AN ENHANCED DISCRETE ARTIFICIAL BEE COLONY ALGORITHM FOR COMMUNITY DETECTION IN SOCIAL NETWORKS

THET THET AUNG

UNIVERSITY OF COMPUTER STUDIES, YANGON

JUNE, 2022

An Enhanced Discrete Artificial Bee Colony Algorithm for Community Detection in Social Networks

Thet Thet Aung

University of Computer Studies, Yangon

A thesis submitted to the University of Computer Studies, Yangon in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

June, 2022

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

29.6.2022

………………………………… ………………………………………

Date Thet Thet Aung

ACKNOWLEDGEMENTS

Firstly, I would like to thank the Union Minister, the Ministry of Science and Technology for giving me the opportunity to study Ph.D. course by providing support that all me to perform the research in University of Computer Studies, Yangon.

Secondly, I would like to express my deepest gratitude to Dr. Mie Mie Khin, Rector of the University of Computer Studies, Yangon (UCSY), for allowing me to develop this thesis and giving me general guidance during the period of my study. Further, I would like to offer my special thanks to Dr.Mie Mie Thet Thwin, former Rector, the University of Computer Studies, Yangon, for her guidance and encouragement during the Ph.D. life study.

I sincerely would like to express my greatest pleasure and the deepest appreciation to my supervisor Dr. Win Lelt Lelt Phyu, Professor, Faculty of Computer Science, University of Computer Studies, Yangon (UCSY). She believed in me and gave the chance to do this research. Her excellent ideas, guidance, caring, patience, excellent help and creative ideas have assisted me in broadening my research skills.

I am very grateful to the external examiner, Dr. Myo Min Than, Pro-Rector of University of Computer Studies (Lashio), for helpful advice.

Especially, I would also like to extend my respectful gratitude to Dr. Thin Lai Lai Thein, Professor, Course-coordinator of Ph.D 10th batch, the University of Computer Studies, Yangon, for her valuable advice, moral and emotional support, encourage which are valuable to me. Her constant support has been invaluable in the successful completion of this research.

I am very grateful to Daw Aye Aye Khine, Associate Professor, and Head of Department of English, University of Computer Studies, Yangon (UCSY), for her valuable support throughout of my Ph. D. course work, doing this research and editing my thesis from the language point of view.

In addition, I would like to extend my appreciation and thanks to the board of examiners for valuable comments and recommending my thesis. I would like to extend my respectful appreciation to Dr. Khine Khine Oo for her encouragement and guidance during the deployment of this research and my former supervisor, Dr. Thi Thi Soe Nyunt for her patient supervision and providing me with valuable ideas throughout the study of this research. I also thank all of my Ph.D 10th batch members for their encouragement during the Ph.D. student life.

Lastly, I would like to express my appreciation to my parents, my aunt, my husband and friends who have supported me with their love and encouragement. They always support me during my period of studies, especially for this Doctorate Course. I would like to thank every person who directly or indirectly contributed towards the success of this thesis.

ABSTRACT

Real world social network is represented as a graph structure which has a collection of non-empty set of nodes and set of connections of nodes. The nodes in a social graph are people, and edges represent the connection between them. People in the society are naturally grouped to form communities in their workplace, friendship and family. The study of complex social network topology has triggered the interest of many scientists in recent years. Some important features of a social network need to be understood in terms of node density in cluster, degree of node, reachability, size of network, geodesic distance and diameter. Thus, the idea of community within a social network has emerged as an educational tool for social network analysis that lays the foundation for some high-level initiatives. Finding community structure is the key to research the structural components of relationships in the social network. Community detection aims at grouping nodes based on the links between them to form strongly link sub groups of graph from the whole graph.

Many community discovery techniques have been borrowed on inspired from the problem of hierarchical clustering, graph partitioning in modern graph theory, as well as the graph clustering or dense sub-graphs discovery problem in the graph mining area. These algorithms were designed to reveal the mesoscopic nature of various networks. It is not known how good the algorithm is in terms of computational time, efficiency of these algorithms and accuracy, which remains still open. Community detection in social network is one kind of the optimization problems because it cannot get exact solution and only get an optimal solution. So, various types of community detection algorithms attempt to capture the intuitive notation that nodes within the group have higher intra-density and have lower inter-density with other nodes in the other group.

Population-based nature inspired optimization algorithms have attracted extensive research interests over the past decade. Researchers used these algorithms to search the hidden communities in social networks. An enhanced discrete artificial bee colony D-ABC algorithm is proposed for solving community detection problem in social networks. In this thesis, some of the challenges and applications of social network community discovery for analyzing network documents will be explored. The proposed system focuses on undirected and unweighted social networks and considers the connection of nodes for the community detection problem.

The proposed algorithm D-ABC is designed with an effective initialization strategy with label-based solution representation and one way crossover operation of genetic operators is used in the search strategy process. D-ABC selects modularity function as the fitness function of each solution. The efficiency of searching process can be improved by heuristic function which contain genetic algorithm operations. Most of the real social network community detection problem do not know each community size and numbers of community. D-ABC algorithm can also detect communities in social network that does not know real community numbers and sizes. When it detects the communities on a network structure, it uses neighbor nodes similarity measurement. In this thesis, standard social datasets with ground-truth results and various scale real-world networks without ground-truth from the Stanford Large Network Dataset Collection are used. The performance of the proposed system is assessed by accuracy, efficiency and effectiveness. The result indicates good quality with accurate communities as well as their density improvement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1
INTRODUCTION

Humans tend to live in society by grouping together into large or small community structures. A community evolves totally according to religion, gender, interest, hobbies, work category and many others. In recent decades, the advance of internet technology has created virtual groups, communities and societies where people connect and share their information. Despite the distance around the world, people are trying to socialize through social networks. The social networks have grown to the foundation of digital socialization and assisted individuals to share the information with their friends and family members. These networks need to be analyzed from different perspective based on the importance of the networks to find the most prominent personalities on it. People in these virtual networks can be categorized into community structures like our physical cluster or community. Today, the detecting of communities in network has become one of the important functions in social network analysis.

Real social networks are created as the graph structure. Vertices in the network represent individual actors or organizations and link represents the connection between nodes. The resemblance values among vertices and they can be made as contributors to their correlating community. A community can be a collection of relationships between individuals, groups of objects, groups of business, family member relationships, groups of education and occupation groups that behave the same. Usually, the community is set up like a partition of a social network by detecting them. It is also related with the clustering of a social network into various connected subgroups or communities, with strong cohesion or interactions between nodes in subgroups.

Many approaches from various areas, such as telecommunications, physics, computer science, mathematics, biology and sociology have been used to detect communities in social networks. Example of community structure in scientific collaboration helps to discover a community of researchers who work in or have expertise in similar research areas. Community detection in social network also helps to find a person who joins a particular group based on the relationship they have between members of a particular group. It can be affected in information science, geography, sociology and marketing.

1.1 Research Problem Statements

The community discovery problem requires partitioning the network as a group of densely connected nodes where nodes belonging to groups are sparsely connected. Social network community detection has two computational complexities. They are the size and structure of social networks that evolve over time. There are many different types of detection algorithms for community. The most concerning problem in social networks community detection is the validity of communities. Traditional community detection algorithms cannot detect effectively community results in various scale of networks. Large-scale network community detection is very difficult and has not yet been sufficiently solved. Over the last few decades, scientists worked the great effort on detecting community structure in various networks.

Social network community detection has become NP-hard optimization problem. If the number of nodes in the networks increases, the networks computational complexity will grow exponentially. The issue of choosing a particular community detection algorithm (CDA) is important when the goal is to disclose the network community structure. Some algorithm can effectively solve in large scale networks and other can only solve small scale networks. An effective community detection algorithm is needed that can solve various size of networks. Then, the choice of a given methodology should influence the outcome of the experiments. Researcher used swarm intelligence-based algorithms which are proposed to solve the optimization problems. In optimization-based methods, the results quality of network community are measured by using the quality functions. The efficiency of algorithm is based on the result of quality function and search ability of the algorithm. The proposed system evolves the background theory of nature inspired based Artificial Bee Colony (ABC) algorithm, to find the optimal solution in community detection problems. It contains a number of possible solutions. They are numerically encoded and then calculated fitness function to get optimal solution.

Finding communities in complex network have many challenges. Network can be huge: millions of actors and billions of interactions among actors. When existing community detection method applied directly to network this size, it will fill. Next challenges are to find algorithms that generate valid results within reasonable computing time.

1.2 Motivation of the Research

Social networking sites are effectively building societies and communities. These societies are viewed as a network of nodes that represent the depiction of people and edges. It is a great and efficient way to store, model and display complex systems which have many data. Therefore, an effective and quick community detection method is needed to analyze the various size of social networks. Community detection is very important in many applications for understanding and extracting information from complex systems. It has been extensively studied over the last few years. This research work contains the studies of existing community detection methods and proposes a new community detection method by evaluating these algorithms on the real datasets.

Social networks contain cluster groups based on people's interest, location, occupation and other similar factors. A useful tool for analyzing large and complex network is community discovery. The problem that community detection tries to fix for identifying the groups of nodes that are more closely connected than any other network. Community detection in various network has led to important findings in various domain area. It shows that community is a meaningful organization unit and that provides new insights into the structure and function of the entire network under investigation. The result of community can be used in other areas such as link prediction [54], recommendation, customer segmentation, influence analysis and vertex labeling.

1.3 The Objectives of the Research

The objectives of doing the research on this field are to detect the quantitative community structure for the given network dataset and to check the accuracy of members in each community. To achieve this purposes, the research works are done as follows:

- Finding the structure of social network
- Making sense out of social network that is to extract information about individuals from the network they participate in.
- Discovering complex community patterns of social network
- Showing how to create graph using GraphX in Spark for large network
- Evolving the effective community detection algorithm (CDA) for various scale social network
- Taking out the important information from each community

- Identifying the validity of members in the communities.
- Understanding the data of social network based on nodes and its relation with others.
- Developing effective iterative community detection algorithm on Spark

1.4 Contribution of the Research

This research proposed the efficient and effective community detection algorithm called an enhanced discrete artificial bee colony algorithm (D-ABC) for various kinds of social networks. The major contribution of this research is to propose the effective community detection algorithm (CDA) for well-connected communities in different sized networks. It is bases on Artificial Bee Colony (ABC) Algorithm and is redesigned to fit community problems. Label propagation based effective initialization is used for population initialization, and an effective strategy is proposed to generate the initial population and to reach an optimal solution quickly. GraphX is used to store graph structure especially for large network because graph data is too large to fit on a single machine for large scale networks. The proposed DABC algorithm is experimented on Spark Framework for mining community structure in various-scale complex social networks, especially for large networks.

1.5 Organization of the Research

The dissertation is structured with six chapter. Chapter 1 describes about the introduction of community detection in social network, motivation, objectives and contribution of the research work. Chapter 2 reviews the literature survey on different kind of community detection algorithms for different networks and their usage in many different application areas. Different types of community detection algorithm (CDA) in various kind of networks are discussed in detail such as hierarchical based community detection algorithms, community detection algorithms based on graph partitioning, clustering and modularity values. Quality metric functions of results community are also discussed in this chapter.

Chapter 3 represents the relative methods of graph representation, artificial bee colony (ABC) and other related theory for this research. Chapter 4 presents the proposed system detailed explanations of an enhanced version of discrete type Artificial Bee Colony Algorithm called Discrete Artificial Bee Colony (D-ABC) Algorithm and it is tested on Sample networks and Karate Club network. There are three core steps: population initialization, find best solution using D-ABC, and produce the best solution as an output. Chapter 5 explains about dataset, the detailed analysis social networks that contains real results and other various kinds of networks. Then, evaluate the result quality using quality metrics for algorithm’s efficiency, accuracy and effectiveness.

Lastly, Chapter 6 describes the conclusion together with discussion and research findings, and future works. The most important contents of the proposed solution are brief and the future research direction is suggested in the last Chapter.

CHAPTER 2
LITERATURE REVIEWS AND RELATED WORKS

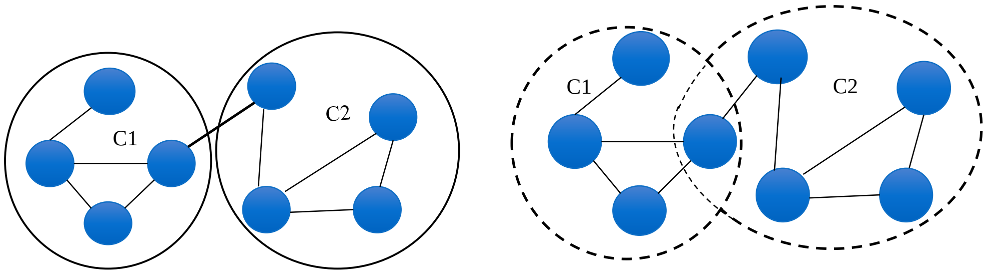
In recent years, user interactions on social network have grown tremendously regardless of time and location. These increment interactions have caused huge datasets with vast range to extract interesting user behavior and understand the network. Today, internet technology is developing and people can connect each other using this technology. Many information about the connections between users can be obtained through analyzing network relation which can complement other kind of user's detail information. The analysis of this data can provide basic information that are important for tasks such as recommendation and help in network visualization and navigation. The rapid growth of social networks makes it essential to identify the structure of the community. Firstly, community detection task in disjoint and overlapping communities among different type of social network is reviewed in this chapter. Secondly, various types of related community detection algorithms will be described in this chapter. Thirdly, evaluation metrics for community detection is presented and its application areas are described.

2.1 Social Network Community Structure

Social networks are constructed as a graph structure G (V, E), which contain a set of nodes or vertices (V) that corresponds individuals or organization and a set of two node connection or set of edges (E) . The community detecting process is to discover the clusters or cohesive groups in the network. The network community structure is represented as $C=\{c_1, c_2,\dots,c_n\}$ and each community contains a set of vertex groups with high edge density between vertices and low edge density between different groups. Example of social network communities are an organization based on same job group, scientists' coauthors who published papers in a same area, citation networks based on research topic and metabolic network based on functional groupings. Social network community detection can provide knowledge about how the topology and functionality of the network affect each other [24]. Discovering the community structures in social network make members in this network can easily interact with people with similar and comparable result. Today, the enormous growth rate of social network requires intensive research from recent research done to identify community partitions in social networks. Vertices in networks are often grouped into tightly packed cluster with high density of edges within a cluster and low edge density among other groups.

2.2 Disjoint and Overlapping Community Detection

Community detection is the task of separating a set of network vertices into subset, with the connections in these subsets being denser than the connection between them. These subsets are called communities, and each community is a subset of vertices that are more closely connected within the group than the other group members in the graph. Community structure C on the graph G = (V, E) is determined as C = {c_1, c_2, …, c_n} where $c_i \cap c_j = \emptyset$ and $U_{i=1}^{nc}$ c_i=V. where c_i is a non-empty subset of v and nc denotes the number of communities in c. The input of the community detection problem is the graph structure and output is a community structure. Figure 2.1 shows the disjoint or non-overlapping and overlapping community structure for a sample network.



(a) Disjoint Community Structure (b) Overlapping Community Structure

Figure 2.1 Two Types of Community Structure

Disjoint community structure means every node has one and one community means a single vertex belongs to at most one community. Example is in karate club friendship, vertex is a karate club member and edge is member friendship. Each member has only one community. That understand the interactions among members and help the network visualization. Overlapping community structure aids, each node can be owned by one or more groups, means a single vertex belongs to possibly multiple communities [21]. Categorization is based on algorithm design ideas such as clique percolation, link partition, and local expansion, dynamic and statistical inference [48].

The first overlapping community detection algorithm (CDA) was proposed by Palla et al. [41]. The algorithm used k-cliques based approach in which clique was a subset of nodes. Each node was adjacent to every other node. K-clique defines size of clique such as 3-cliques represents each sub cluster contains three nodes. Node becomes a member of more than one community. Shen et al. [51] proposed another clique based overlapping community detection approach. The proposed algorithms used overlapping modularity measure based on number of maximal clique. In 2010, Chen et al. proposed overlapping CDA for weighted network. This algorithm used local algorithm to expand a partial community starting from a unique node [8].

Yaozu Cui and other coauthor described an overlapping community detection in complex network. The author found that cluster coefficient and maximal sub graph between two neighboring communities. All the subgraphs are eliminated from the original networks and then they combined these networks based on cluster coefficient of two adjacent maximal sub graphs. The new extended modularity method was used to enhance the proposed overlapping algorithm [10]. Overlapping community detection was the difficult task in social network analysis.

2.3 Community Detection using Graph Partitioning Algorithms

This section highlights the literature reviews on the graph partitioning method used in community detection problem. Last year, many community detection methods have been developed for social network. Researchers used graph-partitioning methods. In graph-based clustering, cut the graph into several partitions. Partitioning algorithm process a graph partitioning into a predefined cluster number of smaller components that contain specific properties. A common property that needs to be minimized is called the cut size. A cut is a set of graph vertices that are divided into two separate subsets, and the size of the cut is the number of edges between the components. A multi-cut is the set of links that divides a graph into two or more components. For graph partitions, it must specify not only the number of clusters to retrieve but also the size of the components. Otherwise, it may be useful to place the lowest order vertices in one component and the remaining vertices in another component. The graph partitioning method is not suitable for detecting communities in such cases, as the number of communities is generally unknown in advance. For a given graph, divide the nodes of graph into subsets no larger than a given maximum size, as to minimize the total cost of the edges cut.

The Kernighan-Lin algorithm was one of the first techniques for solving graph partitioning problems. It divides the graph nodes into subsets of given size to reduce the total costs on all edges cut. A main disadvantage of this algorithm is that the number of groups must be predetermined. W. Kernighan and S. L in 1970 [27] proposed the Kernighan-Lin algorithm or KL algorithm that originated from graph partitioning method. The two most important methods of community inference based on its degree-corrected variant or the stochastic block can be mapped onto version of the familiar version of minimum-cut graph partitioning problem. This is demonstrated by modifying the Laplacian spectrum method to estimate the community. The algorithm maximizes different modular value between the number of links within the components and the number of edges between the components. KL algorithm is one kind of heuristic based optimization algorithm that partitions the network into two given size communities. It starts with initial solution then repeats the process to find the candidate solution. It finds local optimal solution rather than global solution. It needs to know the number of community and size of communities for a given network. It may be very sensitive to preliminary conditions and it is able to get low efficiency for some complicated networks.

Graph partitioning is to divide the graph into multiple parts while minimizing an objective function. It divides the nodes of k clusters of pre-defined size, such that the edges that present among the groups is minimal. It needs to know the number of partitions is necessary. One kind of spectral bisection method is the well-known graph partitioning algorithm. In the past years, it has been applied to clustering analysis of complex networks.

The authors Luca Donetti and Miguel A. Munoz proposed improved spectral algorithm to detect network communities [13]. It used quadratic optimization techniques to reduce the given “cut” function that receive partitions with the minimum “cut” was considered the most appropriate. According to matrix theory, the spectral bisection method transforms a minimum “cut” problem into constrained optimization problem and an approximate optimal solution of quadratic optimization problem can be attained by computing the second smallest Laplace matrix eigenvector of the graph structure network [14].

M.E.J Newman et al. took into consideration three distinct problems related with the network structure community detection through maximization modularity by statistical inference and normalized-cut graph partitioning [36]. They solved the problem by using spectral algorithm that employ eigenvectors of the network matrix representations. They described those specific options for the independent parameters that appear in the spectral algorithms. In these three problems, they computed the principal eigenvector of the normalized Laplacian matrix and separated the nodes according to the sighs of the vector elements. They showed that the algorithm in these three problems becomes the same. The algorithm described that it differ from the standard spectral algorithm for maximizing modularity described in previous literature, but at least in the spectral formulation, there is no difference between these three problems. Computer-generated network and real-world network are used, the results for these networks appear to perform better than the traditional graph partitioning algorithms.

The graph partition algorithm for community detection is not suitable for large graphs because it requires to define cluster numbers and the size of each group. Therefore, an effective algorithm is required that is capable of yielding this information in its result.

2.4 Hierarchical based Community Detection

In this part, the community detection algorithm (CDA) based on hierarchical algorithm is pointed out. There are two types of hierarchical community detection algorithm. They are divisive algorithm and agglomerative algorithm. It is very common in the analysis of social networks, biology, telecommunication and computer science engineering, marketing and etc. Real world network is a hierarchical structure. It shows the group of members in various levels, with little modules included in big groups, which are gain part of larger groups and so on.

Divisive algorithm is top-down based approach in which the community iteratively divided into two portions by deleting links between the nodes that have the least similarity. Agglomerative algorithm is bottom-up approach that merges two communities if they are similar enough. It begins with each vertex as its unique community and finally end up with the entire graph as one network community if they may be no longer stopped at a few earlier levels. The hierarchical algorithms have some advantages than graph partitioning algorithm. It is unnecessary to know the information about the number of cluster and size of each cluster. However, it does not allow to divide among many clusters and to select one or more partitions that represent the best community structure.

Most of the existing hierarchical agglomeration community detection algorithms have either unsatisfied community detection results or high computational complexity. Mingwei Leang et al. proposed Community Detection Algorithm via Community Similarity Measures (CDACSM) algorithm that is one of the new hierarchical agglomeration community detection algorithms. The proposed algorithm produce quantitative community structure and required low computational complexity. Connected nodes are clustered as a groups, then make community measure and combined community with the community merging rule. Then adjusted some nodes by using SHARC that is an improve label propagation algorithm. Finally, it combined communities based on similarity measure of community [31]. It used single node community similarity measure and merging rule. The main purpose was to solve the low accuracy and longtime complexity of agglomeration algorithm. It was suitable for a large network.

Divisive algorithm has some demanding situations: how to efficiently identify external links, how to correctly eliminate external link and how to end a divisive algorithm without help is predefined parameters. Therefore, Xiaoyu Ding et al. suggested a concept of the vulnerable link and autonomous partitioning. They proposed the concept of weak links, a link-break strategy and an autonomous divisive algorithm and broken link strategies are based on the weak link. No parameters, non-topological information and community definition are needed [16].

2.5 Community Detection using Modularity-based Algorithm

Modularity-based CDA is also called Girvan Newman algorithm. Newman and Girvan have proposed modularity function to calculate the strength of clustering in the network. Modularity (Q) is well known quality measure function for community detection. It attempts to achieve the cluster problem in its compact form. Most of the cluster algorithm that used modularity are described in this section. Modularity is designed to measure an objective quality for each split of a network into communities (cluster or modules). It is used as an optimization method for discovering networked community structures. In a highly modular network, the connections between nodes within modules are tight but the connections between nodes on different modules are sparse.

Newman proposed a greedy method for modularity maximization. The greedy algorithm based on agglomerated clustering in 2004 [37]. It included the clusters of nodes which are consecutively connected to larger group. Thus, the modularity value increased when the merging process finished. Clauset et al. updated Newman algorithm that is due to the lack of adjacency matrix and it contained an excess number of unnecessary processes [9]. This greedy modularity optimization tends to create quickly large communities with small expenditures, often provides low values of maximum modularity. Danon et al. proposed the normalization of the modularity Q variability obtained by adding the two communities with the fraction of edges that happened to one of the two communities in favor of small groups [11].

The high values of modularity mean the algorithm can detect better community results for a given network. Network partition quality basically is based on maximizing modularity value. Therefore, the maximizing modularity is the important for an input graph. Ehsan Moradi et al. described a new parallel CDA based on modularity for large network. The author proposed algorithm is based on the Louvain's hierarchical processing method that allocated threads to compute the aggregation modularity of two adjacent communities. It prevents from thread interface and parallelization overhead more than the previous parallel algorithm [17].

The aim of network community detection is to ensure nodes in the network have higher intra-connection and lower inter-connection. In big data era, community detection process has become one of the challenging issues in large scale social network. Ranjan Kumar Behera et al. described the efficient community detection algorithm based on modularity. It proposed that will have relatively less time complexity as compared to other CDA. To reduce time complexity, k value is set in the range from 2 to lama that rely on the social network size. Before algorithm execution, Lama is a threshold value need to set. For small network, the value of k is 2 and for huge network k must be larger than 2. Each node in the network is assigned one of k value and using eigen vector corresponding to highest eigen value of modularity matrix. This process is repeated recursively until modularity value is no longer increased [4].

2.6 Community Detection using Nature Inspired based Algorithms

Nature inspired algorithms mimic various process observed from nature. These algorithms have become promising solution for solving different real-world problem in a suitable way and have found a position among researchers to solve different real-world problems that do not have exact solution but can be find an approximate solution. They get effective capacity that provide acceptable solution to many real complex problems. This section provides nature inspired based optimization algorithms to detected communities in different types of networks. Evolutionary algorithms have produced a good solution for different kind of optimization problem. They produce many possible solutions for community detection problem for social networks, which come from a type of nature inspired algorithms such as bee colony algorithm, particle swarm optimization, bat algorithm, genetic algorithm (GA), ant colony algorithm, chicken swarm optimization algorithm, artificial bee colony (ABC) algorithm and etc.

Community detection has become an optimization problem based on the various type objective quality functions that captures the better intra connection than the inter connection. Ahmed Ibrahem Hafez and his coauthors modified ABC algorithm to detect community in various networks. The proposed optimization algorithm applied different quality functions are in the optimization process in order to evaluate their performance [23]. Guoqiang et.al proposed GA algorithm to detect community in complex network, they used modularity density as the fitness function of proposed algorithm [22].

The large amount of data in social network is necessary to analyze and find hidden connections between nodes. Ramadan Babers et al. presented community detection problem as an optimization case and explained how to divide the network into subgroups. Effective optimization technique called Lion optimization algorithm is used to discover social network community [3]. The algorithm used with two different optimization functions, modularity and community fitness to get more accurate result and optimal community structure. The authors would like to consider some criteria to improve the qualitative accuracy result and scalability of community detection problem and will use nature based inspired algorithm to detect communities of social network.

In 2016, discrete bat algorithm for community detection in complex network has been proposed by Anping Song et al. [53]. This algorithm automatically defines the number of communities and can search global optimal solution than the previous algorithm. The traditional bat algorithm is proposed by Yang in 2010. This algorithm can be used to solve continuous problems. Community detection problem is one of the discrete type problems. Therefore, authors modified the original bat algorithm and proposed a novel discrete algorithm for CD various complex network. Modularity is used as the objective function and this algorithm has the challenges for large network.

In 2007, Mursel Tasgin et al. used GA in complex social networks community detection which is based on network modularity optimization. It does not need to know the community numbers. The genetic based algorithms have several parameters like the number of chromosomes, mutation rate, crossover rate and other. A chromosome contains N genes and that represent a candidate solution. The number of nodes in the given network is represented as N and new choromsome are create by using one-way crossover operation. The algorithm provides proof of concept in genetic algorithm and capables of producing results consistent with previous algorithms and doesn't aim to find concert the parameter of genetic algorithm [57].

To get efficient community result, Clara Pizzuti presented the efficient community detection algorithm using nature inspired based genetic algorithm called GA-Net. The algorithm specialized variation operators that are aimed at only considering the interactions between nodes, reducing the possibility of finding possible solution and improving the convergence of the methods. The algorithm uses the locus-based adjacency representation. This approach contains the concept of community score to measure the result community structure result of a network and search optimal solution by maximizing the community score [43]. Halalai et.al also used the genetic algorithm with distributed nature to detect the community in complex network [46].

Researcher also proposed distributed approach for large social network community detection using nature inspired based algorithm. Raluca Halai et al. proposed distributed approach based on genetic algorithm for identifying scalable community structure in complex network. They used the know community structure given network and turn genetic parameters. Authors proposed a distributed genetic algorithm on Apache Mahout Framework. They used locus-based representation for graph representation, modularity and community score are used for fitness function and chose uniform crossover operation and repaired mutation operator employed [30].

A parallel based algorithm for CD in complex network using evolutionary approach are presented by Marius Joldos and Camelia Chira. A coarse gained evolutionary algorithm is proposed and several populations of potential solutions are advanced in parallel. Same individual representation is used in the parallel populations. Each solution in the population can be evolved using different fitness function. Authors will continuously develop these parallel community detection approach to reduce time in larger network [26].

Forge large scale network, some researchers use parallel algorithm based on spark framework. The authors tested parallel particle swarm optimization algorithm for CD in large network on spark framework. This parallel PSO algorithm are designed effective solution representation and specific updating strategy of discrete PSO for parallel computing. A new Spark API for graph, GraphX is used to improve the ability of handling large complex network [59].

2.7 Conventional Community Detection Algorithms in Igraph's Library

Researchers proposed various types of algorithms to detect community in social network. The following algorithms are famous community detection algorithms that can be used by installing igraph library package in python and R.

Edge Betweenness Algorithm: It was performed by computing the edge betweenness of the graph and cutting the edges with the highest edges betweenness score. Then recalculated edge betweenness of edges and removed the edge with the highest score [19].

Fast Greedy Algorithm: It is modularity optimization algorithm (MOA) to detect community structure in given networks. First, every node is considering a community and all relationships are removed in this algorithm. Then, each link is inserted to this network and the two vertices is merge to increase modularity. This process is done repeatedly until the highest modularity value is achieved [9].

Infomap Algorithm: It begins with encoding the network into groups. It then sends the signal to a decoder through a capacity limited channel, the message is decoded by decoder and creating a set of possible candidates for the original graph [49].

Label Propagation Algorithm: Raghaven et.al proposed label propagation algorithm. Each vertex in the network is initialized one of k label. Then each node assigns the label of its majority neighborhood node. This process will end when every node has the same label as the majority of its neighbors [45].

Leading Eigenvector Algorithm: Newman proposed this algorithm which is top down hierarchical modularity optimizing function that divide the graph into two subgroups in each step. The division itself increase the modularity range appropriately [38].

Louvain Algorithm: This algorithm is described by Vincent D. Blondel [6]. This algorithm finds small communities by optimizing modularity locally on all nodes, then each small community is grouped into one node. It is based on modularity measure and a hierarchical approach.

Optimal Algorithm: It calculates the community optimal structure in a given social network in terms of maximal modularity score [38].

Spinglass Algorithm: It is based on an approach from statistical physics. Edges should connect the nodes in the same spin state (current context community), but nodes of different states (belonging to different communities) need to disconnect [48].

Walktrap Algorithm: Walktrap algorithm is proposed by Pon and Latapy. Start with an ungrouped partition and calculate the distances value between all adjacent nodes. Then select two adjacent communities to merge them and update the distance between communities. Its results are used to merge each community from bottom to up [44].

2.8 Evaluation Metrics in Community Detection

This part outlined community assessment measure topology and mentioned some popular validation measures used to compare the proven community with the fundamental truth structure. Several metrics can be used to evaluate social structure quality. If the society knows the network connection, it will be easier to assess the communities identified by comparing them with community structures. Therefore, an organized and detailed survey of the proposed metrics is needed for detection and evaluation of the community structure.

Newman and Girven, 2004 to measure result quality of the communities, proposed modularity metric. It became the most popular and widely used metrics for CD [39]. In 2007, the authors presented resolution limit problem in modularity metrics. They found that modular optimization may fail to identify smaller-scale modules that depend on the total network size and the degree of interconnectivity of the modules, even in cases where unambiguous modules are defined [18].

Benjamin H. Good et al. showed other problems of modularity called degeneracies: this measurement allows for a high modular but structurally different solution for a single graph. They also studied the range of behavior to maximize the likelihood of a model pattern of augmented modular networks and show that the important of the network size and components number [20].

Several versions of modularity resolution have been proposed to solve the resolution limit problem of modularity by Arenas, Fernández, Fortunaton and Gómez, in 2008. They discovered the modularity by replacing edges with motif. And then Lambiotte suggested that different types of quality functions with multi-resolution to solve the problem. The important values for the resolution parameter are selected by using complementary measurement of the robustness of uncovered partitions [29].

The Rand Index (Rand, 1971) is a method of comparing modified classifier based on a pair of classified properties. If one solution places both objects in the same groups or one solution in different groups, it is called that two solutions correspond in a few objects [47]. Purity measurement (Manning, Raghavan and Schütze, 2008) was the first used in the context of community detection [34].

The comparing different social structures problem can be solve by measuring the Normalized Mutual Information (NMI) metric. Danon et al. proposd NMI to find community detection problem in networks and it can be used in many work [12]. It based on the confusion metrics in which the rows represent the real communities and the columns represent the detected community. It calculates normalize mutual information of two community structures in network analysis. NMI score can be calculated based on the real and the generated partition. The NMI value varies between zero and one. Zero value means that the detect community coincide with ground truth (GT) and one refers to a perfect match.

Orman, Labatut and Cherifi found that the measurements do not address network analysis because they do not consider the network connection. They proposed modified version of these measurement. In this case, poor placement with high grades would result in greater punishment compared to low cuts. Modified version of NMI, ARI and Purity are weighted versions, namely W-NMI, W-ART, WPU. Slightly different from community detection work is the study of the development over time. This issue has attracted a lot of research due to its great application to world conditions [40].

2.9 Application Areas of Community Detection

This section presents application area of community detection and how this community information can be applied to the development of various types of application are discussed. There is a growing trend in using network science algorithms including community detection methods, in various area of healthcare. These areas are protein networks, healthcare fraud detection, drug prescriptions and drug abuse. The authors try to find physician's real medical specialties based on their prescription history [52]. Community discovery in medical field can be used as an alternative for missing values of the healthcare database and can help scientists and researchers to get more reliable and accurate results. Using big data and community discovery tool by interviewing experts in the field. Authors attempted to explore a CD approach in an innovative way to identify the real medical experts, concerning the history of their prescriptions. Other researchers have not adopted this innovation approach. This is the first study to identify the most specialized physicians by using prescription data to implement community detecting methods.

Udrescu et al. [58] presented a new approach for analyzing drug-drug interaction networks by adopting clustering and topological community detection techniques such as the modularity based and energy model layout of community detection algorithms. In addition, Landon etc. [30] explained a network science based CD algorithm for identifying groups of physicians who are likely to have pre-established relationships. Sun et al. proposed an outlier based fraud detection approach in a heterogeneous network of physicians and drugs. This approach divides the doctor into different communities and looks for outliers in each community [55]. A bipartite graph if drugs and proteins linked by drug-target binary associations is created by Yildirim et al. [62]. Using a powerful local clustering operation, similar types of drags were categorized according to Anatomical Therapeutic Chemical classification.

Real world social network is represented as complex network. One of the characteristics of social networks inherently contain a community structure. Many researchers proposed community detection algorithm with different disciplines. It is used to identify criminal user groups. Pinheiro et al. [42] studied CD of fraud events on telecommunication network in order to identify customer behavior and investigate outliers by using community detection.

Community detection result groups can be created from real person account or bot account. They support or spread criminal ideas or terrorism-like activates. The authors in [50] performed CD to detect the community in criminal networks, then they also performed some manual analysis. CD is also used for politics to find political ideologies or individual politicians on a particular social group. It can be specialized to track evolution of this effect over time. This effect is typically created by influence or astroturfing bots. Bots try to give a fake impression on real grassroots to support a policy, individual, and product campaign [5].

Link prediction is used to assess the likelihood between members of a network and to determine missing links, fake links, and future links. The underlying network structure is discovered via a CD algorithm, then the possibility of being a link between two members is evaluated. Soundarajan and Hopcroft [54] showed that the information of community obtained by CD methods improve the accuracy of similarity based link prediction.

2.10 Summary

Community detection (CD) is a common problem in analyzing graph data from the graph theory perspective. Graph community detection means that vertices in the graph are grouped into tightly connected clusters, with high density within group edges and low density between groups. Especially, community identification in large networks is an important issue for many scientific domains. Researchers developed different types of community detection algorithm to examine the structural feature of real-world complex networks. This chapter presents related research of disjoint or overlapping community detection in large social network. Firstly, it discusses the community detection in social network and then reviews the two types of community detection in social network such as disjoint community detection and overlapping community detection. In the next part, it studies the graph partitioning based algorithm for community detection. Then discusses the community detection in the point of hierarchical algorithm. Moreover, the traditional community detection algorithms are combined with quality function, modularity to get more effective community results. Therefore, the community detection using modularity approach are also learnt in this chapter.

Most of the previous CD algorithms have scalability and time complexity challenges for large scale network. The exact community detection result can not get for large scale network and algorithms can produce optimal result. The problem can not solve polynomial time. Therefore, community detection become NP-hard problem. So, researcher proposed natured-inspired based optimization algorithm for uncovering community in large social network. Thus, nature inspired based community detection algorithms problem are also discussed.

Various kind of scoring functions for community detection is discussed in one parts of this chapter. Researcher can use different kinds of evaluation function consider on the community detection domain. The final section reviews the different application areas that use the effective community structure information to get their valuable information.

CHAPTER 3
THEORETICAL BACKGROUND

Many real-world networks are represented as a graph structure. One of the most important cases of a graph analysis is the structure or grouping of the community. Detecting the communities is a great important for computer science, biology, sociology, etc., in which these systems are represented as graph. The communities can be viewed as a summary of the entire network, and it is easy to understand.

Community means a collection of nodes in a group that have many edges. They are connected to the same set of nodes and relatively small number of edges connected to the nodes of the other groups. Finding community structure in a network can be viewed as a relatively independent graph partition. When developing a community detection algorithm in social network, it needs to consider two aspects such as effectiveness and efficiency. The effectiveness of the algorithm means all the steps necessary to obtain the result must be feasible with the available resources. Efficiency is gained by using distance functions and clustering algorithm that can analyze particular community structure in an efficient way. Many clustering algorithms have been proposed by considering these aspects, which can be widely categorized in two parts: data adaptation and algorithm adaptation. This chapter describes the background theory of community details, graph based representation, community detection framework for nature inspired algorithm, famous Artificial Bee Colony (ABC) algorithm for solving optimization problem, quality metrics to measure the result community quality and theoretical background of Spark architecture.

3.1 Graphs

Graphs used to model pair wise relationships between objects. Graph is a collection of nodes and edges. Many real-world networks can be constructed as graphs structure. Person or organization may be node and their connection may be edges. Social network is shown as a graph structure with an undirected and unweighted graph structure that means there is no distinction between the two nodes associated with each edge.

This section introduces about graph notation, type of graphs and property graph. A graph structure is indicated by G= (V, E) in which a graph G is composed by V and E. The notation V = $\{v_1, v_2, \ldots, v_n\}$ are the set of nodes or vertices in the graph G and the notation $E=\{(v_i,v_j) \mid vi,v_j \in V\}$ are the connection or links between nodes. Graph may be undirected or directed or unweighted or weighted. The notation of graph structure can be varied based on the graph nature. An undirected graph is a graph in which all the edges have bidirectional that means edges are unordered pairs of vertices. In directed graph, all the edges point to single direction. A weighted graph represents that each edge in the graph is assigned a weight or cost but unweighted graph does not include this value. Figure 3.1 describes four different types of graph structure as an example.

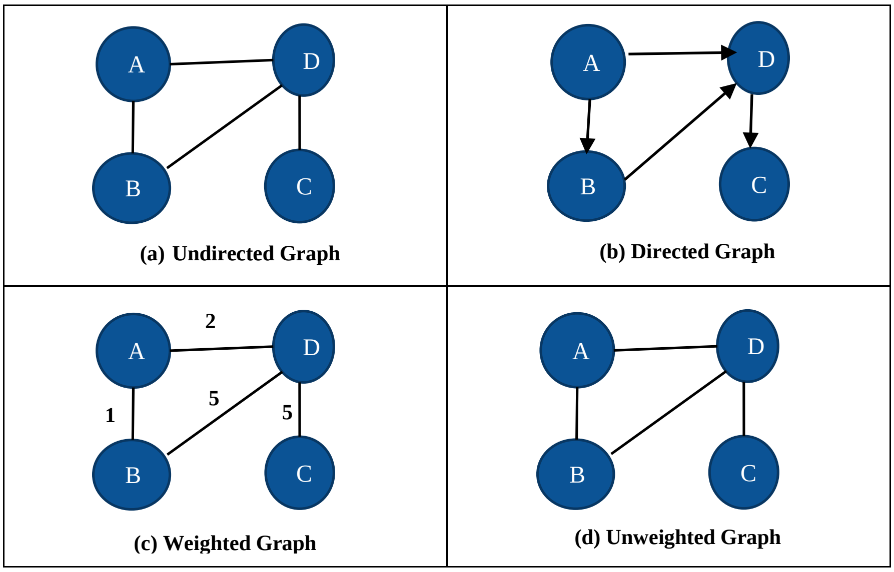


Figure 3. 1 Different Types of Graphs

3.1.1 Graph Representation

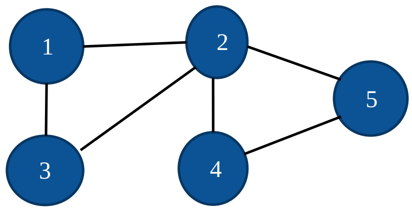
Graph representation can give many different perspectives and can make a problem much similar such as more accurately and providing the right tools to solve problem. The graph contains two components such as nodes and edges. These components have natural correspondences to problem elements. In the theory of graph, a graph representation is a way to store graph into the community memory. To represent a graph, it needs a set of nodes and its neighbors that are vertices directly connected by edges. For weighted graphs, weights are associated with each edge. There are several ways of graph representation that depend on the density of the graph edges, the type of operations and ease of use [66]. The most useful graph representations are adjacency list and matrix.

Adjacency matrix is used to represent which nodes are adjacent to each other and it is a sequential representation. There are edges that connect the nodes to the graph.

This representation uses n-by-n matrix N. If one node i to node j have a connection, the corresponding element N, N_{ij}=1, otherwise N_{ij}=0.The adjacency matrix use Boolean vales 0 or 1 (means adjacent or not). Figure 3.2 demonstrates the adjacency matrix for graph representation with undirected graph structure.

Nodes	**1**	**2**	**3**	**4**	**5**
1	0	1	1	0	0
2	1	0	1	1	1
3	1	1	0	0	0
4	0	1	0	0	1
5	0	1	0	1	0

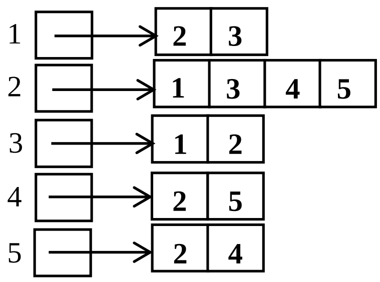
(a) Adjacency Matrix



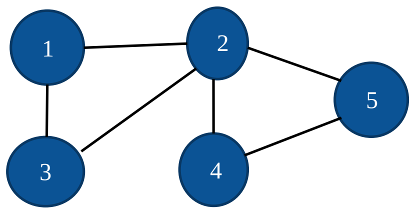
(b) Adjacency Matrix Graph Structure

Figure 3. 2 Adjacency Matrix-based Graph Representation

Adjacency list based on a linked representation. In this representation, the graph contains each vertex with a list of its neighbors. For each vertex i, store and vertices array adjacent to it. So, ever vertex in the graph contains a list of adjacency vertices. Vertices in an array is indexed by vertex number. The corresponding array element points to singly linked list, close to v. Figure 3.3 shows an adjacent list representation of the undirected social network graph.



(a) Adjacency List



(b) Adjacency List-based Graph Structure

Figure 3.3 Adjacency List-based Graph Representation

The information acquired from social network depends on the representation method used and the type of analysis required to be carried out in a certain social network. Figure 3.4 shows the existing popular product for Graph System. According to the figure, neo4j, TITAN and jena can be used for graph storage, GIRAPH, Dato, GraphX, Flink/Gelly are frameworks for graph processing, boost, ScaleGraph and blazegraph can be applied for analytics and Gephi, graphviz and D3 are visualization products.

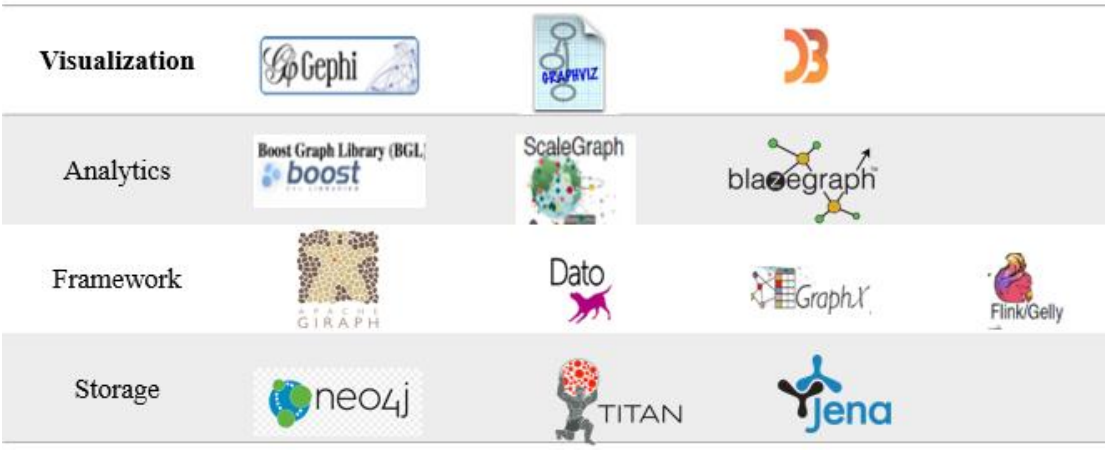


Figure 3.4 The Existing Products for Graph System

3.1.2 Node, Its Degree and Neighborhood of Network

In the network G, the degree of a node i is the number of nodes adjacent to i. If a node that has many neighbor nodes, it will be the important node for a given network. If the two vertices v and u have a connection, they will be neighbor. Let N_i be the neighborhoods set of vertices i in a graph that is the set of vertices in which these set of vertices are directly connected to i with a link.

Table 3.1 Node, Its Degree and Neighborhood of Sample Network

Node	Neighbor Node	Degree
1	2, 3	2
2	1,3,4,5	4
3	1,2	2
4	2,5	2
5	2,4	2

If two nodes have many common neighbors, they are more likely to share a link in the future [56]. This will be using the number of common neighbors as a pair wise proximity measure. The following Table 3.1 describes the sample node degree and its neighborhood for Figure 3.2 (b). The three columns describe the node name as **Node**, neighborhood of each node as **Neighbor Node** and degree of each node as **Degree**.

3.2 Community Detection in Social Network

Today, the analysis of social networks has become one of the key areas of research. Community detection is one kind of social network analysis tasks. Detecting of social network community has become a key issue for effective advertisements, how to track change and recommender systems. The community structure of social network has real world communities and virtual communities. The group of users with social, economic and political interests is the real world and a community created on social sites is a virtual community. The emergence of communities in the graph structure of social networks is reflected in this system.

The community is essential to the network and is vital for both exploring networks and predicting connections that have not yet been observed. Social network clustering structure is shown in figure 3.5. The simple social network contains two community structure C = {C1, C2}. Node A, B, C and D are in the same community C1 and node E, F, G reach in the same community C2. A graph or network is one of the most common and intuitive technique to represent the individual or objects from a dataset and also includes their interactions and relationships. Communities means group of densely connected nodes, while the relationships between communities are sparse.

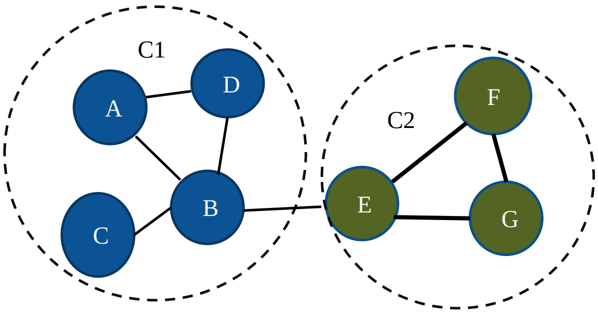


Figure 3.5 Community Structure of Simple Social Network

Discovering community structure is a process that discover the group structure of vertices in a given network and helps to better understand the architecture of the network. Estimating all clustering probabilities is a difficult NP-hard problem. Therefore, this is not an easy task and many algorithms were proposed using different strategies and used different kinds of objective functions. Finally, the result communities' structure can be measured using evaluation metrics and these metrics are discussed in the next part of this chapter.

3.3 Framework using Nature Inspired-Based Community Detection Algorithm

Many different approaches and algorithms are proposed to analyze the community structure of complex real world network. These algorithms use techniques, method and principles in physics, graph theory, mathematics and artificial intelligence. These algorithms are not satisfactorily solved the community detection problem in some large networks. The community detection in social network has become an optimization problem.

Many heuristic search algorithms are applied to solve the optimization problem. The nature-inspired based algorithm is a heuristic method that is based on the optimization of quality metrics. Therefore, researcher applied various nature-based heuristic algorithms to evolve optimal communities for social network and then showed to outperform the previous community detection method in terms of computation times [2]. Nature inspired-based algorithm is population-based optimization algorithm. It can get acceptable solution when solving many real life complex problems. Figure 3.6 shows the step by step procedure of the nature inspired-based algorithm

Algorithm 3.1 : Nature Inspired-based Algorithm

1. Initialization all individuals in the population or search agents
2. Evaluate the population and find the best solution
3. Repeat until the stopping criteria is met
    i. Compute the fitness function for all solutions in population or search agents
    ii. Find the best optimal fitness value
    iii. Calculate the new fitness value obtained against the previous one
    iv. If the new fitness value is better than the old one, then replaces the fitness value with the newly obtained value
    v. Return the best fitness value
4. Exit

Figure 3. 6 Pseudo Code of Nature Inspired-based Algorithm

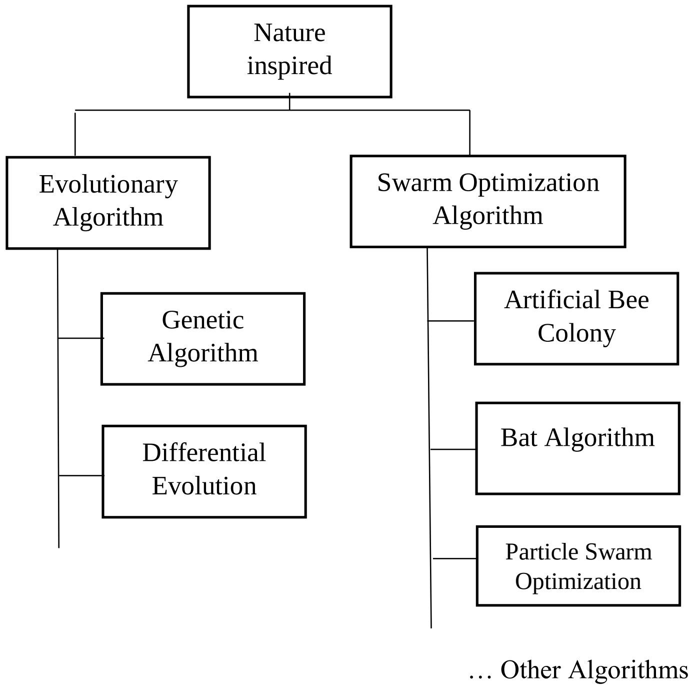


Figure 3.7 Categories of Nature Inspired-Based Algorithm.

There are two classes of nature inspired algorithm such as evolutionary algorithms and swarm optimization algorithms. The hierarchical structure of nature-based algorithm is described in Figure 3.7. Evolutionary based algorithms use crossover and mutation process that enables the solution to the problem to be found. Swarm-based algorithm observes movement or search strategies to optimize a population. The complexity of framework depends on optimization algorithm's complexity and quality function for the application domain. This session presents a framework to apply heuristic search optimization strategies to detect community by using various type of nature inspired algorithms. This framework composed with four main parts such as **analysis, initialization, evolution,** and **result generation phases**.

Analysis, initialization and result generation phases process are similar in all proposed algorithm. Evolution phase is modified for developing the evaluation of new proposed nature based algorithm or existing algorithm to detect community structure in social networks. Phases in this framework are independent each other by modifying only one evolution phase which makes the design flexible to test different algorithm. Researcher used this framework for community detection problem as the flexible platform. The Figure 3.8 shows the overview architecture of each phase and support modules [3]. Its supports module contains common functions such as quality metrics, cluster form and genetic operator such as crossover and mutation can be used to support the execution of these phases.

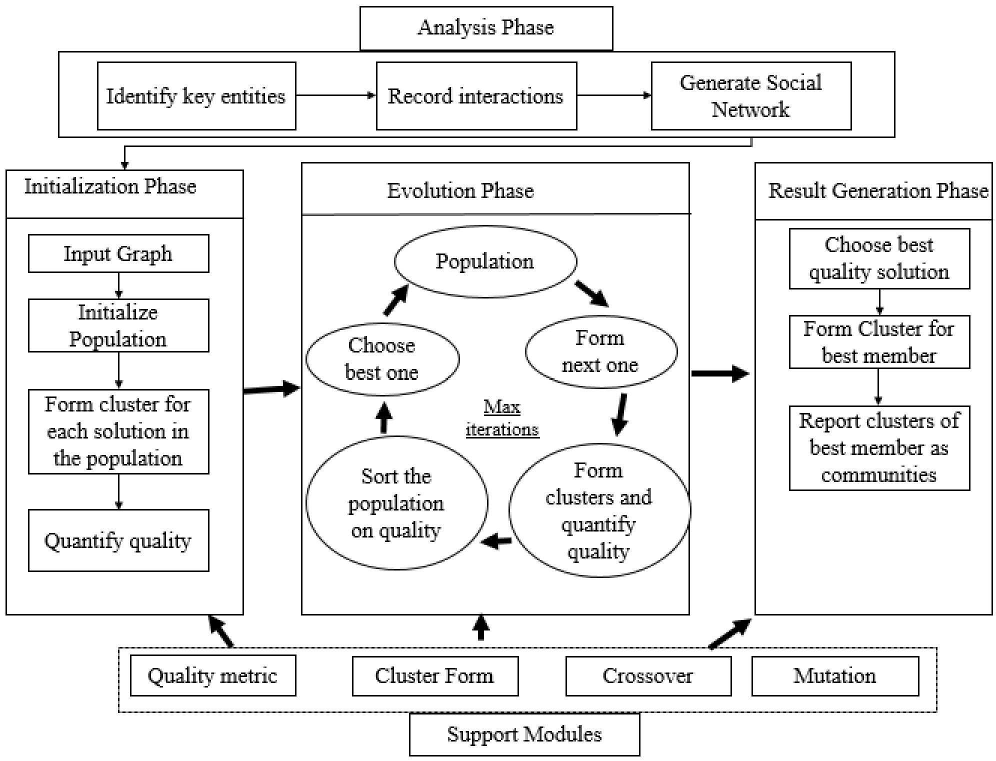


Figure 3.8 Framework for Community Detection using Nature-based Algorithm

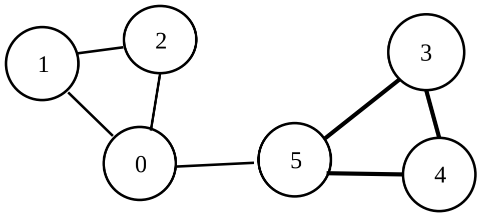
3.3.1 Analysis Phase

In this phase, the relationships and main entities in the organization are searched. Main entities identification done for creating social graphs and collecting the need of organization for future evaluation. After finishing the key entities identification, their relationships are analysis to construct social graph. These will be recorded and tracked to create the relationship matrices, which are constructed into a social network graph. Figure 3.9 shows the example of creation of a social network in which six actors in an organization. Adjacency matrix is used for social network graph representation. In the matrix position N_{ij} (describes in graph representation section) stores the number of interactions between node i and node j.

The diagonal matrix will be zero because there is no individual connection to himself. When creating graph, the total interaction between node i and node j exceeds acceptable distance (this example defines 4 for demonstration). Here, the social interaction graph shows as an undirected unweighted graph for simplicity. Therefore, the entries in the matrix of above triangular matrix and the related things in the lower triangular matrix are average. If the value between the two nodes is larger than the thread limit, the edge will be create between the two nodes.

Nodes	0	1	2	3	4	5
0		6	5	2	1	4
1	4		4	3	3	1
2	5	6		3	2	2
3	3	3	3		8	7
4	0	1	2	6		5
5	9	3	3	4	5	

(a) Social Interaction Matrix



(b) Matrix Based Social Graph

Figure 3.9 Example of Creating Social Graph

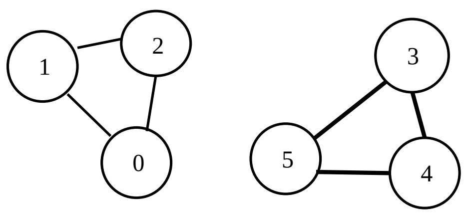
3.3.2 Initialization Phase

In this phase, a set of randomly initialized individual that represent the popular generations that will be optimized in the next phase. The analysis phase interaction graph is delivered as the input of this phase. For the number of nodes V and the set of edges E in graph G = {V, E}, the set of individual solution in population are generated by using various type of initialization strategies. Disjoint community detection strategy (node based initialization) is presented in this system. For community detection problem, there are two famous type of population initialization such as locus-based solution representation and label-based solution representation. In two schemas, a vector of V positions represents individuals.

In locus based initialization, individual vector is represented as X = {X_1, X_2, X_3, …, X_N}. Each i^{th} vector in X is composed as X_i = { x[0], x[1], …., x[v]}, in which v is equal to the number of nodes. Figure 3.10 describes the locus based initialization example. Each index position in the vector and its corresponding value is nodes of the corresponding value. A value x [4] =3 means that node 4 and node 3 have a connection or edge. This value is chosen randomly from the neighbors set of node 4. After initialization, it decodes the result cluster. Get two node clusters 0-1-2 and 3-4-5.

Node	0	1	2	3	4	5
value	1	0	1	4	3	4

(a) Node vector



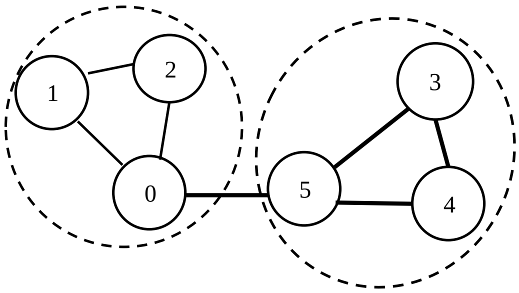
(b) Communities of nodes

Figure 3.10 Locus-based Initialization

In label based initialization, individual vector is represented as $X = \{X_1, X_2, X_3, \ldots, X_N\}$ and each i^{th} vector can also be described as $X_i = \{x[0], x[1], \ldots., x[v]\}$. Each vector position is an integer value in the range $\{1,2,3\ldots, k\}$ in which k means the number of communities for the network. Each j^{th} value of vector array specifies the cluster label they belong. If the index values of array are same, these nodes reach in the same cluster. Figure 3.11 describes the label based initialization example. In which node 0, 1 and 2 are in the same community label and node 3, 4 and 5 reach in the same community.

Node	0	1	2	3	4	5
value	1	1	1	3	3	3

(a) Node Vector



(b) Community of Nodes

Figure 3.11 Label-based Node Initialization

The proposed nature inspired based community detection algorithm can apply various initialization strategies depending on the community requirements of the application domain. After initializing population, the quality of result cluster of each solution can be calculated by using various kind of evaluation metrics. The evaluation metrics are discussed in the next sections. A better quality of community result means that each subgraph has strong intra connection and weak inter connection to other sub graphs. For further optimization, this is given to evolution phase through different refining methods and selections.

3.3.3 Evolution Phase

After getting the population from the initialization phase, the new or existing nature inspired based algorithm is used to get a certain number of generations of the cluster results. To evolve the fitness quality of each individual for next generation, individual solutions are chosen to perform the operations such as crossover, mutation or search space. The operators can differ based on the optimization strategy of the proposed algorithm and it helps to improve the individual vector's fitness. Evolutionary based algorithm work crossover and mutation and swarm based algorithm use space search. If the vectors contain the better quality result, they will choose for next generation. Individuals are generated using repetitive refinement with different operator that improves the fitness of communities and helps in evolving next generation.

3.3.4 Result Generation

After evaluating the maximum iteration for evolution population by creating a new population, the clusters result with the best fitness value in the final generation is selected as an output.

The above framework can be abstracted for testing and evaluating new nature-based algorithms of varied complexities for community detection problem. The quantitative community result relies on the efficiency of solution initialization and evolving strategies during evolution.

Various kind of nature inspired algorithm have been used to evolve optimal communities in many real life complex network. Most of applied algorithm for community detection are Particle Swarm Optimization, Group Search Optimization, Genetic algorithm, etc. It is used to improve the convergence and optimality of communities.

3.4 Artificial Bee Colony Algorithm

Since the last two decades, analyzing the behavior of the swarm system to describe new intelligent approaches has been the ongoing research for researches. Swarm intelligence algorithm use the collective behavior of animals or inserts such as bat, fish, ant, lion and etc. The researchers tend to solve the optimization problem and decide to apply many heuristics based search algorithm. Among them, Artificial Bee Colony (ABC) algorithm is one of the most widely applied algorithms to solve numerical optimization problem. It is a swarm based metaheuristic algorithm proposed by Karaboga in 2005 [15].

This algorithm simulates foraging behavior of honey bee in nature. Many numerical optimization problems have been solved by ABC algorithm such as clustering, energy minimization, routing, process scheduling, travelling salesman problem and other various optimization process. This method became popular and it can widely use, because of its good convergence properties. ABC algorithm consists of three set of bees such as employed bees, onlooker bees and scout bees. The main steps of the artificial bee colony algorithm are described in Algorithm3.2 that are described in figure 3.12.

Algorithm 3. 2: Artificial Bee Colony (ABC) Algorithm
1: Initialization
2: Evaluation
3: Repeat until maximum cycle number
i. Employed bee phase
ii. Calculate probability for onlooker
iii. Onlooker bee phase
iv. Memorize the best solution
v. Scout bee Phase
4. Exit

Figure 3. 12 Artificial Bee Colony Algorithm

Firstly, ABC algorithm produces a randomly distributed initial population (the set of solutions Ns). Ns is also the number of food sources and each food source means the possible solution for the optimization problem. Each solution in SN has D dimensional vector, in which D is the number of optimization parameter. Then evaluate the quality of each food source using one kind of objective function. The quality of food source means the fitness value of each solution. After initialization, the population is exposed to an iteratively maximum cycle number of two major steps such updating new solution in three bee phases and select the best solution in each iteration to get optimal solution.

In the employed bee phases, each employed bee chooses a new food source based on the neighborhood of previous food source position. It generates a change to the position in its memory and checks the quality of the candidate source. Then employed bee memorized the higher quality food source. If new food source's is better than that of old one, the old food source is replaced by new candidate food source. After completing the search process, all employed bees share the fitness value their new food sources to the onlooker bees on the dance area. The new food source is found by the following equation 3.1.

$$v_{ij} = x_{ij} + \emptyset_{ij}(x_{ij} - x_{kj})$$

Equation 3.1

where v_{ij} is a new food source modified by the previous food source x_{ij} , x_{kj} is the randomly selected position from neighbour food source, indexes k and j are chosen randomly k ∈ {1, 2,..., SN} and j ∈ {1, 2,...,D}, k is determined randomly (k != i), φ_{ij} is a random number between [-1, 1]. The fitness value of each food source is essential for finding the global optimal. After selection the candidate food sources by using greedy selection, the fitness function is calculated by using Equation 3.2.

$$\mathrm{fit_i} = \begin{cases} \dfrac{1}{1+\mathrm{f_i}} & , \quad \text{if } \mathrm{f_i} \geq 0 \\ 1 + \mathrm{abs(f_i)} & , \quad \text{if } \mathrm{f_i} < 0 \end{cases}$$

Equation 3.2

where fit_i is fitness value of i^{th}food source and f_i is the objective function value which is defined according to the problem.

An artificial onlookers use probability based selection process to select relevant food sources. If the food source has the higher nectar amount, probability of a food source for selected by onlooker bees will increase. The quality of food source is evaluated based on Equation 3.3 probability function.

$$\mathrm{P_i} = \frac{\mathrm{fit_i}}{\sum_{\mathrm{n-1}}^{\mathrm{SN}} \mathrm{fit_n}}$$

Equation 3.3

Where P_i is the probability of food source i, fit_i is the fitness value of the food source i, that is calculated in equation 3.3. After the onlooker bee selects a food source, it finds a new candidate food source from a randomly search in the neighborhood of food source using equation 3.1. Then, the algorithm memorizes the best food source in all food source.

In scout bee process, if the fitness value of the food source position cannot improve or reach predefined limit, it will be abandoned and replaced by a new food source position. So, scout bee randomly searches new food source (solution) by using the Equation 3.4.

$$\mathrm{x_i^j} = \mathrm{x_{min}^j} + \mathrm{rand}(0{,}1)(\mathrm{x_{max}^j} - \mathrm{x_{min}^j})$$

Equation 3.4

Where $x_{max}^{j}$is upper bound of food source position in dimension j, x_{min}^{j}lower bound of food source position in dimension j and rand (0,1) is a random number in the range [0,1]. The number of iterations is controlled by using maximum cycle number.

The process will repeat until the output of the objective reaches a defined threshold value or the number of iterations equals to maximum cycle number.

In ABC, each food source represents a possible solution for optimization problem and the nectar amount of a food source correlate with fitness value of each solution. The number of employed bees is equal to the number of food source and there is only each employed bee on each food source. Employed bees and onlooker bees performed local selection process called greedy selection process. In this process, the nectar amount of the candidate source is better than the old one, the bee forgets the current food source and memorizes the candidate food source. Else the bee remains the present one in the memory. A random selection process is performed by scout bees. In this process employed bees and onlooker bees carry out the exploitation process in the search space and scouts control the exploration process. Figure 3.13 shows the flow diagram of the artificial bee colony algorithm.

Comparative studies indicated that ABC performance is compared with other population based algorithm such as Particle swarm optimization, genetic algorithm (GA) and differential evolution algorithm (DE). The general idea of ABC begins with random solutions and then attempts to find better solution for searching the neighborhoods of the current best solutions and abandoning unpromising solutions. Advantage of ABC algorithm is that it has few control parameters than other nature inspired algorithms. It has population size, limit and maximum cycle number MCN. ABC is the fast convergence speed, simple, flexible, robust, contains both exploration and exploitation and it can easily be combined with other type of optimization algorithms. The disadvantage is the limitation of search space by initial solution because normal distribution should use in initialize step.

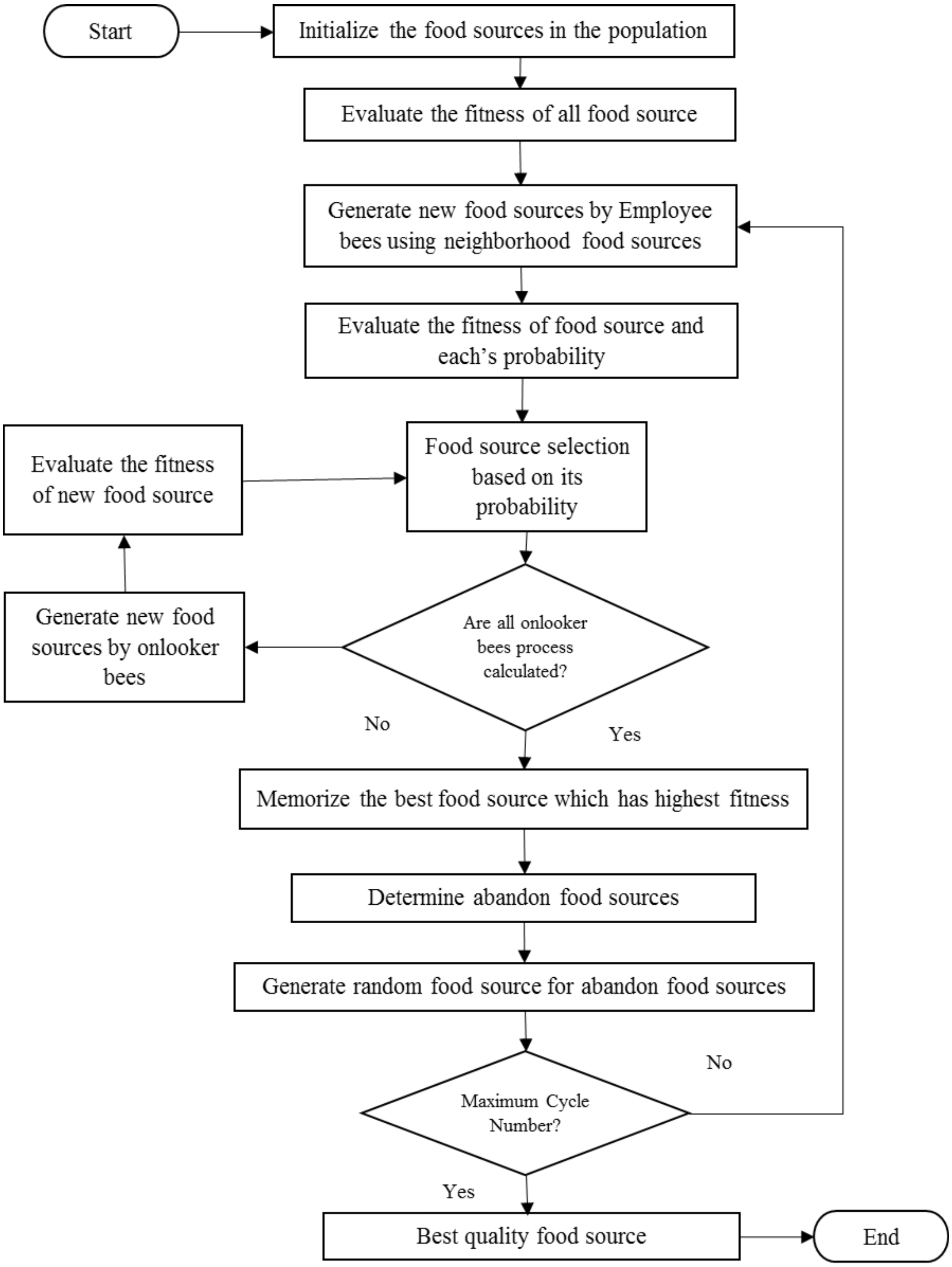


Figure 3.13 Flow Diagram of the Artificial Bee Colony Algorihtm

3.5 Evaluating of Community Detection Methods

Detecting community is a fundamental task of network science that finds to describe the giant networks structure by partitioning its vertices into communities based on their connection to partners. It is similar to the graph clustering and both attempts to identify meaningful groups within a dataset. After getting the result of communities, it needs to map with the set nodes back to the real world to see whether they appear to make intuitive sense whether it is a possible social community. For a given network, different algorithms usually give different results. Thus, evaluating the performance and finding the best one is very important for these algorithms.

Generally, when evaluating the final result of community detection algorithm, there are two main parts. The first thing is to make sure the output results are accurate. The next part is to assess the performance in term of processing time and memory usage. For community detection problem, it is not clear the first evaluating because of the nature of problem and dataset under study. So, most of the evaluation measure is used to compare the goodness of its output. The testing strategies of community detection algorithms depend on the following cases:

1. The problem is that there is no knowledge of the output result and therefore use quality measurements to evaluate methods.

2. If unable to find similarity, possible measurements are used to adjust similarities about how the output result of that method are possible.

3. There may be more than one solution, similarity and quality measure can be used in evaluation process.

The following section discusses the evaluation metrics to test the quality of community detection methods.

3.5.1 Evaluation Metrics

A good quality measure of a community needs to be able to distinct a community such as a group of nodes in a network from just randomly connected a group of nodes. When assessing the quality of detected communities, if there is no ground truth (GT) for communities, it will use quality function that characterizes how community like is the connectivity structure of a given set of nodes. All metrics are based on the intuition that communities are set of nodes with many connections to the cluster and few connections to the rest of the other clusters.

In quality measurement, it considers two important characteristics of community such as internal edges connectivity and external edges connectivity. Internal edge connectivity represents the number of connections between nodes in a community and this value should be high in order the group to be a good community structure. External edge connectivity means the number of edges where node in the cluster is connected to the other external clusters and this measure should be minimized be become a good quality community structure [7]. Most of these evaluation methods based on four kinds of quality function such as internal connection, external connection, both internal and external connection or model of network. These metrics are discussed in the following parts.

Symbol meanings of evaluation metrics are described as follows:

- For an undirected graph G =(V,E) with $n = |V|$ nodes, the set of nodes n and $m = |E|$ edges, the set of edges e.
- C={S_1,S_2,…,S_g) is the network community structure that contain g number of communities, S_i be the a community selected from C
- n_s is the count of nodes in S, n_s = |S|, m_s is the count of edges in S, $m_s = |\{(u, v) \in E: u \in S, v \in S\}|$;
- c_s, the count of edges on the boundary of S, $c_s = |\{(u, v) \in E: u \in S, v \notin S\}|$;
- $d(u)$ is the degree of node u;
- f (S) is the quality measure function that estimate the quality of given community S.

Metrics on Internal Connectivity

Internal density is one method of internal connection-based method. This method defines the amount of ratio for all edges in network graph and the intra community interactions. One of the simplest measures for community quality that is biased towards coarse-grained communities. Equation 3.5 is the equation of internal density metric.

$$f(S) = \frac{m_s}{n_s\,(n_s - 1)/2}$$

Equation 3.5

The larger amount of f (S) means better for internal density and this amount measures the edges fraction that appears among the nodes in S.

Edges Inside is the count of edges between the members of S. Edge Inside equation is described in Equation 3.6.

$$f(S) = m_s$$

Equation 3.6

Average Degree is the members' internal degree average of S. Its equation is expressed in Equation 3.7.

$$f(S) = \frac{2m_s}{n_s}$$

Equation 3.7

Fraction Over Median Degree (FOMD) is the nodes fraction of S that contain internal degree larger than dm, where dm means the value of median value $d(u)$ in V, where a function $f(S)$ is the notion of quality of a network community S, n_s is the number of nodes in S, $n_s = |S|$; $d(u)$ is the degree of node u and m_s the number of edges in S.

$$f(S) = \frac{|\{u: u \in S, |\{(u, v): v \in S\}| > dm\}|}{n_s}$$

Equation 3.8

Triangle Participation Ratio (TPR) is the nodes fraction in S that possess to a triad, where $f(S)$ function characterizes that like the community is the nodes connectivity in S.

$$f(S) = \frac{|\{u: u \in S, \{(v, w): v, w \in S, (u, v) \in E, (u, w) \in E, (v, w) \in E\}6 = 0\}|}{n_s}$$

Equation 3.9

Cluster coefficient (CC) purposes on the community members' tendency to cluster together. So, high coefficient of clustering defines high probability that the connections in the observed communities are same. An undirected network's global clustering coefficient can be calculated as

$$CC = \frac{1}{cn}\sum_{i=1}^{cn}\left(\frac{1}{|C_i|}\sum_{v \in C_i}\frac{2|\{e_{ts}: v_t, v_s \in N(v) \cap C_{i,e_{ts}} \in E\}|}{d(v)(d(v)-1)}\right)$$

Equation 3.10

Where $N(v)$ consists of the neighbors of node$v$, $d(v)$ is the degree of$v$, cn is the quantity of community, and C_i is the i^{th} community.

Metrics on External Connectivity

Cut Ratio is external connectivity based function. The fraction of existing edges leave from the cluster is defined as

$$f(S) = \frac{c_s}{n_s\,(n - n_s)}$$

Equation 3.11

The smaller value of external density is the better for the external connectivity.

Expansion computes the count of edges per node that point outside the cluster. Lower value of expansion has higher quality function. Equation 3.12 is the equation of expansion.

$$f(S) = \frac{c_s}{n_s}$$

Equation 3.12

Metrics on Internal and External Connectivity

Conductance is the internal and external based metrics. This function is used to measures the ratio of the total number of edges of S pointing to the other external communities and the total number of edges connected with S. Its equation is defined as

$$f(S) = \frac{c_s}{2m_s + c_s}$$

Equation 3.13

Better communities should have a lower conductance value.

Normalized Cut:

$$f(S) = \frac{c_s}{2m_s + c_s} + \frac{c_s}{2(m - m_s) + c_s}$$

Equation 3.14

Normalized Cut measures the fraction of total edge volume that points outside the cluster and the number of nodes which is the number of edges on the boundary of S. The Normalized Cut can be calculate by using Equation 3.14.

Maximum-ODF (Out Degree Fraction):

$$f(S) = \max_{u \in s} \frac{|\{(u, v) \in E : v \notin S\}|}{d(u)}$$

Equation 3.15

Out Degree Fraction is the edges' maximum fraction of a node in S that focus outside S. Equation 3.15 is the equation of Maximum-ODF .

Average-ODF:

$$f(S) = \frac{1}{n_s} \sum_{u \in s} \frac{|\{(u, v) \in E : v \notin S\}|}{d(u)}$$

Equation 3.16

Average-ODF means the edges' average fraction of nodes in S that focus out of S. Average-ODF can be calculated by using Equation 3.16.

Flake-ODF:

$$f(S) = \frac{|\{u : u \in S, |\{(u, v) \in E : v \in S\}| < d(u)/2\}|}{n_s}$$

Equation 3.17

Flake-ODF means the nodes' fraction in S that consist fewer edges focusing inside than to the outside of the cluster. Flake-ODF can be determined by using Equation 3.17.

Network Model

Modularity (Q) is network model based evaluation metrics. Modularization measures the strength of each partition taking into account the degree distribution and defined as having many internal edges and few external edges. Equation 3.18 is the equation of modularity function.

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{d_i d_j}{2m}] \delta(c_i, c_j)$$

Equation 3.18

Where m is the number of edges in the graph, d_i is the degree of vertex i, d_j is the degree of vertex j. c_i is the community to vertex i, c_j is the community to vertex. δ Function is 1 if i and j belong to the same community, otherwise it is equals to 0.

Modularity measures the number of relationships between the probability of having an edge joining two sites and the fact that the sites are in same community. The modularity value is a scalar value, its range is between -1 and 1. It can be used to compare the partitions quality obtained from various community detection methods. It is also as an objective quality measure for each split of a network into communities.

Accuracy Metrics

Another strategy is to compare the method output with a ground truth of the input network using similarity measure. Some datasets have ground truth result. In which case the accuracy of result community structures quality can be measure using normalize mutual information function.

The accuracy of the proposed method is calculated in specification of normalized mutual information (NMI) applied on datasets with ground-truth by determining to what extent the communities distributed by the proposed method are constant with the real ones.

Normalize mutual information (NMI) is validation based quality metric. Calculating the network partitions' quality has become an important issue if the given network has truth community result. NMI can be widely used to measure the accuracy of community detection result. NMI calculates the similarity between actual partitions and detected partitions. To calculate the performance of community detection algorithm by their ability to observe so called ground truth communities. A is the real partition of given network A = $\{A_1, A_2, \ldots, A_r\}$ and B is the output partition of an proposed algorithm B = $\{B_1, B_2, \ldots, B_n\}$, r and n are community number of A and B. Normalized mutual information NMI (A, B) can be measure using equation 3.19.

$$NMI(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} C_{ij}\log\left(\frac{C_{ij}N}{C_iC_j}\right)}{\sum_{i=1}^{C_A} C_i\log\left(\frac{C_i}{N}\right) + \sum_{j=1}^{C_B} C_j\log\left(\frac{C_j}{N}\right)}$$

Equation 3.19

Where, N is the total number of vertices, a confusion matrix is C, the count of communities in partition A is C_A and the number of communities in partition B is C_B, C_{ij} is the set of nodes present in community i of A (community j of B), C_i is the count of members in row i (or column j).

If the true similarity and detected communities contain the larger value of NMI, the result communities will achieve the better performance quality result. For truth and detected communities being equal, the NMI (A, B) will be 1 and if they are different communities, NMI (A, B) will be 0. These metrics not only used to perform clustering quality but are also used as objective functions in optimization based clustering methods.

3.6 Architecture of Apache Spark

Apache Spark is an open source cluster computing system that is growing rapidly. The ecosystem of libraries and frameworks that enable advanced data analysis is expanding [25]. It provides a faster runtime than map reduce based analytics and provide in-memory distributed computing. There have java, Scala, Python and R APIs. The ecosystem of Apache Spark is show in Figure 3.14.

Spark Core: Spark Core is the heart of Spark and does the core functionality. It is responsible for task scheduling, storage systems interaction, fault recovery, and memory management.

Spark SQL: The Spark SQL is a new module in Spark for structure data processing. It supports JDBC and ODBC connections that establish relationships between Java objects and existing databases, data warehouses and business intelligence tools. It also supports variety of data sources such as JSON, Hive tables and Parquet.

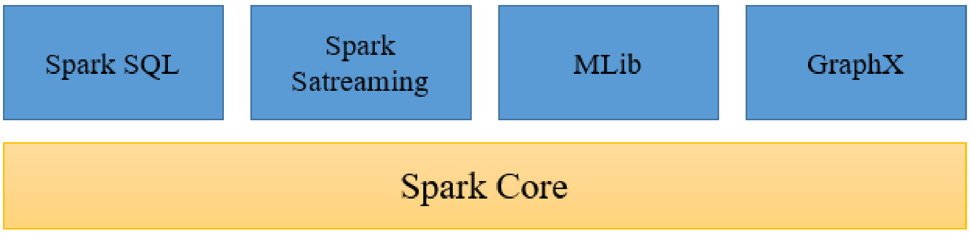


Figure 3.14 Spark Ecosystem

Spark Streaming: It is a Spark component that supports scalable and fault-tolerant processing of streaming data. It performs streaming analytics using Spark Core's fast scheduling capability. It accepts data in mini-batches and performs RDD transformations on that data. Its design allows you to reuse application designed for streaming data to analyze batches of historical data with minor changes. The log files created by web servers can be considered as a real-time data stream.

MLib: The MLib is Spark’s machine learning library with various machine-learning algorithms. It consists algorithms and utilities such as classification, clustering regression, dimensionality reduction and other analyzing components. It is nine times faster than using Apache Mahout.

GraphX: The GraphX is a library use to manage graphs and perform graph-parallel calculations. GraphX facilitates to create a directed graph with arbitrary properties attached to each node and edge. It supports various fundamental operators to manipulate graph like subgraph, join Vertices, and aggregate Messages [60].

Apache Spark is designed on two main abstractions called resilient Distributed Dataset RDD and Directed Acyclic Graph DAG. It can serialize RDD structures for storage on a hard drive. Apache Spark, a distributed data processing framework, can be used to solve the problem of iterative processing by providing efficient computations on Resilient Distributed Datasets (RDDs). It follows the master slave architecture. Its cluster consists of a single master and multiple slaves. The basic architecture of Apache spark is shown in Figure 3.15.

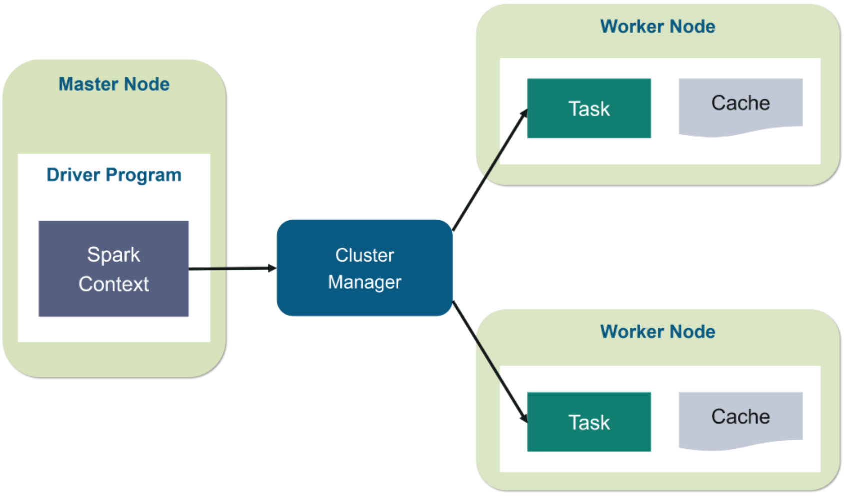


Figure 3.15 Apache Spark Architecture

Spark Driver is called main program and it creates Spark Context. A Spark Context includes all the basic functionalities. It has the various components responsible for translating the user written code into jobs that are performed on the cluster. Spark Driver and Spark Context watch the job execution within the cluster. Spark Driver works with the cluster manager to manage various other jobs and Cluster Manager does the resource allocation work. Then, the job is split into multiple smaller tasks which are further distributed to worker nodes.

Whenever an RDD is created in the Spark Context, it can be distributed across many worker nodes and can also be cached there. Worker nodes called slave nodes perform the tasks assigned by the Cluster Manager and return the Spark Context. An executor is responsible for the execution of these tasks. The lifetime of executors is the same as that of the Spark Application. If the number of workers is increased, then the job can be divided into more partitions and execute several tasks. It will be a lot faster [65].

3.6.1 Spark GraphX

GraphX provides a platform for implementing a graph processing system using the underlying Spark environment. It reuses the features of Spark Core, it is a component that give primary functions like scheduling, distribution of data and monitoring. The basic abstraction provided by Spark is RDD (Resilient Distributed Dataset). It is a partitioned data structure that is processed by the computing primitives provided by Spark Core. Each Spark application can be developed in a cluster. The application is running on a logically dedicated driver node and typically many numbers of worker nodes that process tasks in parallel [60].

GraphX extends the Spark RDD by presenting a new graph abstraction: a directed multigraph with properties attached to each vertex and edge. GraphX realized the distributed storage and graph processing efficiently due to the inherent nature of RDD and it could be applied to the large-scale graph computing. Compare to the MapReduce, GraphX improves and optimizes the performance of computing significantly. The optimization in storage of vertex and edge information make it close to or reach the performance of professional graph platform like GraphLab. Graph abstraction in Spark is as follows;

```
abstract class Graph(V,E){
val vertices: vertexRDD[V]
val edges: EdgeRDD[E}
val triplets: RDD[EdgeTriplet[VD,ED]].
```

3.7 Summary

In the era of big data, large real world social networks are currently available to make the issue of the efficiency of community detection algorithm. The basic concepts of community detection based on the graph and nature inspired-based framework have been highlighted in this chapter. Firstly, it discusses basic knowledge of graph. Moreover, the communities in the social network are also discussed. The various kinds of graph structure are also presented in this chapter. Continually, from the nature inspired based algorithm framework for community detection, four important phases are explained.

After that, this chapter also discusses what type of categories should be considered to evaluate the quality of community detection algorithms. Then community detection metrics are presented that are based on various kinds of connectivity. Then, NMI is discussed to test the accuracy of result community structure of the given network. If the proposed algorithm gets more accurate result in the ground truth social network, the guarantee can be obtained to test the proposed algorithm in various size of datasets. So various kind of measurements of community goodness metrics are also discussed in this chapter. Finally, the architecture of Spark for faster computation is presented and then to test the parallel graph processing for large network, spark GraphX is also discussed.

CHAPTER 4
THE PROPOSED SYSTEM ARCHITECTURE AND IMPLEMENTATION

The previous chapters describe the knowledge of community detection and some current trends and relevant methods to provide additional functionality to express the community detection in social network. The optimization algorithms based on population find near-optimal solutions to the difficult optimization problems by motivation from nature. In this chapter, different consideration of community detection algorithm based on algorithm of artificial bee is proposed. In this algorithm, label-based solution representation and enhance artificial bee colony operation for discrete nature problem are also mentioned.

Firstly, the proposed algorithm considers to get maximizing modularity value result communities. Population initialization is one kind of important phase for nature inspired based algorithm. Thus, novel population representation with label-based solution initialization considers the various perspectives. Then, discrete nature of artificial bee colony algorithm (D-ABC) is proposed for community detection problem. The backbone of this algorithm is imitated and then converted this algorithm to solve community detection problem. Because numerical comparison results demonstrated that artificial bee colony algorithm gets effective result than other nature inspired based algorithm. And then, it is a simple and easy implementation. One point crossover operation based initialization strategy is considered and added in the proposed discrete artificial bee colony algorithm (D-ABC). Spark framework is also used to implement this algorithm because the ABC algorithm is one kind of iterative algorithm.

4.1 Discrete Artificial Bee Colony Algorithm

The proposed system is based on the node connection of a given network in undirected and unweighted network. According to the definition of communities in social network that nodes in the same group have denser connection. Firstly, the edge lists dataset is accepted to create graph and then creating graph using GraphX in Spark. Secondly, finding the community structure using discrete artificial bee colony algorithm and the best community structure is produced finally. Figure 4.1 describes the overview of the proposed system architecture. In this system, input dataset is a network dataset that contains edges lists with text file structure. These files are read and graph structure is created using graph representation using adjacency list.

Before operating search strategy the proposed algorithm, initialize population is evaluated using label based solution representation method. And calculate the fitness function using modularity function. In the search strategy of proposed discrete artificial bee colony algorithm, iteration is calculated until the maximum cycle number. In each iteration, the algorithm produces best fitness value and its solution. When the search strategy reaches the maximum cycle number, the algorithm produces the best solution with best fitness value as an output. Finally, the system reports the cluster of best member as community structures.

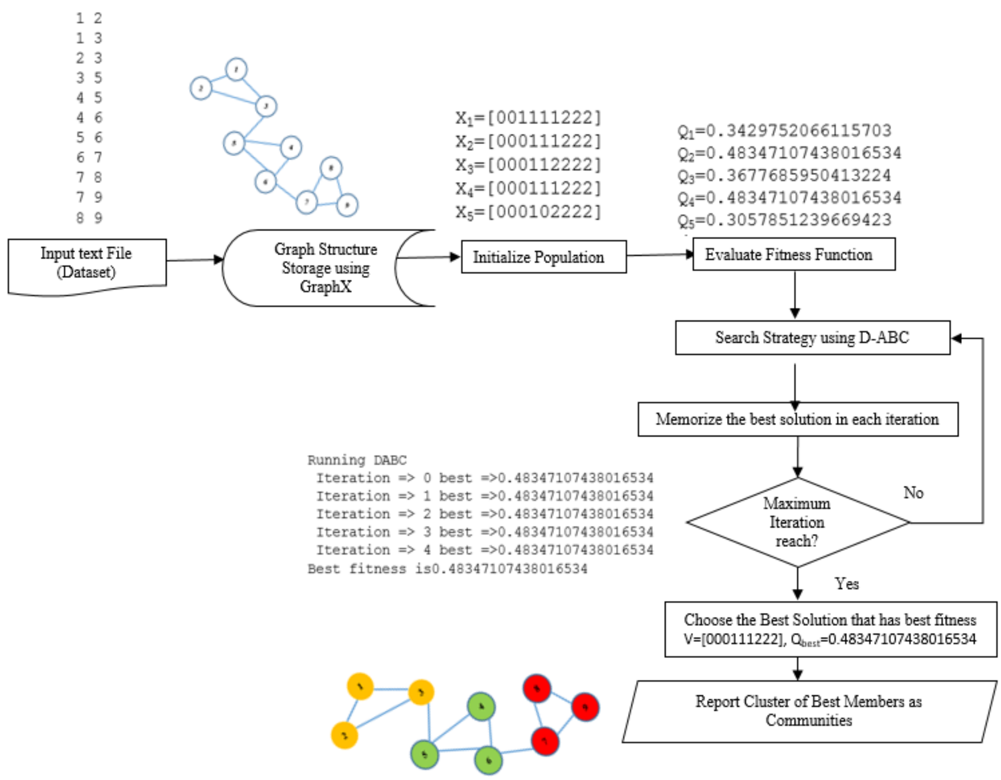


Figure 4. 1 Overview of the Proposed System

The D-ABC algorithm includes two core steps such as initialization and artificial bees search strategy. In the first step, there are two sub-steps, such as solution representation and fitness computation. To represent the solutions, the specific data structure is being used in D-ABC. It illustrates the structure of community in a network with proper format. Then the solutions are initialized randomly by using this representation. For each food source or solution, modularity is calculated as the fitness function. Inspired by the traditional ABC bees' search strategy, each food source is evolved around the search space through several iterations, to optimize the objective function. The proposed initialization strategy describes in Algorithm 4.1, the proposed D-ABC algorithm is presented in Algorithm 4.2.

4.1.1 Load Graph

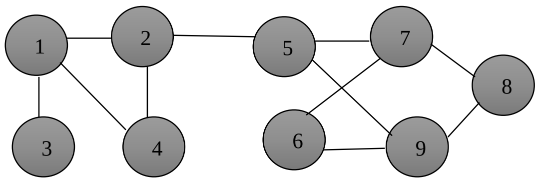
In data preprocessing, the system drops the direction of the edges and treats each network as an unweighted, undirected static graph to represent all network consistently. Because group members may be disconnected in the network, it considers each connected component of the group as a separate community.

The edge list dataset contains edges lists of graph from a text file which is loaded in Spark RDD. This data is loaded in Spark RDD to make use of fault tolerance in cluster and Spark RDD provides efficient use for cluster memory. Efficiency is achieved by parallelizing the processing of a file across multiple nodes in the cluster. These edge list file will be broadcast to all worker node and will be available to each worker. Once the data is loaded into an RDD. Then creating graph, Graph<VD, ED>, VertexRDD<long> and EdgeRDD<Long> in GraphX and the graph contains vertices set and edge set. Then create internal graph as collection set that contains vertices and its neighbor vertices list as adjacency list.

4.1.2 Population initialization and Solution Representation

In meta-heuristic algorithms, population initialization of candidate solutions play an important role. The population will be used as a start point to search for other candidate solution. The good initial population generates better solutions and speed up the convergence of the algorithm in searching process. The D-ABC algorithm uses label-based solution representation as its representation scheme. This representation is sometimes called string-based solution representation. Each position in the vector is an integer value. Each index value in the vector specifies the community label to which they belong. Label based solution representation strategy is described in section 3.3.2 in the previous chapter. Each solution has an array of N genes, each of which belongs to a node in the graph. Each gene owns the community label of each node. Label based encoding schema can automatically define the numbers of community in network without decoding the solution representation.

Figure 4.2 (a) shows the graph structure of the sample network. It contains nine nodes and eleven edges. Each node represents each actor and edge represents the relationship between two nodes. Figure 4.2 (b) illustrates an example of label-based solution representation for a network with 9 nodes. The arbitrary solution represents by the label-based solution representation and the translation of this solution to the graph is nodes 1, 2, 3, 4, 5 with the same community label and other nodes reach in the other community. Therefore, nodes 1, 2, 3, 4, 5 are in the cluster label 2 and nodes 6, 7, 8, 9 are in the cluster label 1. Cluster labels can only be used to define the same cluster of densely connected nodes. It can be seen that each gene of the solution takes its own community label.



(a) Graph Structure of Sample Network

NodeID	1	2	3	4	5	6	7	8	9
CID	2	2	2	2	2	1	1	1	1

(b) Solution Representation for Sample Graph

Figure 4. 2 Graph Structure and Solution Representation of Sample Graph

The individual solutions in the population are created randomly and they may be the possible solutions for the given network. In fact, the randomly generated individual may have the same index value, they reach in the same cluster number and they will have relationship among them.

The set of solutions in the population are initialized using label based solution representation. The population initialization is one of the important roles for nature inspired-based optimization problem. The numbers of food sources represent the number of solutions in the population. The algorithm accelerates the convergence rate during searches in the search space, if initial population produces better solutions. The population initialization strategy is defined as follows:

1. Initialize the set of solutions $S = \{S_1, S_2, S_3, \ldots, S_p\}$ in the population , where p is the population size.
2. Every solution $S_i = \{S_{i1}, S_{i2}, S_{i3}, \ldots, S_{iv}\}$, in this vector v represents the number of vertices in the particular network. Firstly, initialize the random number of community label to all nodes in the network.

3. Choose random nodes for each solution. Then, assign it neighbor nodes' label with its community label. If the randomly selected node has no neighbors, select the next other node.

The set of solutions in the population (number of food sources) of Figure 4.2 are {1 1 2 1 2 2 2 2 2}, {2 1 2 2 1 1 1 1 1}, {1 1 1 1 1 2 2 1 1}, {1 1 1 1 1 2 2 2 2}, {1 1 1 1 2 2 2 2 2}. In food source {1 1 1 1 2 2 2 2 2}, nodes 1, 2, 3, and 4 contain in the same community and node 5,6,7,8 and 9 reach in the other community. It got two community structures. After initialization, evaluate the fitness of each food source using objective modularity function. Algorithm 4.1 represents the initialization strategy of the proposed algorithm for solving community detection problem in social network analysis that are demonstrated in figure 4.3.

Algorithm 4.1 : The Population Initialization Algorithm

Input: Network structure as graph: G= (V, E), Population Size: p, $\alpha < 0.3$, number of community: NC
Output: the set of solutions in the population
Begin
  Generate a population of solutions $S= \{S_1, S_2, S_3..., S_p\}$, each food source
  $S_i= (S_{i1},S_{i2},S_{i3}\ldots,S_{iv})$ in which (v=no of vertices), i= {1,2, 3…, p}
  For each food source in S do
    tc= 0;
    For each element in S_i do
        S_i=Random number of NC;
    End for
    Repeat
      Randomly select a vertex S_{ik};
      label=community label of S_{ik};
      S_{ik} and its neighbor vertices = label;
      tc=tc+1;
    Until tc== (int) (α* n)
  End for
End

Figure 4. 3 The Proposed Population Initialization Algorithm

4.1.3 Fitness Calculation

Modularity function is used as the fitness function for D-ABC. In nature-inspired algorithm, the quality metrics have the role of biological fitness. They can be used to measure the result community quality of individual. Good community structures have higher modularity value. For example, fitness values of {2 2 2 2 2 1 1 1 1} is 0.31.

The modularity calculation of the solution 2 2 2 2 2 1 1 1 1 is described in the equation 3.18.

Where *m* is the number of link for the sample network =11,

i=1, 2, 3, 4, 5, 6, 7, 8, 9

j=1, 2, 3, 4, 5, 6, 7, 8, 9

Degree of 1, d_1=3

Degree of 2, d_2=3

Degree of 3, d_3=1

Degree of 4, d_4=2

Degree of 5, d_5=3

Degree of 6, d_6=2

Degree of 7, d_7=3

Degree of 8, d_8=2

Degree of 9, d_9=3

Community, c = (2, 2, 2, 2, 2, 1, 1, 1, 1) # node 1, 2,3,4,5 reach in same community and node 6,7,8,9 reach in the same community.

$$Q = \frac{1}{2*11} [(A_{11} - \frac{d1d1}{2*11})\delta(c1,c1) + (A_{12} - \frac{d1d2}{2*11})\delta(c1,c2) + (A_{13} - \frac{d1d3}{2*11})\delta(c1,c3) +$$

$$(A_{14} - \frac{d1d4}{2*11})\delta(c1,c4) + (A_{15} - \frac{d1d5}{2*11})\delta(c1,c5) + (A_{16} - \frac{d1d6}{2*11})\delta(c1,c6) + \ldots +$$

$$(A_{97} - \frac{d9d7}{2*11})\delta(c9,c7) + (A_{98} - \frac{d9d8}{2*11})\delta(c9,c8) + (A_{99} - \frac{d9d9}{2*11})\delta(c9,c9)]$$

$$= \frac{1}{22} [(0-3*3/22) *1+(1-3*3/22) *1+(1-3*1/22) *1+(1-3*2/22) *1+(0-3*3/22) *1+$$

$$(0-3*2/22) *0+. . . +(0-3*3/22) *1+(1-3*2/22) *1+(0-3*3/22) *1]$$

$$= \frac{1}{22} [-0.409+0.59+0.86+0.72-0.409+0+. . . -0.409+0.72-0.409]$$

= 0.31

4.1.4 Search Strategy

Three bees process the search strategy for the population. Each bee moves to find the best in the locality and moving it toward the global best of swarm. Genetic operators such as crossover and mutation based operations are used to move each bee to the best positions.

Solution initialization strategy employed by the D-ABC purpose to get effective solution quality. Each of these evaluation cycles will affect the most appropriate result. Community detection is one type of discrete optimization problem. When employed bees create new food source, they choose neighborhood food source in the population. Firstly, choose a random genetic index based on the current food source. Then swap the indexes values of current food source with values in the same indexes of neighborhood food source. Example of creating a new food source is described as follow. For the sample network figure 4.2 (a), the original food source and neighborhood food source are made crossover operation. The randomly selected node 4 in the original food source. Then select the nodes which have the same community label with these nodes. Node 1, 2, 3 and 5 have the same community label as node 4. Then swaps the original food source gene value with the same gene index value in an adjacent food source. Then calculate the fitness of candidate food source. The fitness value of candidate food source is greater than the current food source. Finally, the best food source is chosen as the new food source. In this example, the candidate food source's fitness value is greater than the current food source. It chooses the candidate food source as the new food source with best fitness 0.39.

Current food source

Node ID	1	2	3	4	5	6	7	8	9
Community Label	2	2	2	2	2	1	1	1	1

Neighbor food source

Node ID	1	2	3	4	5	6	7	8	9
Community Label	2	2	2	2	1	1	1	1	1

Candidate food source

Node ID	1	2	3	4	5	6	7	8	9
Community Label	2	2	2	2	1	1	1	1	1

After finding the candidate solution then evaluate the fitness of candidate solution. Select the best solution between old and new solution that contain best the modularity result. Then each food source probability is calculated in the population.

Employed bees give the quality information about the solution to onlooker bees. An onlooker bee select its attractive food source depend on its probability value. The food source probability can be calculated by using the equation 4.1. In which, fit_i is fitness value of X_i and maxfit is the maximum fitness value of best solution.

$$P_i = (0.9*(fit_i /maxfit)) +0.1$$

Equation 4.1

Onlooker bee select a solution based on the probability value of each solution. Then neighborhood food source is selected and new food source is created as employed bee. If the solution (food source) is greater than trial limit, the solution is abandoned and scout bee find new random solution to get global optimal solution.

Each cycle remembers the best solution found in the population and iterates up to the maximum number of cycles. The proposed D-ABC algorithm select the best member of food source in the population. Finally, it provide output community with its members. For Figure 4.2 (a) shows two communities (1, 2, 3, 4) and (5, 6, 7, 8, 9) that contains the highest fitness value 0.3925 by using the D-ABC algorithm. Figure 4.4 presents graph structure of the result community for sample network which is the best quality food source. Step by step procedure of Discrete Artificial Bee Colony algorithm is described in algorithm 4.2 that are illustrated in figure 4.5.

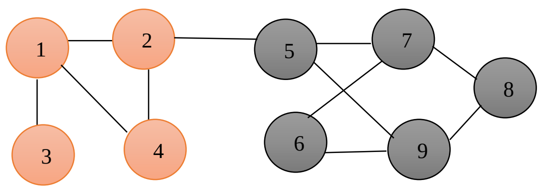


Figure 4.4 Detected Community Structure of Sample Network

Algorithm 4.2 : Discrete Artificial Bee Colony Algorithm (D-ABC)

Input: A network G= (V,E), MCN, Pz, Limit
MCN: maximum cycle number
Pz: population size, Limit: trial count
Output: Community membership assignments for network nodes $C=\{C_1,C_2\ldots,C_k\}$, k is the number of communities.
Algorithm:
1. Read input network and initialize the position of food source $X=\{X_1, X_2,X_3\ldots,X_p\}$ using string-based encoding schema.
   1.1 For each solution X_i, compute fitness of food source using Modularity
   1.2 Keep the best food source (BestFS) with highest fitness in the population
2. Cycle=1
3. while(Cycle<MCN) {
   3.1 Employed Bees Phase
   For each employed bee
   3.1.1 Generate new solution V_i by applying the neighborhood of X_i using one point crossover operation
   3.1.2 Calculate the fitness of new solution V_i using Q
   3.1.3 If (fitness of V_i> fitness of X_i) then replace old solution with new one and reset trial limit of new solution; else increase trial limit count of old solution by 1
   End for
   3.2 Compute the probability P_i for each X_i
   3.3 Onlooker Bees Phase:
   For each onlooker bee
   3.3.1 Generate new solution V_i by applying neighborhood of X_i using one point crossover, depend on the probability of P_i of food source
   3.3.2 Calculate the fitness of new solution V_i using Q
   3.3.3 If(fitness of V_i>fitness of X_i) then replace old solution with new one and reset trial limit of new solution; else increase trial limit count of old solution by 1
   End for
   3.4 Scout Bee Phase
   3.4.1 If (trial count of X_i> limit) then reset the limit and generate a new solution; else continue
   3.5 If (fitness of BestFS < fitness of X_i) then assign the best solution food source X_i as the global best; else continue
   3.6 Cycle++} //end while
4. Output the best solution that contain the community membership assignments of network's node

Figure 4. 5 The Proposed Discrete Artificial Bee Colony Algorithm

4.2 Case Study 1: Community Detection on Sample Network

Internal graph, for sample network (figure 4.2 (a)) is created by using D-ABC algorithm. The text file includes the edges lists of the network and can be input of the system. These text files are loaded into the algorithm and create internal graph using GraphX of Spark Framework. Firstly create graph G = (V, E), Group the set of vertices and the set of edges form the input text file using

Graph<long,long> and JavaRDD<Edge<long>>. Finally, the proposed system got the set of vertices and the set of edge in EdgeRDD<long> edge, VertexRDD<long> vertices. Then create internal graph using Graph<long,long> and JavaRDD<Edge<long>>. Table 4.1 presents the internal graph structure of the sample network dataset. The first column is the vertex name and the last column is the neighbor of these vertex (for the first row, the neighbors of node 1 are node 2, node 3 and node5 or node 1 have three connections with node 2, node 3 and node 5). Then community structure detection is performed using proposed D-ABC algorithm. According to the algorithm 4.2, firstly create the set of solutions for initialization.

Table 4.1 Creating Internal Graph for Sample Network Dataset

Vertex	Neighborhood
1	2,3,5
2	1,4,5
3	1
4	2,8,9
5	1
6	8,9
7	8,9
8	4,6,7
9	4,6,7

Table 4.2 shows the sample set of possible solutions in the population in initialization stage. Initialize population considers the node degree centrality. Population size is set 10 for the sample network dataset. The first column in the table shows the solutions in the population, in which 2 and 1 mean the community label of

nodes (for the first row, solution 112122222 means node 1, node 2 and node 4 reach in the same community label 1 and the remaining nodes 3,5,6,7,8,9 are in the same community label 2).

The second column is about the fitness values of each solution in the population (Modularity is used as the fitness function of each solution). This column shows the quality of each solution in the population. Among them, the algorithm stores the solution with the best modularity result as the best solution. In metaheuristic algorithm, the number of solutions in the population, the iteration number and limit amount can be adjusted based on the network size. If one solution in the population has the highest modularity value, it will be stored as the best solution and its modularity results are considered as the best fitness. In next iteration, the best solution will change based on the solution results of next iteration.

Table 4.2 Initialize the Set of Solutions in the Population for Sample Network

Population	Modularity Value
1 1 2 1 2 2 2 2 2	0.28
2 1 2 2 1 1 1 1 1	0.21
1 1 1 1 1 2 2 1 1	0.07
2 1 2 2 1 1 1 1 1	0.21
1 1 1 1 1 2 2 2 2	0.31
1 2 1 2 2 2 2 2 2	0.11
1 2 1 2 2 2 2 2 2	0.11
2 2 1 2 2 1 1 1 1	0.22
2 1 2 2 1 1 1 1 1	0.21
1 1 1 1 2 2 2 2 2	0.39

In the employed bee phase, each solution is modified with other solution in the population and compared their fitness values. The number of solutions is equaled to the number of employed bees. The bees produce the new food sources and calculate the fitness of candidate food sources then compare old one and choose best food source as the new food source. Table 4.3 demonstrates the solutions in the population in first iteration after running the employed bee phase.

Table 4.3 Set of Solution in Employed Bee Phase

Population	Modularity Value
1 1 2 1 2 2 2 2 2	0.28
2 1 2 2 1 1 1 1 1	0.21
1 1 1 1 2 2 2 2 2	0.39
1 1 2 1 2 2 2 2 2	0.21
1 1 1 1 1 2 2 2 2	0.28
1 2 1 2 2 2 2 2 2	0.11
2 2 2 2 1 1 1 1 1	0.39
2 2 1 2 2 1 1 1 1	0.22
2 1 2 2 1 1 1 1 1	0.21
1 1 1 1 2 2 2 2 2	0.39

Table 4.4 Set of Solution in Onlooker Bee Phase

Population	Modularity Value
1 1 1 1 2 2 2 2 2	0.39
2 1 2 2 1 1 1 1 1	0.21
1 1 1 1 2 2 2 2 2	0.39
1 1 2 1 2 2 2 2 2	0.21
1 1 1 1 1 2 2 2 2	0.28
1 1 1 1 2 2 2 2 2	0.39
2 2 2 2 1 1 1 1 1	0.39
2 2 1 2 2 1 1 1 1	0.22
2 2 2 2 1 1 1 1 1	0.39
1 1 1 1 2 2 2 2 2	0.39

In onlooker bee’s phase, food source selection is made based on the probability of food source from the employed bee. Onlooker bee produces the new food source based on the probabilities of the solution and calculates the fitness of candidate food source. Then, compare the fitness between old and candidate food source. If the new food source is better than the previous value, replace the fitness value of better food source and then select the best one. The changes of some food source in the solution set based on the probabilities of the solutions can be seen in Table 4.4.

If one solution reaches the trial limit, scout bee produces the new food source using the solution initialization strategy and discards old food source. The repetition process is done until the stropping criteria is reached or the maximum iterations are completed. After calculating the search process until maximum iterations, the proposed algorithm produces the community structure for the sample network.

The proposed system use the input value of community is 2, Maximum iteration is 10 and Population Size is 10 are used for community detection in Sample dataset network. The final best solution is 1 1 1 1 2 2 2 2 2 and it has the best fitness value 0.39. Table 4.5 shows the result of the best community structure for the sample network which have modularity 0.39. Then, calculate the accuracy of the result community structure by using Normalize Mutual Information (NMI). The NMI value gets 1 for a sample dataset. It means the detected community structures are same as the real community structures. Then calculate the quality of community results by using internal connectivity, external connectivity, both internal and external connectivity, and network model metrics. The following results show the quality values of community result in each metric.

Internal Connectivity Metric:

Internal density=0.31666

Internal and External Connectivity Metrics

Normalized cut =0.3456

Conductance=0.1888

External Connectivity Metrics

Expansion=0.45

Cut Ratio=0.1

Network Model Metric:

Modularity=0.39

Finally, the system produces community structures of the best solution for the given network. Table 4.5 shows that node 1, 2, 3 and 4 reach in same community number and node 5,6,7,8 and 9 group in same community number 2.

Table 4.5 Detected Community Structure for Sample Network

Community No.	**Node**	**No. of Members**
Community 1	1,2,3,4	4
Community 2	5,6,7,8,9	5

4.3 Case Study 2: Community Detection on Karate Club Network

By using the proposed algorithm, karate club dataset is also detected to get the effective community structures. It is a friendship network between the 34 members of a karate club at a US University. It contains 34 nodes and 78 edges. The neighborhood information of each vertex is stored as the link list structure that is shown in Table 4.6.

Initialize population of karate dataset is described in Table 4.7. This table contains the ten solutions in the population.

Table 4.6 Creating Internal Graph for Karate Dataset

Vertex	**Neighborhood**
1	2,3,4,5,6,7,8,9,11,12,13,14,18,20,22,32,
2	1,3,4,8,14,18,20,22,31,
3	1,2,4,8,9,10,14,28,29,33,
4	1,2,3,8,13,14,
5	1,7,11,
6	1,7,11,17,
7	1,5,6,17,
8	1,2,3,4,
9	1,3,31,33,34,
10	3,34,
11	1,5,6,
12	1,
13	1,4,
14	1,2,3,4,34,
15	33,34,
16	33,34,
17	6,7,

18	1,2,
19	33,34,
20	1,2,34,
21	33,34,
22	1,2,
23	33,34,
24	26,28,30,33,34,
25	26,28,32,
26	24,25,32,
27	30,34,
28	3,24,25,34,
29	3,32,34,
30	24,27,33,34,
31	2,9,33,34,
32	1,25,26,29,33,34,
33	3,9,15,16,19,21,23,24,30,31,32,34,
34	9,10,14,15,16,19,20,21,23,24,27,28,29,30,31,32,33,

Table 4.7 Initialize Set of Solutions for Karate Dataset

Population	**Modularity Value**
2 2 1 2 2 2 2 2 2 1 2 2 2 2 1 1 2 2 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1	0.329
1 1 1 1 1 1 1 1 2 2 1 1 1 1 2 2 1 1 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2	0.371
1 1 1 1 1 1 1 1 2 2 1 1 1 1 2 2 1 1 2 2 2 1 2 2 1 2 2 2 2 2 2 2 2 2	0.32
1 1 2 1 1 1 1 1 2 2 1 1 1 1 2 2 2 1 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2	0.33
2 2 2 2 2 2 2 2 1 1 2 2 2 2 1 1 2 2 1 2 1 2 1 1 1 1 1 1 1 1 2 1 1 1	0.345
1 1 2 1 1 1 1 1 1 2 1 1 1 1 2 2 1 1 2 1 1 1 2 2 2 2 2 2 2 2 1 2 2 2	0.307
2 2 1 2 2 2 2 2 1 1 2 2 2 1 1 1 2 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1	0.313
2 2 1 2 2 2 2 2 1 1 2 2 2 1 1 1 2 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1	0.313
1 1 2 1 1 1 1 1 2 2 1 1 1 1 2 2 1 1 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2	0.359
1 1 2 1 1 1 1 1 1 2 1 1 1 1 2 2 1 1 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2	0.329

After finishing the search process of the proposed D-ABC algorithm such as employed bee, onlooker bee and scout bee phases produces the best solution with highest fitness value in the population. The best solution is {1 1 1 1 1 1 1 1 2 2 1 1 1 1 2 2 1 1 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2}. The network model metric, modularity value is 0.37146.

NMI result for karate club dataset is one. The internal density of result community structure is higher than the external connectivity metrics, cut ratio. Good community structures have higher internal density and lower external connectivity. Modularity and internal density metrics are the higher, the better and the other four metrics are the lower the better. The following results show the quality measurement of result community structures of karate club dataset.

Internal Connectivity Metric:

Internal density=0.1259

Internal and External Connectivity Metrics:

Normalized cut =0.460

Conductance=0.2565

External Connectivity Metrics:

Expansion=1.180

Cut Ratio=0.069

Network Model Metric

Modularity=0.37146

Table 4.8 is the final best solution in the population. In one community contains nodes 9,10,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34 and the next community contains 1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22. It is equal to the ground truth community structure. Karate club split according to the conflict between Admin John (node 34) and instructor Mr.Hi (node 1). Therefore, half of karate members reach in John's group and the last members found in a new club with Mr.Hi.

Table 4.8 Detected Community Structure for Zachary's Karate Club Dataset

Community No.	Node	No. of Members
Community 1	9,10,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33,34	18
Community 2	1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22	16

4.4 Summary

This chapter provides the architecture and implementation of proposed D-ABC algorithms in details. The experimental results are also described by using the standard real-world datasets. Intelligence-based swarm algorithm's computation is an important approach to address optimization problem, primarily focusing on the collective behavior of decentralized, self-organized system. The proposed method is called Discrete Artificial Bee Colony (D-ABC) algorithm that is inspired by the intelligence behaviors of the bee swarms. The global optimization process of D-ABC is accelerated by positive feedback of bee colony optimizing.

D-ABC contains three main steps: creating graph using GraphX, food source initialization, and three heuristic searching process. Algorithm 4.1 includes the initialization strategy for the solutions in the population which has used the label-based solution representation.

The ABC algorithm has influenced the intelligent behavior of real bee colonies introduced by Karaboga for numerical optimization problems. The ABC algorithm provides a high performance and accurates solution when the solution problem has continuous nature. However, if the solution space of the problem is discrete, then the underlying ABCs need to be modified to solve this type of optimization problem. Community detection problem is one of the most widely studied combinatorial optimization problems. It is a NP-hard problem whose computational complexity rises exponentially when the number of nodes increasing. This study applied discrete version of ABC with neighborhood operator for community detection in social network that is described in Algorithm 4.2. Therefore, different discrete neighborhood operators are replaced with the basic ABC's solution updating equations.

CHAPTER 5

EXPERIMENTAL ANALYSIS AND EVALUATION

This chapter describes the experimental analysis and evaluates the implementation of the proposed system on various social networks. Some social networks with ground truth communities and some dataset without ground truth are used for the experiment. The task of detecting communities in social networks has become an interesting research task in the big data era. However, if there are no identical, unique communities in the network, this can be expected to be easy.

Some important properties of social network need to understand network size, node density, degree if each node, edges reachability, strength and width of these network. Social network community detection has become a fundamental tool for social network analysis that lays the fundamental for several higher level analytics. Nowadays, automation of community detection has become a favorite research topic in computational intelligence field. It is a long-standing research topic in sociology, as it related to urbanization, social marketing, criminology and other various research field. It needs to measure the results quality of isolated community structure. Three well-known perspective approaches called accuracy, efficiency, and effective are used to make the quality measurement.

The efficiency of the proposed system is executed based on experimental and theoretical attitude. Most of the datasets have no ground truth result. Therefore, the effectiveness of algorithm is examined by using the evaluation metrics based on internal connectivity, external connectivity, both internal and external connectivity and network model based metrics. To measure the accuracy of result community structures, NMI is used to compare the similarity between the detected and real community structure. This chapter discusses the experiment results using proposed algorithm called discrete artificial bee colony algorithm (D-ABC) and then compared with other algorithm using various quality metrics.

This chapter provides a detailed evaluations for the proposed research work, which covers the above three measurements. The datasets are introduced and detailed evaluation methods and results are presented. The findings are summarized and the proposed work is ultimately intuitively evaluated. All experiments are executed on a computer running Intel Core i5-5200U CPU@ 2.20GHZ, Window 10, 16G RAM and 64 bit Operating System.

5.1 Social Network Datasets

For implementation of all consistent social networks, each network will be considered as an undirected and unweighted static graph. Two datasets are used to provide better and more deeply evaluation results than the current approaches. All of the real network datasets are available in UCINET and Stanford Large Network Dataset Collection. The first type of network is considered as real-world small scale networks with ground truth cluster results.

Karate Club Dataset: It is the standard graph dataset which is a famous social network for community detection example. The college karate club social network that studied by Wayne W.Zachary for three years [63]. It gathered 34 karate club members and documented links between pairs of members who interacted outside the club as US University in the 1970s. The graph structure of the karate club network has been shown in Figure 5.1. It has become a popular example of networked community structure in 2002. Due to the conflict between administrator John A (node 34) and his coach, Mr. Hi (node 1), the club is split in two. So, half of the members formed a new club with Mr. Hi, and the other part is found in a new instructor of karate dataset. Figure 5.2 illustrates the real community structures of karate club dataset with two different colors. The network has two ground truth communities and its members are listed in the Table 5.1.

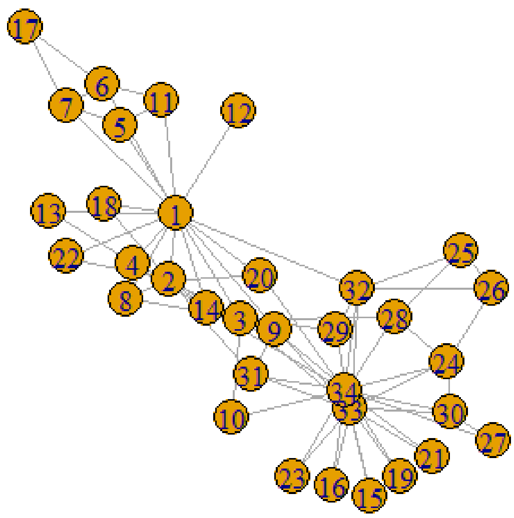


Figure 5. 1 Graph Structure of Karate Club

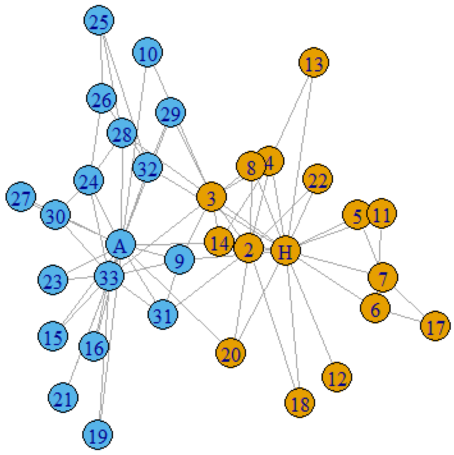


Figure 5. 2 Ground Truth Structure of Karate Club

Table 5. 1 Ground Truth Community Structure of Karate Club Dataset

Community No.	Node	No. of Members
Community 1	34,9,10,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33	18
Community 2	1,2,3,4,5,6,7,8,11,12,13,14,17,18,20,22	16

Bottlenose Dolphin Network: This network is an undirected bottlenose dolphin network that simulates the behavior of bottlenose dolphins seen during a seven-year period from 1994 to 2001. These dolphins live in Doubtful Sound, New Zealand, as arranged by Lusseau et.al [33]. In this network, each node is a member of bottlenose dolphin two nodes are connected if the corresponding dolphins are observed together more often than expected. There are 62 nodes, each node represents dolphins and 159 edges, each node represents associations between pairs of dolphins which occasionally expected to occur more frequently and it showed in figure 5.3. The vertices in this network can be divided into two ground truth community structures. Table 5.2 describes the ground truth clusters of dolphin network dataset.

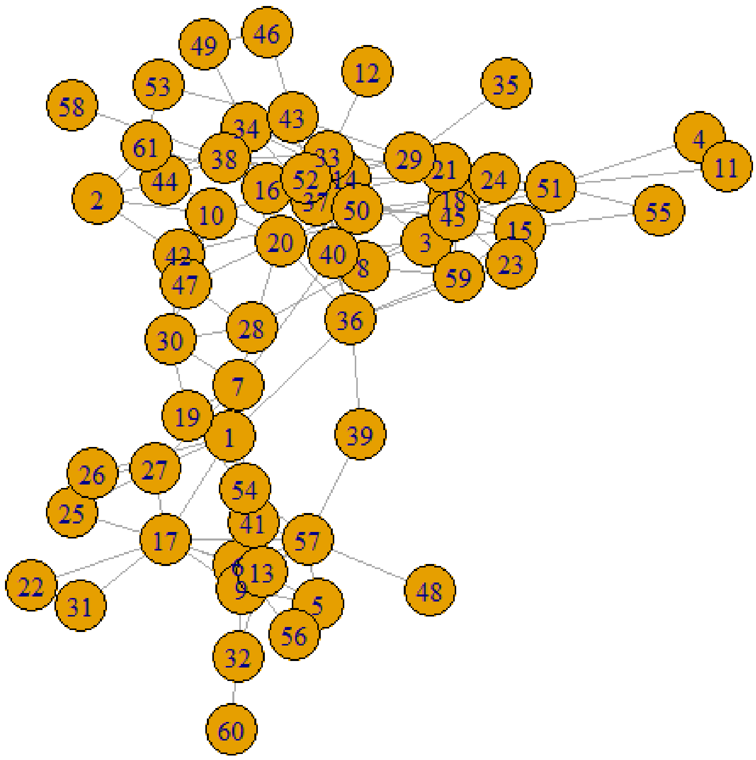


Figure 5. 3 Graph Structure of Dolphin

Table 5. 2 Ground Truth Community Structure of Dolphin Dataset

Community No.	**Node**	**No. of Members**
Community 1	2,6,7,8,10,14,18,20,23,26,27,28,32,33,42,49,55,57,58,61	20
Community 2	1,3,4,5,9,11,12,13,15,16,17,19,21,22,24,25,29,30,31,34,35,36,37,38,39,40,41,43,44,45,46,47,48,50,51,52,53,54,56,59,60,62	42

Polbook: It is a book network of US politics published around the period of the 2004 presidential election and sold by the online book seller Amazon.com. The network was compiled by V.Krebs and is unpublished, but can be found on Kreb's web site. This undirected graph has 105 vertices and 441 edges, described in Figure 5.4, in which nodes represent books and edges represent connection between books and buyers that are frequently co-purchase of book by the same buyers [28]. Table 5.3 describe the ground truth community structure of PolBook.

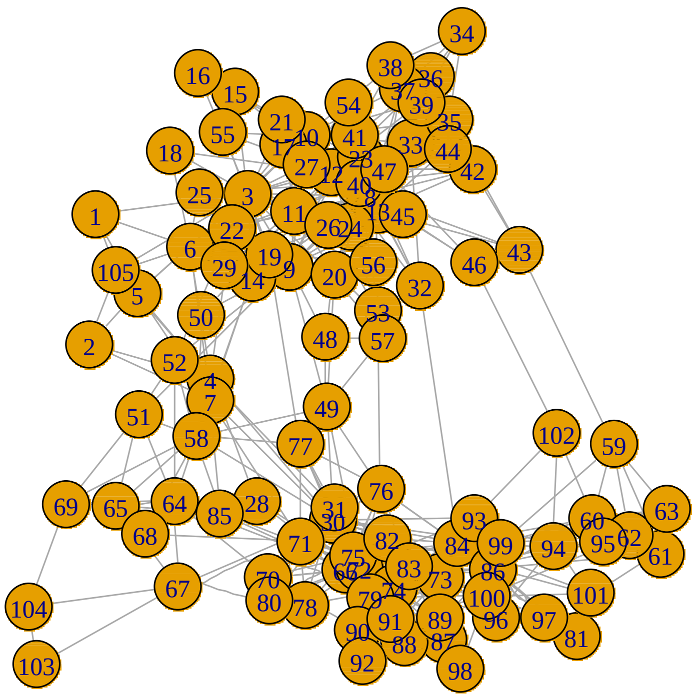


Figure 5. 4 Graph Structure of PolBook

Table 5. 3 Ground Truth Community Structure of PolBook Dataset

Community No.	**Node**	**No. of Members**
Community 1	1,5,7,8,19,29,47,49,52,70,77,104,105	13
Community 2	31,32,60,61,62,63,64,65,66,67,68,69,71,72,73,74,75,76,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100,101,102,103	43
Community 3	2,3,4,6,9,10,11,12,13,14,15,16,17,18,20,21,22,23,24,25,26,27,28,30,33,34,35,36,37,38,39,40,41,42,43,44,45,46,48,50,51,53,54,55,56,57,58,59,78	49

American College Football Network: It is about the football games between American colleges during a regular season in fall 2000. There are 115 nodes and they represent teams; 613 edges indicate the season game between vertices in the year. That grouped in 12 communities. Football that denotes matches played between US college football teams in year 2000 [19].

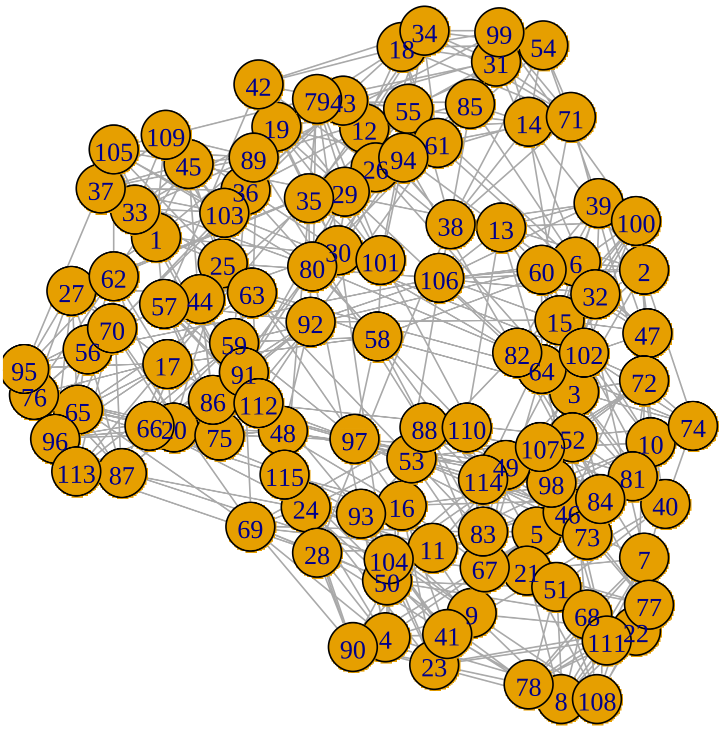


Figure 5. 5 Graph Structure of Football

Table 5. 4 Ground Truth Community Structure of Football Dataset

Community No.	Node	No. of Members
Community 1	1,5,10,17,24,42,94,105	8
Community 2	13,15,19,27,32,35,39,44,55,62,72,86,100	13
Community 3	2,26,34,38,46,90,104,106,110	9
Community4	8,9,22,23,52,69,78,79,109,112	10
Community 5	20,30,31,36,56,80,95,102	8
Community 6	3,7,14,16,33,40,48,61,65,101,107	11
Community 7	37,43,81,83,91	5
Community 8	12,25,51,60,64,70,98	7
Community 9	18,21,28,57,63,66,71,77,88,96,97,114	12
Community 10	4,6,11,41,53,73,75,82,85,99,103,108	12
Community 11	45,49,58,67,76,87,92,93,111,113	10
Community 12	29,47,50,54,59,68,74,84,89,115	10

The graph structure of football is shown in Figure 5.5. The ground truth structure of football dataset is shown in Table 5.4.

Employee (Michael, 1997) is a dataset that concern the impressive sawmill related informal communication. Nodes are associated with 24 employees of wood working company and edges of the network accommodated frequent discussion about the bias between pairs of colleagues. Three communities in this dataset comprise as four Spanish speaking, nine young English speaking workers under 30 and eleven older English speaking workers [35]. Figure 5.6 illustrates the graph structure of employee dataset and its ground truth community structures are shown in table 5.5.

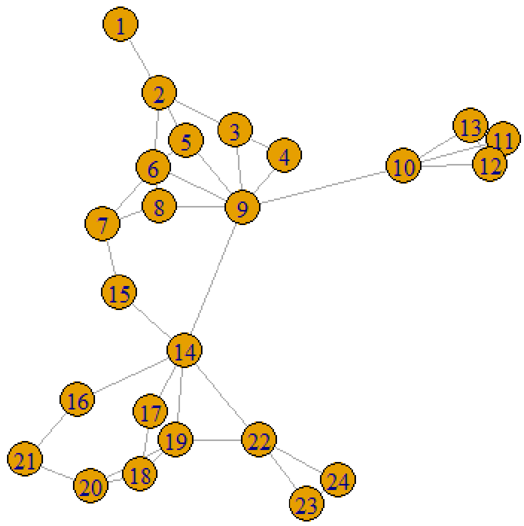


Figure 5. 6 Graph Structure of Employee

Table 5. 5 Ground-Truth Community of Employee Dataset

Community No.	Node	No. of Members
Community 1	10,11,12,13	4
Community 2	1,2,3,4,5,6,7,8,9	9
Community 3	14,15,16,17,18,19,20,21,22,23,24	11

Table 5. 6 Social Network Datasets with Ground-Truth Community

Dataset	Node	Edge	Community Number
Karate Club	34	78	2
Dolphin	62	159	2
PolBook	105	441	3
Football	115	613	12
Employee	24	36	3

The experimental results of the proposed system are based on both quality and quantity of result community structure. Therefore, the network used for evaluation needs to meet certain criteria. The results of interpretation are displayed and the network should be exposed to facilitate method and algorithm verification. The statistical information of these networks is described in Table 5.6. Dataset means the network dataset name and node is the number of nodes and edge is the number of edges and community number is ground truth community number of each dataset.

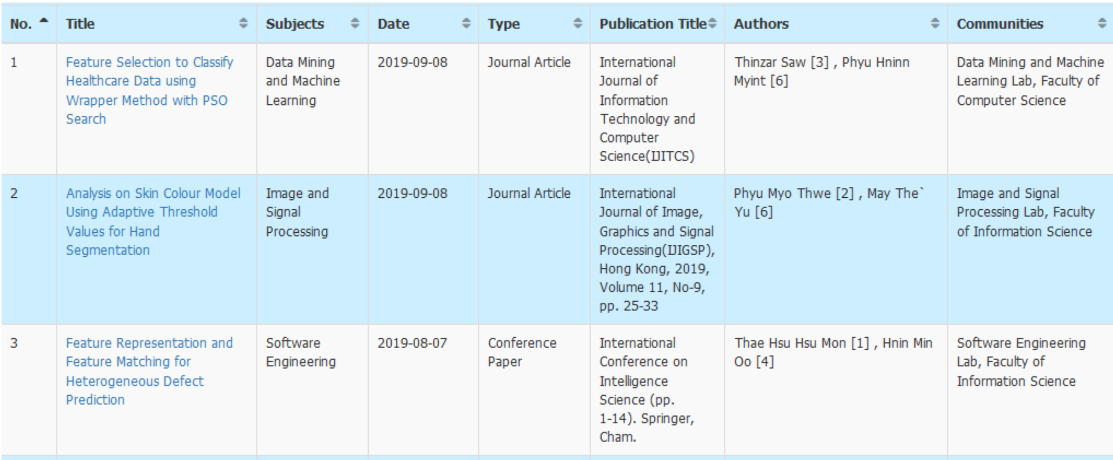
Then, the second networks employ four real world networks that does not have ground truth community results. Table 5.7 reports the information of these datasets.

Email Network: The network was built using e-mail data from a large research institute in Europe. We have anonymous information about all incoming and outgoing emails between members of the Research Institute [64].

Facebook: This dataset contains the friends' lists from Facebook that can be get from snap datasets. These data are collected from survey participant using Facebook app. The dataset contains node profile and ego networks. For all users, anonymity is a new value by replacing the internal Facebook ID. Also, although feature vectors have been provided from this dataset, the interpretation of these features is unclear. The dataset can be obtained from SNAP Stanford Large Network Dataset Collection. The dataset contains 4039 nodes and 88234 edges [32].

DBLP network: This dataset describes the authors of scientific papers using collaboration graph. The relationship between two authors denotes a common publication. Edges are marked with the date of the publication [61].

UCSM Coauthor Dataset: UCSM coauthor network has been taken from the publication list of UCSM research repository from website. Web crawler crawls this data and collects information of co-authors. Co-author information can be obtained from the publication lists. The publication information on the UCSM webpage is shown in Figure 5.7. This data was retrieved from UCSM webpage in 2019. There are 80 authors ad 105 author relationships involved in creating the UCSM co-author network. This network will ignore the publication paper if the treaty is intended to be written by only one author. If the publication has three authors such as a, b and c, their relationships will be a-b and a-c [67].



No.	Title	Subjects	Date	Type	Publication Title	Authors	Communities
1	Feature Selection to Classify Healthcare Data using Wrapper Method with PSO Search	Data Mining and Machine Learning	2019-09-08	Journal Article	International Journal of Information Technology and Computer Science(IJITCS)	Thinzar Saw [3] , Phyu Hninn Myint [6]	Data Mining and Machine Learning Lab, Faculty of Computer Science
2	Analysis on Skin Colour Model Using Adaptive Threshold Values for Hand Segmentation	Image and Signal Processing	2019-09-08	Journal Article	International Journal of Image, Graphics and Signal Processing(IJIGSP), Hong Kong, 2019, Volume 11, No-9, pp. 25-33	Phyu Myo Thwe [2] , May The` Yu [6]	Image and Signal Processing Lab, Faculty of Information Science
3	Feature Representation and Feature Matching for Heterogeneous Defect Prediction	Software Engineering	2019-08-07	Conference Paper	International Conference on Intelligence Science (pp. 1-14). Springer, Cham.	Thae Hsu Hsu Mon [1] , Hnin Min Oo [4]	Software Engineering Lab, Faculty of Information Science

Figure 5. 7 UCSM's Research Repository Page

Table 5. 7 Real-world Networks Dataset Without Ground-truth

Network Name	No. of Node	No. of Edge
Email	1005	25571
Facebook	4039	88234
DBLP	425957	1049866
UCSM Co-author	80	105

5.2 Experiments on Community Detection Algorithm Capability

The ability of the proposed system to generate correct community detection results was tested on Employee, Zachary's Karate, Dolphin, and PolBook networks with ground truth datasets and some other datasets such as small and large network datasets. If the dataset has known community numbers, the proposed algorithm will use these community numbers to get an exact community structure. For each network, the proposed D-ABC algorithm has run 10 times and recorded the best modularity result over the 10 runs of the algorithm. The result is described in the following parts. For employees' connection dataset, after finding community structure in this network using the proposed method, the achievement results of three detected communities are illustrated in Table 5.8. The results show that the proposed D-ABC gets the exact result with ground truth community structure.

Table 5. 8 Detected Community Structure of Employee Dataset

Community No.	Node	No. of Members
Community 1	14,15,16,17,18,19,20,21,22,23,24	11
Community 2	1,2,3,4,5,6,7,8,9	9
Community 3	10,11,12,13	4

The ground truth community structure of Zachary's karate club dataset consists of two community structures, one with 18 members and the other with 16 members. Karate club dataset are detected by using the proposed algorithm and gets the exact community structures as the ground truth communities. The obtained members in each community are presented in Table 5.9.

Table 5. 9 Detected Community Structure of Zachary's Karate Dataset

Community No.	Node	No. of Members
Community 1	34,15,16,19,21,23,24,25,26,27,28,29,30,31,32,33	16
Community 2	1,2,3,4,5,6,7,8,9,10,11,12,13,14,17,18,20,22	18

Lusseau's dolphin social network is also used to detect community for validation. The nodes are divided into two large groups and the obtaining result is demonstrated in Table 5.10. In ground truth community result, one community has 20 members and other community has 42 members. The results of the community are slightly different but the quality of fitness value has the optimized result.

Table 5. 10 Finding Community Structures in Dolphin Network

Community No.	Node	No. of Members
Community 1	2,6,7,8,10,14,18,20,23,26,27,28,29,31,32,33,40,42,49,55,57,58,61,	23
Community 2	1,3,4,5,9,11,12,13,15,16,17,19,21,22,24,25,30,34,35,36,37,38,39,41,43,44,45,46,47,48,50,51,52,53,54,56,59,60,62	39

The obtaining result community structures of PolBook dataset is described in Table 5.11. The size of communities for book dataset are 13, 43 and 49. The detected communities are 18, 41 and 46. The result community structure and ground truth result community structure have a little different. However, the network modularity result of PolBook dataset gets the effective result when using the proposed algorithm.

Table 5. 11 Obtaining Community Structures of PolBook Dataset

Community No.	Node	No. of Members
Community 1	14,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24 ,25,26,27,28,30,33,34,35,36,37,38,39,40,41,42,43, 44,45,46,47,48,49,50,51,54,55,56,57,58,	46
Community 2	29,31,32,60,61,62,63,64,67,71,72,73,74,75,76,77,7 8,79,80,81,82,83,84,85,87,88,89,90,91,92,93,94,95 ,96,97,98,99,100,101,102,103,	41
Community 3	1,2,3,5,6,7,8,52,53,59,65,66,68,69,70,86,104,105	18

The detected community structure of Football dataset is illustrated in Table 5.12. The real community number of Football dataset is 12 and the proposed D-ABC identified ten communities in this dataset. The modularity value of results community is over 0.8 for football dataset.

Table 5. 12 Detecting Community Structures of Football Dataset

Community No.	Node	No. of Members
Community 1	1,5,10,17,24,42,94,105	8
Community 2	13,15,16,19,27,32,35,39,44,55,62,72,86,100,101	15
Community 3	2,26,34,38,46,90,104,106,110	9
Community4	8,9,22,23,43,52,64,69,78,79,109,112	12
Community 5	20,30,31,36,56,80,81,83,95,102	10
Community 6	3,7,14,33,40,48,61,65,107	9
Community 7	18,21,28,57,63,66,71,77,88,96,97,114	12
Community 8	4,6,11,41,53,73,75,82,85,99,103,108	12
Community 9	12,25,29,37,45,49,51,58,60,67,70,76,87,91,92,93,98,113	18
Community 10	47,50,54,59,68,74,84,89,111,115	10

5.3 Parameter Adjustment on Datasets

Assignment definition parameters in the metaheuristic algorithms can be adjusted according to the size and nature of network. The parameters setting in sample network dataset are changed with different possible values. The best solution reaches the best modularity value where the number of iteration and solution size are adjusted. If the number of population is too big, the swarm will move away from the region and lose good solutions. If the number of iteration and population size are small, the swarm will not reach the optimal solution. Therefore, the parameter number and iteration setting are difficult to obtain the suitable partitioning.

Table 5.13 is the result community structure of the sample network dataset based on various parameter values. The size of this data is small and the network connection of each group is clear. And nodes in the group have higher connectivity and lower connection to the other community member. Therefore, the result will be same when using the proposed algorithm. The result community number is 2 for all experiments. The modularity value is not greater than 0.39 in all experiments.

The experiment shows that, if the network has densely connected in each community, it will not get other optimal solution and reach only one global solution in optimization. Therefore, changing the parameter value does not effected the results of community structure and modularity value. The population size starts from 5 because search processes require many food sources to exchange the each gene values. This test will make until population size 100. So, population size is set to a minimum of 5 for this network and iteration number can set any value from 5 to 50. If the population size increases in small connectivity employee network, calculation time also increases and the best result will not be changed.

Table 5. 13 Testing Different Parameters in Sample Dataset

Iteration	Population Size	Modularity	Community Number
5	5	0.39256198347107457	2
5	10	0.39256198347107457	2
10	10	0.39256198347107457	2
20	10	0.39256198347107457	2
20	50	0.39256198347107457	2
50	70	0.39256198347107457	2

Table 5.14 describes how modularity Q values evolved in Zachary club using the empirical parameters in the proposed algorithm. The number of iterations start at 20 to 100 and the number of solution in the population starts form 10 to evolve till 100. The result of modularity increases when the number of population size and iteration number increase. And also community number increased depends on the iteration and population size. This testing purposes to get effective modularity result, it does not consider to get the ground truth community number.

Table 5. 14 Testing Different Parameters in Karate Dataset

Iteration	Population Size	Modularity	Community Number
10	10	0.2679158448389211	2
20	20	0.35996055226824286	2
20	50	0.39907955292570496	3
50	50	0.4022846811308338	4
100	100	0.41740631163708036	4
200	200	0.41978961209730375	4

For the ground truth dataset, the proposed system used the given number of community to reach optimal point rapidly. If the dataset doesn't have ground truth community number, the proposed system sets the number of community as square root of the number of nodes. The objective of the community detection is the number of nodes in the group are densely connected and sparsely connected to the other nodes in other groups. The algorithm finds the best solution with high quality modularity value.

Parameters are adjusted based on the network size and nature. For the small dataset, the number of clusters formed as small, but if the number of parameter increase, the community structure will not uncover at the finer granularity. The figure 5.8 shows the effective parameters to reach optimal modularity value for each network in this system. For the small network the iteration is enough lower 100 and population size is 50. But the large scale networks that contain over thousand nodes and edges, it require more iteration and population size is suitable above 100. The other well know meta-heuristic optimization algorithm requires nearly 500 iteration and 500 population size.

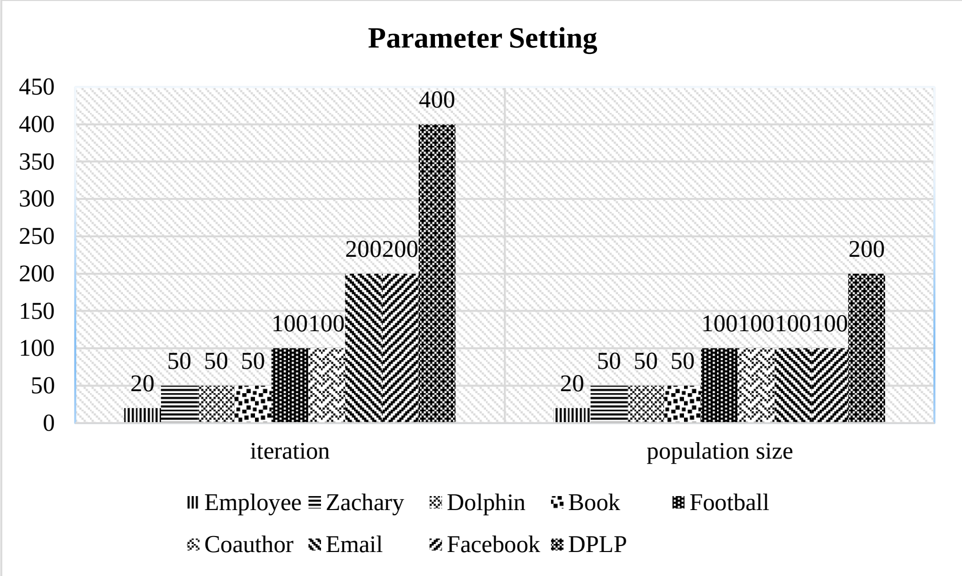


Figure 5. 8 Parameter Setting on Various Datasets

For employee dataset, maximum cycle number MCN or iteration 30 and population size 50 is suitable. MCN 50 and population size 70 are effective for Zachary karate club dataset to get ground truth community result. For Dolphin and PolBook datasets, population size is set to 100 and MCN is set 100 to be the most closely related member as ground truth community results.

5.4 Experiment on Accuracy

The accuracy of the proposed system is estimated by using NMI (Normalized Mutual Information) metric that is measure the similarity between the detected result communities and ground truth communities. The proposed system delivers communities that are consistent with the real ones in some dataset. Table 5.15 shows the NMI metrics score for five datasets. According to the results of accuracy metric, it demonstrates that Employee and Karate club datasets performs accurate result as the ground truth result.

The dolphin network can be detected over 70% of NMI. For PolBook network, NMI can be detected over 50%. And Football community results get over 80% accuracy. Based on the community result, it can be concluded that NMI is good in the communities which have strong connections within internal nodes in the community and weak relationships to nodes in the other rest groups. The accuracy of the proposed algorithm compares with the accuracy of Label Propagation Algorithm LPA [45]. LPA gets over 80% accurate results in four datasets and 54% accuracy in PolBook dataset.

Table 5. 15 Accuracy Measurement of Results Communities Obtained by Two Algorithm

Datasets	D-ABC(NMI)	LPA(NMI)
Employee	1	0.864
Zachary	1	0.836
Dolphin	0.835	0.88
PolBook	0.552	0.54
Football	0.89	0.86

5.5 Efficiency

This part expresses the proposed algorithm efficiency base on practical and theoretical result. The proposed algorithm does not traverse to all node in each solution. Set , the population size is p, maximum cycle is N, the number of nodes is n, and the number of edge is e. Theoretically, the first algorithm for solution initialization spends O(pn) time and the search algorithm using proposed algorithm spends O(Np $\log_2 n$) in evolving communities. The detailed calculation of the proposed two algorithm complexity are presented below.

<u>For Population initialization: Algorithm 4.1</u>

<u>Initialized population</u>

Time Complexity= p*$\frac{n}{3}$=O(pn)

<u>For Search Strategy: Algorithm 4. 2</u>

For maximum iteration in Loop

Time Complexity= O (N)

For employed bee

Time Complexity= p*$\log_2 n$=O (p*$\log_2 n$)

For onlooker bee

Time Complexity=p* $\log_2 n$=O (p*$\log_2 n$)

Total Time Complexity for search strategy=O (N) *(2p* O ($\log_2 n$))

= O(Np $\log_2 n$)

Therefore, the time complexity of overall system is O(pn)+ O(Np $\log_2 n$).

5.6 Experiment on Effectiveness

The effectiveness of the result communities for some kinds of real social datasets such as employee, karate, dolphin, book and football network are tested by using the community scoring metrics. Four types of quality functions are used such as network model, internal connectivity, external connectivity, both internal and external connectivity for the effectiveness of the result communities. For internal connectivity measurement, internal density metrics is used. For external connectivity measurement, expansion and cut ratio are used and then conductance and normalized cut is used to measure on both external and internal connectivity. Modularity values are network model based quality metrics, it is also used as the fitness function of the proposed system.

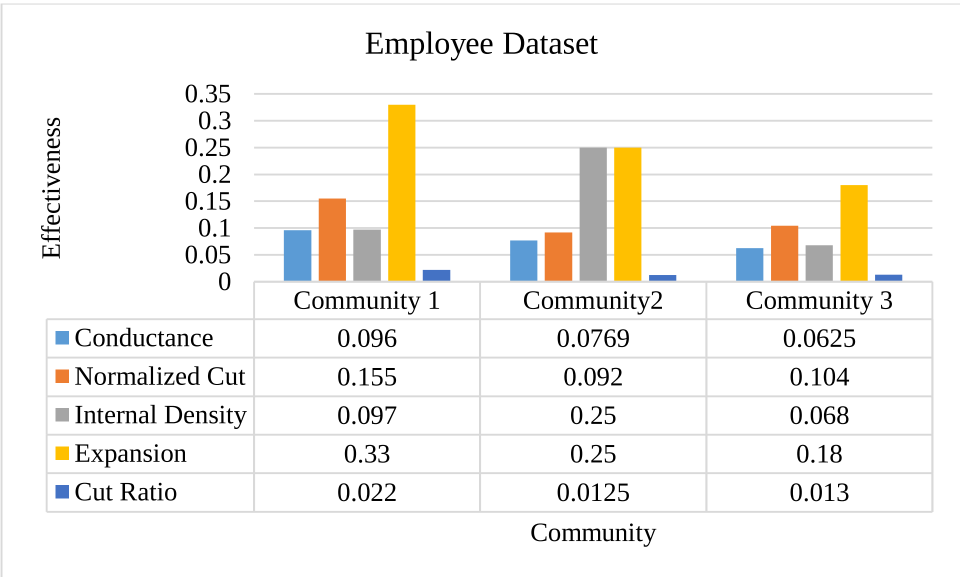


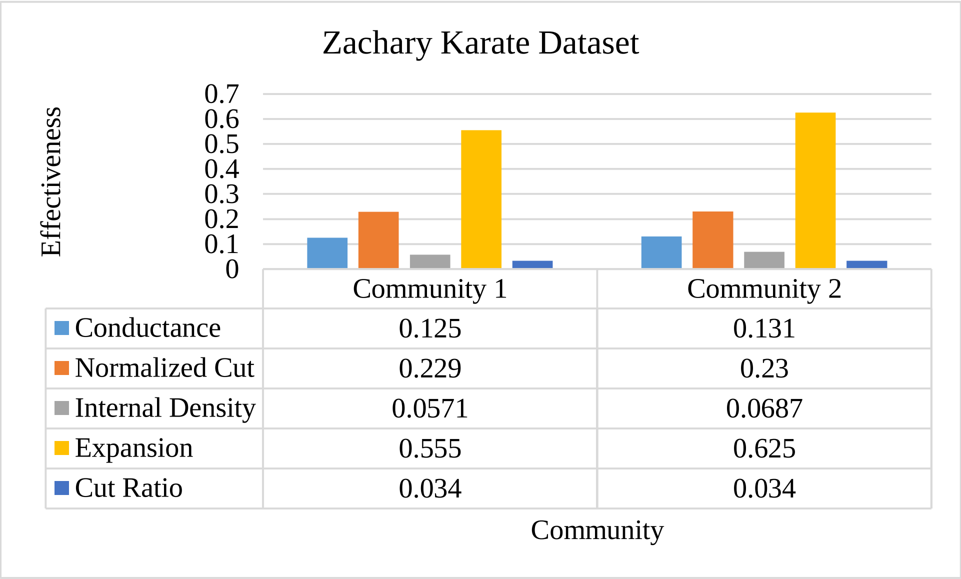
Figure 5. 9 Effectiveness of Each Community for Employee Network

When measuring the quality of result community structures of employee dataset, it gets the effective quality results in each community structure. The quality results are shown in figure 5.9. From the conductivity measurement point of view, the faction of total edge volume that points outside the community decreased to less than 0.1 that indicates the community result can be effectively detected. The calculation of expansion based on the average number of edges that point outside the community and if the expansion value is small, the result will be more effective. Therefore, it can be concluded that expansion value is also significant for community density connectivity.

For internal connectivity, community 2 has higher internal density than other two community structure. The higher the internal density, the best detected communities are. For external connectivity, community 3 has lower expansion value than other two community structure. The higher expansion and cut ratio, the worse detected communities are. For the employee dataset, it gets higher internal density and lower cut ratio (external connectivity) in all community structures,

The quality measurement results of Zachary karate dataset are illustrated in figure 5.10. It shows that the dataset get effective results on the five metrics. It shows that the internal density of detected communities values have increased and external density cut ratio values have grown down in all communities. Since, the expansion that measures the number of edges per node pointing to the other cluster, the two community structure get nearly expansion values in all community structures. The two community structures have suitable quality results from five quality metrics.

In conductance measure, the fractions of total edge volume outside the community have decreased lower than 0.15 that show the result of good detected community. Overall it can be seen the significant effectiveness on these networks.



	Community 1	Community 2
Conductance	0.125	0.131
Normalized Cut	0.229	0.23
Internal Density	0.0571	0.0687
Expansion	0.555	0.625
Cut Ratio	0.034	0.034

Figure 5. 10 Effectiveness on Zachary Karate Club Network

Figure 5.11 demonstrates the significant result of dolphin network that is internal density and cut ratio. The internal density of two clusters is higher than the cut ratio result of two communities. Although the value of expansion is high in community 1 with 0.318 spreading edges connected to different communities. Expansion value of community 2 is 0.175 and this value is smaller than the expansion result of community 1.

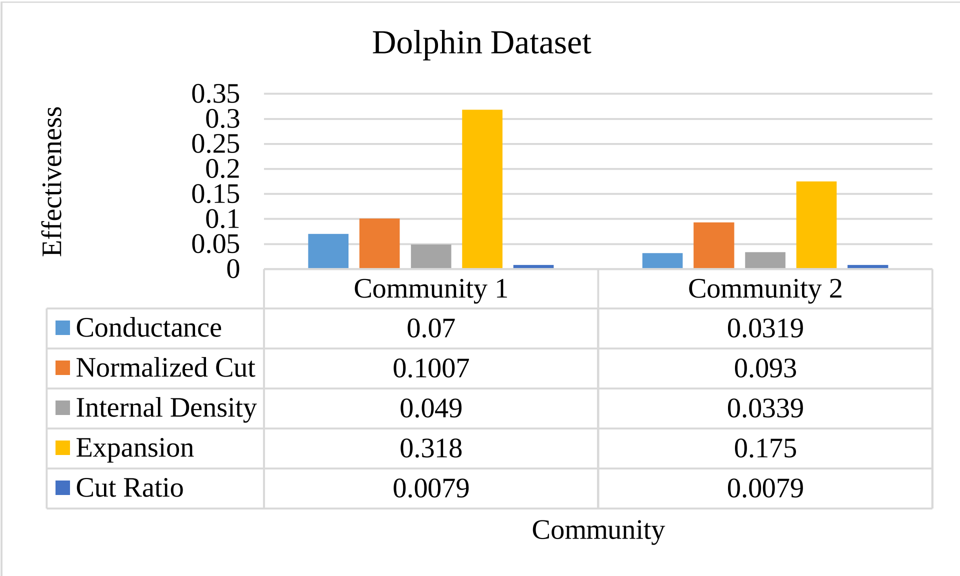


Figure 5. 11 Effectiveness on Dolphin Network

Figure 5.12 demonstrates the US Political Book network dataset. It shows that the expansion value 1.7 reaches the highest value in community 3. The external connectivity cut ratio get effectiveness results nearly 0 because they have sparsely connections to other external communities. Community 3 has the highest expansion value than the other communities and all communities have effective internal density result. Conductance value reaches 0.27 and normalize cut value reaches 0.31 in community 3. Both external and internal connectivity metrics of community 3 is higher than the other two communities.

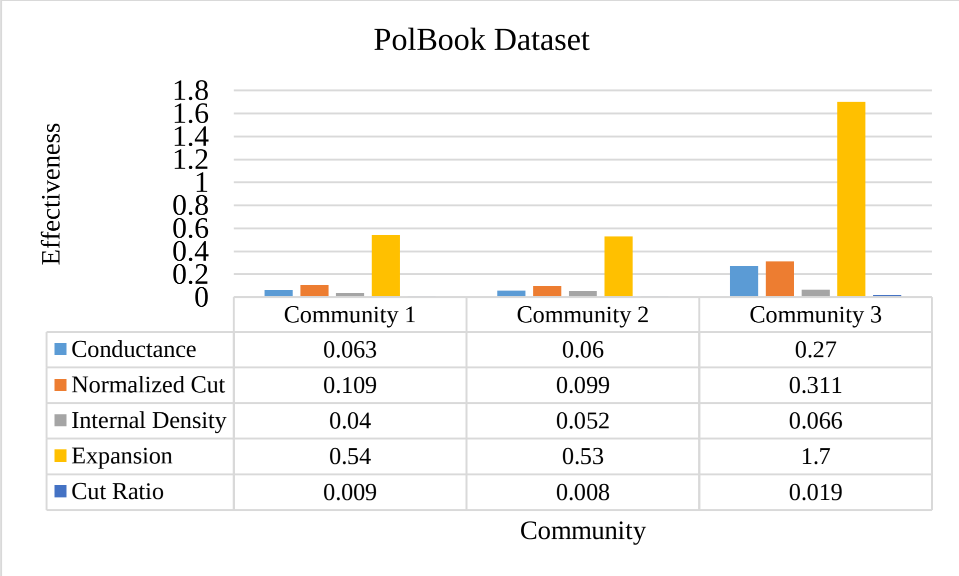


Figure 5. 12 Effectiveness on Book Network

Figure 5.13 illustrates the results of the modularity, and NMI for four real networks are achieved by using the proposed algorithm. This shows that the node measurements collected on the set of nodes have close relationship between the nodes in the community. The modularity Q value reaches the value between minus one to one, and modularity value is generally 0.3~0.7 range in real world network. The greater value of modularity, community structure is obvious. NMI values is one for first two datasets, Employee and Zachary Karate club dataset. The proposed algorithm gets suitable accuracy result in Dolphin and Book datasets. Figure 5.14 shows the modularity results of four dataset compared with LPA and ground truth.

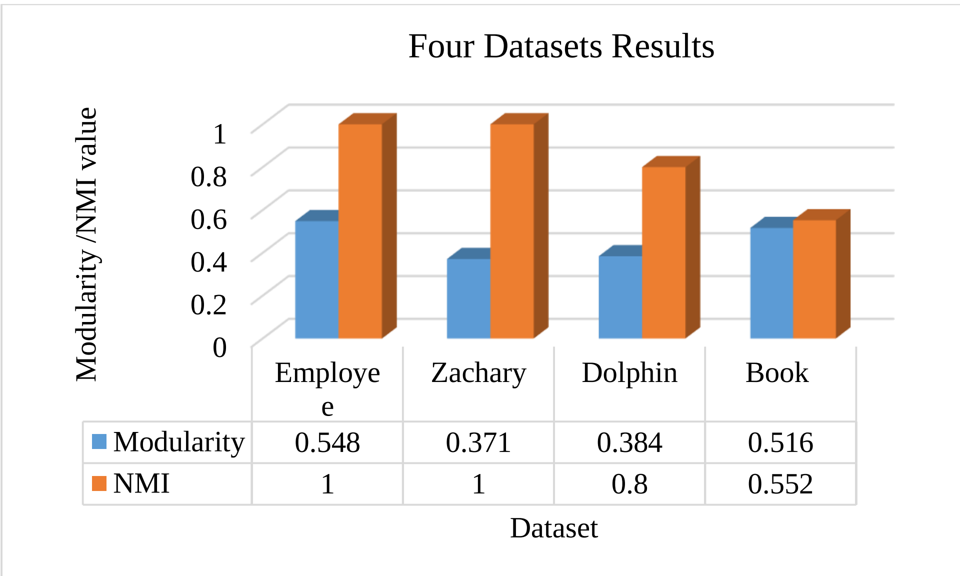


Figure 5. 13 Modularity and NMI results of Four Dataset

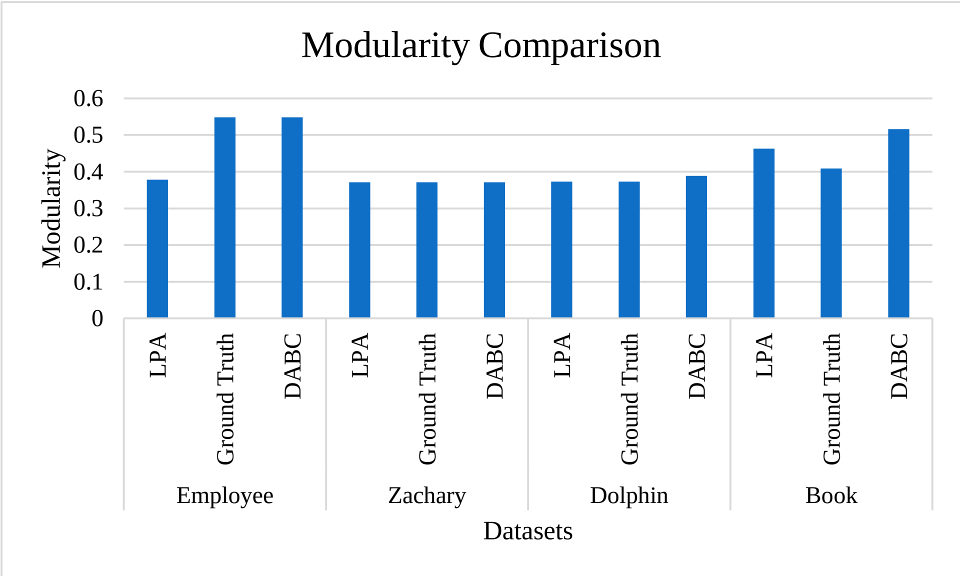


Figure 5. 14 Modularity Comparison on Four Network Datasets with Three Algorithm

The proposed algorithm experiment is also compared with the well-known community detection algorithm LPA (Label propagation algorithm) and ground truth result. The comparisons are made on the four types of scoring function such as network model, internal connectivity, external connectivity, and internal connectivity.

Table 5.16 shows the comparison on the network model function, modularity. The standard quality measurement of community structure is modularity (Q). The greater the modularity value, the better the result of communities in the sub group. The network model-based objective function needs to maximize. In this result, the proposed algorithm gets the same modularity result as ground truth and LPA in the first two dataset Employee and Zachary karate dataset. In Dolphin and Book dataset, it gets the more effective modularity result than the others.

Combined internal and external connectivity metrics such as conductance and normalized cut is used to describe the effective of the proposed algorithm. The group of this objective functions requires to be minimized includes conductance and normalized cut. These functions calculation depend only on the network. According to the result of Table 5.17, D-ABC detected the community structure with effective result of conductance and normalized cut in Employee and Zachary karate datasets exactly equal to the ground truth dataset.

Table 5. 16 Comparisons of Network Model with Ground Truth Result and LPA Algorithm

Dataset	Algorithm	Modularity
Employee	LPA	0.378
	Ground Truth	0.378
	D-ABC	0.378
Zachary	LPA	0.371
	Ground Truth	0.371
	D-ABC	0.371
Dolphin	LPA	0.373
	Ground Truth	0.373
	D-ABC	0.389
Book	LPA	0.463
	Ground Truth	0.409
	D-ABC	0.516

Table 5. 17 Both Internal and External Connectivity Metrics Comparisons on Four Dataset

Dataset	Algorithm	Conductance	Normalized cut
Employee	LPA	0.234	0.35
	Ground Truth	0.236	0.352
	D-ABC	0.236	0.352
Zachary	LPA	0.256	0.46
	Ground Truth	0.256	0.46
	D-ABC	0.256	0.46
Dolphin	LPA	0.091	0.174
	Ground Truth	0.091	0.174
	D-ABC	0.113	0.215
Book	LPA	0.046	0.65
	Ground Truth	0.977	1.19
	D-ABC	0.393	0.519

LPA algorithm gets same results as ground truth in Zachary and Dolphin dataset. For dolphin dataset, the result of proposed algorithm is less than others. But the algorithm gets better results than other algorithms in PolBook network dataset. When seeing overall results, the proposed algorithm produces more suitable results for all datasets. For external and internal connectivity metrics, the higher the conductance and normalize cut, the worse the detected communities are.

Comparisons on external connectivity quality functions to measure all clusters quality are discussed in Table 5.18. The quality of external connectivity of each community can define as the less, the better. According to the results, D-ABC get better values on the three datasets as ground truth values except dolphin network.

Table 5. 18 Comparisons on External Connectivity Metrics

Dataset	Algorithm	Expansion	Cut ratio
Employee	LPA	0.75	0.048
	Ground Truth	0.76	0.048
	D-ABC	0.76	0.048
Zachary	LPA	1.176	0.069
	Ground Truth	1.18	0.069
	D-ABC	1.18	0.069
Dolphin	LPA	0.442	0.014
	Ground Truth	0.442	0.014
	D-ABC	0.552	0.0178
Book	LPA	3.51	0.042
	Ground Truth	6.324	0.08
	D-ABC	2.77	0.036

The modularity result of D-ABC is compared with the modularity accuracy results of FG (Fast Greedy) [9], (BOCD) bi-objective Community Detection [1] and EA (Evolutionary Algorithm). The following Table 5.19 describes the Modularity and NMI quality values of each dataset. D-ABC achieves more accurate results on all datasets because it uses the ground truth community numbers of three networks. It makes the algorithm more targeted and can improve the detection accuracy of the communities. Bi-Objective CD provides more modular results than other algorithms in karate club and dolphin dataset. Its community number differs slightly from the real community number. FG and EA receive the appropriate modular value and NMI results in all datasets. D-ABC gets better modularity value and NMI result in football dataset.

Table 5. 19 Modularity and NMI Results Comparison of Four Algorithms using Three Datasets

Algorithm	Karate		Dolphin		Football	
	Modularity	NMI	Modularity	NMI	Modularity	NMI
D-ABC	0.371	1	0.38	0.835	0.59	0.89
FG	0.38	0.69	0.49	0.55	0.54	0.7
BOCD	0.419	0.695	0.507	0.615	0.577	0.878
EA	0.38	0.83	0.46	0.78	0.56	0.76

As demonstrated in Figure 5.18, the proposed approach shows the effectiveness of modularity values on different datasets. These datasets cannot get ground truth data. Therefore, their qualities are measured by using network modularity result. The experiments make map reduce procedure of Spark framework on standalone computer. According to the table, the proposed algorithm gets effective modularity result in all datasets.

In Figure 5.18, Infomap, label propagation and the proposed algorithms show the modularity results by comparison on four networks that have different sizes. LPA algorithm gets effective community results in Facebook dataset, but it cannot detect effective community in email dataset. Infomap algorithm detects suitable communities in all datasets and cannot detect dblp dataset that contains million nodes and billion edges. According to modularity results, the proposed algorithm effectively detects the community structures with optimized modular values in all datasets.

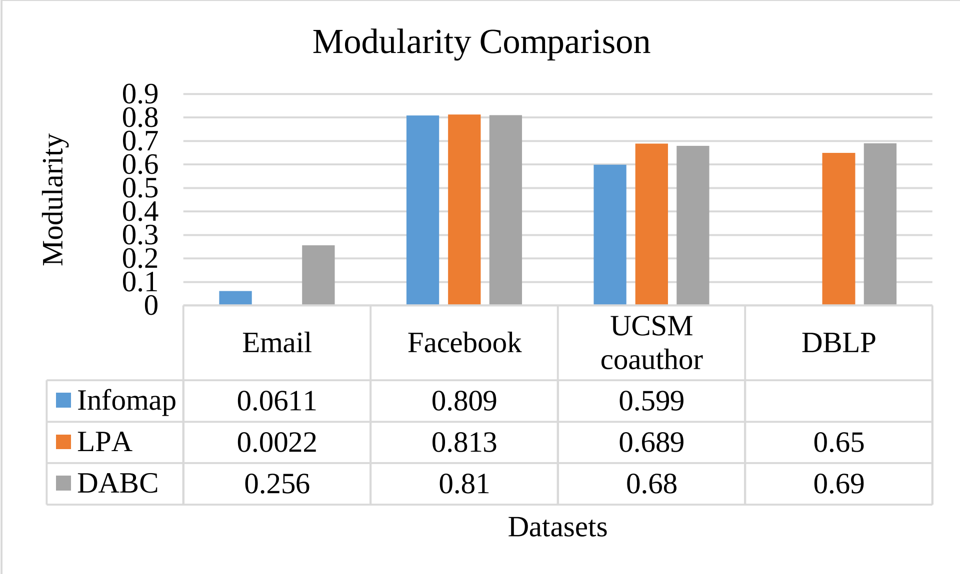


Figure 5. 15 Comparison of Modularity Results on Four Datasets

5.7 Comparison with Other Well-Known CD Algorithms

To show the proposed algorithm's effectiveness, the researcher has compared the modularity results of the proposed algorithm with other traditional algorithms and well-known meta-heuristic algorithms named as Genetic Algorithm (GA) and Artificial Bee Colony Algorithm (ABC) on some dataset.

Firstly, the finding results based on ground truth data are compared with ABC algorithm. Figure 5.16 shows the modularity and accuracy measurement results comparison using two algorithms on some datasets. According to the result ABC gets more effective result on small size networks and D-ABC gets more effective modularity on football network. Moreover, D-ABC gets higher accuracy than ABC algorithm in all datasets. The proposed algorithm uses ground truth community number to get more effective community results. When the datasets have actual community number, D-ABC will get the suitable modularity result and can also get more accurate NMI results on all datasets.

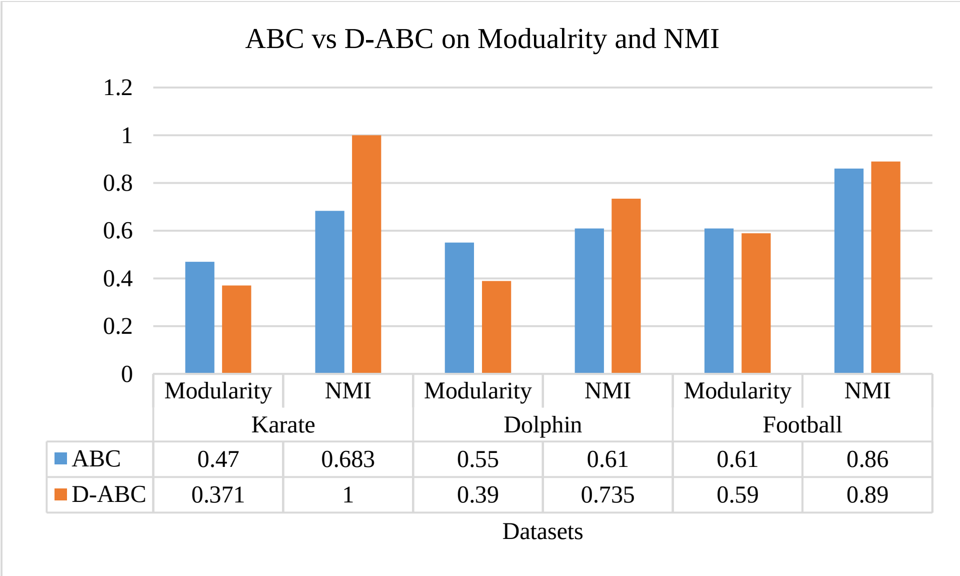


Figure 5. 16 Modularity and NMI Companions between ABC and D-ABC

ABC algorithm used graph-based solution representation that is also called locus based solution representation. This representation requires encoding and decoding step for the result community for each solution.

The datasets that do not know the community number, the proposed system calculates the number of community using label based encoding schema. But to reach more optimal solution quickly, the proposed system calculated the community number by using square root of the number of nodes. This calculation will be used if the number of node is greater than 16. For below 16 members, the number of community sets the half of the node number to reach optimal solution quickly. The maximum number of clusters is not greater than the number of nodes. The purpose of the community detection algorithm is to get better modularity result. However, the proposed system clusters the datasets with better modularity results.

Table 5.20 compares the modularity result of four datasets with other types of community detection algorithms. D-ABC gets more modular values in all datasets. Louvain dataset gets better modular value in karate club dataset and Infomap gets effective modularity result in Dolphin dataset. All algorithms get similar modularity results in football dataset except Leading Eigenvector.

Table 5. 20 Modularity Values Comparison of Six Algorithms on Four Datasets

Algorithm	Karate	Dolphin	Football	Book
Infomap	0.402	0.52	0.598	0.496
Label propagation	0.402	0.51	0.598	0.498
Leading eigenvector	0.393	0.491	0.492	0.465
Louvain	0.419	0.519	0.598	0.515
Walktrap	0.353	0.488	0.598	0.501
DABC	0.419	0.526	0.598	0.523

To compare the Genetic Algorithm, the proposed system set population size=100, iteration100 as GA algorithm on four datasets such as Karate, Dolphin, Football and Employee networks. Table 5.21 is the comparison of GA and D-ABC for four datasets based on modularity value. In this table, Com-No is the detected community number of each dataset.

Table 5. 21 Modularity Comparison of Two Algorithms on Four Datasets

Algorithm	Karate	Com-No	Dolphin	Com-No	Football	Com-No	Employee	Com-No
GA	0.4092	4	0.5342	4	0.5695	7	0.591	4
D-ABC	0.419	4	0.526	4	0.598	8	0.563	4

Two algorithms use 100 iteration size and solution is 100 in all dataset. According to the results, D-ABC gets effective result than GA algorithms on two datasets such as Karate and football.

Figure 5.17 and figure 5.18 demonstrate the modularity result of Zachary karate club dataset in each iteration. Figure 5.17 is the result of modularity using D-ABC algorithm. It shows that modularity result rapidly increase at iteration number among 8 to 15. Iteration number 12 to 35 has slightly increased. When reaching the iteration number over 35, the modularity result is stable, it does not change until iteration number is 100. In Figure 5.18, modularity values of karate dataset using GA has changed before iteration number 10. Then the result is stable until maximum iteration.

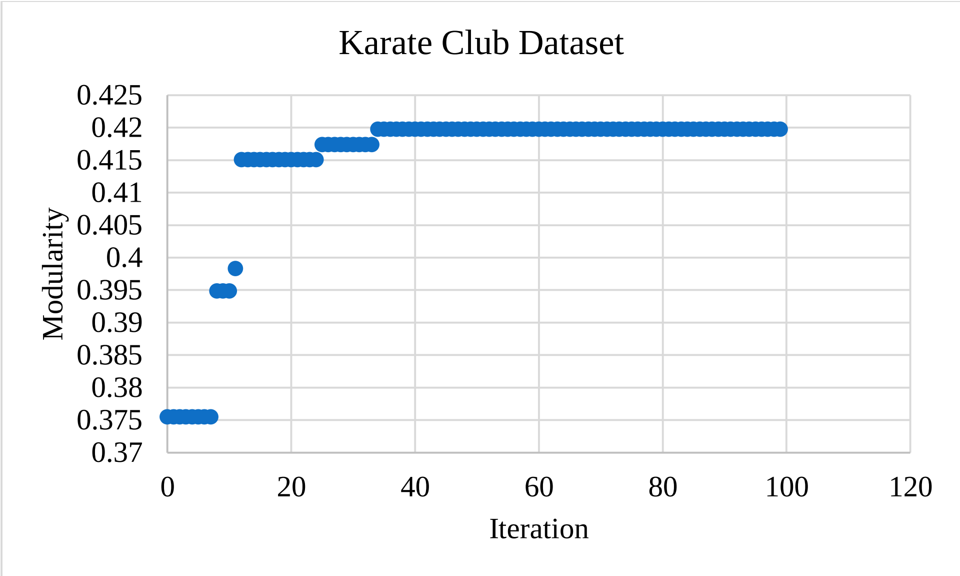


Figure 5. 17 Modularity Results of Karate Dataset in Each Iteration Using D-ABC

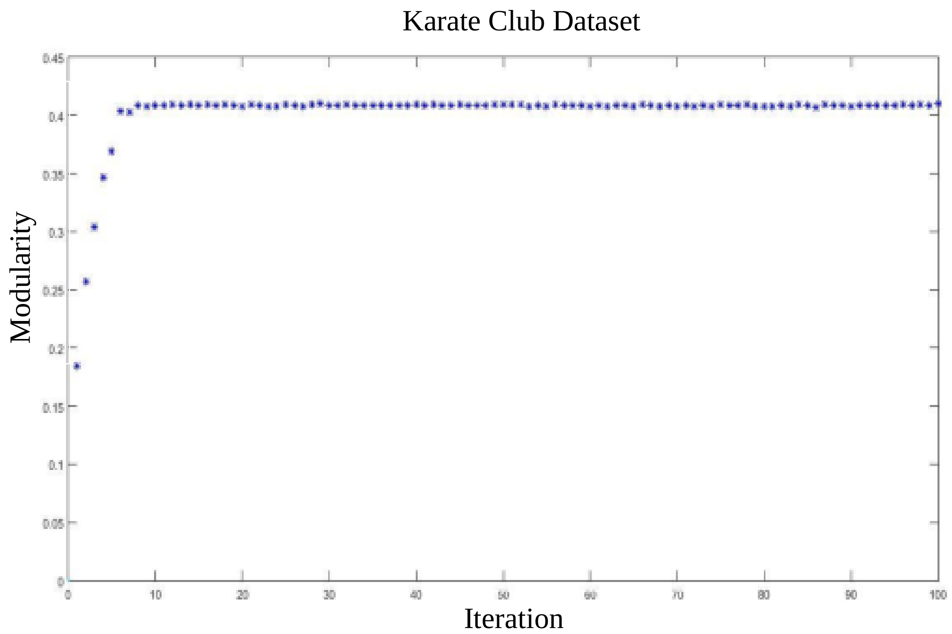


Figure 5. 18 Modularity Results of Karate Dataset in Each Iteration Using GA

The result variation of Dolphin dataset is also described in Figure 5.19 and Figure 5.20. The modularity values change in each iteration before reaching the iteration number 60 in the proposed system. Then modularity value 0.52 is stable until the maximum iteration. The modularity values change in GA algorithm before iteration number 20 and fix result are reached after iteration number 20.

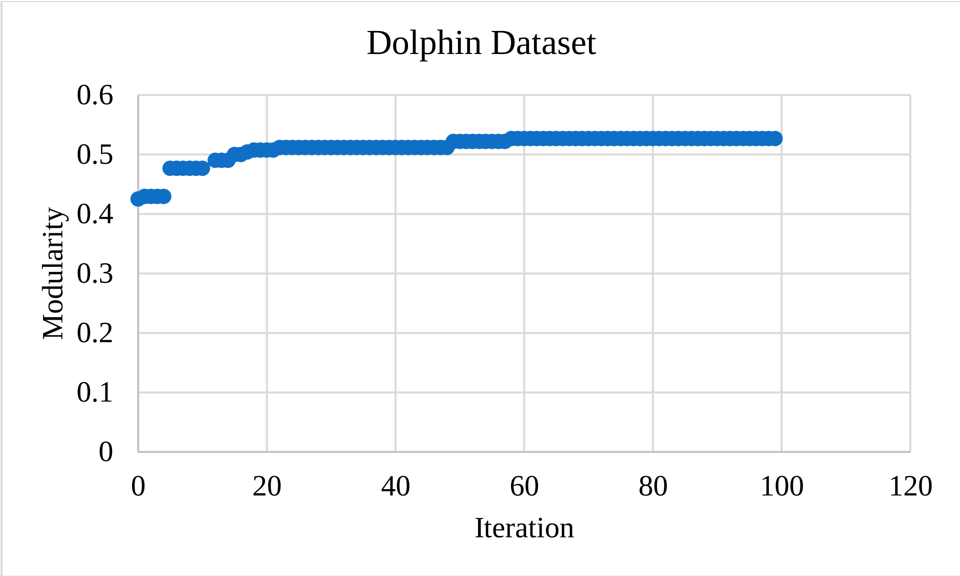


Figure 5. 19 Modularity Results of Dolphin Dataset in Each Iteration Using D-ABC

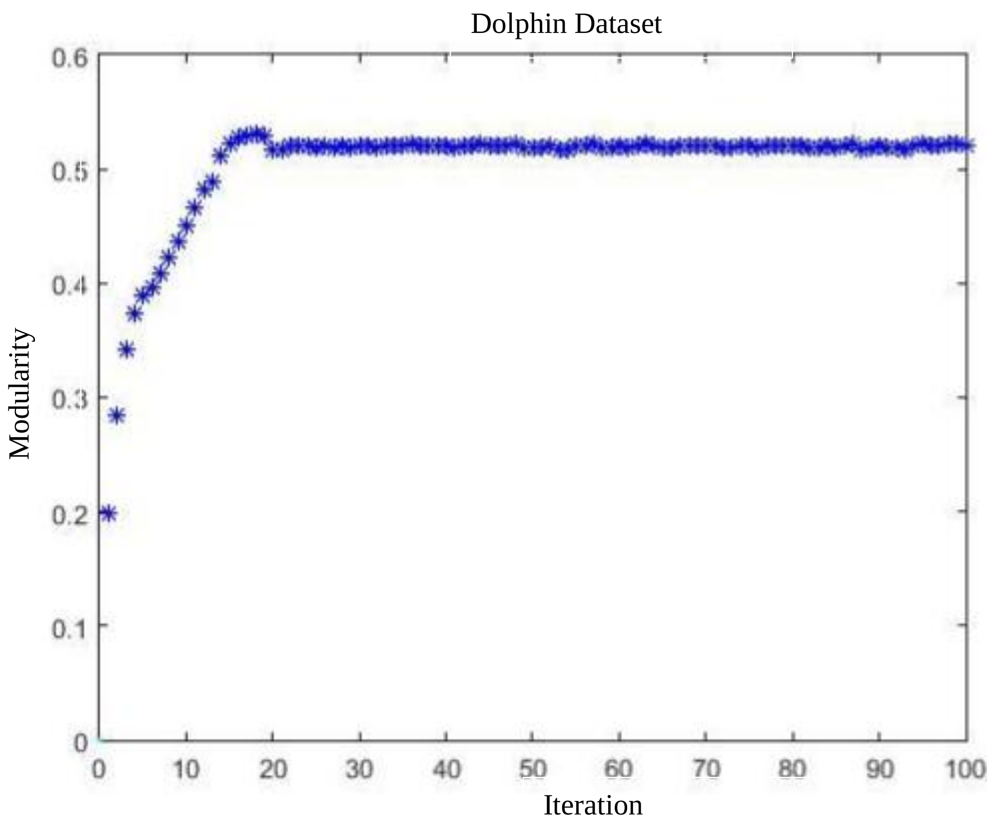


Figure 5. 20 Modularity Results of Dolphin Dataset in Each Iteration Using GA

Figure 5.21 shows the evolving modularity result in each iteration of datasets using D-ABC and 5.22 tests on GA iteration on football dataset. When using D-ABC on football dataset, each iteration gets different modularity result. After reaching iteration number 60, the value is stable in 0.59825. When using GA, the modularity values evolves before iteration number 20, the value is stable at 0.5695

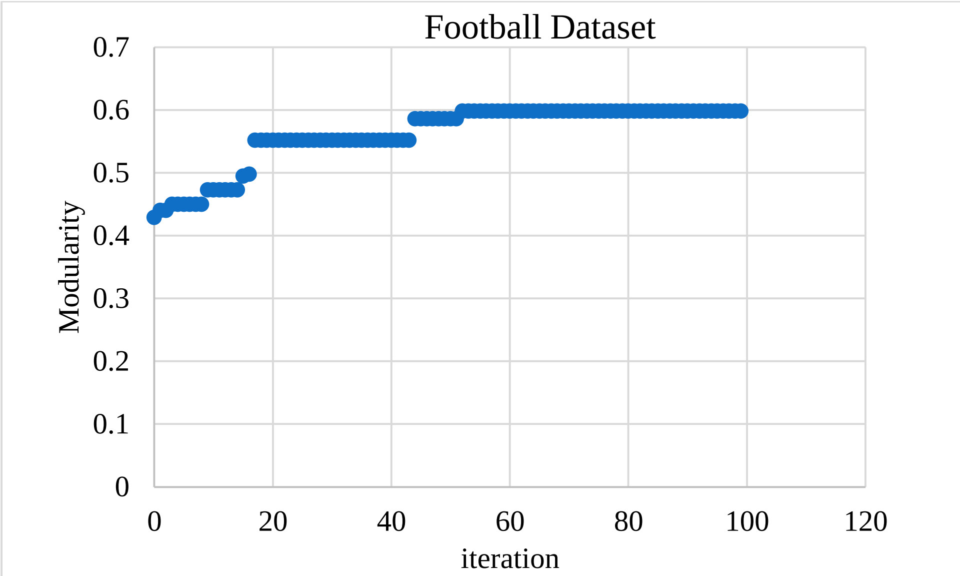


Figure 5. 21 Modularity Results of Each Iteration on Football Dataset Using D-ABC

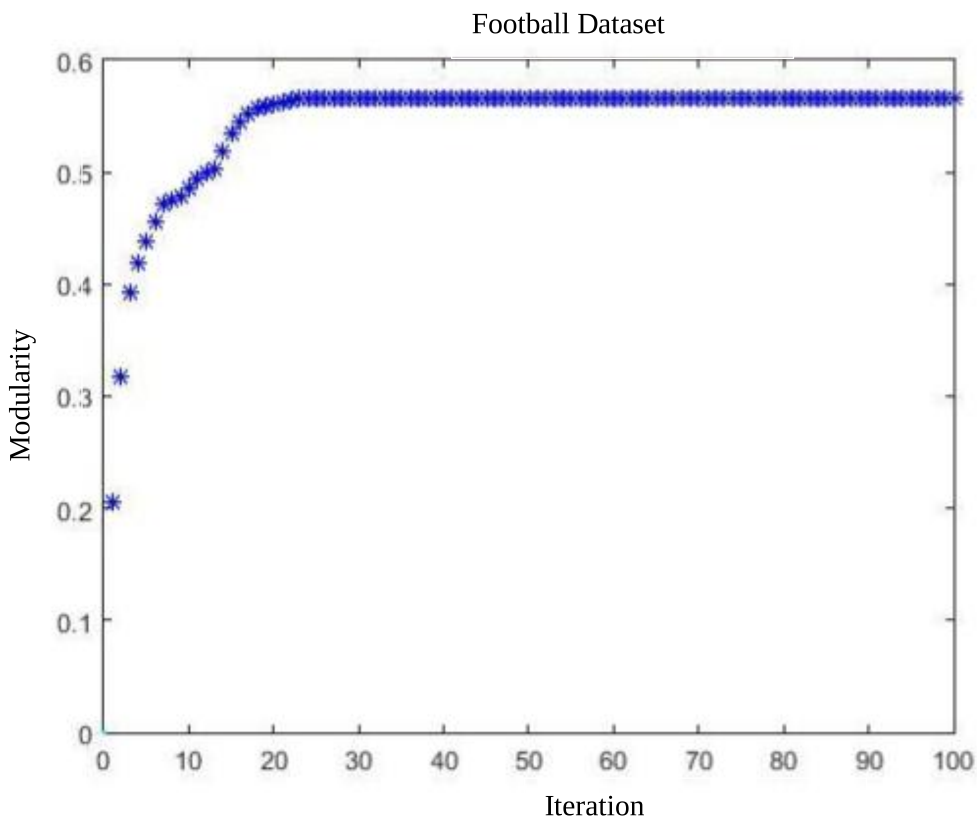


Figure 5. 22 Modularity Results of Each Iteration on Football Dataset Using GA

In Employee dataset, the modularity value evolution can be seen before iteration number 10 in both algorithms. When using D-ABC, the modularity value reaches 0.561

shown in Figure 5.23. The Figure 5.24 shows the modularity result in each iteration by using GA. The modularity value of GA also stables before iteration number 10.

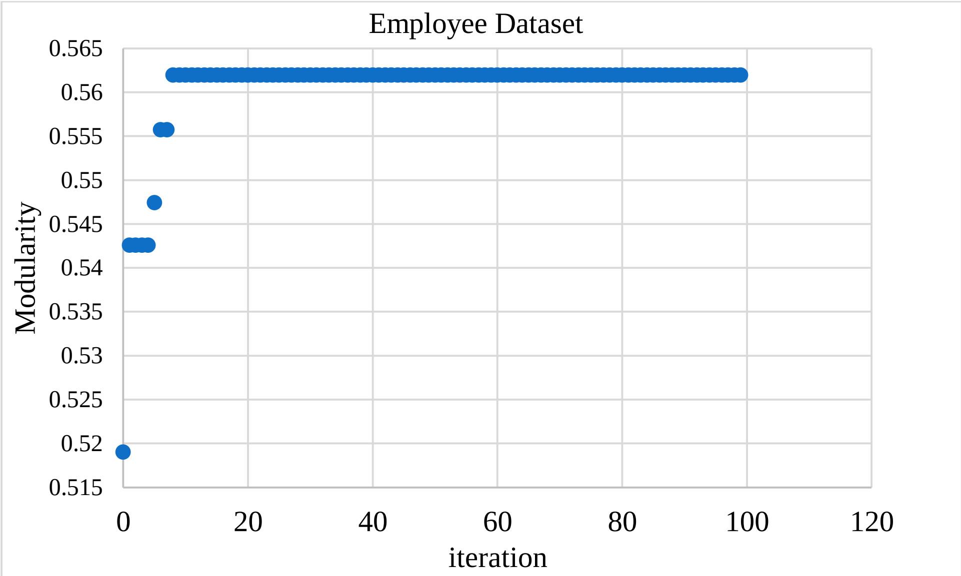


Figure 5. 23 Modularity Results of Each Iteration on Employee Dataset Using D-ABC

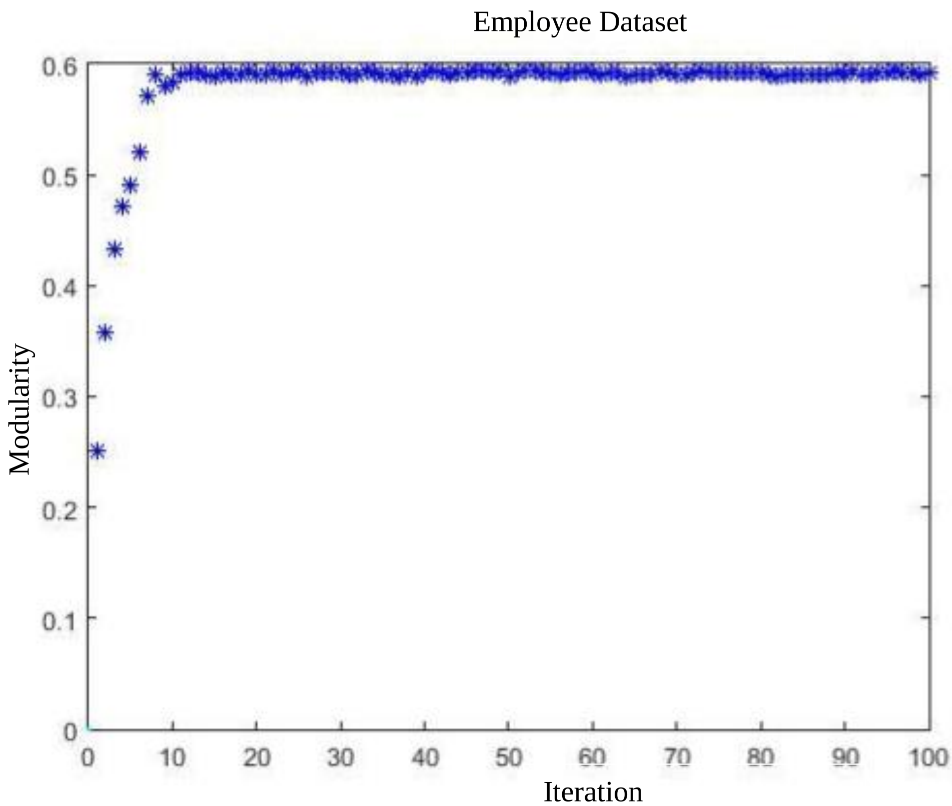


Figure 5. 24 Modularity Results of Each Iteration on Employee Dataset Using GA

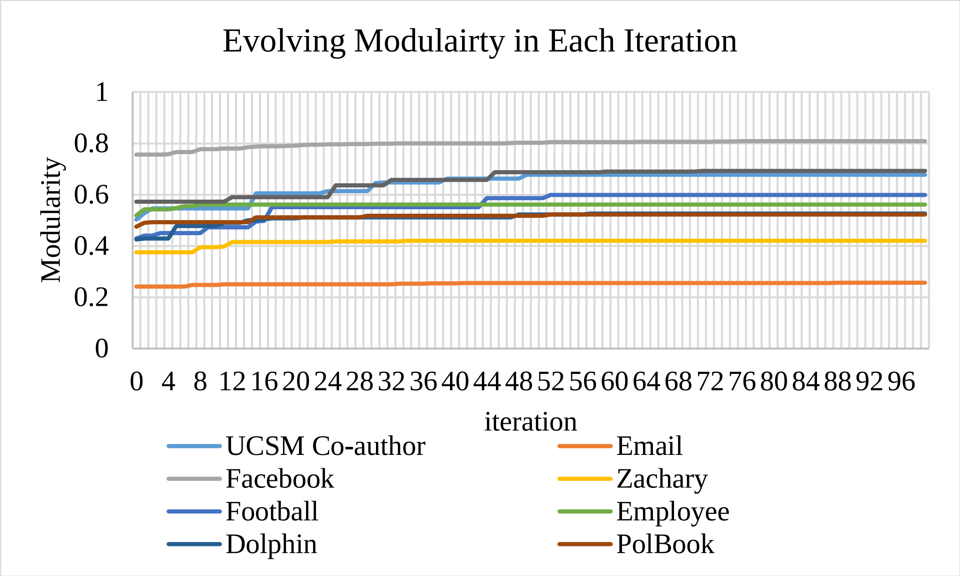


Figure 5. 25 Evolving Modularity Results in Each iteration for Nine Datasets

Evolving modularity values in each iteration are shown in figure 5.25 for nine datasets. All of the datasets use maximum iteration 100 and population size 100. After reaching iteration number 50, most of the datasets get the optimal result and their modularity results are slightly changed. The time taken for each dataset is different. Small datasets take under an hour to reach the optimal solution and large datasets take more than one hour to reach the optimal solution.

5.8 Findings and Discussions

The proposed work in this system was accomplished by thoroughly evaluating the well-known community detection algorithms such as LPA, ABC,BOCD, GA ,infomap and many other algorithms. Based on the above detailed assessment which covers many aspects of community identification, that the proposed approach has its strengths and weaknesses. The algorithms have their own efficiency and can be adapted to various performance measure and different characteristics of datasets. The system needs ground truth result to measure the accuracy of the algorithm when using NMI metric. The results show that the proposed algorithm works well with NMI.

There are many kinds of quality measure functions for community detection problem, among them the proposed system used inter cluster density, Expansion, Cut Ratio, conductance, normalized cut and modularity. Finally, the results demonstrated on different kinds of networks show the valuable performance which covers the accuracy, efficiency and effectiveness of the proposed algorithm. The proposed system uses the ground truth community number to reach optimal solution quickly, if the dataset has known community number. Based on this community, the system detects the community structures of the given network to get effective modularity results and clusters. The number of community is calculated based on the number of vertices of the network when the dataset does not know the exact community number. It also get the effective results in four networks as GA algorithm. The system tested various types of population initialization. The proposed initialization is the more suitable for the proposed discrete artificial bee colony algorithm. The algorithm purposes that the output community structures of the given network have the high modularity result. Therefore, population size and iteration are also adjusted to get better result.

The proposed algorithm detects more effective community results on densely connected networks such as Employee and karate club datasets. It also detects the community structures of other datasets compared with other researcher's proposed algorithms that can get from their research papers and Github. According to experimental results, the proposed algorithm benefits to the performance which covers the efficiency, effectiveness and accuracy.

5.9 Summary

This section tests the experiment on certain social network datasets with ground truth information. Then, it discusses measurement of the accuracy called NMI score. In this measurement, how accurate the proposed algorithm's work must be, when the result is matched with the ground truth result of dataset. It also describes the experiments on the other real network datasets that does not know ground truth results.

It then discusses the community results comparison with other optimization based community detection algorithms and the proposed algorithm. It also described six well known evaluation methods which are used to assess the effectiveness of the outcome community structure. These methods are modularity, internal density, expansion, cut ratio, normalized cut and conductance. This chapter presents the effectiveness of the proposed system by evaluating the experiment on some kinds of ground truth networks and other networks.

CHAPTER 6

CONCLUSION AND FUTURE WORK

People in society have always preferred to stay in community by forming small or large communities. Community may develop based on race, color, religion, interest, gender, occupation and hobbies. Communities in a network are collection of nodes that have more connected to each other in a cluster than nodes in the other parts of the network. Today, social community discovery becomes a main topic in the field of computational intelligence and it has long been a topic of social research as it relates to urban studies, social marketing, criminology and many other disciplines.

Several previous community detection techniques were developed to find community in complex social network by using the hierarchical relationships within the network. In general, those methods cannot obtain the exact solution. So, community detection problem becomes optimization problems where the objective function is the maximization of some quality index. In this thesis, artificial bee colony optimization based algorithm is proposed to detect community in social network. It also depends on modularization of the network. The performance of the algorithm directly influence the quality function of the problem. The proposed algorithm was tested on various types of real world networks. Experimental results confirm the validity, efficient and the efficiency of this method. The proposed algorithm called Discrete Artificial Bee Colony (D-ABC) Algorithm is an optimization technique that works effectively for the community detection problem according to the experimental results.

The optimal results demonstrate that the performance of the proposed approach give the promising result in terms of accuracy, and successfully finds the most suitable modularize community structure. In this chapter, the main contents of the proposed system are summarized and future research direction is also suggested.

6.1 Research Conclusion

The efficient community detection algorithms is very important for social network analysis that contains millions of nodes and edges. Discovering community also known as the graph partition in social network reveals hidden relationships between nodes in it. The members in each cluster have dense relationships than other members in the network. In this thesis, it considers on the undirected and unweighted graph. Adjacency list and matrix are used for graph creation in traditional graph algorithm. When the graph size is large, they will not create a graph structure representation. Therefore, GraphX in Spark Framework is used to represent graph structure of real world network. Feature analysis and predicting the behavior of complex networks is the attractive research areas in the theories and behavioral implications for topology analysis. Finding the structure of the community can give many effectiveness for the above works. Results can be used in recommendation system, link prediction, and influence user analysis and customer segmentation.

In classification and cluster analysis, accuracy is a commonly used measure. It uses the real category information of the data as a benchmark to measure how many data are divided into the correct groupings. They can get exact one solution. But community detection problem in social network does not solve to reach the exact solution, it can produce optimal solution. The researcher tries to get optimal solution for community detection in social network problem.

Modified version of traditional artificial bee colony algorithm is used to detect community structure in graphs based on Spark Framework. ABC is used in numerical optimization problem. D-ABC can also be used use in discrete type problem such as community detection. GraphX in Spark is used to create graph instead of different type of graph representation such as adjacency list, matrix. It does not need the information of how many communities are to be generated and it is not given as an input. The applicable initialization scheme and search strategy of bees in ABC is proposed to solve discrete problem. To evaluate the success of the algorithm, Modularity Q is used as fitness function. For ground truth dataset, normalize mutual information is used to test the algorithm's accuracy. If datasets which doesn't have ground truth results, community scoring functions will be used. They are calculated based on external connectivity, internal connectivity, both external and internal connectivity and network model. These techniques will be used to judge community quality.

The proposed system also evaluates effective modularize results of output community in the real world social network. Finally, six kinds of quality functions such as modularity, conductance, normalized cut, expansion, cut ratio and internal density is used to measure the quality of result community. The more the modularity value, the better the result. For the other metrics results, the less the better.

To demonstrate the effectiveness of the proposed algorithm, various test on five small real world networks with ground-truth communities are carried out. At the same time, test the performance of the proposed algorithm on different size of real networks assume non ground truth dataset. It does not need any prior knowledge about the number of communities and community size. It turns out that this work is done in a modular way compared to the other well-known optimization algorithms such as artificial bee colony algorithm, genetic algorithm and many other CD algorithms. By doing this, the proposed system achieved the good community list for the target network.

6.2 Findings and Discussions

Generally, it can be divided into two kinds of community detecting algorithm. These are the algorithm of disjoint and the overlapping community division. The proposed system used on disjoint community detection. Firstly, the raw graph data will be loaded. If the data format of graph is directed, the proposed system treat directed graph as undirected graph with each edge represents two directed edges. This corresponds to the parameter of graph structure information. This step makes the proposed algorithm suitable for Spark platform.

Secondly, initialize the food source of D-ABC. Two types of encoding schema such as locus based and label based can be used for initialization step. Locus-based encoding schema needs the decoding step to know the community members. Therefore, label based solution initialization is chosen for population initialization. Then calculate the fitness of each food source using chosen fitness function and then search the best food source using three kind of bees. There are many kinds of fitness function to adjust the result community structure. All metrics have pros and cons. Modularity function has resolution limit problem. A paper using ABC algorithm proposed by Ahmed Ibrahem Hafez and co-author used many kinds of objective functions as the fitness function of the algorithm. According to the results, some metrics get effective results in some dataset, but not in other datasets. Averagely, modularity function gets proportional result in all dataset. So, modularity is used as the fitness function of the proposed algorithm.

In Initialization procedure, that generates the food source and initialize the position according to the ID of each node so that all these food sources are the same as other. Node degree centrality is used to find the similar nodes or connectivity nodes in a community. If the node has the most interaction to other node, it will be the most important node in a group. For one third of each food source, randomly choose node and change its cluster to the majority clusters of its neighbors. This process is repeated several times. The effective initialization strategy makes the convergence of algorithm faster. The system shows that string-based solution with the appropriate modifications is a more reliable and efficient method in detecting communities in large-scale networks than the popular CD methods.

If the dataset does not have ground truth community, label-based encoding schema can produce the number of community for the given network. It does not need to know the group size and group number. To reach quickly the convergence result, the number of community is calculated by using the square root of the number of nodes. This calculation is useful for sample test datasets. The number of iterations and population sizes are also important for evolutionary algorithm. For some dataset such as Zachary karate club, employee and dolphin dataset needs population size 50 and iteration 50 to get optimal modularity result. At this iteration and population size, the proposed algorithm gets optimal community result. The rest datasets require population size 100 and iteration require more than 100 depend on the network size to get optimal result. It does not need to set same iteration number and population size.

In the search strategy, the equations proposed in traditional artificial bee colony algorithm can be used for classification and clustering for continuous problem. Therefore, the nature of ABC is changed to solve discrete problem. ABC has poor exploitation and is good at exploration. Exploration is made in scout bee and exploitation is made in employed and onlooker bees. When finding the new food source in employed and onlooker bee phase, one point crossover based operation is used to swap the food source and adjust to get better exploitation result.

If the external connectivity among communities is lower than the internal connectivity, then the system can detect the significant community. The low values of the internal, external connectivity and the large values of modularity are achieved by using this system.

Some nodes which are tightly connected to their neighbors may be comparable with the most centrality node whose neighbors are sparsely connected with each other and the communities may be deviated from the ground truth. The lowest similarity among communities can be clearly separated from the significant community, however; the more similarity values between the communities may yield the higher deviation of communities from ground truth. The separation power among communities of the proposed system achieved the slightly high conductance. The values for the number of edges per node that points outside the cluster are also slightly high in expansion.

In the efficiency of this work, all food sources are evolved in two phases such as employed and onlooker bee phase because of the nature of the artificial bee colony algorithm. Calculation take twice time to evolve solution in this search strategy. The proposed system take O (log2cpn). Generally, all of the optimization algorithm take O (cpn) time to get optimal solution. It does not consider the time taken of the objective functions calculation. Time taken for computation is slightly high because of single machine memory resource and processor.

The accuracy of the proposed system may decrease than the method without considering the number of ground truth community. But the modularity may increase than some methods and the high value of internal density that shows in the experiments. Therefore, all of the parameter will be adjusted to produce the well-connected community in social networks.

6.3 Limitations of Research

This study describes several important parts of the experiment and aware of some limitation and bias of this work. One limitation is the usage of network dataset and another is that the system only considers the edge connection between nodes and doesn't consider the attribute value of nodes. Therefore, the accuracy is not accurate in some datasets.

The research only considers the edge information of the network not consider the attribute and weight of the network. The community structure of some social network overlaps but the system considers only finding unrelated communities in complex social network. The system will not stop when it does not evolve the solutions that the solution does not change in many iterations. It will run until the maximum cycle is reach. The fitness function, modularity has resolution limit problem [18]. If the number of community increases in small size dataset, the proposed algorithm produce many communities and some clusters will contain only one member. The proposed system will not produce the effective community structures in this time.

6.4 Future Research Direction

The results of this proposed system are superior to other algorithms. But there are still issues that require further investigation and discussion. In future, this system will need to consider the weight value of nodes and the important node attribute value on various social network community detection. One way to improve the proposed system is to consider the important node that has many degrees, which may reach more optimal results in community detection.

Future research development in parallel computing, distributed architecture, big data, supercomputing will be making it really easy to study, inspect and evaluate social networks efficiently. Future work will be considered to detect community on large scale social graph, and GraphX on Apache Spark will also be used for graph parallel processing.

To make the community detection algorithm more efficient, it can combine measurements similar to node weights and the importance of detailed information about the impact of nodes on community detection in complex networks.

The future work may be focused on setting some criteria to improve the accuracy and the scalability of community detection problem. Future works will also calculate the fitness functions on multi-objective function for community detection problem to improve more quality of the results. Future research will be implemented on distributed Spark framework for multi-objective optimization community detection on large scale networks.

AUTHOR'S PUBLICATIONS

[P1] Thet Thet Aung and Thi Thi Soe Nyunt, "Scalable Community Detection using Island based Artificial Bee Colony Algorithm", 16th International Conference on Computer Applications, 2018, February, 486-493

[P2] Thet Thet Aung , " Community Detection in Social Network using Artificial Bee Colony with Genetic Operator", The 2nd International Conference on Advanced Information Technologies, 2018, November, 45-50

[P3] Thet Thet Aung and Thi Thi Soe Nyunt, "Community Detection in Social Graph using Nature inspired based Artificial Bee Colony Algorithm with Crossover and Mutation", 2019 IEEE 4th International Conference on Computer and Communication Systems,Singapore, 2019, February, 213-217

[P4] Thet Thet Aung and Thi Thi Soe Nyunt, "Discrete Artificial Bee Colony Algoritym for Community Detection in Social Networks", The 17th International Conference on Computer Applications, 2019, February, 53-59

[P5] Thet Thet Aung and Thi Thi Soe Nyunt, " Modularity based ABC Algorithm for Detecting Communities in Complex Networks", 2019 3rd asia conference on machine learning and computing(ACMLC 2019), 2019 December

[P6] Thet Thet Aung and Thi Thi Soe Nyunt, "Community Detection in Scientific Co-Authorship Netowrks using Neo4j", The IEEE 18th International Conference on Computer Application 2020, February, 29-35

[P7] Thet Thet Aung and Thi Thi Soe Nyunt, "Modularity based ABC Algorithm for Detecting Communities in Complex Networks", International Journal of Machine Learning and Computing, 2020, February, Vol.10 NO 2, 323-329

BIBLIOGRAPHY

[1] R.Agrawal, “Bi-objective community detection (BOCD) in networks using genetic algorithm”, Communications in Computer and Information Science, vol 168. Springer, Berlin, Heidelberg

[2] N. Arora, H. Banati,"A Framework to Mine Communities Using Nature Inspired Algorithms", Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), doi: 10.1109/ICRCICN.2016.7813656, ,2016

[3] R. Babers; A. E. Hassanien; N. I. Ghali. “A Nature-Inspired Metaheuristic Lion Optimization Algorithm For Community Detection”, Computer Engineering Conference (ICENCO), 2015 11th International, IEEE (2015). 978-1-5090-0275-7/15/2015.

[4] R.K. Behera, S.K. Rath, “An Efficient Modularity based Algorithm for Community Detection in Social Network”, International Conference on Internet of Things and Applications (IOTA), 2016, 978-1-5090-0044-9/16/$31.00 ©2016 IEEE.

[5] A. Bienkov, "Astroturfing: what is it and why does it matter?", https://www.theguardian.com/commentisfree/2012/feb/08/what-isastroturfing.

[6] V. D. Blondel, J. Guillaume, R. Lambiotte, E. Lefebvre: “Fast Unfolding Of Communitiesi in Large Networks.” J. Stat. Mech,2008,P10008.

[7] T. Chakraborty, A. Dalmia, A. Mukherjee, N. Ganguly. “Metrics for Community Analysis:A Survey”, ACM Comput. Surv. 50, 4, Article 54 (August 2017), 37 pages, doi : 10.1145/3091106.

[8] D. Chen, M. Shang, Z. Lv, and Y. Fu, “Detecting Overlapping Communities Of Weighted Networks Via A Local Algorithm,” Physica A: Statistical Mechanics and its Applications, vol. 389, no. 19, pp. 4177–4187, 2010.

[9] A. Clauset, M. E. J. Newman, Cristopher Moore. “Finding Community Structure in Very Large Networks”. Physical Review E 70, 066111, 2004.

[10] Y. Cui, X. Wang and J. Li, “Detecting Overlapping Communities in Networks Using The Maximal Subgraph And The Clustering Coefficient”, Physica A, vol. 405, pp. 85–91, 2014.

[11] L.Danon, A. Diaz-Guilera, A. Arenas "The Effect Of Size Heterogeneity On Community Identification In Complex Networks", Journal of Statistical Mechanics: Theory and Experiment, p.P11010 , 2006.

[12] L. Danon, J. Duch, A. Diaz-Guilera, A. Arenas, “Comparing Community Structure Identification", J Stat Mech P09008, 2005.

[13] L. Donetti, M. A. Munoz, “Detecting Network Communities: a new systematic and efficient algorithm”, J. Stat. Mech. (2004) P10012, arXiv:cond-mat/0404652, 25 Oct 2004.

[14] L. Donetti, M. A. Muñoz, “Improved Spectral Algorithm For The Detection Of Network Communities”, AIP Conference Proceedings 779, 104 (2005); https://doi.org/10.1063/1.2008598.

[15] D. Karaboga, "AN IDEA BASED ON HONEY BEE SWARM FOR NUMERICAL OPTIMIZATION”,TECHNICAL REPORT-TR06,Erciyes University, Engineering Faculty, Computer Engineering Department 2005.

[16] X. Ding , J. Zhang ,J. Yang , Y. Shen , “ An Autonomous Divisive Algorithm for Community Detection Based On Weak Link And Link Break Strategy”, Hindawi, Mathematical Problems in Engineering, 15 January 2018, Volume 2018.

[17] E.Moradi, M. Fazlali, H.T. Malazi , “Fast Parallel Community Detection Algorithm Based on Modularity”, 18 th CSI international Symposium on Computer Architecture and Digital Systemm, 2015, 978-1-4673-8023-2/15/$31.00 ©2015 IEEE.

[18] S. Fortunato, M. Barthélemy,"Resolution Limit in Community Detection”. PNAS January 2, 2007 104 (1) 36-41; https://doi.org/10.1073/pnas.0605965104

[19] M. Girvan and Mark E. J. Newman. “Community Structure in Social And Biological Networks”, Proc. Natl. Acad. Sci. U.S.A., 99(12):7821–7826, 2002.

[20] B. H. Good, Y.A. D. Montjoye, A. Clauset, “The Performance Of Modularity Maximization In Practical Contexts”, arXiv:0910.0165, 1 Apr 2010, https://doi.org/10.48550/arXiv.0910.0165 .

[21] M.K. Gottimukkla, Dr H. Seetha, “Overlapping and Disjoint Community Detection in Social Networks- A Review”, Global Journal Of Engineering Science And Researches,ICRTCET-2018, ISSN 2348-8034.

[22] C. Guoqiang, G. Xiaofang, “A Genetic Algorithm Based on Modularity Density for Detecting Community Structure in Complex Networks", 2010 International Conference on Computational Intelligence and Security, 2010.

[23] A.I Hafez, H. M. Zawbaa, A.E Hassanien, A. A. Fahmy, “Networks Community Detection Using Artificial Bee Colony Swarm Optimization,” Advances in Intelligent Systems and Computing, vol. 303, pp. 229-239, 2014.

[24] R. Hosseini, A. Rezvanian. "AntLP: ant-based label propagation algorithm forcommunity detection in social networks", CAAI Transactions on Intelligence Technology, 2020.

[25] J.Ilamchezhian, V.Cyril Raj ,”Spark Framework: A Review”, internaltional journal of Pure and applied Mathematics, volume 118, No18 2018, 4815-4821.

[26] M.Joldos , C.C Technical, "A Parallel Evolutionary Approach To Community Detection In Complex Networks”, 2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE Xplore: 23 November 2017, DOI: 10.1109/ICCP.2017.8117011.

[27] B. W. Kernighan, S. Lin," An Efficient Heuristic Procedure For Partitioning Graphs". The Bell System Technical Journal (Volume: 49, Issue: 2, Feb. 1970), DOI: 10.1002/j.1538-7305.1970.tb01770.x.

[28] J. Kunegis. "KONECT – The Koblenz Network Collection.” , In Proc. Int. Conf. on World Wide Web Companion, pages 1343–1350, 2013.

[29] R. Lambiotte, “Multi-scale Modularity in Complex Network”, 2010, arXiv:1004.4268v1 [physics.soc-ph] 24 Apr 2010.

[30] B.E. Landon, et al., “Using Administrative Data To Identify Naturally Occurring Networks of Physicians”, Med. Care 51 (8) (2013) 715–721.

[31] M. Leng, J. Wang, P. Wang, X. Chen, “Hierarchical Agglomeration Community Detection Algorithm via Community Similarity Measures”, TELKOMNIKA Indonesian Journal of Electrical Engineering, Vol.10, No.6, October 2012, pp. 1510~1518.

[32] J. Leskovec , A. Krevl. "SNAP Datasets: Stanford large network dataset collection", June 2014. http://snap.stanford.edu/data/index.html.

[33] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. “The Bottlenose Dolphin Community of Doubtful Sound Features A Large Proportion of Long-Lasting Associations.” Behavioral Ecology and Sociobiology, 54:396--405, 2003.

[34] C.D. Manning, P. Raghavan and H. Schutze, “Text Classification And Naive Bayes.” Introduction to information retrieval, 1(6), 2008.

[35] J.H.Michael," Labor Dispute Reconciliation in A Forest Products Manufacturing Facility". Forest products journal, 47(11/12), p.41, 1997.

[36] M. E. J. Newman, “Spectral Method For Community Detection And Graph Partitioning”, PHYSICAL REVIEW E 88, 042822, 2013.

[37] M. E. J. Newman, "Fast Algorithm For Detecting Community Structure In Networks", Physical review E, 69(6), p.066133, 2004.

[38] M. E. J. Newman. “Finding Community Structure In Networks Using The Eigenvectors Of Matrices”, Physical Review E 74, 036104 (2006)[30]

[39] M. E.J. Newman, M. Girven, “ Finding and Evaluating Community Structure In Networks”, March 2004, Physical Review E 69(2 Pt. 2):026113

[40] G.K. Orman, V. Labatut, H. Cherifi." Towards Realistic Artificial Benchmark for Community Detection Algorithms Evaluation". International Journal of Web Based Communities, 9(3), pp.349-370.

[41] G. Palla, I. Der´enyi, I. Farkas, and T. Vicsek, “Uncovering The Overlapping Community Structure Of Complex Networks In Nature And Society”, nature, 435(7043), pp.814-818.

[42] C. A. R. Pinheiro, “Community Detection To Identify Fraud Events In Telecommunications Networks” SAS SUGI proceedings: customer intelligence, 2012.

[43] C.Pizzuti, “GA-Net: A Genetic Algorithm for Community Detection in Social Netowrks”, @Springer-Verlag Berlin Heidelberg 2008 , G. Rudolph et al. (Eds.): PPSN X, LNCS 5199, pp. 1081–1090, 2008.

[44] P. Pons, M. Latapy. “Computing Communities in Large Networks Using Random Walks”. In Computer and Information Sciences-ISCIS 2005, 284–293 (Springer, 2005).

[45] U.N. Raghavan, R. Albert, S. Kumara. “Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks”. Physical Review E 76, 036106 (2007).

[46] R. Halalai, C. Lemnaru, R. Potolea." Distributed Community Detection in Social Networks with Genetic Algorithms". In Proceedings of ICCP 2010. Cluj-Napoca, Romania. August 2010.

[47] W. Rand, "Objective criteria for the evaluation of clustering methods." Journal of the American Statistical association, 66(336), pp.846-850, 1971.

[48] J. Reichardt , J. &Bornholdt. “Statistical Mechanics Of Community Detection”, Physical Review E 74, 016110 (2006).

[49] M. Rosvall, D. Axelsson, C. T. Bergstrom, “The map equation”, The European Physical Journal Special Topics 178,13–23 (2010).

[50] H. Sarvari, E. Abozinadah, A. Mbaziira, and D. Mccoy, "Constructing And Analyzing Criminal Networks", in 2014 IEEE Security and Privacy Workshops , 2014, pp. 84-91.

[51] H.-W. Shen, X.- Q. Cheng , and J.-F. Guo, “Quantifying And Identifying The Overlapping Community Structure In Networks”, Journal of Statistical Mechanics: Theory and Experiment, vol. 2009, no. 07, p. P07042, 2009

[52] S.Shirazia, A. Albadvi, E. Akhondzadeha, F. Farzadfarb, B. Teimourpour, “A New Application Of Community Detection For Identifying The Real Specialty Of Physicians”, International Journal of Medical Informatics 140 (2020) 104161

[53] A. Song, M. Li, X. Ding, W. Cao, K. Pu, “Community Detection Using Discrete Bat Algorithm”, International Journal of computer science, 43:1, IJCS_43_1_05, 2016.

[54] S. Soundarajan, J. Hopcroft, "Using Community Information To Improve The Precision Of Link Prediction Methods", in Proc. 21st Int. Conf. on World Wide Web,2012, pp. 607-608.

[55] C. Sun, Q. L. B, H. Li, S. Zhang, Y. Zheng, "Community Outlier Based Fraudster Detection", (2017), pp. 410–421.

[56] L. Tang, H. Lui . "Community Detection and Mining in Social Media." Synthesis lectures on data mining and knowledge discovery, 2010, Vol. 2, No. 1 , Pages 1-137.

[57] M. Tasgin, A. Herdagdelen,H. Bingol, “Communities Detection In Complex Networks Using Genetic Algorithms”, oai:arXiv.org:0711.0491v1

[58] L. Udrescu, L. Sbârcea, A. Topîrceanu, A. Iovanovici, L. Kurunczi, P. Bogdan & Mihai Udrescu , "Clustering Drug-Drug Interaction Networks With Energy Model Layouts: Community Analysis and Drug Repurposing", Scientific Reports, (June) (2016) 1–10, DOI: 10.1038/srep32745 6 .

[59] S.Wang, M. Gong, Y. Wu and Xiaolei Qin, "Parallel Particle Swarm Optimization for Community Detection in Large-Scale Networks." In: Shi Y. et al. (eds) Simulated Evolution and Learning. SEAL 2017.

[60] R. S. Xin, J. E. Gonzalez, M. J. Franklin, I. Stoica , “GraphX: A Resilient Distributed Graph System on Spark”, (GRADES 2013), June 23, 2013, New York, New York, USA

[61] J. Yang , J. Leskovec , "Defining and Evaluating Network Communities based on Ground-truth" ,Proceedings of 2012 IEEE International Conference on Data Mining (ICDM), arXiv:1205.6233, ICDM, 2012.

[62] M.A. Yildirim, K. Il Goh, M.E. Cusick, A.L. Barabási, M. Vidal, "Drug-Target Network." , Nat Biotechnol, 2007 Oct;25(10):1119-26. doi: 10.1038/nbt1338.

[63] W. Zachary, “An Information Flow Model for Conflict And Fission in Small Groups”. J. of Anthropological Research, 33:452--473, 1977.

[64] https://snap.stanford.edu/data/email-Eu-core.html

[65] https://www.javatpoint.com/apache-spark-architecture

[66] https://www.mygreatlearning.com/blog/representing-graphs-in-data-structures

[67] https://www.ucsm.edu.mm/

LIST OF ACRONYMS

ABC	Artificial Bee Colony Algorithm
CC	Cluster Coefficient
CD	Community Detection
CDA	Community Detection Algorithm
CDACSM	Community Detection Algorithm via Community Similarity Measure
DE	Differential Evolution
DAG	Directed Acyclic Graph
DAG	Directed Acyclic Graph
D-ABC	Discrete Artificial Bee Colony Algorithm
E	Edge
FOMD	Fraction Over Median Degree
GA	Genetic Algorithm
GT	Ground Truth
KL	Kernighan Lin
LPA	Label Propagation Algorithm
MCN	Maximum Cycle Number
Q	Modularity
MOA	Modularity Optimization Algorithm
V	Node/ Vertice
NMI	Normalize Mutual Information
RDD	Resilient Distributed Dataset
RDDS	Resilient Distributed Datasets

SHARC	Sharper Heuristic for Assignment of Robust Communities
TPR	Triangle Participation Ratio
UCSM	University of Computer Studies, Mandalay