Journal of Theoretical and Applied Information Technology
15th June 2024. Vol.102. No. 11

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

A NEW ALGORITHM FOR COMMUNITY DETECTION IN COMPLEX SOCIAL NETWORKS

HICHAM SADIKI[1], **MEROUANE ERTEL**[2], **AZEDDINE SADQUI**[3], **SAID AMALI**[4]

[1,2,3]Informatics and Applications Laboratory (IA), Faculty of Sciences, Moulay Ismail university, Morocco
[4]Informatics and Applications Laboratory (IA), FSJES, Moulay Ismail university, Morocco
E-mail: [1] h.sadiki@umi.ac.ma, [2] m.ertel@edu.umi.ac.ma , [3] a.sadqui@umi.ac.ma , [4] s_amali@yahoo.com

ABSTRACT

This article introduces a refined approach to identifying community structures in complex social networks. With a focus on accuracy and efficiency, our algorithm takes into account the complex nature of social networks by enhancing traditional methodologies to accurately capture community patterns. Central to our approach is the "community score," a pivotal metric gauging community partition quality. We've tailored variation operators, including a new crossover operator, to strengthen this foundation, improving both convergence and precision. A notable innovation is the dynamic determination of community count. Unlike fixed assumptions, our approach adapts the count based on network structure, adeptly detecting communities of diverse sizes and shapes. Moreover, we highlight border nodes' significance as community connectors. Weighted interactions involving these nodes improve community transition detection, refining partitions and spotlighting boundary-critical nodes. Through extensive experiments on synthetic and real-world datasets, the superiority of our algorithm over conventional methods becomes evident. Improved modularity and precision metrics validate our approach's efficacy.

Keywords: *Genetic Algorithm, Community Detection, Social Network Analysis, Graph Partitioning, Clustering, Complex Networks, Community Structure, Genetic Operators.*

1. INTRODUCTION

In recent years, the investigation of community structure detection in intricate networks has garnered significant attention due to its relevance in numerous real-world scenarios. Networks, ranging from social networks to biological systems, exhibit a fascinating property known as community structure (*the division of the network into densely interconnected clusters with sparse connections between them*) [1]. This property offers insights into the way individuals and entities interact within the network. Various methods have been proposed to tackle this challenging problem, aiming to accurately identify these community structures [2][3][4].

This paper presents a novel genetic algorithm-based approach, referred to as **GASNET**, designed to efficiently detect community structures within social networks. Inspired by the challenges posed by the intricate nature of these networks, our approach optimizes a fitness function specifically designed to uncover densely connected groups of nodes while considering sparse inter-group connections. Our method builds upon the foundation of genetic algorithms, a powerful optimization technique [5].

Intriguingly, the distinctiveness of our algorithm lies in its ability to adaptively determine the number of communities within a network, eliminating the need for a predefined count. This adaptability is achieved through the progressive refinement of a "community score", a global quality measure of the network partitioning. Moreover, we introduce specialized variation operators to improve convergence by focusing on the inherent correlations among nodes.

This paper unfolds as follows: Section 2 provides an overview of existing community detection algorithms. In section 3 we present the necessary background to formalize the problem and introduces the "community score". Section 4 delves into the representation and variation operators employed in our approach. The last section offers insights into the experimental results of our approach on both synthetic and real-world datasets, highlighting its prowess in accurately detecting network structures.

2. RELATED WORK

The exploration of community detection in social networks has fostered a plethora of methodologies, encompassing both traditional and modern computational techniques. In this section, we provide an overview of significant contributions in the realm of community detection, with an emphasis on genetic algorithm-based approaches.

2.1. Traditional Approaches

Early methods for community detection primarily revolved around graph partitioning algorithms. Newman and Girvan introduced the concept of modularity, a measure of the quality of network partitions based on the density of edges within communities compared to the expected density [1][6]. Their Newman-Girvan algorithm was among the first to formalize this notion and has been widely employed in early community detection studies. The algorithm operates by iteratively removing edges with high betweenness centrality, gradually revealing the community structure of the network. However, these methods face challenges in terms of scalability and efficiency when applied to large networks due to their computational complexity.

2.2. Evolutionary and Genetic Algorithm-based Approaches

Genetic algorithms (GAs) have emerged as potent tools inspired by natural evolution processes for solving optimization problems. These algorithms maintain a population of potential solutions and iteratively evolve them through selection, crossover, and mutation operations. In the context of community detection, GAs have been harnessed to efficiently uncover latent structures in social networks.

Mazur et al. proposed a GA-based approach for community detection that optimizes a fitness function considering both intra-community density and inter-community connectivity [7]. Each individual in the GA population corresponds to a network partition, and through successive generations, the algorithm refines these partitions to enhance the accuracy of community detection.

Drawing inspiration from genetic evolution, G. Bello-Orgaz et al. introduced a Multi-Objective Genetic Algorithm (MOGA) for community detection [8]. By concurrently considering multiple objectives such as modularity, conductance, and community size, the MOGA algorithm offers a holistic approach to identifying diverse community structures that exhibit varying levels of internal cohesion and external separation.

Y.-C. Chiu et al. proposed a hybrid methodology that combines a genetic algorithm with a simulated annealing process. This hybrid approach aims to accelerate convergence while maintaining solution quality. By capitalizing on the complementary strengths of genetic algorithms and simulated annealing, this method addresses the trade-off between exploration and exploitation in the optimization process [9].

Haritha Akkineni et al. focus on the importance of community detection in social networks, especially in online social networks. It discusses the limitations of traditional algorithms and structures used for community detection and proposes a method that overcomes these drawbacks by identifying communities on social networking sites. The proposed method involves the use of the DBSCAN algorithm to detect outliers and improve the quality of detected communities[10].

2.3. Recent Advances

Continual advancements in genetic algorithm-based community detection have aimed to overcome the shortcomings of traditional techniques. E. Akachar et al. proposed an algorithm named "ACSIMCD" to overcome the problem of detecting communities by focusing only on the modified parts of the network and taking into account previously known information. The algorithm updates the community structure locally, rather than recalculating it from scratch at each snapshot. This approach is more efficient and scalable, allowing community structure to be detected and updated in real time [11].

Furthermore, the landscape of genetic algorithm-based community detection has witnessed the emergence of hybrid algorithms that combine genetic algorithms with other optimization techniques. For instance, particle swarm optimization (PSO) and ant colony optimization (ACO) have been synergistically integrated with genetic algorithms to enhance community detection accuracy and computational efficiency [12].

In summation, the array of genetic algorithm-based methodologies for community detection underscores their versatility and potential to uncover intricate structures within social networks.

3. PROBLEM FORMULATION

The fundamental problem of community detection in social networks revolves around the complexity of identifying inherently coherent groupings within complex network structures. A social network SN can be modeled as a graph G = (V, E), where V is a set of objects, referred to as nodes or vertices, and E is a set of links, known as edges, connecting two elements from V (see figure 3.1). A community (or cluster) within a network constitutes a group of vertices exhibiting a high density of edges within the group, and a lower edge density between groups (see figure 3.2). The challenge lies in detecting k communities within a network, where k is unknown, by partitioning nodes into k subsets that are highly intra-connected and sparsely inter-connected.

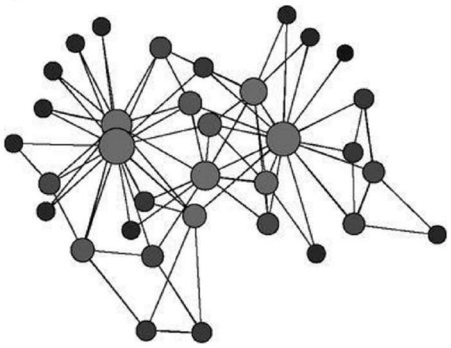


Figure 3.1: Friendship network between members of a club. This social network from a study conducted in the 1970s shows the pattern of friendships between the members of a karate club at an American university. The data were collected and published by Zachary [13].

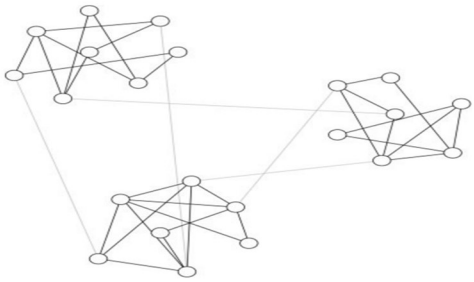


Figure 3.2: A schematic representation of a network with community structure. In this network there are three communities of densely connected vertices, with a much lower density of connections between them [14].

In the context of graphs, the adjacency matrix is frequently employed to address this challenge. For a network with N nodes, the graph can be represented by the N × N adjacency matrix A, where the entry at position (i, j) is 1 if there exists an edge between node i and node j, and 0 otherwise. The community detection problem can then be reformulated as the task of finding a partition of A into k sub-matrices that maximize the sum of the sub-matrices' densities. A naive density measure for an N x N sub-matrix is the count of ones (i.e., interactions) it contains. However, this interaction count fails to provide insights into interconnections among nodes.

To address this limitation, a density measure based on volume and row/column means was introduced in [3] and applied to identify co-clusters within sparse binary matrices. Co-clustering, also known as bi-clustering [15], diverges from clustering by simultaneously grouping both object and attribute dimensions in a dataset. The identification of sub-matrices can be seen as a specific instance of co-clustering where both dimensions represent the same concept, i.e., the nodes of the graph.

To formulate this problem, we will follow the following steps:

- ✓ Let $S = (I, J)$ be a sub-matrix of A, where I is a subset of rows $X = \{I_1, \ldots, I_N\}$ of A, and J is a subset of columns $Y = \{J_1, \ldots, J_N\}$ of A.
- ✓ Let a_{iJ} denote the mean value of the ith row of S:

$$a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij}$$

- ✓ Let a_{Ij} denote the mean value of the jth column of S:

$$a_{Ij} = \frac{1}{|I|} \sum_{i \in I} a_{ij}$$

- ✓ The volume v_S of sum-matrix $S = (I, J)$ is the number of 1 entries a_{ij} :

$$v_S = \sum_{i \in I, j \in J} a_{ij}$$

- ✓ The power mean of $S$of order r, denoted as $M(S)$ is defined as:

$$M(S) = \frac{\sum_{i \in I} (a_{iJ})^r}{|I|}$$

A metric relying on the volume and the mean of rows and columns, enabling the identification of dense and maximal sub-matrices, can be formulated as described as follows:

- ✓ Consider a sub-matrix denoted as $S = (I, J)$.

- ✓ Let $M(S)$ the power mean of S with an order denoted by r.
- ✓ The score of S is defined as $Q(S) = M(S) \times v_S$.
- ✓ The community score of a partition $\{S_1, \ldots, S_k\}$ of matrix A is then defined as:

$$CS = \sum_{i}^{k} Q(S_i)$$

The community identification task can be defined as the community score CS maximization goal. It is important to point out that higher values of the exponent r tend to shift the CS towards matrices with fewer null elements. This phenomenon stems from the fact that higher r values amplify the influence of densely interconnected nodes while decreasing the contribution of less connected nodes when calculating the community score. In the experiments section, we demonstrate that when the modular structure of the network is not well defined, opting for higher values of r helps to detect communities efficiently.

4. GASNET ALGORITHM

In this section, we delve into a comprehensive exploration of our algorithm, which stands as the cornerstone of our approach for community detection. We present a detailed description of our algorithm, outlining its essential components, the adopted representation method for partitioning the network, and the specific genetic operators that facilitate the evolutionary process.

4.1. Algorithm Overview

Taking into account the principle of an evolutionary algorithm (EA) can be succinctly described (Figure. 4.1). When considering a combinatorial optimization problem, a population comprises a set of points in the solution space, each of these points being referred to as an individual. Each individual possesses a distinct genetic makeup that characterizes and differentiates it from other individuals; these genes essentially constitute the elementary blocks that define a solution. Typically, an individual can be represented as a list of integers for combinatorial problems, a vector of real numbers for numerical problems in continuous spaces, or a sequence of binary numbers for Boolean problems. If necessary, these representations can be combined within complex structures. Our GASNET algorithm also encapsulates a strategic genetic algorithm approach tailored for community detection in complex networks and explores the vast solution space through successive generations, with the primary objective of discovering meaningful community structures inherent to the network.

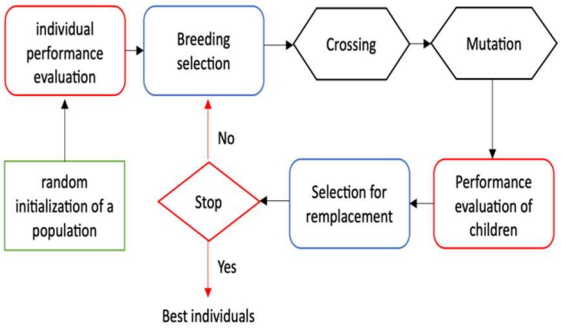


Figure 4.1: Principle of a standard evolutionary algorithm [16]

4.2. Genetic Representation

The approach employed by our clustering algorithm involves utilizing the locus-based adjacency representation, which was adopted by [17] for multiobjective clustering tasks. Within this graph-based framework, an individual within the population is composed of N genes denoted as $\{g_1, \ldots, g_N\}$, with each gene capable of adopting allele values j from the range $\{1, \ldots, N\}$. These genes and alleles serve as representations of nodes within the graph $G = (V, E)$, which effectively models the social network SN. Notably, assigning a value j to the ith gene implies a connection between nodes i and j within V. Consequently, in the derived clustering solution, nodes i and j are grouped together in the same cluster.

However, an additional step, known as decoding, is essential to identify all the constituent components of the corresponding graph. Nodes that participate within the same component are assigned to a single cluster. As highlighted in [17], this decoding process can be executed in linear time. A noteworthy advantage of this representation lies in its automatic determination of the number of clusters k. This determination is rooted in the quantity of components encapsulated within an individual, a process facilitated by the decoding step.

To illustrate this concept, consider a network as depicted in Figure 4.2, featuring eleven nodes. This network can be effectively divided into three distinct groups, each differentiated by varying colors and node shapes. Among the numerous possible

genotypes, the configuration shown in Fig 4.3, representing the optimal solution, translates into the graph structure visualized in Fig 4.4. Each connected component serves to unite nodes that correspond to the partitioning observed in Figure 4.2.

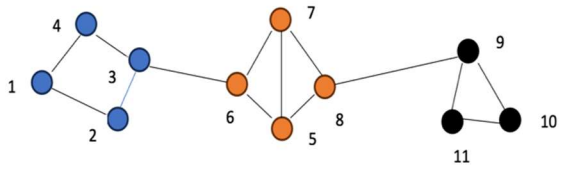


Figure 4.2: A network represented as a graph

Position	1	2	3	4	5	6	7	8	9	10	11	12
Genotype	1	2	1	5	6	3	8	10	6	11	8	11

Figure 4.3: the genotype's locus-based representation

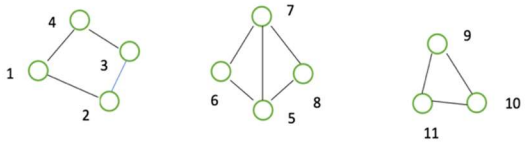


Figure 4.4: the genotype's structure represented in a graph-based format

4.3. Objective Function

As previously explained, the decoding of an individual yields a varying number of components $\{S_{1,}, \ldots, S_k\}$ into which the graph is partitioned. Our aim is to identify a partitioning that maximizes the community score, as discussed in the preceding section. This approach ensures the formation of communities with strong intra-connections and limited inter-connections. Consequently, the objective function is defined as $CS = \sum_i^k Q(S_i)$, where k signifies the number of components in the partitioning and S_i represents each individual component.

4.4. Initialization

Our initialization process takes into consideration the actual connections among nodes within the social network. A purely random generation of individuals could result in components that are disjointed in the original graph. For instance, a randomly generated individual might assign an allele value j to the ith position, yet there may be no connection between nodes i and j, meaning that the edge (i, j) is absent. In such instances, grouping nodes i and j into the same cluster would be an inaccurate choice.

To address this limitation, we implement a repair mechanism following the generation of an individual. This repair process involves a check to verify the existence of a valid link between a gene at position i and the allele value j. If the edge (i, j) exists, the value j is retained. However, if the edge is absent, j is replaced with one of the neighbors of i. This guided initialization approach biases the algorithm towards decomposing the network into interconnected groups of nodes. We refer to an individual that generates this type of partitioning as "safe" because it prevents the creation of uninteresting divisions that involve disconnected nodes. The inclusion of safe individuals enhances the convergence of the method, as it constrains the space of potential solutions.

4.5. Uniform Crossover

Crossing, or crossover, aims to enrich the population diversity by manipulating the structure of chromosomes. Typically, crossovers involve two parents and generate two child individuals. To develop a crossover operator, three steps are required:

- ✓ Selection of two surviving chromosomes, which are chosen through the reproduction procedure.
- ✓ Random selection of a location to cut these two chromosomes.
- ✓ Joining the chromosome segments back together by crossing them. As a result, the initial two chromosomes have exchanged segments of genetic code.

During this operation, two chromosomes exchange parts of their strings to create new chromosomes. These recombinations can be simple or multiple.
In the first case (Figure 4.5), the two chromosomes intersect and exchange gene portions at a single point.

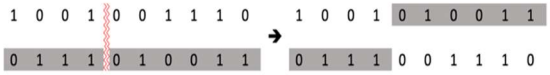


Figure 4.5: One-point Crossover

In the second case (Figure 4.6), there are multiple crossover points (2 or 3 may suffice), known as multi-point crossover. This operation is more prevalent.

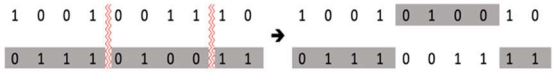


Figure 4.6: Multi-point crossover

When multiple locations are selected, the bits of the string are compared between the two parents. The bits are exchanged with a fixed probability, usually 0.5. This type of crossover is called "uniform crossover"(Fig 4.7).

Figure 4.7: Uniform Crossover

We have chosen to employ uniform crossover due to its ability to ensure the preservation of effective connections among nodes within the social network in the resultant child individual. This is particularly relevant given the biased initialization approach. Each individual within the population possesses a "safe" attribute, signifying that if a gene i contains value j, then the corresponding edge (i, j) exists.

Consequently, when two safe parents are available, a random binary vector is generated. During uniform crossover, genes are selected from the first parent where the vector bears a value of 1, while genes from the second parent are chosen where the vector is 0. These selected genes are then combined to form the child individual. In this process, each position i of the child contains a value j derived from one of the two parents. Thus, the edge (i, j) is assured to exist. This implies that a safe child is produced from two safe parents.

4.6. Mutation

The mutation operator, which randomly alters the value j of an i-th gene, leads to futile exploration of the search space due to the aforementioned observations regarding node connections. Consequently, the potential allele values are constrained to the neighbors of gene i. This remedied mutation process ensures the generation of a mutated child that is secure, wherein each node is exclusively linked to one of its neighbors.

Initiated with a randomly initialized population that has undergone repairs to ensure safety, GASNET commences its operation within the context of a network denoted as SN and represented by the graph G. Each individual generates a graph structure, wherein every component represents a connected subgraph of G. Over a fixed number of generations, the genetic algorithm computes the fitness function for each solution member and employs specialized variation operators to yield the updated population.

In the experimental results section, we demonstrate that the fitness function effectively steers the genetic algorithm towards successfully discerning the optimal partitioning of SN, ultimately converging to a solution within a limited number of iterations. Prior to presenting the experimental findings, the subsequent section provides an overview of the primary approaches to community detection.

5. EXPERIMENTAL RESULTS AND DISCUSSION

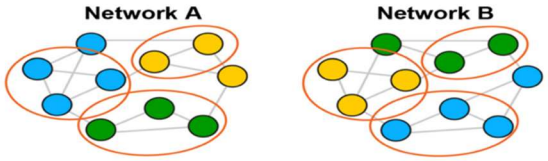
In this section, we assess the efficacy of our approach using synthetic data. Additionally, we conduct a comparative analysis between the outcomes yielded by GASNET and those documented by Girvan and Newman in [14] concerning real-world networks with established community partitions. In both scenarios, we demonstrate the proficiency of our genetic algorithm in accurately identifying network structures while maintaining competitive performance against Girvan and Newman's methods.

To evaluate our approach's proficiency in effectively identifying the community structure of a network, we employ the benchmark introduced by Girvan and Newman in [14]. This network comprises 128 nodes partitioned into four communities, each consisting of 32 nodes. Random edges are established between pairs of vertices, while ensuring that $Z_{in} + Z_{out} = 16$, where Z_{in} and Z_{out} represent the internal and external degree of a node concerning its community. When $Z_{in} > Z_{out}$, nodes within a community have more connections to their fellow community members than to nodes from other communities, making it essential for a robust algorithm to uncover these relationships. We generated 50 distinct networks across a range of Z_{out}values from 0 to 8. To assess the similarity between the actual partitions and the detected ones, we utilized the Normalized Mutual Information (NMI) metric. The reliability of NMI has been demonstrated in [18]. Given two partitions A and B of a network into communities, the confusion matrix C is constructed (Fig 5.1), where each element C_{ij} denotes the count of nodes from community i in partition A that also belong to community j in

partition B. The normalized mutual information is defined as follows:

$$NMI(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij}\log(\frac{N_{ij}N}{N_{i.}N_{.j}})}{\sum_{i=1}^{C_A} N_{i.}\log(\frac{N_{i.}}{N}) + \sum_{j=1}^{C_B} N_{.j}\log(\frac{N_{.j}}{N})}$$

Where A and B represent the respective community structures of these graphs. C_A corresponds to the number of communities in partition A, while C_B denotes the number of communities in partition B. N stands for the total number of nodes, which remains consistent across both community structures.



Confusion matrix	C1	C2	C3	Total
C1	0	3	1	4
C2	1	0	2	3
C3	3	0	0	3
Total	4	3	3	10

Figure 5.1: An illustration of calculation of Normalized Mutual Information (NMI) quantifying similarity between two community structures. Colors of the nodes represent assigned community and red circles indicate overlaps in communities. Confusion matrix is used to measure the overlap between the two community structures. In the above figure, Community C1 in Network A corresponds to Community C2 in Network B. The NMI is calculated with the confusion matrix, and gets high when the overlap between the communities is high [19]

The variable N_{ij} signifies the overlap between the i-th community in partition A and the j-th community in partition B. This overlap quantifies the number of nodes shared between these communities. $N_{i.}$ represents the total count of nodes in the i-th community of partition A, while $N_{.j}$ denotes the total count of nodes in the j-th community of partition B. It's important to note that the formula accounts for the case where an overlap leads to $0 \times \log(0)$, which results in 0. This calculation adheres to standard conventions. The resulting value obtained from this calculation is known as the Normalized Mutual Information (NMI), which ranges between 0 and 1. A value of 0 indicates complete independence between the community structures, while a value of 1 indicates their complete identity.

Before delving into the analysis and discussion of the experimental outcomes, let's begin by outlining the algorithms, datasets, and performance metrics that were employed for comparison with our approach.

In our investigation, we undertake a comparison of GASNET against seven diverse algorithms encompassing various approaches, including static methodologies. Our selection includes two static techniques: the initial one (Louvain) operates solely on structural information, whereas the second one (I-Louvain) incorporates both content and structure data. Additionally, we have incorporated five dynamic algorithms, among which (NEIWalk) is tailored for content-based networks. A brief outline of these methods is provided in the subsequent subsection.

- ✓ The DynaMo (Dynamic Multilayer Optimizer) method is a dynamic community detection algorithm that operates on multilayer networks. It focuses on optimizing a quality function by considering both the structure and content information of the network as it evolves over time [20].
- ✓ The FacetNet method is a dynamic algorithm designed for community detection in content-based networks. It takes into account both the network structure and node attributes to identify meaningful communities [21].
- ✓ The I-Louvain method is an enhanced version of the Louvain algorithm that combines both content and structural information for community detection [22]. It leverages both node attributes and network structure to improve the accuracy of community identification.
- ✓ The Louvain method is a community detection algorithm that focuses on maximizing the modularity of a network [4]. It operates in two phases: a greedy optimization phase, where nodes are iteratively moved to maximize modularity, and an agglomeration phase, where communities are treated as nodes to further refine the structure. While efficient for large networks, it can be sensitive to initial conditions and relies solely on network structure for detection.
- ✓ The MIEN (Maximization of Information in Evolutionary Networks) method is a dynamic algorithm that optimizes the information flow within evolving networks to detect communities [23]. It takes into account both structural changes and attribute information to enhance the accuracy of community detection.

- ✓ The NEIWalk (Neighbor Similarity-based Information Walk) [24] method is a dynamic algorithm designed for content-based networks. It incorporates attribute information and employs

a random walk approach to identify communities as the network evolves.

- ✓ The method proposed by Z. Zhao is a dynamic community detection approach that leverages both structural and content information to uncover evolving communities in dynamic networks [25].

In this section, we have employed two real-world network datasets for our experimentation: the DBLP network and the CORA network. Their descriptions are provided below:

- ✓ DBLP network: [26] This dataset consists of co-authorship connections derived from the DBLP Computer Science Bibliography. It serves as a common benchmark for assessing community detection algorithms within academic collaboration networks. Within this dataset, nodes symbolize authors, while edges connecting nodes signify co-authoring relationships across diverse research papers. The DBLP network encapsulates the interrelationships among authors engaged in collaborative efforts within the realm of computer science. This dataset holds significance in evaluating the efficacy of community detection algorithms in recognizing research communities and comprehending collaborative dynamics within the academic sphere.
- ✓ CORA network: [27] This dataset models a citation network involving scientific papers in the field of computer science. Nodes in the dataset represent individual research papers, and the directed edges signify the citations between these papers. Each paper is enriched with content-based attributes derived from its title and abstract. Renowned for its applications in evaluating algorithms for tasks like link prediction, node classification, and community detection, this dataset offers insights into the dissemination of knowledge and scholarly impact within the realm of computer science.

We evaluated the quality of the obtained communities using the DBLP network. The NMI values of eight approaches for all datasets are presented in Table 5.1, as well as in Figure 5.2.

Table 5.1: The NMI values achieved for each snapshot of the DBLP network.

Snapshots	GASNET	DynaMo	FacetNet	I-Louvain	Louvain	MIEN	NEIWalk	Z.Zhao
1	0,9140	0,9311	0,8992	0,8528	0,8328	0,9324	0,8552	0,8534
2	0,9169	0,9381	0,8532	0,8532	0,8432	0,9267	0,8500	0,8813
3	0,8958	0,8931	0,8678	0,8448	0,8418	0,8787	0,8483	0,8982
4	0,9131	0,8912	0,8512	0,8412	0,8412	0,8959	0,8472	0,9101
5	0,8720	0,9086	0,8714	0,8481	0,8481	0,8891	0,8590	0,9083
6	0,8994	0,8536	0,8612	0,8400	0,8500	0,8910	0,8600	0,8731
7	0,8903	0,8723	0,8837	0,8434	0,8334	0,8803	0,8703	0,8966
8	0,8910	0,8899	0,8512	0,8381	0,8381	0,8610	0,8510	0,8843
9	0,8911	0,8645	0,8504	0,8304	0,8404	0,8631	0,8615	0,8903
10	0,9016	0,8739	0,8645	0,8345	0,8345	0,8627	0,8546	0,8819

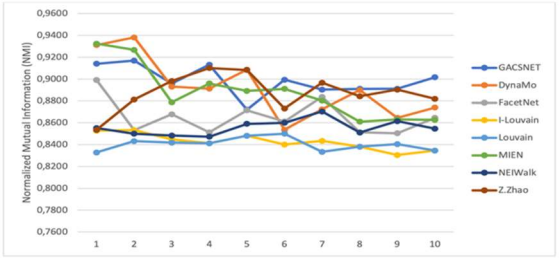


Figure 5.2: Comparison outcomes of NMI between GASNET and the remaining seven algorithms of the DBLP network.

The obtained results present a comparative evaluation of our GASNET algorithm against a selection of well-established community detection methods, including DynaMo, FacetNet, I-Louvain, Louvain, MIEN, NEIWalk, and Z.Zhao. The analysis offers valuable insights into the performances of these methods across ten different snapshots of the network. Importantly, our GASNET algorithm consistently stands out as a high-performing approach, achieving commendable NMI scores for all snapshots. This remarkable consistency underscores its robustness in identifying and capturing significant community structures within dynamic networks.

Interestingly, the data also highlights the dynamic nature of several algorithms. DynaMo exhibits noticeable fluctuations in its NMI scores, indicating its sensitivity to changes in network configurations. This underscores the importance of parameter tuning for methods sensitive to network

dynamics. Furthermore, FacetNet and MIEN show variable performances across snapshots, suggesting their adaptability to diverse network contexts. While I-Louvain and Louvain maintain relatively stable NMI scores, they fail to achieve the consistent performance of GACSNET. This emphasizes GACSNET's ability to adapt to evolving networks and consistently provide accurate community detection results. Overall, the results underscore the prominence of GASNET as an effective tool for dynamic community detection, with its adaptability and robust performance setting it apart from other algorithms evaluated in the study.

On the other hand, to ensure the performance of our algorithm, we evaluated the quality of the obtained communities using the CORA dataset network. The NMI values of eight approaches for all datasets are presented in Table 5.2, as well as in Figure 5.3.

The NMI values obtained for various snapshots of this network reveal that our algorithm consistently achieves remarkable performance, with NMI scores ranging from approximately 0.8388 to 0.9205. This suggests the effectiveness of our approach in identifying significant community structures at different time points. While methods like DynaMo and NEIWalk display competitive performance, their NMI scores exhibit more fluctuations, possibly indicating sensitivity to network changes. FacetNet and MIEN show variable performance, underscoring their adaptability to different network contexts. I-Louvain and Louvain maintain relatively stable NMI scores but fall short of GACSNET's scores, highlighting the consistent capability of our approach to capture evolving community structures in dynamic networks. Overall, the results underscore the robustness and reliability of GASNET for dynamic community detection tasks, making it a trustworthy and high-performing algorithm for such applications.

Table 5.2: The NMI values achieved for each snapshot of the CORA network.

Snapshots	GASNET	DynaMo	FacetNet	I-Louvain	Louvain	MIEN	NEIWalk	Z.Zhao
1	0,8924	0,9217	0,8591	0,8515	0,8599	0,8799	0,9115	0,9217
2	0,9135	0,8246	0,8523	0,8532	0,8229	0,8629	0,8913	0,9099
3	0,8599	0,9019	0,8374	0,8691	0,8071	0,8471	0,9059	0,9048
4	0,9205	0,8009	0,8713	0,8897	0,8187	0,8287	0,8797	0,8375
5	0,9144	0,8089	0,7744	0,8747	0,8044	0,7044	0,8744	0,7989
6	0,8632	0,7987	0,8232	0,8432	0,8132	0,8032	0,8432	0,8123
7	0,8388	0,8834	0,7593	0,8292	0,8192	0,7992	0,8539	0,8857
8	0,8568	0,7890	0,7971	0,8158	0,7958	0,7458	0,8358	0,8013
9	0,8561	0,7999	0,7552	0,8052	0,8062	0,7562	0,8352	0,8164
10	0,8799	0,8699	0,7893	0,8393	0,7771	0,7671	0,8593	0,8338

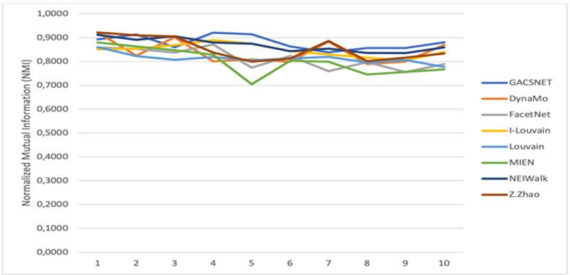


Figure 5.3: Comparison outcomes of NMI between GASNET and the remaining seven algorithms of the CORA network.

In summary, the comparative analysis of our approach's performance against other methods was conducted on two real-world datasets: the DBLP network and the CORA network. The results consistently demonstrated that GASNET achieved high Normalized Mutual Information (NMI) scores for various snapshots of the networks. This underscores its effectiveness in detecting meaningful community structures in dynamic contexts. Overall, GASNET showcased its reliability and efficacy as a leading algorithm for community detection in dynamic networks.

In conclusion, Figure 5.4 illustrates the outcomes of applying our algorithm to the Zackary's Karate Club study. This network was established by Zachary [28], who analyzed the social connections among 34 members of a karate club over a two-year span. The diagram was recreated using the Gephi software [29]. We discerned four distinct clusters, depicted in the figure with varying node colors. However, the two smaller communities operate as subgroups within the two major communities.

Consequently, our algorithm can identify more tightly-knit interactions. For instance, as depicted, the small community with nodes shaded in orange consists of five nodes. Each of these nodes is linked to the larger community with nodes shaded in green solely through their connection to node 1. In reality, a more intimate connection exists among these five nodes.

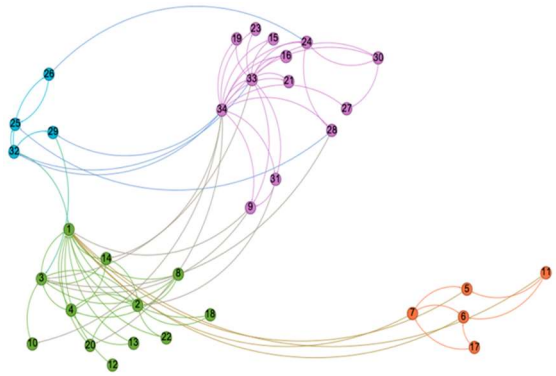


Figure 5.4: Community structure identified through the GASNET algorithm

In contrast, Girvan and Newman [8] identified the two groups into which the karate club was divided. Yet, they misplaced node 3. A similar finding is documented in [2], where node 27 is mispositioned. In comparison, our approach accurately categorizes these two nodes. The results obtained underscore the capability of genetic algorithms to effectively tackle the task of community identification within networks.

6. CONCLUSION

In summary, this article has presented an innovative approach for community detection in complex social networks using the proposed algorithm. We have explored the challenges associated with identifying community structures in both real and synthetic networks and demonstrated the effectiveness of our method through extensive experiments.

Our algorithm has showcased its ability to capture community patterns within networks by leveraging a genetic representation based on network topology. Harnessing the power of genetic algorithms, our approach has successfully navigated the complexities of real-world networks, providing a robust and scalable solution for community detection. Experimental results have indicated that GASNET competes with existing approaches in terms of accuracy and consistency in community detection. We have also showcased the algorithm's flexibility by applying it to various network types, enhancing its relevance for a wide array of real-world applications. The advancements presented in this article carry significant implications across various domains such as social media analysis, network biology, and beyond. By offering an original and high-performing approach to community detection, GASNET opens new avenues for understanding the intricate structures underlying diverse networks.

In conclusion, this article has contributed to the advancement of community detection in complex networks by offering an innovative approach and substantiating its relevance through rigorous experiments. We anticipate that this research will continue to inspire new ideas and applications in the realm of network analysis and data science.

REFERENCES

[1] M. E. J. Newman, « Detecting community structure in networks », *Eur. Phys. J. B - Condens. Matter*, vol. 38, nº 2, p. 321-330, mars 2004, doi: 10.1140/epjb/e2004-00124-y.

[2] C. Pizzuti, « GA-Net: A Genetic Algorithm for Community Detection in Social Networks », in *Parallel Problem Solving from Nature – PPSN X*, vol. 5199, G. Rudolph, T. Jansen, N. Beume, S. Lucas, et C. Poloni, Éd., in Lecture Notes in Computer Science, vol. 5199. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, p. 1081-1090. doi: 10.1007/978-3-540-87700-4_107.

[3] S. Fortunato, « Community detection in graphs », *Phys. Rep.*, vol. 486, nº 3-5, p. 75-174, févr. 2010, doi: 10.1016/j.physrep.2009.11.002.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, et E. Lefebvre, « Fast unfolding of communities in large networks », *J. Stat. Mech. Theory Exp.*, vol. 2008, nº 10, p. P10008, oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[5] John H. Holland, *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

[6] M. E. J. Newman, « Modularity and community structure in networks », *Proc. Natl. Acad. Sci.*, vol. 103, nº 23, p. 8577-8582, juin 2006, doi: 10.1073/pnas.0601602103.

[7] P. Mazur, K. Zmarzłowski, et A. J. Orłowski, « Genetic Algorithms Approach to Community Detection », *Acta Phys. Pol. A*, vol. 117, nº 4, p. 703-705, avr. 2010, doi: 10.12693/APhysPolA.117.703.

[8] G. Bello-Orgaz, S. Salcedo-Sanz, et D. Camacho, « A Multi-Objective Genetic

Algorithm for overlapping community detection based on edge encoding », *Inf. Sci.*, vol. 462, p. 290-314, sept. 2018, doi: 10.1016/j.ins.2018.06.015.

[9] Y.-C. Chiu, L.-C. Chang, et F.-J. Chang, « Using a hybrid genetic algorithm–simulated annealing algorithm for fuzzy programming of reservoir operation », *Hydrol. Process.*, vol. 21, nº 23, p. 3162-3172, nov. 2007, doi: 10.1002/hyp.6539.

[10] H. Akkineni, M. M. Bala, V. Takellapati, M. Nallamothu, et S. Yadlapati, « MEASURING RESEARCH INTEREST SIMILARITY AMONG AUTHORS USING COMMUNITY DETECTION », . *Vol.*, nº 11, 2022.

[11] E. Akachar, B. Ouhbi, et B. Frikh, « ACSIMCD: A 2-phase framework for detecting meaningful communities in dynamic social networks », *Future Gener. Comput. Syst.*, vol. 125, p. 399-420, déc. 2021, doi: 10.1016/j.future.2021.06.056.

[12] R. J. Kuo, Y. J. Syu, Z.-Y. Chen, et F. C. Tien, « Integration of particle swarm optimization and genetic algorithm for dynamic clustering », *Inf. Sci.*, vol. 195, p. 124-140, juill. 2012, doi: 10.1016/j.ins.2012.01.021.

[13] M. E. J. Newman, *Networks: an introduction*. Oxford ; New York: Oxford University Press, 2010.

[14] M. Girvan et M. E. J. Newman, « Community structure in social and biological networks », *Proc. Natl. Acad. Sci.*, vol. 99, nº 12, p. 7821-7826, juin 2002, doi: 10.1073/pnas.122653799.

[15] S. C. Madeira et A. L. Oliveira, « Biclustering algorithms for biological data analysis: a survey », *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 1, nº 1, p. 24-45, janv. 2004, doi: 10.1109/TCBB.2004.2.

[16] Johann Dréo , Alain Pétrowski , Patrick Siarry , Eric Taillard, *Métaheuristiques pour l'optimisation difficile*, 1ère édition. 2003.

[17] J. Handl et J. Knowles, « An Evolutionary Approach to Multiobjective Clustering », *IEEE Trans. Evol. Comput.*, vol. 11, nº 1, p. 56-76, févr. 2007, doi: 10.1109/TEVC.2006.877146.

[18] L. Danon, A. Díaz-Guilera, J. Duch, et A. Arenas, « Comparing community structure identification », *J. Stat. Mech. Theory Exp.*, vol. 2005, nº 09, p. P09008-P09008, sept. 2005, doi: 10.1088/1742-5468/2005/09/P09008.

[19] F. Taya, J. De Souza, N. V. Thakor, et A. Bezerianos, « Comparison method for community detection on brain networks from neuroimaging data », *Appl. Netw. Sci.*, vol. 1, nº 1, p. 8, déc. 2016, doi: 10.1007/s41109-016-0007-y.

[20] D. Zhuang, J. M. Chang, et M. Li, « DynaMo: Dynamic Community Detection by Incrementally Maximizing Modularity », *IEEE Trans. Knowl. Data Eng.*, p. 1-1, 2019, doi: 10.1109/TKDE.2019.2951419.

[21] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, et B. L. Tseng, « Facetnet: a framework for analyzing communities and their evolutions in dynamic networks », in *Proceedings of the 17th international conference on World Wide Web*, Beijing China: ACM, avr. 2008, p. 685-694. doi: 10.1145/1367497.1367590.

[22] D. Combe, C. Largeron, M. Géry, et E. Egyed-Zsigmond, « I-Louvain: An Attributed Graph Clustering Method », in *Advances in Intelligent Data Analysis XIV*, vol. 9385, E. Fromont, T. De Bie, et M. Van Leeuwen, Éd., in Lecture Notes in Computer Science, vol. 9385. , Cham: Springer International Publishing, 2015, p. 181-192. doi: 10.1007/978-3-319-24465-5_16.

[23] T. N. Dinh, Ying Xuan, et M. T. Thai, « Towards social-aware routing in dynamic communication networks », in *2009 IEEE 28th International Performance Computing and Communications Conference*, Scottsdale, AZ, USA: IEEE, déc. 2009, p. 161-168. doi: 10.1109/PCCC.2009.5403845.

[24] C.-D. Wang, J.-H. Lai, et P. S. Yu, « NEIWalk: Community Discovery in Dynamic Content-Based Networks », *IEEE Trans. Knowl. Data Eng.*, vol. 26, nº 7, p. 1734-1748, juill. 2014, doi: 10.1109/TKDE.2013.153.

[25] Z. Zhao, C. Li, X. Zhang, F. Chiclana, et E. H. Viedma, « An incremental method to detect communities in dynamic evolving social networks », *Knowl.-Based Syst.*, vol. 163, p. 404-415, janv. 2019, doi: 10.1016/j.knosys.2018.09.002.

[26] « DBLP network ». [En ligne]. Disponible sur: http://www.arnetminer.org/citation

[27] « CORA network ». [En ligne]. Disponible sur: https://people.cs.umass.edu/~mccallum/data.html

[28] W. W. Zachary, « An Information Flow Model for Conflict and Fission in Small Groups », *J. Anthropol. Res.*, vol. 33, nº 4, p. 452-473, déc. 1977, doi: 10.1086/jar.33.4.3629752.

[29] « Gephi ». [En ligne]. Disponible sur: https://gephi.org/