

Optimizing Modularity for Community Identification Using a Genetic Algorithm

Harithashree K. S.*  and Prerna Sidana 

Glocal School of Technology and Computer Science, Glocal University, Saharanpur, Uttar Pradesh, India

Corresponding author: **Harithashree K. S.** | E-mail: ksharitha.mbl@gmail.com

Citation: Harithashree K. S. and Prerna Sidana (2025). Optimizing Modularity for Community Identification Using a Genetic Algorithm. *Discover Engineering: An International Journal*. DOI: <https://doi.org/10.51470/DE.2025.6.1.24>

Received 17 January 2025 | Revised 18 February 2025 | Accepted 19 March 2025 | Available Online 20 April 2025

Abstract

One popular metric for gauging the robustness of a community's structure in community discovery is the modularity function. Using a genetic algorithm and a random walk model, we present a new way to encode network community structures and an approach to community detection that maximizes modularity. To begin, the initial adjacency matrix was weighted using the results of applying the random walk model to the calculation of node similarity. A weighted network link topology was subsequently encoded into a chromosome in accordance with the closest neighbor-based similarity representation approach. By adjusting the structure of a set number of chromosomes and then computing their modularity, the genetic algorithm was able to find the optimal modularity value, which is related to the number of communities and their structure. On several actual social networks, we put our technique to the test. The innovative approach achieved higher modularity value in comparison to multiple state-of-the-art methods. The results obtained using the suggested strategy are more realistic because it does not necessitate a predetermined number of communities when community partitioning begins. In this case, the ideal community layout and number of communities are calculated automatically.

Keywords: Community structure; Genetic Algorithm; random walk model.

1 INTRODUCTION

A network connection diagram can show real-life links between items or individuals. These networks are complicated. Examples include biological networks, the Web, and social collaborative networks. Complex networks are everywhere. Community structure discovery, which divides large networks into modules based on internal links, is crucial to analysis. Researchers from several domains study community detection. Due to the demand for social network analysis, many community structure detection algorithms have been created. Girvan and Newman introduced edge betweenness-based community finding [1]. Newman proposed an assessment function for modularity to assess fragmented communities' efficacy [2]. Divide a network into a set number of communities to determine its modularity. A higher modularity number indicates a better split. [5], values between 0.3 and 0.7 indicate a highly modular network. [3] report that modularity became the partition standard in community detection. The novel community finding technique SKSC uses modularity Q value as the division evaluation function and is based on kernel spectral clustering. It debuted recently. Splitting a network around the most modular node is best. In recent years, community identification investigations concentrating on innovative modularity approaches have gained popularity. [4] proposed a convex technique to community finding modularity optimization.

This optimization issue appears to have no optimum solution.

Network partitions present a new optimization difficulty. Several external optimization methods establish modularity values. The genetic algorithm, an evolutionary algorithm that is simple and effective, was the most common community discovery method. This work introduces GAcut, a unique community discovery method that uses the random walk model and a genetic algorithm to optimize modularity Q for community partitioning. Starting with the random walk model, we calculate node transition probabilities. To make every node's nearest neighbor unique, a network's adjacency matrix is weighted. We present a weighted network community structure expression. One recommended encoding genes using weighted similarity based on neighbors' positions. One can modify chromosomal shape with the genetic algorithm. We categorize networks by maximizing modularity. It doesn't require a community number like other methods. GAcut partitioning surpassed other cutting-edge community identification and modularity optimization approaches in modularity value on real-world networks with the ground truth. To evaluate the strategy, we tested it on additional large actual networks with unknown community patterns. Researchers found that GAcut can independently recognize community structure.

2. Methodology

Using a mix of modularity maximization and similarity metrics based on random walk distance, this study suggests a way for detecting communities in social networks. The primary objective is to optimize the modularity score Q with a genetic algorithm in order to improve the efficiency and accuracy of community detection. Important parts of the process include a genetic algorithm, a representation of the network, and a representation of the chromosomes, a random walk distance, and a comparison to previous approaches.

Network Representation

Community detection originates from graph partitioning techniques, where a network is represented by a graph

$G = (V, E)$, where:

- V is the set of nodes (vertices).
- E is the set of edges connecting nodes in V .

The network is typically represented using a symmetric adjacency matrix $A = [n \times n]$, where:

- $n = |V|$ is the number of nodes in the network.
- If a connection exists between nodes v_i and v_j ,
- then $A_{ij} = A_{ji} = 1$; otherwise, it is 0

The goal of community detection is to divide the set of nodes V into k communities = $\{C_1, C_2, \dots, C_k\}$, where intra-community connections are maximized, and inter-community connections are minimized. Within community detection, the objective is to partition the set of nodes V into k communities, denoted as $\{C_1, C_2, \dots, C_k\}$, with the objective of maximizing intra-community connections and minimizing inter-community connections.

Modularity Maximization

The goal is to maximize Q , which reflects the best division of communities. In terms of modularity, Q is a popular way to measure how well a community is structured. Here is the definition:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$$

The adjacency matrix is denoted by A . The degrees of the nodes i and j , are represented by k_i and k_j , respectively. The sum of all edges is represented by m . The Kronecker delta function, denoted as $\delta(C_i, C_j)$, is 1 when nodes i and j are members of the same community and 0 otherwise. Maximizing Q , which represents the optimal division of communities, is the objective.

Random Walk Distance for Similarity

To enhance the community detection process, a random walk distance is used to convert the adjacency matrix into a weighted matrix. Given the adjacency matrix A , the transition probability from node i to node j is defined as:

$$p_{ij} = \frac{A_{ij}}{D_i}$$

Where:

- D is the degree matrix $D = \text{diag}(d_1, d_2, \dots, d_n)$.

The t -step transition matrix P^t is computed as:

$$P^t = P^{t-1} \times P$$

The random walk distance between nodes i and j is:

$$d(i, j) = \sum_{k=1}^t p_{ij}^k$$

Using this distance metric, the similarity between nodes i and j is:

$$s_{ij} = \sqrt{\frac{\sum_{k \neq i, j}^n [d(i, k) - d(j, k)]^2}{n - 2}}$$

This similarity matrix S is symmetric, and the smaller the value of s_{ij} , the more similar the nodes.

Chromosome Representation

A chromosome represents a possible community structure. For each node i , we identify its nearest instance, if node 1's nearest neighbor is node 3, the chromosome representation will show a link between them.

Figure 1 illustrates the chromosome encoding process:

- (A) shows the original network.
- (B) shows the nearest neighbor-based similarity matrix and chromosome encoding.
- (C) depicts the reduced network structure after chromosome decoding, which clearly divides the nodes into communities.

Genetic Algorithm for Modularity Maximization

The genetic algorithm (GA) is used to optimize the community structure by maximizing the modularity Q . The steps are as follows:

- Initialization: Generate m initial chromosomes, where each chromosome represents a potential community division.
- Evaluation: Calculate the modularity Q for each chromosome using the modularity function.
- Selection: Choose two chromosomes with the highest modularity as parents.
- Crossover: Perform uniform crossover by randomly selecting gene positions in both parent chromosomes and swapping their values to form new chromosomes.
- Mutation: Introduce random changes in the chromosomes by selecting random positions and altering community assignments to avoid premature convergence.

Iteration: Repeat the crossover and mutation steps to generate a new population, and continue until the maximum modularity Q is obtained

Comparison with Existing Methods

The proposed method is compared with Ga-Net, a community detection method based on genetic algorithms. While Ga-Net uses a locus-based adjacency representation and focuses on binary networks, the proposed method utilizes the random walk model to transform the adjacency matrix into a weighted matrix, allowing for more accurate detection of community structures in weighted networks. Additionally, the proposed method does not require the number of communities to be pre-specified, unlike Ga-Net, and provides a more flexible and scalable approach.

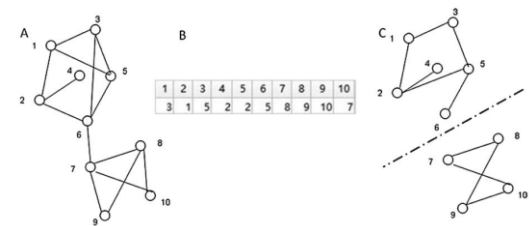


Figure 1: shows the process of chromosome encoding and decoding. (A) A diagram depicting the initial network connections. The first line represents the chromosome's position, while the second line represents its genotype; this representation is based on nearest neighbor similarity and has ten nodes. (C) It is possible to encode the chromosome and divide it into two modules.

3 Experimental Results

Table 1.1. Top outcomes for modularity and the number of communities for various algorithms

Network	Algorithm	Q_max	Q_avg	NC (Number of Communities)
Karate	GAcut	0.4198	0.4189	4
	Fast_mo	0.4198	0.4158	4
	Fast_newman	0.3974	0.3974	4
	NMF	0.3990	0.3280	2
	Qcut	0.4198	0.4198	4
Dolphin	HQcut	0.4198	0.4198	4
	GAcut	0.5277	0.5269	5
	Fast_mo	0.5273	0.5193	5
Football	GAcut	0.6046	0.6033	10
	Fast_mo	0.6046	0.6038	10
Erdős	GAcut	0.4541	N/A	31
Email	GAcut	0.4541	N/A	41
Yeast	GAcut	0.4725	N/A	132

The proposed GAcut algorithm was tested on several standard network datasets and compared with related algorithms in community detection literature, such as Fast_mo, Fast_newman, NMF, Qcut, and HQcut. The datasets included social networks like the Karate club, Dolphin, and Football, as well as larger networks like Erdős, Email, and Yeast.

- GAcut outperformed other algorithms across most datasets in terms of modularity (Q_max) and the number of communities (NC).
- For instance, in the Karate club network, GAcut achieved the highest modularity value (Q_max = 0.4198), detecting 4 communities, which was also the highest among all compared algorithms.
- In the Dolphin network, GAcut also performed best, with a Q_max of 0.5277, surpassing other algorithms.
- For larger networks such as Erdős, Email, and Yeast, GAcut yielded higher modularity values than other algorithms like Ga-net and SKSC, which indicates better community detection and a stronger network modularity structure.

Despite the impressive performance, the resolution limit issue remains in larger networks, where the detection of smaller communities is challenging. GAcut's higher modularity indicates that it detects communities more effectively, but further studies are needed to address resolution limits in large networks.

4. Conclusion

We present a partitioning algorithm that determines the logical division of a community's structure automatically by applying novel methods of community structure encoding and decoding, which are based on genetic development of modularity maximization. Community detection results are improved when the random walk model is used in conjunction with it. This approach is proven to be an unsupervised community partitioning method since it does not necessitate the specification of a community number during the partitioning operation. Also, the optimal number of communities and best modularity value are both found automatically by the innovative method. Thus, for big real-world networks lacking a priori established architecture, the innovative technique is highly applicable.

References

1. Dhawan, S., Singh, K., & Batra, A. (2021). Defining and evaluating network communities based on ground-truth in online social networks. In *Recent Innovations in Computing: Proceedings of ICRIC 2020* (pp. 151-163). Springer Singapore.
2. Rostami, M., Berahmand, K., & Forouzandeh, S. (2021). A novel community detection based genetic algorithm for feature selection. *Journal of Big Data*, 8(1), 2.
3. You, X., Ma, Y., & Liu, Z. (2020). A three-stage algorithm on community detection in social networks. *Knowledge-Based Systems*, 187, 104822.
4. Su, Y., Zhou, K., Zhang, X., Cheng, R., & Zheng, C. (2021). A parallel multi-objective evolutionary algorithm for community detection in large-scale complex networks. *Information Sciences*, 576, 374-392.
5. Tsung, C. K., Ho, H. J., Chen, C. Y., Chang, T. W., & Lee, S. L. (2020). Detecting overlapping communities in modularity optimization by reweighting vertices. *Entropy*, 22(8), 819.