [10]. For example, communities on the web may represent pages on related subjects [7] and communities in a citation network may signify papers that are related to a single topic [4] .The identification of these communities could help us to comprehend and exploit the networks more efficiently.

The capability of detecting communities in a network can give valuable intuitions to understand how the structure of ties influences people and their links. Communities are of interest for many reasons. For example, the performance of services provided on the World Wide Web can be improved by grouping Web clients with similar interests and geographical proximity as each group could be served by a dedicated mirror server [16]. Similarly, clustering clients with similar interests in purchase networks enables to develop efficient recommendation systems and boost the business prospects of online retailers [32].

Community detection can be formulated as an optimization problem, where the objective function selected maximizes the number of links inside each partition of the network. Genetic Algorithm is one of the most widely employed method to solve complex optimization problems. GA is based on the mechanics of biological processes of reproduction and natural selection to solve for the 'fittest' solutions [14].

The objective function plays a vital role in the evolution process of a Genetic algorithm. In many applications, the real community structure is not known, so, there is a requirement for developing metrics to evaluate identified communities. However, the results obtained from optimization of some of these objective functions fail to depict real community structure [8]. Besides, different objective functions may result in different group allotments. Some may produce communities with coarser granularity (few communities of huge size), though some may give back a finer granularity group structure. Due to lack of a direct way to compare these objective functions based on their definitions, it is important to compare the performance of these objective functions on real networks. Various objective functions have been proposed in literature to capture the notion of communities. In this paper, we evaluated the performance of four well-known objective functions, namely modularity, modularity density, community score and community fitness, proposed in literature, to gauge the quality of partitions of a network. We also compared our results with the two existing algorithms – Fast Modularity and MOGA-Net. Our results showed that the communities' identified by these objectives have different characteristics and modularity density gave the best results to identify community structure. The experimental results give a direction to researchers for selecting the suitable objective function to assess the quality of community structure of the networks.

The remainder of this paper is organised as follows. Section 2 gives some related background about network community detection approaches. Section 3 includes details of the genetic algorithm used for community detection. Section 4 presents the experimental results and conclusion is given in Section 5.

2. Related literature

To precisely analyse the community structure in networks, many community detection methods have been proposed using principles of different domains like physics, artificial intelligence, graph theory etc. This section gives a brief description of some of the pioneer work in community detection and research works in which GA is used for community detection problem.

Girvan-Newman proposed a divisive algorithm that iteratively removes the edges based on highest betweenness value [10]. The edge removal divides the network into communities. The authors also defined a measure known as 'modularity' to estimate the quality of communities found. The modularity has been widely used by researchers to measure the goodness of the partitions obtained from the community detection algorithms. They also proposed a faster method version of previous algorithm [26]. Radicchi F. et al. introduced the concept of strong and weak communities and proposed an algorithm based on the GN algorithm and edge clustering co-efficient [31]. But, these community detection algorithms have large computational complexity and are not suitable for very large networks. A more detailed review can be found in [2].

Among the GA approaches, M. Tasgin and H. Bingol were first to propose community detection based on genetic algorithm using network modularity measure to find the best community structure. An individual or chromosome is represented by N genes, where N is the number of nodes of network. The i-th gene corresponds

to i-th node and its value represents the community id of node i. They also introduced one-way crossover operation. The complexity of proposed algorithm is $O(e)$, where e is the number of edges in the network. Moreover, the algorithm does not require any prior information about the number of communities present in the network [36].

Liu et al. used clustering and genetic algorithm to find the community structure in a network. The graph is repeatedly subdivided into two parts, and a nested genetic algorithm is applied to them. They used modularity metric for successive bipartitions of the graph. A bipartition is accepted only if total modularity of the graph is increased by it [20].

In [8], authors have shown that the optimization of modularity has a major drawback. It cannot find even the well-defined communities smaller than a fixed scale. The scale depends on the network size and the interconnection degree of the modules. This resolution limit is a weakness for all those algorithms that use modularity as the objective function to optimize.

C. Pizzuti introduced a new measure of network partitioning called community score and tried to optimize this measure using genetic algorithm known as GA-Net. The locus-based adjacency representation is used to represent an individual of the population. The algorithm does not require the prior information of number of communities to find. Their experimental results on synthetic and real networks showed the viability of the genetic approach to correctly find communities. But the chromosome representation used requires an additional decoding step [30]. C. Shi et al. proposed a genetic algorithm using locus-based adjacency encoding representation and modularity as the fitness function and validated the performance of algorithms on synthetic and real networks [34]. But the objective function used could fail to uncover community structure of modules less than a particular scale. P. Mazur et al. have compared the results of Pizzuti's genetic algorithm by using two fitness functions: modularity and community score [22].

R. Agrawal proposed a Bi-Objective Community Detection (BOCD) using Genetic Algorithm that employed modularity and community score as fitness function. The results obtained showed enhanced performance over the existing community detection methods [1]. Hafez et al. performed both single-objective and multi-objective GA based optimization for community-detection problems. They used the well-known objective functions proposed in literature and also showed correlation among various objectives [13].

C. Pizzuti proposed a multi-objective approach, named multi-objective genetic algorithm (MOGA-Net), to determine communities in networks using genetic algorithms. The first objective function uses the concept of community score and the second defines the concept of fitness of the nodes belonging to a module. The approach was tested on synthetic and real life networks and showed its ability to appropriately identify communities. However, the multi-objective approach used has a time complexity quadratic of the population size [29]. R. Shang put forward a community detection method based on modularity as the objective function, and used simulated annealing method as the local search. The drawback of algorithm is that it requires prior information about the number of community structures [33]. L. Yun et al. developed a GA based on matrix encoding and nodes similarity and network modularity as fitness function. The matrix encoding needs no ad-

ditional decoding and node's similarity is used to generate an initial population. The proposed approach performed poorly with networks with unclear community structure [18].

Ali Ghorbanian's algorithm optimizes modularity density using genetic algorithm approach. Matrix encoding was used to initialize population and used Adjusted Rand Index (ARI) as performance measure. The encoding scheme used requires additional decoding step [9]. G. chen et al. proposed a multi-objective evolutionary algorithm for dynamic networks community detection based on Modularity density and normalized mutual information and designed a local search operator to improve the quality of community detection. But, the algorithm has not been tested on real life networks [3].

So, it is clear from literature review that many objective functions have been proposed that use Genetic Algorithms (GA) as an effective optimization technique. Therefore, it is important to study the performance of these objective functions and to understand the structural properties of groups' identified by various objective functions. This will help in the selection of the most apt objective with regards to a given network.

3. Community detection using Genetic Algorithm

Genetic algorithms have many advantages such as adaptive heuristic search nature, require limited auxiliary knowledge of the problem, converge a problem to a smaller solution space, and generates near optimal solutions, which makes them a suitable candidate for community detection problem. GA is generally appropriate for the problems with a large solution space and where an extensive search for the optimal solution is unfeasible. Moreover, prior knowledge about the community structure is not easy to obtain in real-world. The GA based community detection methods can automatically determine the number of clusters in a network which makes them useful for real world networks. In GA, a population of chromosomes or individuals is randomly initialized where chromosome represents a possible solution to the problem. Each member of the population is then evaluated and a 'fitness' function is calculated for that individual, which shows the goodness a solution member towards the solution of the problem. To improve the overall fitness of the population, Selection is used to discard the worst solutions and only keeping the best individuals in the population. Thereafter, the population of solutions is modified to generate a new population by applying genetic operators like crossover and mutation and the whole process repeats until the stopping criterion is fulfilled [11].

3.1 Representation and initialization

When GA is applied to the community detection problem, the two dominant methods for representing the individuals are used: the string encoding representation [36] and the locus-based adjacency representation [30,34]. Using string encoding, a partition of the network G is encoded as an integer string $x = \{x^1, x^2, \ldots, x^n\}$, where n denotes number of the vertices and x^i is the integer cluster identifier of vertex v_i, whose value lie between 1 and n. In locus-based or graph-based representation,

each individual G is represented by n genes $\{G_1, G_2, \ldots, G_n\}$ and each G_i can take one of the adjacent nodes of node i. Thus, a value of j assigned to the i-th gene, is then interpreted as a link between nodes i and j; in the resulted partition solution, the two nodes will be in the same community. A decoding step, however, is necessary to identify all the components of the corresponding graph. In this paper, we have used string encoding representation because of its simplicity.

During population initialization, each vertex is assigned a random community identifier. However, it is a typical practice to give the genetic algorithm not a completely random initialization but a biased one in order to accelerate the convergence. To introduce bias, we randomly picked a vertex v_i and assigned its cluster id to all of its neighbours. This operation was repeated αn times for each chromosome in the initial population where $\alpha = 0.2$ is used in this paper.

3.2 Selection

The Tournament selection procedure is used to select parental population for mating in GA. Tournament selection provides selection by holding a tournament among a few individuals that are randomly chosen from the population. The individual with the highest fitness is the winner of the tournament of the S tournament competitors. The winner is then inserted into the mating pool that comprises of tournament winners and hence provides a higher average fitness than the average population fitness [12].

3.3 Fitness function

In this paper, the performance of following objective functions proposed in literature has been evaluated:

Definition 1. Modularity: The network modularity Q is defined as

$$Q = \sum_i (e_{ii} - a_i^2), \tag{1}$$

where i represents index of communities, e_{ii} is the fraction of edges connecting two nodes in a community i, to the total number of edges in the network and a_i is the faction of all the edges with at least one node in the community i to the total number of edges in the network [27]. Although, Modularity maximization is an effective method for community detection but Fortunato and Barthelemy showed the resolution limit problem of this method [8].

Definition 2. Modularity Density: Let N_1 and N_2 be the two disjoint subsets of N, then $L(N_1, N_2) = \sum_{i \in N_1, j \in N} a_{ij}$ and $L(N_1, N_1) = \sum_{i \in N_1, j \in N_1} a_{ij}$ and $L\left(N_1, \bar{N_1}\right) = \sum_{i \in N_1, j \in \bar{N_1}} a_{ij}$, where $\bar{N_1} = N - N_1$. Given a partition $\Omega = \{N_1, N_2, \ldots, N_m\}$ of the graph, where N_i is the vertex set of subgraph G_i for $i = 1 \ldots m$. The modularity density is defined as follows:

$$D = \sum_{i=1}^{m} \frac{L(N_i, N_i) - L(N_i, \bar{N_i})}{|N_i|}. \tag{2}$$

A general modularity density measure is given as

$$D_\lambda = \sum_{i=1}^{m} \frac{2\lambda L\left(N_i, N_i\right) - 2(1-\lambda)L(N_i, \bar{N})}{|N_i|}. \tag{3}$$

By changing the value of λ, more detailed and hierarchical grouping of the networks can be uncovered [19].

Definition 3. Community Score: Let $\mathbf{R} \subset G$ be the sub graph and the degree of i with respect to $\mathbf{R}$ can be written as

$$k_i\left(\mathbf{R}\right) = k_{\text{in}}^i\left(\mathbf{R}\right) + k_{\text{out}}^i(\mathbf{R}), \tag{4}$$

where $k_{\text{in}}^i\left(\mathbf{R}\right) = \sum_{j\in\mathbf{R}} a_{ij}$ is the number of edges connecting i to the other nodes in $\mathbf{R}$ and $\mathbf{A}$ is the adjacency matrix of G.

$k_{\text{out}}^i\left(\mathbf{R}\right) = \sum_{j\notin\mathbf{R}} a_{ij}$ is the number of edges connecting i to the rest of the network.

Let μ_i is the fraction of edges connecting i to the other nodes in $\mathbf{R}$. Then, $\mu_{i=}\frac{1}{|\mathbf{R}|}k_{\text{in}}^i\left(\mathbf{R}\right)$ where $|\mathbf{R}|$ is the cardinality of $\mathbf{R}$.

The power mean of $\mathbf{R}$ of order k, $M\left(\mathbf{R}\right)$ is given by

$$M\left(\mathbf{R}\right) = \sum\nolimits_{i\in\mathbf{R}} \left(\mu^i\right)^k. \tag{5}$$

The volume V_r of a community is defined as the number of edges connecting nodes inside $\mathbf{R}$,

$$V_r = \sum\nolimits_{i,j\in\mathbf{R}} a_{ij}. \tag{6}$$

The score($\mathbf{R}$) of $\mathbf{R}$ is defined as

$$\text{score}\left(\mathbf{R}\right) = M\left(\mathbf{R}\right) \times V_r. \tag{7}$$

The Community score of a clustering $\{\mathbf{R}_1, \ldots, \mathbf{R}_m\}$ of a network is defined as [30]

$$CS = \sum\nolimits_{i=1}^{m} \text{score}(\mathbf{R}_i). \tag{8}$$

Definition 4. Community Fitness: Let $k_{\text{in}}^i(M)$ and $k_{\text{out}}^i\left(M\right)$ be the internal and external degrees of the nodes belonging to a community M. The community fitness $P(M)$ of M is defined as follows:

$$P\left(M\right) = \sum_{i\in S} \frac{k_{\text{in}}^i(M)}{\left(k_{\text{in}}^i\left(M\right) + k_{\text{out}}^i(M)\right)^\alpha}, \tag{9}$$

where α is resolution parameter that controls the size of the communities [17].

3.4 Cross over

Traditional crossover operation is not suitable for community detection problem. In this, we have used two-way crossing over operation that involves random selection of two chromosomes, called x_a and x_b, where x_a corresponds to the source and x_b to the destination chromosome. Then, random pick a vertex v_i, and then find cluster assigned to v_i (i.e. x_a^i) in the x_a chromosome. Finally, all the vertices with this cluster identifier of x_a are also assigned to the same cluster identifier in chromosome x_b. Then, x_b is taken as source chromosome and x_a as target chromosome, and then repeat the steps described. This process returned two new chromosomes x_c and x_d. An example of two-way crossing over is given in Tab. I. Two-way crossing can increase the diversity of chromosomes by generating children carrying features common to the parents [12].

v		x_a		x_b		x_c	x_d		x_a		x_b		v
1		(5)	→	2	→	(5)	5		5		2		1
2		3		6		6	(6)	←	3	←	(6)		2
(3)	→	(5)	→	6	→	(5)	(6)	←	5	←	(6)	←	(3)
4		7		5		5	7		7		5		4
5		2		6		6	(6)	←	2	←	(6)		5
6		(5)	→	3	→	(5)	5		5		3		6
7		3		2		2	3		3		2		7

Tab. I *Two way Crossover operation.*

3.5 Mutation

During mutation, a vertex is picked randomly on the chromosome, then the cluster of the vertex is randomly changed to the cluster of one of its neighbours. This operator introduces random changes in various chromosomes to increase the diversity of the population and speed up the convergence [12].

The algorithm stops when the maximum number of generations have reached. The above methodology shown in Fig. 1 has been applied on four real life benchmark networks which are discussed in next section.

3.6 Time complexity

Community Detection is NP hard problem in which search space is 2^n. GA involves a heuristic approach, therefore, the solution space is not exhaustibly searched and lowers the computational complexity. The fitness function evaluation is the most time consuming process in the algorithm. Calculating the objective function has the complexity of $\mathcal{O}(n)$ where n is length of chromosome. So, the overall complexity of algorithm is $\mathcal{O}(G \times S \times n)$ which is linear with the size of the network. Here, G is the running generation, and S is the population size.

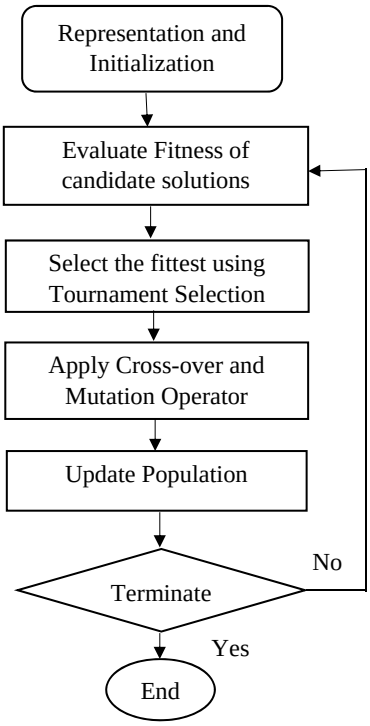

Fig. 1 *Flowchart of GA for Community Detection.*

4. Experimental results

This section gives the simulation results of the algorithm, on four real world benchmark networks, implemented using MATLAB. The parameters employed are given in the Tab. II.

Parameter	Value
Population size, S_{pop}	500
Number of generations	50
Crossover probability, P_{c}	0.8
Mutation probability, P_{m}	0.2
Tournament size	2
Size of the mating pool	$S_{\text{pop}}/2$

Tab. II *Parameter settings for the algorithm.*

4.1 Dataset description

In this paper, four real-world social networks are employed to verify the performance of objective functions. These are well-known benchmark datasets for the community detection problem and have known (true) community structures, which provides ground-truth for validating the results.

- Zachary Karate Club Network: It was generated by Zachary, who examined the fellowship of 34 individuals for a time of two years from a karate club. Over the span of the study, a disagreement developed between the administrator and instructor of the club leading to its split [38]. This network has 34 nodes and 78 edges. The network is shown in Fig. 2.
- Bottlenose Dolphin Network: The network consists of 62 bottlenose dolphins living in New Zealand and was compiled by Lusseau by studying dolphin behaviour for seven years. A link between 2 dolphins was established by their statistically frequent association [21]. The network is shown in Fig. 3.
- American College Football Network: This network, compiled by Girvan and Newman, represents American football games. The vertices in the network are the college football teams and there is an edge between two teams if they played a match during the season. The network consists of 115 nodes and 616 edges grouped into 12 teams. The teams are divided into conferences containing around 8–12 teams each [10]. The network is shown in Fig. 4.
- Books about US politics: The network of political books was compiled by V. Krebs [15]. It consists of 105 books about US politics published in 2004 and sold by the online book retailer Amazon.com. Edges between books represent frequent co-purchasing of books by the same purchasers. Books were divided by Newman according to their political alignment (conservative or liberal) into three communities. The network is shown in Fig. 5.

4.2 Performance measure

Normalized mutual information (NMI) is used to measure closeness between the true community structure and the detected community structure. Given two partitions X and Y of a network, $\mathbf{Z}$ is the confusion matrix. The rows of $\mathbf{Z}$ represent the "real" communities, and the columns represent the "found" communities. The element of $\mathbf{Z}$, z_{ij} is the number of nodes in the real community i that are also present in community j. The similarity measure between the partitions, based on information theory, is then given as [5]

$$\mathrm{NMI}\,(X,Y) = \frac{-2\sum_{i=1}^{C_X}\sum_{j=1}^{C_Y} z_{ij}\log\left(\frac{z_{ij}N}{z_{i.}z_{.j}}\right)}{\sum_{i=1}^{C_X} z_{i.}\log\left(\frac{z_{i.}}{N}\right)+\sum_{j=1}^{C_Y} z_{.j}\log\left(\frac{z_{.j}}{N}\right)},$$

where C_X is the number of real communities, C_Y is the number of found communities and N is the number of nodes. The sum over row i of matrix z_{ij} is denoted by $z_{i.}$ and the sum over column j is denoted by $z_{.j}$. The higher value of NMI indicates the more approximation of detected communities to the true communities.

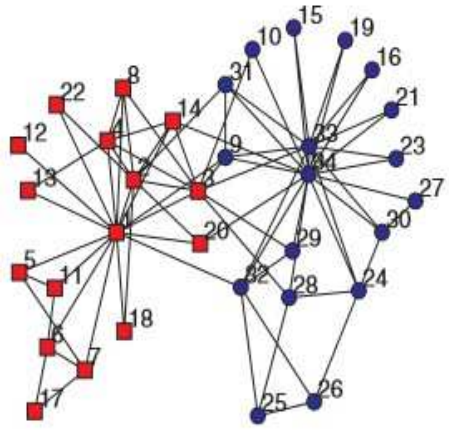

Fig. 2 *Zachary Karate Club Network.*

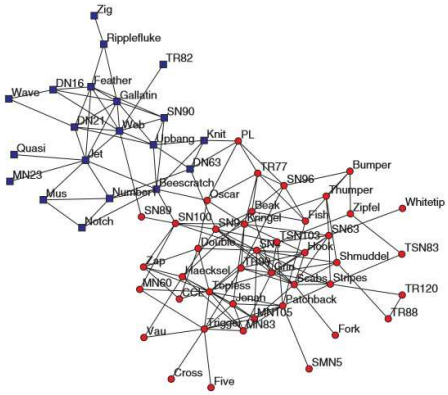

Fig. 3 *Bottlenose Dolphin Network.*

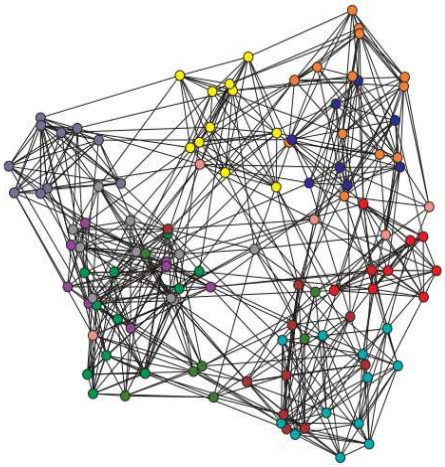

Fig. 4 *American College Football Club.*

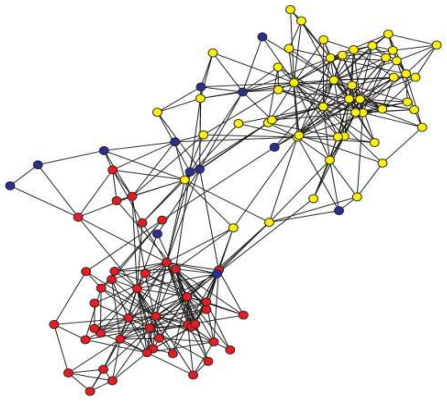

Fig. 5 *Books about US politics.*

4.3 Results and discussion

The following Tab. III shows the result obtained for modularity. For each network, we calculated the normalized mutual information by running the algorithm 10 times and recorded the best normalized mutual information value over the 10 runs of the algorithm.

The results obtained for other objective functions are presented in the Tab. IV. We recorded the results for parameters by choosing different values ranging from 0.2 to 0.8.

The above three objective functions are parameter dependent i.e. by tuning the parameter λ, κ and α, we are able to explore the network at different resolutions. For the smaller values, the number of clusters formed are small, but as we increase the parameters, we can uncover the community structure at the finer granularity.

Dataset	Number of Real Communities	Number of Detected Communities	NMI
Zachary's Karate club	2	4	0.687
Bottlenose Dolphin network	2	5	0.593
American College Football Club	12	11	0.888
Books about US politics	3	4	0.574

Tab. III *NMI obtained for modularity of four networks.*

Dataset	Modularity Density			Community Score			Community Fitness		
	λ	NMI	No. of clusters	κ	NMI	No. of clusters	α	NMI	No. of clusters
Zachary's Karate club	0.2	0.000	1	0.2	0.2259	2	0.2	0.226	2
	0.3	1.000	2	0.3	0.8255	3	0.3	0.6694	3
	0.4	0.6995	3	0.4	0.7730	4	0.4	0.6021	4
	0.6	0.6872	4	0.6	0.7327	5	0.6	0.5921	5
	0.8	0.6369	5	0.8	0.5502	8	0.8	0.4730	8
Dolphin social network	0.2	0.8888	2	0.2	0.8884	2	0.2	0.8888	2
	0.4	1.000	2	0.4	0.4953	14	0.4	0.4724	8
	0.6	0.4778	6	0.6	0.3736	19	0.6	0.4323	10
	0.8	0.4333	11	0.8	0.3635	21	0.8	0.3716	15
American college football	0.2	0.6047	4	0.2	0.5696	6	0.2	0.907	11
	0.4	0.6267	5	0.4	0.9269	12	0.4	0.9234	12
	0.6	0.9323	12	0.6	0.9026	13	0.6	0.9116	13
	0.8	0.9002	13	0.8	0.9260	14	0.8	0.8978	13
Books about US politics	0.2	0.5686	2	0.2	0.5746	3	0.2	0.5979	2
	0.4	0.5734	3	0.4	0.5365	9	0.4	0.5010	7
	0.6	0.5363	6	0.6	0.4688	13	0.6	0.4115	14
	0.8	0.4055	11	0.8	0.4110	18	0.8	0.3691	22

Tab. IV *Results obtained for modularity density, community score and community fitness for four networks.*

From the results obtained, it can be seen that only modularity density achieved a perfect NMI value for two networks namely Karate Club and Dolphin network. That is, the community structure detected is exact of the real community structure. For the American Football Club, no objective function has attained NMI value of 1 but all showed promising result by obtaining a value greater than 0.9. That is, the detected community structure is very close to the real community structure. Similarly, the results obtained for the US Politics Books dataset, the performance of all the three objectives was almost similar and comparable with results obtained by the other state-of the art methods. As shown in Fig. 5, this network is very dense and it is hard to detect real community structure of the network. The Fig. 6

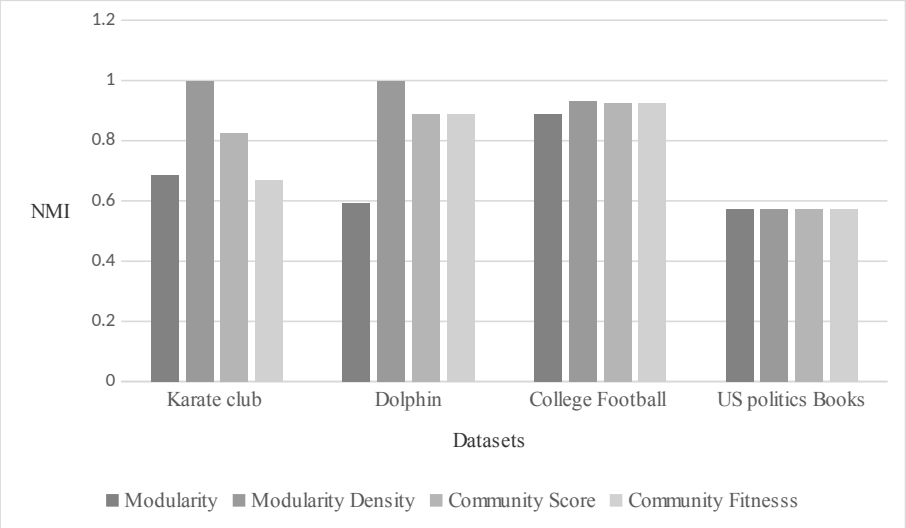

Fig. 6 *Performance of different objective functions based on NMI.*

shows the graphical comparison of objective functions by comparing the best results obtained by them.

The Tab. V shows the results obtained on the basis of number of communities detected by the objective functions. The last column shows the real number of communities. The Modularity Density is able to detect the exact number of communities as compared to other three objectives.

Dataset	Modularity	Modularity Density	Community Score	Community Fitness	Real Communities
Zachary's Karate club	4	2	3	3	2
Dolphin social network	5	2	2	2	2
American college football	11	12	12	12	12
Books about US politics	4	3	3	2	2

Tab. V *Number of communities detected by modularity density, community score and community fitness.*

Fig. 7 shows the comparison of our algorithm based on best NMI value for Modularity Density with MOGA-Net [25] and Fast Modularity Algorithm [15] NMI values for Karate club, Dolphins Network, American Football Club and US Politics Books Datasets. The figure clearly shows the good performance of Modularity Density as compared to other two approaches.

Fig. 8 shows the running time of algorithm w.r.t. different population size and number of generations. The results show that the running time increases almost linearly with the increase in generation and population size.

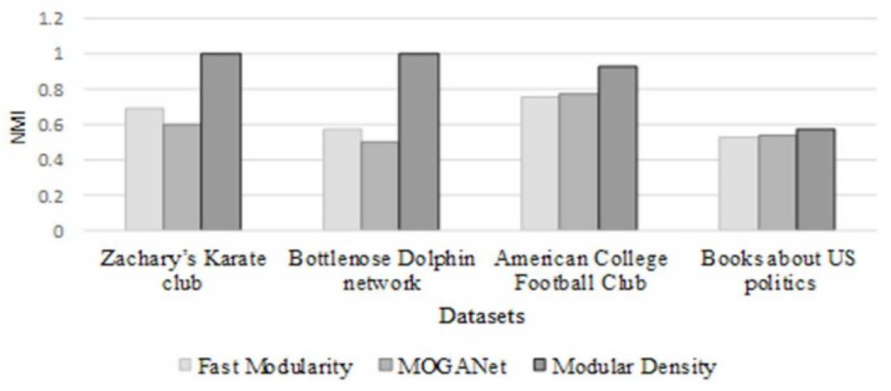

Fig. 7 *Comparison of our algorithm using Modularity Density, GA-Net and Newman's Fast algorithm relative to Normalized Mutual Information for different data sets.*

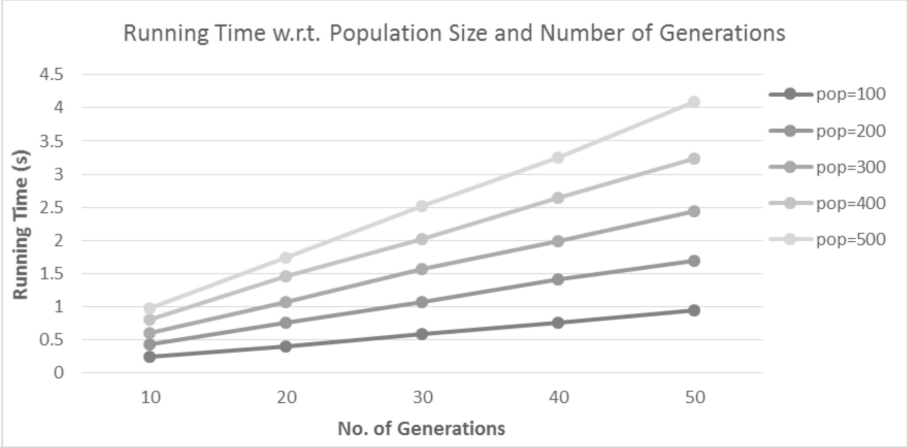

Fig. 8 *Running time of algorithm based on different parameters.*

Tab. VI shows the standard deviation values of NMI recorded over the 10 runs of the algorithm. The small value of standard deviation indicates convergence of the algorithm. Similar results were obtained for the other three datasets.

5. Conclusion

In this paper, we examined the most widely used objective functions for the community detection problem. From the experimental results, it was found that different objective functions result in different community identification. The modularity function, Q suffers from resolution limit problem i.e. it is unable to detect commu-

Dataset	λ	NMI	Std. Dev.(NMI)
	0.200	0.000	0.000
Zachary's	0.300	1.000	0.000
Karate	0.400	0.699	0.102
club	0.600	0.687	0.050
	0.800	0.637	0.030

Tab. VI *Standard Deviation of NMI for Karate Club dataset for modularity density.*

nities of small size and favours network partitions with larger communities. Modularity Density was able to achieve best results both in terms of NMI and number of detected communities. For Karate Club and Dolphin Network, it was able to detect the real community structure. Community Score also attained promising results and for some of networks, its performance is equivalent to Modularity Density. However, the modularity density, Community Score and Community Fitness are parameter dependent objective functions and only by adjusting these parameters ideal community structure can be detected. By varying the value of parameter, we can explore network structure at different granularities. At smaller values, the algorithm tend to divide network into small number of communities of large size (i.e. coarser) and at large values, more detailed community structure can be viewed (i.e. large number of communities with of small size). These findings will be helpful for choosing the most fitting objective function with for a network. Future research will aim at using other evolutionary algorithms to optimize the both single objective and multi objective functions for community detection and to test the algorithms on large scale social networks.

References

[1] AGRAWAL R. Bi-objective community detection (BOCD) in networks using genetic algorithm. *Contemp. Comput.* 2011, 168, pp. 5–15, doi: 10.1007/978-3-642-22606-9_5.

[2] BEDI P., SHARMA C. Community detection in social networks. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 2016, 6(3), pp. 115–135, doi: 10.1002/widm.1178.

[3] CHEN G.Q., WANG Y.P., WEI J.X. A New Multiobjective Evolutionary Algorithm for Community Detection in Dynamic Complex Networks *Math. Probl. Eng.* 2013, doi: 10.1155/2013/161670.

[4] CHEN P., REDNER S. Community structure of the physical review citation network. *J. Informetr.* 2010, 4(3), pp. 278–290, doi: 10.1016/j.joi.2010.01.001.

[5] DANON L., DÍAZ-GUILERA A., DUCH J., ARENAS A. Comparing community structure identification. *J. Stat. Mech. Theory Exp.* 2005, (9), pp. 219–228, doi: 10.1088/1742-5468/2005/09/P09008.

[6] FENG W., ZHANG Z., WANG J., HAN L. A Novel Authorization Delegation for Multimedia Social Networks by using Proxy Re-encryption. *Multimedia Tools and Applications*, 2016, 75(21), pp. 13995–14014, doi: 10.1007/s11042-015-2929-2.

[7] FLAKE G.W., LAWRENCE S., LEE GILES C., COETZEE F.M. Self-organization and identification of web communities. *Computer (Long. Beach. Calif).* 2002, 35(3), pp. 66–71, doi: 10.1109/2.989932.

[8] FORTUNATO S., BARTHELEMY M. Resolution limit in community detection. *Pnas.* 2007, 104(1), pp. 36–41, doi: 10.1073/pnas.0605965104.

[9] GHORBANIAN A. A Genetic Algorithm for Modularity Density Optimization in Community Detection. *International Journal of Economy, Management and Social Sciences.* 2015, 4, pp. 117–122.

[10] GIRVAN M., NEWMAN M.E.J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 2002, 99(12), pp. 7821-7826, doi: 10.1073/pnas.122653799.

[11] GOLDBERG D. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading, MA, 1989, doi: 10.1609/aimag.v12i1.889.

[12] GONG M., FU B., JIAO L., DU H. Memetic algorithm for community detection in networks. *Phys. Rev. E – Stat. Nonlinear, Soft Matter Phys.* 2011, 84(5), pp. 1–9, doi: 10.1103/PhysRevE.84.056101.

[13] HAFEZ A.I., GHALI N.I., HASSANIEN A.E., FAHMY A.A. Genetic Algorithms for community detection in social networks. *Intell. Syst. Des. Appl. (ISDA).* 2012 , pp. 460–465, doi: 10.1109/ISDA.2012.6416582.

[14] HOLAND J.H. *Adaptation in Natural and Artificial Systems.* Univ. of Michigan Press,Ann Arbor, MI, 1975.

[15] KREBS V. http://www.orgnet.com [Unpublished].

[16] KRISHNAMURTHY B., WANG J. On network-aware clustering of Web clients. *ACM SIGCOMM Comput. Commun. Rev.* 2000, 30(4), pp. 97–110, doi: 10.1145/347059.347412.

[17] LANCICHINETTI A., FORTUNATO S., KERTÉSZ J. Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.* 2009, (11), doi: 10.1088/1367-2630/11/3/033015.

[18] LI Y., LIU G. A genetic algorithm for community detection in complex networks. *J. Cent. South Univ.* 2013, 20(5), pp. 1269–1276, doi: 10.1007/s11771-013-1611-y.

[19] LI Z., ZHANG S., WANG R.S., ZHANG X.S., CHEN L. Quantitative function for community detection. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* 2008, 77(3), pp. 1–9, doi: 10.1103/PhysRevE.77.036109.

[20] LIU X., LI D., WANG S., TAO Z. Effective Algorithm for Detecting Community Structure in Complex Networks. *Comput. Sci.* 2007, pp. 657–664, doi: 10.1007/978-3-540-72586-2_95.

[21] LUSSEAU D., SCHNEIDER K., BOISSEAU O.J., HAASE P., SLOOTEN E., DAWSON S.M. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: Can geographic isolation explain this unique trait? *Behav. Ecol. Sociobiol.* 2003, 54(4), pp. 396–405, doi: 10.1007/s00265-003-0651-y.

[22] MAZUR P., ZMARZŁOWSKI K., ORŁOWSKI A.J. Genetic Algorithms Approach to Community Detection. *Symp. A Q. J. Mod. Foreign Lit.* 2010, 117(4), pp. 703–705, doi: 10.12693/APhysPolA.117.703.

[23] MCPHERSON M., SMITH-LOVIN L., COOK J.M., Birds Of A Feather?: Homophily in Social Networks, *Annual Review of Sociology.* 2001, 27, pp. 415–444. doi: 10.1146/annurev.soc.27.1.415.

[24] MISLOVE A., MARCON M., GUMMADI K.P., DRUSCHEL P., BHATTACHARJEE B. Measurement and analysis of online social networks. *Proc. 7th ACM SIGCOMM Conf. Internet Meas.* 2007, pp. 29–42.

[25] NEWMAN M.E.J. Detecting community structure in networks. *Eur. Phys. J. B.* 2004, 38(2), pp. 321–330, doi: 10.1140/epjb/e2004-00124-y.

[26] NEWMAN M.E.J. Fast algorithm for detecting community structure in networks. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* 2004, 69(62), pp. 1–5, doi: 10.1103/PhysRevE.69.066133.

[27] NEWMAN M.E.J. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.* 2006, 103(23), pp. 8577–82, doi: 10.1073/pnas.0601602103.

[28] NEWMAN M.E.J. The structure and function of complex networks. *SIAM Rev.* 2003, 45(2), pp. 167–256, doi: 10.1137/S003614450342480.

[29] PIZZUTI C. A multiobjective genetic algorithm to find communities in complex networks. *IEEE Trans. E Comput.* 2012, 16(3), pp. 418–430, doi: 10.1109/TEVC.2011.2161090.

[30] PIZZUTI C. GA-Net: A Genetic Algorithm for Community Detection in Social Networks. *Parallel Problem Solving from Nature – PPSN X*. 2008, pp. 1081–1090, doi: 10.1007/978-3-540-87700-4_107.

[31] RADICCHI F., CASTELLANO C., CECCONI F., LORETO V., PARISI D. Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. U. S. A.* 2004, 101(9), pp. 2658–2663, doi: 10.1073/pnas.0400054101.

[32] REDDY P., KITSUREGAWA M. A graph based approach to extract a neighborhood customer community for collaborative filtering. *Databases Networked Inf. Syst.* 2002, pp. 188–200, doi: 10.1007/3-540-36233-9_15.

[33] SHANG R., BAI J., JIAO L., JIN C. Community detection based on modularity and an improved genetic algorithm. *Phys. A Stat. Mech. its Appl.* 2013, 392(5), pp. 1215–1231, doi: 10.1016/j.physa.2012.11.003.

[34] SHI C., YAN Z., WANG Y., CAI Y., WU B. A Genetic Algorithm for Detecting Communities in Large-Scale Complex Networks. *Adv. Complex Syst.* 2010, 13(01), pp. 3–17, doi: 10.1142/S0219525910002463.

[35] STROGATZ S.H. Exploring complex networks. *Nature*. 2001, pp. 268–276, doi: 10.1038/35065725.

[36] TASGIN M., BINGOL H. Community Detection in Complex Networks using Genetic Algorithm, 2006, pp. 1–6.

[37] WASSERMAN S., FAUST K. *Social Network Analysis: Methods and Applications*. 1994, p. 825, doi: 10.1017/CBO9780511815478.

[38] ZACHARY W.W. An Information Flow Model for Conflict and Fission in Small Groups. *J. Anthropol. Res.* 1977, 33(4), pp. 452–473.

[39] ZHANG Z., WANG K. A Formal Analytic Approach to Credible Potential Path and Mining Algorithms for Multimedia Social Networks. *Comput. J.* 2015, 58(4), pp. 668–678, doi: 10.1093/comjnl/bxu035.

[40] ZHANG Z., WANG K. A trust model for Multimedia Social Networks. *Soc. Network Anal. Min.* 2013, 3(4), pp. 969–979, doi: 10.1007/s13278-012-0078-4.