Multi-Objective Community Detection Applied to Social and COVID-19 Constructed Networks

A thesis submitted for the degree of Doctor of Philosophy by

Jenan Moosa Ahmed

Department of Computer Science

Abstract

Community Detection plays an integral part in network analysis, as it facilitates understanding the structures and functional characteristics of the network. Communities organize real-world networks into densely connected groups of nodes. This thesis provides a critical analysis of the Community Detection and highlights the main areas including algorithms, evaluation metrics, applications, and datasets in social networks.

After defining the research gap, this thesis proposes two Attribute-Based Label Propagation algorithms that maximizes both Modularity and homogeneity. Homogeneity is considered as an objective function one time, and as a constraint another time. To better capture the homogeneity of real-world networks, a new Penalized Homogeneity degree (PHd) is proposed, that can be easily personalized based on the network characteristics.

For the first time, COVID-19 tracing data are utilized to form two dataset networks: one is based on the virus transition between the world countries. While the second dataset is an attributed network based on the virus transition among the contact-tracing in the Kingdom of Bahrain. This type of networks that is concerned in tracking a disease was not formed based on COVID-19 virus and has never been studied as a community detection problem. The proposed datasets are validated and tested in several experiments. The proposed Penalized Homogeneity measure is personalized and used to evaluate the proposed attributed network.

Extensive experiments and analysis are carried out to evaluate the proposed methods and benchmark the results with other well-known algorithms. The results are compared in terms of Modularity, proposed PHd, and accuracy measures. The proposed methods have achieved maximum performance among other methods, with 26.6% better performance in Modularity, and 33.96% in PHd on the proposed dataset, as well as noteworthy results on benchmarking datasets with improvement in Modularity measures of 7.24%, and 4.96% respectively, and proposed PHd values 27% and 81.9%.

Table of Contents

Abbreviations

ABLP: Attribute-Based Label Propagation

ALP: Asynchronous Label Propagation

ARI: Adjusted Rand Index

BBO: Biogeography Based Optimization

COVID-19: Coronavirus

E: Edge

G: Graph

GN: Girvan Newman

IoT: Internet of Things

LFR: Lancichinetti–Fortunato–Radicchi

MOEA: Multi-Objective Evolutionary Algorithms

MOH: Ministry Of Health

NMI: Normalized Mutual Information

PCR: Polymerase Chain Reaction

Q: Modularity

RI: Rand Index

SARS-CoV-2: Severe Acute Respiratory Syndrome coronavirus 2

SNA: Social Network Analysis

V: Node

VI: Variation of Information

WHO: World Health Organization

Mathematical Abbreviations

k: number of communities in a network

att: number of attributes

n: number of nodes in a community

n_{att}: number of nodes with a certain attribute in a community

N_{att}: number of nodes with a certain attribute in a network

P: Penalty for penalizing the homogeneity measure

Hd: Homogeneity degree

PHd: Penalized Homogeneity degree

MAWPHd: Multi-Attribute Weighted Penalized Homogeneity degree

AAv: Average Attribute value

Q(C: H): Modularity Constrained with Homogeneity

List of Figures

List of Tables

1 Introduction

1.1 Overview

Living in an era of technology, a dramatic amount of data is being produced; according to the World Economic Forum, the data found in the digital realm at the beginning of 2020 was 40 times more than observable stars in the universe [1]. And with the great expansion of social media networks, rich social media content is being generated every day. As a result, even a bigger massive amount of data will be created in the coming years. Having all this amount of data will not only help in performing the task it was created for, as it can also facilitate in executing new functions and outcomes; when first created, the data is raw and senseless. However, instead of neglecting it, raw data and its' relationships can be transformed into meaningful information and structures when applying the appropriate machine learning tools. This chapter discusses the problem statement, research questions raised, the original contribution of thesis in addition to its structure.

1.2 Definitions

Machine learning include the mining of different structured data, images, text, audio, video, or graphs. Graph mining is a contemporary multidisciplinary field, which is basically a form of data mining that deals with graphs instead of normal data, [2] it intends to discover repetitive sub-graphs and interesting patterns that occurs in the input graph. As networks can be outlined in graphs, and the graph is composed of a set of nodes which can be individuals or entities, and edges that represent the connections and interactions between the nodes [3], and hence, graph mining can be utilized on this type of data. In Graph G(V,E), where V donates the set of nodes, and E is the set of edges between the nodes. So, if a link exists between the nodes, the entry of G is represented by 1, and represented as 0 on the contrary. This problem can be classified as NP-Hard (Nondeterministic Polynomial time), a various number of algorithms was proposed and applied for the identification of communities in social networks [4].

And with the increased use of networks in today's world (such as social networks, biological networks, etc.), it has become an important task to study and understand them. Graphs are used to represent networks, which is can be an ideal way to analyse a network, because it is the closest formulation to the real-world situation [5]. Constructing a network to a graph can simplify the analysis as it provides different points of view, in addition to the various tools that can be used to solve the problem. A social network can also be represented as a graph G(V,E); where V (the set of vertices) is used to represent users in the network, and E (the set of edges) visualized any sort of relationships between the users [6]. In other words, in a social network's graph, the users are denoted by nodes, and since there are relationship between the users, the nodes are connected, and the relationships are denoted by the edge and this graph representation helps analysing the social network.

As social Networks today play an important role of our daily life, Social Network Analysis (SNA) has emerged out of the need to better understand social networks, by exploring how people are connecting and determining the network's strength and weakness, which will help in improving and growing more robust networks [7]. SNA is the study of the patterns of relations that comprise social structures, treating these relations as networks of connections among the individuals and groups that enter them [8], it refers to the study of relations and connections between nodes, and it includes the analysis of social structures, positions, roles and many others.

Social Network Analysis played an important role in the field of marketing influence maximization [9], fraud detection [10] or recommender systems [11], and tourism management [12]. Researchers also linked the social network analysis with the pandemic of COVID-19; it was used to analyze the discussion on Twitter related to COVID-19 and to investigate the sentiments toward the virus [13], while another study aimed to develop an understanding of the drivers of the 5G COVID-19 conspiracy theory and strategies to deal with such misinformation [14].

Structuring a social network by a graph can also result in performing a number of tasks[5], such as centrality analysis, position/role analysis, network modeling, information diffusion, network classification and outlier detection, viral marketing and link prediction, and community detection. Community detection recognizes the communities formed by the actors in a social network, by studying the network topology [15]. Community detection is an important tool for analyzing complex networks by studying the structures and functional characteristics of the networks[3]. The detection of communities has remarkable results in various fields, e.g., social science, bibliometrics, marketing and recommendations, biology, etc.[15].

1.3 Aim and Objectives

The aim of this research is to study the community detection problem in social networks experimentally, and analytically. The findings can potentially result in more accurate and effective detection of densely connected group of nodes in social networks, in terms of Modularity, Homogeneity, and accuracy measures.

In more detail, the objectives of this thesis are as follow:

1. Review and analyze the existing community detection methods, applications, and datasets.
2. Investigate the evaluation measures used to evaluate the detected communities.
3. Examine the homogeneity measure in attributed social networks.
4. Utilize the community detection tool to elucidate the COVID-19 pandemic.

1.4 Problem statement

To formulate and analyze a network, the most accurate method is to represent it as a graph, this can simplify the problem as it provides different points of view. One important filed in network analysis is the Community Detection, it can help in understanding the structures and functional characteristics of the network. Communities represent a principled way of organizing real-world networks into densely connected groups of nodes, as a community is a cluster of nodes that are strongly connected to each other in a subnetwork than to the rest of the network. These nodes most likely share the same taste, interest, preferences, or choices. The main problem of community detection professedly includes the results of detected communities that achieve the best measures in terms of accuracy and efficiency evaluation. Nonetheless, attributed networks contain fruitful information and can be employed to detect more accurate and homogenous communities, this area needs to be investigated further in terms of algorithms design, evaluation metrics and homogeneity maximization. And although Community Detection has remarkable results in various fields, e.g., social science, bibliometrics, marketing and recommendations, biology etc. It was not utilized to study the distribution of a certain virus or disease in a social network.

1.5 Research Questions

By investigating the Community Detection problem, this thesis formulates 3 research questions:

- Does the attributes consideration enhance the community detection results in terms of Modularity and accuracy (Normalized Mutual Information, Rand Index, Adjusted Rand Index and Variation of Information) ?
- Can homogeneity measure be improved and used to detect more accurate communities in real-world networks?
- How can the contact tracing data generated from COVID-19 contact tracing applications be utilized for research purposes?

1.6 Original Scientific Contribution

This thesis centers around theory and practice of community detection, and its application in analyzing the structure in social networks.

The specific novel contributions of this thesis include:

1. Develop an improved Label Propagation algorithm (Attribute-Based Label Propagation ABLP) that considers the nodes' attributes to achieve a fair homogeneity value, while maintaining high Modularity, Accuracy measures (Normalized Mutual Information, Rand Index, Adjusted Rand Index, Variation of Information) and Privacy measure (Split-Join). [Published in The International Conference on Agents and Artificial Intelligence ICAART 2022, will be held on 3-5 February 2022]

2. Formulate a Penalized Homogeneity degree measure, which is an adaptive homogeneity measure, with penalty and weight modulation, that can be utilized in consonance with the user's requirements. [Published in The International Conference on Agents and Artificial Intelligence ICAART 2022, will be held on 3-5 February 2022]

3. Based on the literature review, a research gap of employing homogeneity as a constraint in Community Detection was identified, and accordingly, homogeneity as a constraint in Modularity based methods is investigated, and algorithm is developed. [Published in The International Conference on Agents and Artificial Intelligence ICAART 2022, will be held on 3-5 February 2022]

4. The first effort to link the COVID-19 contact tracing with the community detection problem, it is based on finding communities of virus infected nodes and study their behavior to limit the spread of the virus. This is applied through creating two COVID-19 networks, represented by graphs. The datasets are published on (IEEE DataPort DOI: https://dx.doi.org/10.21227/8zxr-vh55) and a paper was submitted on 09-12-21 to IEEE Transactions on Emerging Topics in Computational Intelligence.

5. A dataset based on COVID 19 distribution formed by tracing the transmission of the virus among the world countries, based on the first cases of an infected person who had a travelling history from Wuhan, China to their hometown. This experiment demonstrates the spread of COVID-19, by implementing several community detection algorithms and evaluating the results using the Modularity measure. [Dataset used in a conference paper published in European, Asian, Middle Eastern, North African Conference on Management and Information Systems (EAMMIS) 2021 [16]]. [Dataset published on IEEE DataPort DOI: https://dx.doi.org/10.21227/8zxr-vh55]

6. A novel dataset constructed on COVID-19 contact tracing in the Kingdom of Bahrain is created, to help identify communities of infected persons and study their attributes' values. [Submitted on 09-12-21 to IEEE Transactions on Emerging Topics in Computational Intelligence], [Dataset published on IEEE DataPort DOI: https://dx.doi.org/10.21227/8zxr-vh55]

7. Algorithms' performance comparative study on proposed COVID-19 World Countries dataset was conducted, results were evaluated in terms of Modularity and Normalized Mutual Information. [Published in European, Asian, Middle Eastern, North African Conference on Management and Information Systems (EAMMIS) 2021 [16]].

8. A systematic review to study the existing methods of community detection tools in graph mining was summarized and conducted [Published in SSRN Electronic Journal as Proceedings of the Industrial Revolution & Business Management: 11th Annual PwR Doctoral Symposium (PWRDS) 2020] [17].

Scientific contributions can be illustrated in Figure 1-1

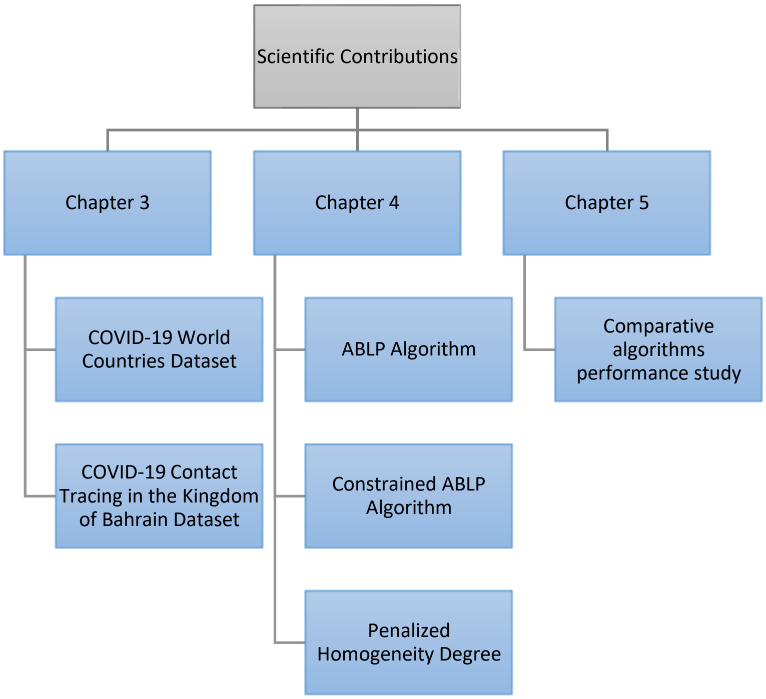

Scientific Contributions
Chapter 3
Chapter 4
Chapter 5
COVID-19 World Countries Dataset
COVID-19 Contact Tracing in the Kingdom of Bahrain Dataset
ABLP Algorithm
Constrained ABLP Algorithm
Penalized Homogeneity Degree
Comparative algorithms performance study

Figure 1-1: Original Contributions of thesis

1.7 Thesis Structure

The thesis is organized into the following chapters:

Chapter 2 provides a summary of the literature and theoretical background into social network analysis and community detection problem. It includes the community detection algorithms, constrained algorithms, evaluation metrics and objective functions, applications and datasets.

Chapter 3 proposes two novel datasets based on the contact tracing of COVID-19. The first dataset is based on the travelling history of infected persons, the tracing is formed into a network of world countries. The other dataset is an attributed network, based on COVID-19 infected cases and the contact tracing in the Kingdom of Bahrain, a network of infected persons and their relationships with other infected persons is formed.

Chapter 5 proposes two community detection algorithms for attributed networks: Attribute Based Label Propagation Algorithm for Community Detection in Social network, with homogeneity as an objective function once, and as a constraint the second. The algorithms' concept considers the attributes' values when distributing the nodes into communities. In addition to proposing a new measure for homogeneity and a Penalized Homogeneity degree, which can be personalized based on the user's requirements. Experiments of both algorithms are conducted on the COVID-19 contract tracing in the Kingdom of Bahrain dataset in addition to two recognized attributed datasets. The results are evaluated, visualized, and discussed in terms of Modularity, Normalized Mutual Information, Rand Index, Adjusted Rand Index, Variation of Information, And Split-Join.

Chapter 5 is a comparative study of algorithms' performance on the first proposed dataset (COVID-19: World Countries), and a dataset from the literature (Zachry's Karate Club) performed on a number of well-known community detection algorithms (Louvain, Greedy, Spectral Clustering, Girvan Newman, Asynchronous Label Propagation, and Kernighan-Lin). The results are evaluated by the Modularity measures.

Chapter 6 is the final chapter that summarizes the findings within this research as well as future work.

Thesis structure is presented in Figure 1-2

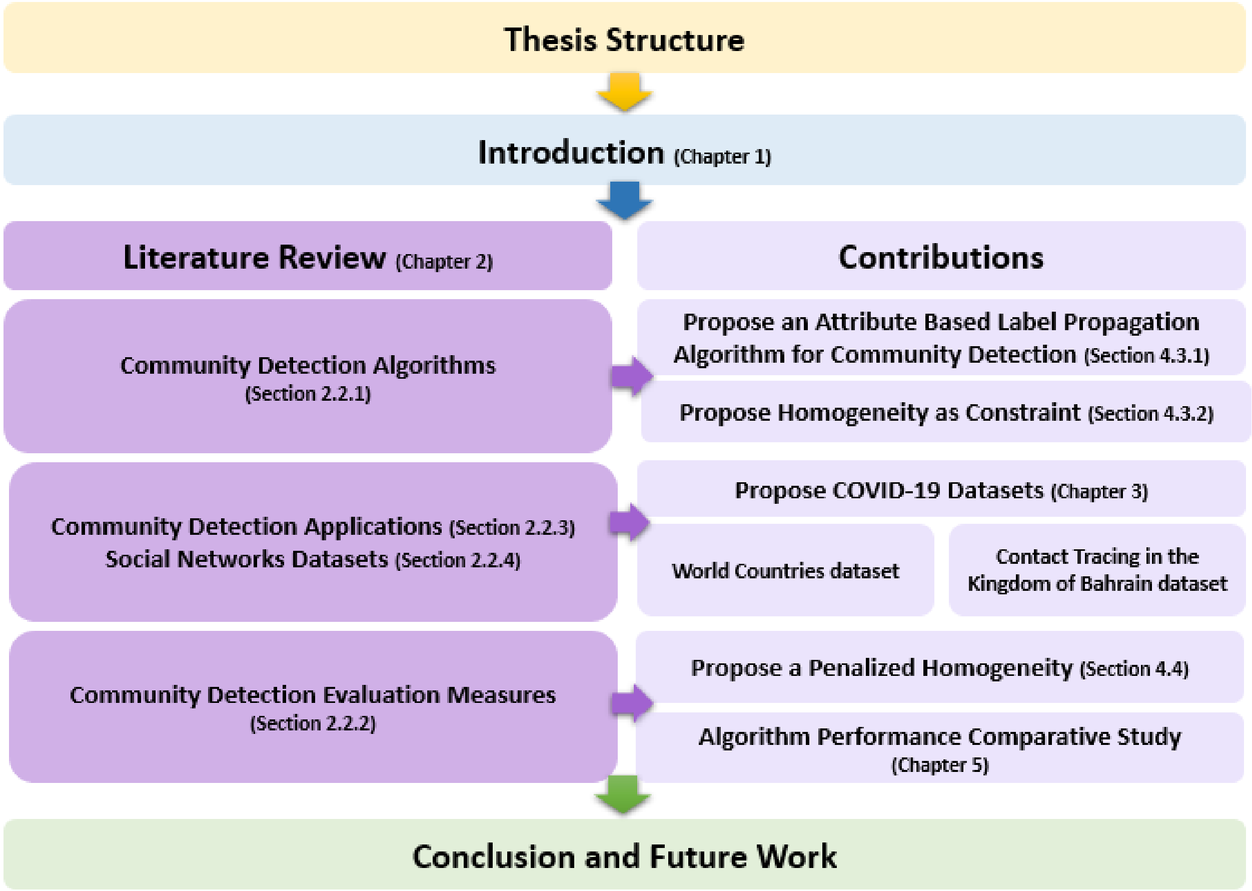

Thesis Structure
Introduction (Chapter 1)
Literature Review (Chapter 2)
Contributions
Community Detection Algorithms (Section 2.2.1)
Propose an Attribute Based Label Propagation Algorithm for Community Detection (Section 4.3.1)
Propose Homogeneity as Constraint (Section 4.3.2)
Community Detection Applications (Section 2.2.3)
Social Networks Datasets (Section 2.2.4)
Propose COVID-19 Datasets (Chapter 3)
World Countries dataset
Contact Tracing in the Kingdom of Bahrain dataset
Community Detection Evaluation Measures (Section 2.2.2)
Propose a Penalized Homogeneity (Section 4.4)
Algorithm Performance Comparative Study (Chapter 5)
Conclusion and Future Work

Figure 1-2: Thesis Structure

2 Literature Review and Theoretical Background

2.1 Overview

A network is consisted of connected communities that are created by the communicated individuals, and a community is a densely connected group of nodes that is also connected to the rest of the network [18]. Bedi and Sharma (2016) stated that any network can be outlined in a graph, [3] the graph is composed of a set of nodes which can be individuals or entities, and edges that represent the connections and interactions between the nodes.

A social network is a great example of that, structuring a social network by a graph can result in performing several tasks [5], such as centrality analysis, position/role analysis, network modelling, information diffusion, network classification and outlier detection, viral marketing and link prediction, and community detection.

In different research fields, a community is also referred to as group, cluster, cohesive subgroup, or module. Communities are subjective and can be defined based on the context. The main goal of community detection is to detect and discover the strongly connected members as compared to the rest of the members, and thereby, nodes in the same community, outline similar or comparable tastes, opinions, properties, and functions, which makes community detection a significant tool in scientific inquiries and data analytics [19]. The extracted communities can considerably analyze building ontology for semantic web, detecting topics in tagging systems, analyzing the behavior of a community, and personalized search and recommendations[2].

Community detection in social networks is being widely studied, and different algorithms and techniques have been proposed so far, which includes traditional algorithms [20] [21], evolutionary algorithms [22][23], heuristic [24], [25] hierarchical clustering [26],[27], spectral clustering [28], label propagation [29], neural networks [30], etc.

This chapter presents a review of related literature used in this study; it includes community detection algorithms, evaluation metrics, community detection applications and datasets.

2.2 Related Work

Several research studies were conducted to solve the community detection problem; each has used a different approach. Graph partitioning methods have been proposed in a number of researches[31]; in a graph, each subset of strongly connected nodes is called a partition, and the partitioning cut size should be set to the number of edges whose adjacent vertices are located in different partitions, and the edges are called cut edges [32].

The social networks studied in the community detection problem varied; some researchers focused on complex, large-scale, or dynamic networks, whereas others examined attributed networks. And while many algorithms were designed to discover disjoint communities, assuming that nodes in each network belong to exactly one community, in real world, a node (or a person) in a certain network, does not necessarily belong to exactly one community; as a person can be in more than one community at a time, and here is where overlapping occurs- which is illustrated in Figure 2-1 [33]. Originally, a community consists of densely connected nodes, so intuitively, when a node connects multiple communities with similar strength, it is more likely to be an overlapping node [34]. So, if a certain node has both links with community *x* and *y*, then this node as an overlapping node. For example, in Figure 2-1, nodes a, b, c, and d are overlapping nodes in the network, as they belong to more than one community.

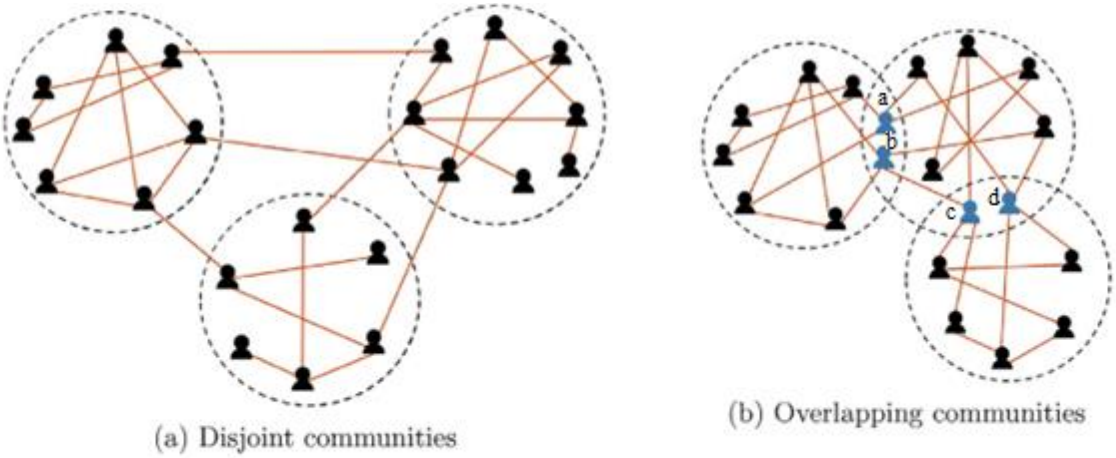


Figure 2-1: (a) Disjoint vs (b) Overlapping communities (courtesy of [33]**)**

The main difference between disjoint and overlapping community detection algorithms is that the latter can discover overlapped communities, where nodes can belong more than one community at the same time as a node can share the same taste or preferences of more than one group of people.

2.2.1 Community Detection Methods

The **Traditional Graph Partitioning** is "the problem of dividing the network into a defined number of parts of given sizes, in such that the cut size (the number of edges running between parts) is minimized" [35].

Many optimization methods to detect the community structure have been proposed. The most known method is the modularity optimization that was proposed by [36]; the network is broke up into communities by removing the edges via using a divisive technique, one of a possible "betwennness" measure is used to determine the edges to be removed. And after each removal, the measures are recalculated.

Spectral Methods were used to detect communities in dynamic networks, a model that is used to derive a maximum a posteriori estimator for community detection was proposed [37], it is based on a constrained spectral clustering problem. Particularly, the transition probabilities for each community modify the graph adjacency matrix at each time point, this formulation provides a relationship between statistical network inference and spectral clustering for dynamic networks. And to solve the overlapping community detection problem, a spectral method called Local Expansion via Minimum One Norm (LEMON) was proposed, it uses short **Random Walks** to approximate an invariant subspace near a seed set, which is referred to as local spectra. Local spectra can be viewed as the low-dimensional embedding that captures the nodes' closeness in the local network structure [38].

Random-walk method was also used in a number of researchers to reveal communities in dynamic networks [39]. Okuda et al. proposed a method based on the assumption that a random walker, upon starting the walk, usually shifts in the community with the starting vertex. restrained random-walk similarity method. Subsequently, these vertices which are passed by the random walker are alike and can be formed into a community. Another research employed the random walk technique in dynamic social networks to detect the closest nodes for each user, it could detect the overlapping communities, and an adaptive version forced the impacted nodes to detect the new closest nodes and the changed communities adaptively [40].

Genetic Algorithm (GA) is a widely used algorithm that gives high quality solutions for both constrained and unconstrained problems, and community detection is considered as an optimization problem; GA has also been used as method to detect the communities [41], [42], [43]. Genetic Algorithms were excessively used in complex networks; a method that adopts matrix encoding to enable traditional crossover between individuals was introduced [44]. The nodes similarity was used to produce initial populations, this resulted in improving the diversity while keeping a decent accuracy value. In another study in complex networks, a proposed method was used to enable a flexible and adaptive analysis of the characteristics of a network from different levels of detail according to an analyst's needs [45]. Moreover, a Clustering Coefficient-based Genetic Algorithm (CC-GA) was proposed for detecting communities in social and complex networks, it was based on the generation of the initial population and the mutation method, which were employed to enhance the efficiency and accuracy of results [46]. A local search strategy called Enhanced Multi-Objective Genetic Algorithm for Community Detection (EMOGACD) was introduced recently, it aims to detect communities in complex networks by using the vector-based method [47]. In large-scale networks, Genetic Algorithms were also used [1]. In addition to using them to detect dynamic communities [48]. And to address the overlapping problem, genetic algorithms were utilized in several research; a Multi-Objective Genetic Algorithm was designed using measures related to the network connectivity, it uses a phenotype-type encoding based on the edge information [49]. GA was also used to form weak cliques which consist of similar neighbors, and the shared cliques by nodes can then form communities [50].

In general, **Evolutionary Algorithms** (EAs) have become common as they do not demand differentiability of functions and constrains, and do not require multiple runs to produce a set of different possible solutions, they are also compatible for practical applications because they automatically set the complex networks' clusters [51]. In EA methods, the community detection task is framed as different optimization problems, which leads to deigning a proper metaheuristic to handle them. Gong et al. proposed an algorithm that maximizes the density of internal degrees, and minimizes the density of external degrees simultaneously, it generates a collection of possible solutions that represent various divisions to the networks at different hierarchical levels.

As community detection problem is complicated, it is hard to be addressed by a single-objective methods, furthermore, most optimization algorithms employ only single optimization criteria [52] For this reason, many **Multi-Objective Evolutionary Algorithms** were proposed to overcome this problem. The optimization-based methods measure the network partitions and hence, optimize the objective functions, the efficiency of optimization and community detection rely on the algorithm's searching efficiency and the employed functions.

A multi-objective algorithm that employs genetic algorithm was proposed by Pizzuti [43], optimizes two objective functions that can determine the strongly connected nodes which have inter connections. The first objective function uses the community score concept to measure the quality of division in the network's communities; the clustering is denser when the community score is high. The other objective uses the fitness concept, it defines modules that have the highest sum of node fitness (community fitness), and the number of external links is minimized when the sum attains its farthest value. A swarm based meta-heuristic optimization method was proposed by [53]; it simulates the behavior of artificial bee colony. It is consisted of three types of bees; experienced/employee forager: this bee keeps a food source in her mind and shares this information with onlookers, onlooker: selects a food source and tries to improve this source, and scout bees: flies in a dimensional search space to find the optimum solution. The ratio of scout, forager and experienced forager are usually determined manually. A recent discrete Biogeography Based Optimization (BBO) algorithm was proposed by Rehanian et al. [54], the proposed method employs the Pareto-based approach, it solves the community detection problem by maximizing the objective functions separately and its output is a set of non-dominated solutions. However, the BBO cannot find overlapping communities or detect communities in dynamic networks.

Other well-known methods were also employed to solve the problem of community detection; **Fuzzy Methods** played an important role in dividing the nodes [55], [56] [57], [58], [59]. An evolutionary multi-objective optimization based on Fuzzy method was proposed by Tian et al., it was designed to find an appropriate fuzzy threshold for each node, so that diverse overlapping community structures can be uncovered [60]. Fuzzy methods were mainly implemented in large-scale networks to detect overlapping communities, recent research proposed a fast fuzzy modularity maximization method that exploits iterative equations to calculate modularity and reduces the complexity for large networks by breaking them into multiple sub-networks and then applying the method to detect the overlapping communities [61]. And for the same purpose, Naderipour et al. designed a fuzzy model based on an algorithm in complex networks, it based on both resources of data related to the nodes' attributes and nodes' structure. [62]

On the other hand, **Neural Networks** also contributed to the community detection problem, a method based on deep learning of ground-truth communities, with the aim of revealing community structure in large real-world networks was recently proposed [63]. It is based on an edge-to-image model that transfers the edge structure to an image structure, and the edges are classified into two categories which the same community edges and others between different communities. This makes it easier to obtain local views of network communities by breadth-first search based on edge classification, it also applies the merging preliminary communities with local modularity, making it easy to optimize the community structure and obtain the final community structure of given networks. Another study aimed to utilize deep learning techniques on heterogeneous network community detection [64], however, it was not able to perform in large-scale networks.

To brief the literature review, a number of related works was filled in (Table 2-1 to Table 2-5), to summarize some of the community detection methods used based on the problem statement in terms of algorithms, type of network, and aim of research.

Table 2-1 represents the research studies that used genetic algorithms and swarm intelligence to solve the community detection problem. It is evident that genetic algorithms as part of evolutionary algorithms were extensively used to solve the community detection problem over the past years.

It is noted that there isn't a method that can deal with all types of networks, as different networks have different space-time properties. Therefore, the special characteristics of the network need to be considered before designing the method. This will result in more accurate results. The main problem is the detection of accurate communities in a network. Different genetic algorithm approaches were used to address this problem, some has focused on the type of targeted network (such as complex, and weighted networks) while others were focused on overlapping problem, and probability distribution. To assess the overlapping, fuzzy techniques were used as the node can belong to more than one community at the same time, this can be summarized in Table 2-2. And to better capture the networks' structure, node importance was considered in several research studies Table 2-3.

Table 2-1: Genetic Algorithms and Swarm Intelligence in Community Detection

Source/ year	Algorithm	Outcome
[65] 2020	Markov chain-based algorithm based on a Genetic algorithm	Enhance the transition probability in the dynamic process of Markov chain-based algorithm. It uses a hybrid algorithm that can adaptively search for a better combination of parameters
[66] 2019	Combining the heuristic operator of ant colony optimization and the multi-objective evolutionary algorithm based on decomposition	Solve complex network community detection problems
[46] 2018	Clustering Coefficient-based Genetic Algorithm for detecting community structures in a network by optimizing the modularity	Detect communities in social and complex network by generating the initial population and the mutation method
[67] 2018	Multi-objective optimization method that uses all Pareto fronts to detect overlapping communities	Use all Pareto fronts to detect overlapping communities
[68] 2018	Multi-objective community detection method based on a modified version of particle swarm	Solve the graph clustering problem by detecting the structures which are closer to real ones
[69] 2017	Genetic algorithm k-means clustering algorithm	Identify community structure in a multi-relational network through relational learning
[45] 2017	Generational genetic algorithm that includes efficient initialization methods and search operators	Enable a flexible and adaptive analysis of the characteristics of a network from different levels of detail according to an analyst's needs
[70] 2017	A modified Genetic Algorithm that with alleles encoding and half uniform crossover	Detect communities in social networks
[44] 2016	K-path initialization method which makes full use of the topological information	Find whether a K-path initialized generic algorithm can bring significant increase in Modularity value
[71] 2016	Multi-objective optimization based on genetic algorithm for weighted networks	Detect communities in weighted networks
[72] 2015	An improved multi-objective quantum-behaved particle swarm optimization based on spectral-clustering is proposed to detect the overlapping community structure in complex networks.	Solve the multi-objective optimization problem to resolve the separated community structure in the line graph which corresponding to the overlapping community structure in the graph presenting the network
[73] 2014	Artificial Fish Swarm optimization has been used as an effective optimization technique	Detect communities in terms of accuracy and successfully finds an optimized community structure based on the quality function used
[53] 2014	Artificial bee colony optimization has been used as an effective optimization technique to solve the community detection problem	Solve the community detection problem in terms of accuracy and find an optimized community structure
[74] 2013	Genetic algorithm for community detection in complex networks	Detect communities by using matrix encoding that enables traditional crossover between individuals
[75] 2013	Extended compact genetic algorithm in complex networks	Use statistical learning mechanism to build a probability distribution model of all individuals in a population

Table 2-2 groups the fuzzy methods used to solve the community detection problems. It is intelligible that fuzzy logic was used to overcome the overlapping problem, considering that fuzzy logic is originally based on “degree of truth” rather than the usual “true or false” [76].

Table 2-2: Fuzzy Algorithms and Problem Statement in Community Detection

Source/ year	Algorithm	Outcome
[59] 2019	The Fuzzy C-Mean -based algorithm	Improve the performance of both the Fuzzy C-Mean -based and the K-Mean based algorithms using Graphics Processing Units
[60] 2019	An evolutionary multi-objective optimization fuzzy for overlapping community detection	Overlapping community detection
[77] 2018	Multi-mode multi-attribute fuzzy subtractive clustering algorithm	Detect overlapping communities in location-based social networks with respect to user check-ins and the attributes of venues and users
[58] 2018	A fuzzy agglomerative community detection algorithm	Community detection that iteratively up- dates membership degree of nodes
[78] 2018	Fuzzy clustering algorithm based on the nonnegative matrix factorization method	Overlapping fuzzy Community detection in large scale social networks

Nevertheless, several research studies were conducted based on the node importance, as this technique is employed to enhance the node order of label updating and the structure of label selecting when multiple labels are contained by the maximum number of nodes [79]. So Table 2-3 summarizes some algorithms that were developed based on node-importance, and their problem statements. It is observed that considering the nodes’ importance in the network facilitates the distribution of nodes in the communities, as the communities might change over time according to the nodes’ structures.

Table 2-3: Node-Importance Algorithms and Problem Statement in Community Detection

Source	Algorithm	Outcome
[79] 2021	An Improved Label Propagation Algorithm Based on Modularity and Node Importance	Combine the modularity function and node importance with LPA by using node importance to improve the node order of label updating
[80] 2020	Global and local node influence-based algorithm	Identify the most influential nodes which are considered as cores of communities
[81] 2018	An improved overlapping community detection algorithm, Label Propagation Algorithm with Neighbor Node Influence	Detect overlapping community structures by adopting fixed label propagation sequence based on the ascending order of node importance and label update strategy
[82] 2018	Multi-objective Attributed community detection algorithm with Node Importance Analysis	Detect communities, incorporate nodes’ attribute information and estimating nodes’ importance
[83] 2018	A local approach based on the detection and expansion of core nodes	Detect all the graph’s communities in a network using local information as well as identifying various roles of nodes (core or outlier)

On the other hand, Table 2-4 categorizes the community detection algorithms based on the type of networks. As different research studies were conducted to target a various range of networks, e.g., complex, dynamic, large scale, signed, weighted, and unweighted networks. The type of targeted network in the community detection affects the method used, for example, in large-scale networks, an algorithm used a reduction-based approach in which the size of networks is recursively reduced as the evolution proceeds [84]. Whereas in complex networks a neighbour-based mutation was proposed to preserve the variations and avoid the restriction in the local optima [85].

Table 2-4: Community Detection Algorithms based on type of networks

Source/ year	Algorithm	Outcome
[86] 2021	Overlapping community detection by constrained personalized PageRank	Reduce the problem of redundant diffusion by using a constrained personalized PageRank method for community expansion
[87] 2021	A Parallel multi-objective evolutionary algorithm for community detection in large-scale complex networks	Detect communities in large-scale networks, where the communities associated with key network nodes are detected in parallel
[88] 2021	A spectral clustering method based on the singular decomposition of the adjacency matrix	Detect community in directed stochastic block model
[23] 2020	A semi-supervised algorithm (sE-Autoencoder)	Extend the typical nonlinear reconstruction model to the dynamic network by constructing a temporal matrix to overcome the effects of nonlinear property on the low-dimensional representation
[89] 2018	Evolutionary algorithm	Solves the community detection problem in imbalanced signed networks
[84] 2018	Reduction-Based Multi-objective Evolutionary Algorithm	Detect communities in large-scale complex networks
[90] 2018	Quantum inspired evolutionary algorithm	Discover communities in complex social networks by optimizing modularity
[91] 2017	Evolutionary algorithm	Find the community structure that maximizes the modularity in complex networks
[92] 2017	Parallel quantum-inspired evolutionary algorithms	Detect high-quality communities for varied datasets and work well for both weighted and un-weighted networks
[85] 2016	Multi-objective optimization using discrete teaching–learning-based optimization with decomposition	Solve community detection problems for complex networks

Diverse types of networks have been an area of interest for many researchers, and different algorithms were used to target each network. In addition, attributed networks were also an intent for some research studies as the attributed graphs demonstrate the peculiarities about nodes and relationships among them, which can lead to more precise community detection results [93].

Some of the research studies that were involved in detecting communities in attributed networks are presented in Table 2-5. Many community detection methods have been recently extended to handle attributed networks, as node's attributes are considered as additional topological information [94].

Table 2-5: Community Detection Algorithms in attributed networks

Source/ year	Algorithm	Outcome
[95] 2021	An overlapping community detection algorithm based on an augmented attribute graph	Improve weight adjustment strategy for attributes to help detect overlapping communities more accurately. And enhance the algorithm to automatically determine the number of communities by a node-density-based fuzzy k-medoids process
[96] 2020	A method that incorporates both the topology of interactions and node attributes	Help domain experts to investigate attributes and to better interpret the resulting communities, while boosting performance in terms of edge prediction
[93] 2019	Online community detection using a technique of keyword search over the attributed graph	Advance the community detection problem by using keyword search method, which allows personalized and generalized communities
[97] 2018	Algorithm based on a newly designed higher-order feature termed Attribute Homogenous Motif	Integrate both node attributes and higher-order structure of the network in a seamless way
[98] 2017	Louvain-AND- Attribute (LAA) and Louvain-OR-Attribute (LOA) methods	Analyze the effect of using node attributes with modularity

The attributed networks need to be studied further as the nodes in real networks are connected by their correlating attributes, and not limited to the network's structure. The importance of nodes' attributes needs to be addressed and utilized to solve the community detection problem.

These categorizations show that researchers have classified community detection algorithms in several ways depending on the aim of their research, however, the main aim is to detect communities or structures in social networks. While some focused on large scale, complexed or weighted networks, other researchers focused on detecting overlapped communities without taking into consideration the type or size of the network. It can also be observed that most of them used the evolutionary algorithms such as genetic algorithms, particle swarm and neural networks. Other approaches were also used such as quantum algorithms and neural networks.

It is noticeable that various research studies were conducted to solve the community detection problem. Different types of algorithms, and hybrid techniques were employed to address this problem. Extensive studies were based on multi-objective evolutionary algorithm, as to solve the community detection problem, many aspects need to be considered. Table 2-6 summarizes the community detections algorithm used over the last decade.

Table 2-6: Community Detection algorithms

Source/ year	Evolutionary Algorithm	Swarm Intelligence	Fuzzy	Node-Importance	Spectral Clustering	Mutli-Objective
[95] 2021			✓			
[79] 2021				✓		
[86] 2021				✓		
[87] 2021	✓					✓
[88] 2021					✓	
[23] 2020				✓		
[80] 2020				✓		
[65] 2020	✓					
[59] 2019			✓			
[66] 2019		✓				✓
[60] 2019			✓			✓
[90] 2018	✓					
[77] 2018			✓			
[84] 2018	✓				✓	
[89] 2018	✓					
[81] 2018				✓		
[82] 2018				✓		✓
[83] 2018				✓		
[58] 2018			✓			
[78] 2018			✓			
[46] 2018	✓					
[67] 2018	✓					✓
[68] 2018		✓				✓
[69] 2017	✓					
[92] 2017	✓					
[45] 2017	✓					
[70] 2017	✓					
[91] 2017	✓					
[44] 2016	✓					
[71] 2016		✓				✓
[85] 2016		✓				
[72] 2015		✓				✓
[73] 2014		✓				
[53] 2014		✓				
[74] 2013	✓					
[75] 2013	✓					✓

It can be observed that evolutionary algorithms were excessively used to solve the community detection problem, in addition to swarm intelligence methods. On the other hand, fuzzy methods and node-importance started to take over since 2018. Whereas spectral clustering techniques have not received as much attention.

Also, a considerable number of algorithms were multi-objective, so a table was used to illustrate the objectives considered. Where NRA is Negative Ratio Association, and RC is Ratio Cut, these two objectives have the potential to balance each other's tendency to increase or decrease the number of communities, and that both two objectives are related to the density of subgraphs to overcome the resolution limit. These features make the two objectives be suitable for revealing community structure in networks. And Community Fitness is concerned about minimizing links between communities, whereas Community Score is about maximizing internal links within communities. In addition, KKM stands for Kernel k-means, is the intra-link density in all communities, and conductance is the ratio between the number of edges inside the cluster and the number of edges leaving the cluster. And finally, Num is the number of key nodes in a community, and these keys are identified according to the local topology structure.

Table 2-7: Objectives considered in Community Detection Multi-Objective algorithms

Source/ year	Overlapping	Modularity	Homogeneity	NRA	RC	Community Fitness	Community Score	KKM	Conductance	Num
[87] 2021									✓	✓
[66] 2019				✓	✓					
[60] 2019	✓	✓			✓			✓		
[67] 2018						✓	✓			
[68] 2018					✓			✓		
[84] 2018					✓			✓		
[82] 2018		✓	✓							
[85] 2016				✓	✓					
[71] 2016	✓	✓								
[72] 2015		✓				✓	✓			
[75] 2013							✓		✓	

While the objectives considered in each research study is different, Ratio Cut and Modularity were the most objectives used, however, the algorithms that considered Ratio Cut did not consider Modularity and vice versa.

On the other hand, the number of key nodes, and Homogeneity were the least objectives assessed in multi-objective algorithms. Followed by Negative Ratio Association and Community Fitness, conductance.

The community detection algorithms can also be distinguished based on the type of network used in the study. The types of networks include dynamic, which are networks that vary over time; their vertices are often not binary and instead represent a probability for having a link between two nodes. In addition, weighted networks where the ties among nodes have weights assigned to them, directed networks where the edges in the network are directed, or pointing in only one direction, and attributed networks where the nodes have additional information as attributes assigned to the nodes.

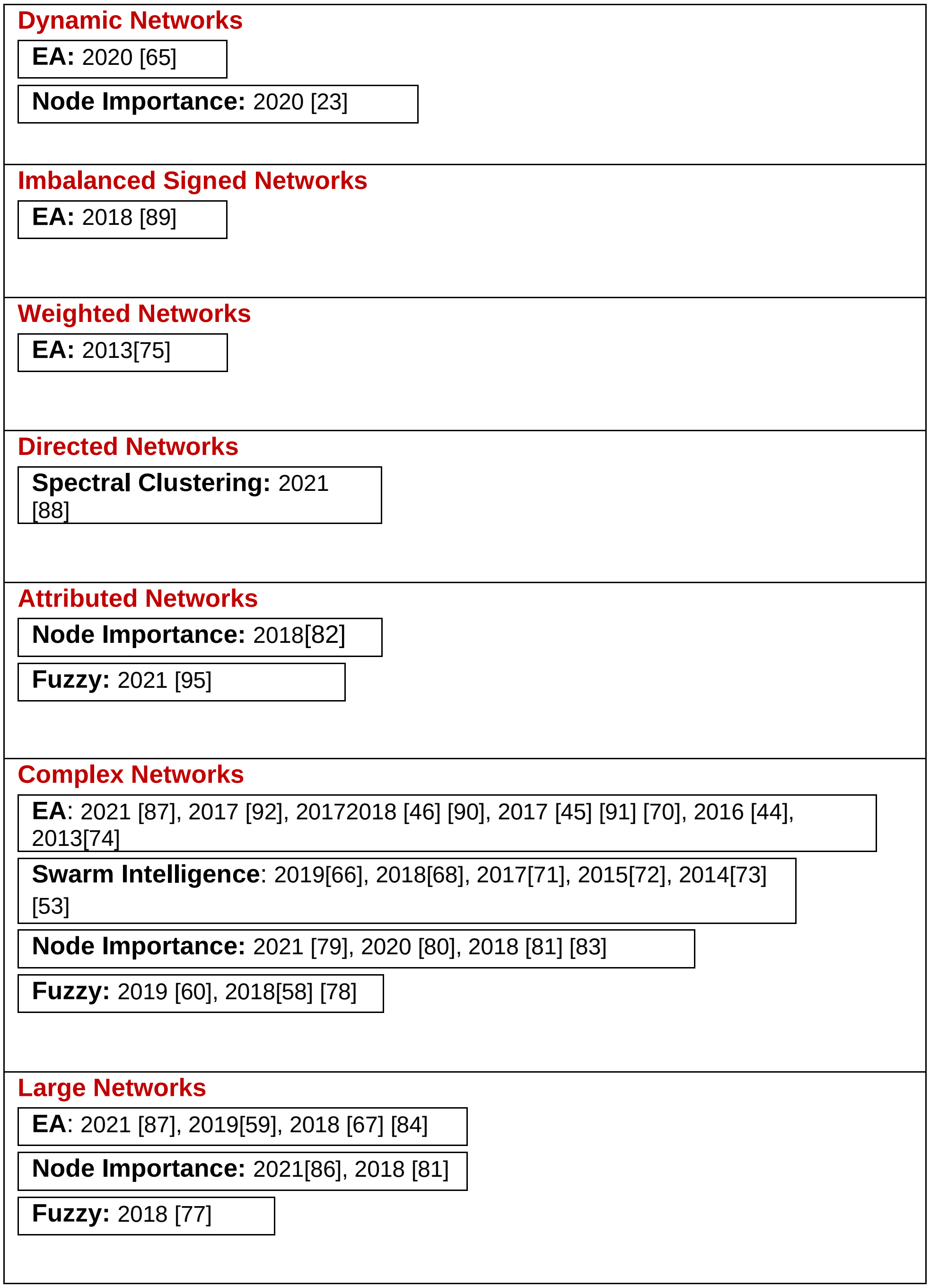
Additionally, complex networks are networks with non-trivial topological features—features that do not occur in simple networks such as lattices or random graphs but often occur in networks representing real systems. In multilayer networks, nodes are organized into layers, and edges can connect nodes in the same layer (intralayer edges) or nodes in different layers (interlayer edges). Signed networks are partitions on nodes such that the intra- community edges are positive and the inter-community edges are negative, and the network is imbalanced when there is no partition such that all the intra-community edges are positive, and all the inter-community edges are negative.

The types of networks used in the community detection problem can be summarized in Table 2-8.Most of the research studies focused on complex and large-scale networks, whereas other types of networks were not employed likewise. Even though attributed, weighted, and directed networks include additional information about the nodes or edges that may assist in the detection of communities in the network.

Table 2-8: Types of networks used in the community detection problem

Author/year	Dynamic	Weighted	Attributed	Complex	Directed	Large	Multi-layer	Imbalanced Signed
[95] 2021			✓					
[79] 2021				✓				
[86] 2021						✓		
[87] 2021				✓		✓		
[88] 2021					✓			
[23] 2020	✓							
[80] 2020				✓				
[96] 2020			✓				✓	
[65] 2020	✓							
[59] 2019						✓		
[66] 2019				✓				
[60] 2019				✓				
[90] 2018				✓				
[77] 2018						✓		
[84] 2018						✓		
[89] 2018								✓
[81] 2018				✓		✓		
[82] 2018			✓					
[83] 2018				✓				
[58] 2018				✓				
[78] 2018				✓				
[46] 2018				✓				
[67] 2018						✓		
[68] 2018				✓				
[69] 2017							✓	
[92] 2017				✓				
[45] 2017				✓				
[70] 2017				✓				
[91] 2017				✓				
[44] 2016				✓				
[71] 2016				✓				
[85] 2016				✓				
[72] 2015				✓				
[73] 2014				✓				
[53] 2014				✓				
[74] 2013				✓				
[75] 2013		✓						

To sum up, the community detection algorithms were linked with the type of networks, as shown in Figure 2-2. Several algorithms were designed to solve the community detection problem in complex and large-scale networks.



Dynamic Networks

EA: 2020 [65]

Node Importance: 2020 [23]

Imbalanced Signed Networks

EA: 2018 [89]

Weighted Networks

EA: 2013[75]

Directed Networks

Spectral Clustering: 2021 [88]

Attributed Networks

Node Importance: 2018[82]

Fuzzy: 2021 [95]

Complex Networks

EA: 2021 [87], 2017 [92], 20172018 [46] [90], 2017 [45] [91] [70], 2016 [44], 2013[74]

Swarm Intelligence: 2019[66], 2018[68], 2017[71], 2015[72], 2014[73] [53]

Node Importance: 2021 [79], 2020 [80], 2018 [81] [83]

Fuzzy: 2019 [60], 2018[58] [78]

Large Networks

EA: 2021 [87], 2019[59], 2018 [67] [84]

Node Importance: 2021[86], 2018 [81]

Fuzzy: 2018 [77]

Figure 2-2: Community detection Algorithms and Networks

The community detection problem has a wide range of techniques. The algorithms discussed did not have restrictions or constraint to employ these various mechanisms. As a result, some constrained algorithms were proposed to address the community detection.

2.2.1.1 Community detection Algorithms with constraints

Constrained community detection approaches are used to take advantage of the existing side information of the network [99]. This aids in generating more efficient and actionable results, and help develop data mining techniques that can handle complex and domain-specified constraints [100]. Ganji, et. al, [100] claims that there are two main motivations for constrained or semi-supervised community detection, which can be summarized in quality solutions and complex problem solving. As for the quality, it means that the available information can be utilized to enhance the quality of results. For instance, the supervision effect has been studied in the presence of noisy links in the network and it has been shown that semi-supervised community detection approaches are usually more robust to noise than topology-based approaches [101]. Table 2-9 presents several constrained community detection methods, along with the evaluation measure used to evaluate the results.

Most of the current community detection methods consider the structural information of networks, but disregard the fruitful information of the nodes, and this results in the failure of detecting semantically meaningful communities[97].

In addition, homogeneity was never studied as a constraint, and was always treated as an objective function, and based on this research gap, homogeneity is studied further. Accordingly, algorithms that maximize Modularity and homogeneity degree are proposed in Chapter 4, in addition, a penalized homogeneity degree is proposed and tested as an objective function, and as a constraint.

Table 2-9: Community detection with constraints where NMI is Normalized Mutual Information

Source	Method	Constraints	Objective function/ Evaluation
[99] 2018	Lagrangian Constrained Community Detection	- Must-Link - Cannot- Link	- NMI - Sensitivity to noise
[100] 2017	semi-supervised community detection based on constraint programming modelling technology	- Global constraints - Community and Instance level	- NMI - Modularity - Run Time
[102] 2017	A semi-synchronous label propagation algorithm with constraints	- Conditions of propagating labels - Exemption of the communities	- NMI - Modularity
[103] 2016	Constrained Label Propagation	The number of links of a node to the nodes in a community	- NMI - NVI (Normalized Variation of Information) - Modularity - Modularity density
[104] 2015	Adding Cohesion Constraints to Models	Cohesion constraints	Modularity
[105] 2013	Constrained fractional set programs	- Volume constraint - Seed constraint	- Normalized cut values - Run time
[101] 2012	a semi-supervised spin-glass model for incorporating pair-wise constraints	- Must-Link ML - Cannot- Link CL	- F-Score - Modularity - Noise Rate
[106] 2010	Size Constrained Greedy Community Detection algorithm	Size of communities	- NMI - corrected NMI - Modularity
[107] 2009	Constrained Label Propagation	Local maxima	- NMI - Modularity

In addition, Table 2-10 sums up the number of iterations executed and whether the number of communities to be generated in pre-defined or not.

Table 2-10: Number of iterations and predefined number of communities in solving Community Detection problem

Source	Year	Number of Iterations/ generations	Predefined number of communities
[79]	2021	100	No
[87]	2021	100	No
[66]	2019	10	No
[60]	2019	100	No
[77]	2018	100	Yes
[58]	2018	5	No
[78]	2018	50	Yes
[68]	2018	100	No
[82]	2018	500	No
[89]	2018	20	Yes
[90]	2018	30	No
[108]	2018	100	No
[84]	2018	20	No
[69]	2017	10	Yes
[2]	2017	40	Yes
[3]	2017	50	No
[54]	2017	200 to 1000	No
[4]	2017	100	No
[45]	2017	200	No
[85]	2016	50	No
[44]	2016	100	No
[109]	2016	200	No

The approximate average number of iterations used for the community detection is 100, and while some research studies defined the number of communities to be generated in advanced, most of them did not, as in real-world situations the number of communities is usually unknown.

2.2.2 Evaluation Metrics

After designing and implementing the method, it needs to be evaluated in terms of efficiency and accuracy. There are different aspects the researchers considered for the method's evaluation. The algorithm can be tested on real-world or synthetic datasets, the evaluation measures include Modularity, NMI, run-time, etc. And as discussed in section 2.2.1, some research studies were interested in finding the overlapping of communities (as a node can belong to more than a community at a time), and thus, specific evaluation measures for overlapped communities have been proposed.

Modularity

Modularity (normally denoted as *Q*), was originally proposed be Newman and Girvan, the Modularity of networks or graph's structure measures the strength of division of a network into communities, it evaluates the goodness of partitions of a graph [36]. Basically, "Modularity is the fraction of edges minus the expected value of their fraction, and the higher modularity measure indicates that the connections are denser. The modularity of a single community is zero, whereby, when the value is 1 or close to it, it indicates that there is a strong community structure" [20]. However, modularity should not be used to compare graphs that have a very different size, as the value increases when the size of graph does [20].

It was noted that in practical situations, an algorithm will be used to find communities within a network for which communities are not known ahead of time, and this calls into doubts: "how do we know if the detected communities are good ones?" And "how good the structure found is?" Moreover, the output of the algorithm represents a hierarchy of possible community divisions, so there is a need to determine the best divisions of the network. Those questions led Newman and Girvan [36] to designate Modularity which is a measure of the quality of a particular division of a network [43].

Considering a particular division of a network into k communities, define k×k symmetric matrix e whose element e_{ij} is the fraction of all edges in the network that link vertices in community *i* to vertices in community *j*. "it is crucial to make sure each edge is counted only once in the matrix e_{ij}—the same edge should not appear both above and below the diagonal. Alternatively, an edge linking communities *i* and *j* can be split, half-and-half, between the *ij* and *ji* elements, which has the advantage of making the matrix symmetric. Either way, there are several factors of 2 in the calculation that must be watched carefully, lest they escape one's attention and make mischief" [36].

The trace of this matrix is $\mathrm{Tr}\,\mathbf{e} = \sum_i e_{ii}$, which gives the fraction of edges in the network that connect vertices in the same community, and clearly a good division into communities should have a high value of this trace. The trace on its own, however, is not a good indicator of the quality of the division since, for example, placing all vertices in a single community would give the maximal value of Tre = 1 while giving no information about community structure at all.

So, a row (or column) is defined $a_i = \sum_j e_{ij}$ which represent the fraction of edges that connect to vertices in community *i*. In a network in which edges fall between vertices without regard for the communities they belong to, there would be $e_{ij} = a_i a_j$.

Therefore, Modularity is measured (Equation 1), which was proposed by [36]

$$Q = \sum_i \left(e_{ii} - a_i^2\right) = \mathrm{Tr}\,\mathbf{e} - \|\,\mathbf{e}^2\,\| \qquad \textbf{(1)}$$

where ||x|| indicates the sum of the elements of the matrix x.

The quantity in (Equation 1) measures the fraction of the edges in the network that connect vertices of the same type (i.e., within-community edges) minus the expected value of the same quantity in a network with the same community divisions but random connections between the vertices [36]. If the number of within-community edges is no better than random, then the Modularity is Zero. However, values approaching the value One, which is the maximum, denotes strong structure of the community. Practically, values of Modularity usually fall in the range from about 0.3 to 0.7.

The expected error on Q can be measured by considering each edge as an independent measurement of the contributions to the elements of the matrix e [36]. In general, Q is calculated for each split of a network into communities as the dendrogram goes down, and look for local peaks in its value, which denotes mostly acceptable splits. Usually, it is found that there are only one or two such peaks, and the height of a peak is a measure of the strength of the community division [36].

It is observed from Table 2-11 that almost all research papers have measured Modularity value to evaluate their algorithm, and then compared the results with other well-known algorithms stating that they tend to achieve higher Modularity value.

In a later research study, Newman stated that a good division of a network into communities is not merely one in which there are few edges between communities; it is one in which there are fewer than expected edges between communities [110]. If the number of edges between two groups is only what one would expect based on random chance, then few thoughtful observers would claim this constitutes evidence of meaningful community structure. On the other hand, if the number of edges between groups is significantly less than expected by chance, or equivalent if the number within groups is significantly more, then it is reasonable to conclude that something interesting is going on.

In general, the concept of Modularity comes from the assumption that true community structure in a network corresponds to a statistically surprising arrangement of edges, can be quantified by using the measure known as modularity [110]. And Modularity can be either positive or negative, with positive values indicating the possible presence of community structure. Thus, one can search for community structure precisely by looking for the divisions of a network that have positive, and preferably large, values of the modularity [110].

Normalized Mutual Information

To compare a modular structure with the proposed algorithm, Normalized Mutual Information (NMI) is used [111]. It measures the similarity of partitions, the value of NMI yield between 0 and 1 where the closer this value to 1 indicates better performance. The Normalized Mutual Information (NMI) is used to evaluate an algorithm or to compare it with other methods, it is based on defining a confusion matrix, where the rows correspond to the “real” communities, and the columns correspond to the “found” communities [111]. NMI can be calculated in (Equation 2) [111]

$$NMI\ (A,B) = \frac{-2\sum_{i=1}^{cA}\sum_{j=1}^{cB} N_{ij}\log\left(\frac{N_{ij}N}{N_{i.}N_{.j}}\right)}{\sum_{i=1}^{cA} N_{i.}\log\left(\frac{N_{i.}}{N}\right)+\sum_{j=1}^{cB} N_{.j}\log\left(\frac{N_{.j}}{N}\right)} \qquad \textbf{(2)}$$

where the number of real communities is denoted c_A and the number of found communities is denoted c_B, the sum over row *i* of matrix N_{ij} is denoted N_i. and the sum over column *j* is denoted N_j.

If the found partitions are identical to the real communities, then *NMI (A,B)* takes its maximum value of 1. If the partition found by the algorithm is totally independent of the real partition, for example when the entire network is found to be one community, *NMI (A,B)* = 0.

The mutual information matches the quantity of common information between the variables. It is symmetric, and the range of NMI is [0,1], with 0 as the lowest bound and 1 is the highest. Furthermore, Danon et al. proclaim that “NMI is more representative of sensitivity if the performance is dubious, since it measures the amount of information correctly extracted by the algorithm explicitly” [111].

It can be perceived from Table 2-11, that the majority of research papers have used NMI to measure and compare their methods.

Overlapping

In real world, a node (or a person) in a certain network, does not necessarily belong to exactly one community; as a person can be in more than one community at a time, and here is where overlapping occurs [52]. For this purpose, an objective function was proposed to guide the search process to find right individuals that can be decoded to both separated and overlapping communities [34].

Formally, given a network ***G=(V, E)***, where ***V***={$v_1, v_2, \ldots, v_n$} is the set of nodes and *E*={(*i*, *j*) | v_i, v_j ∈ ***V*** and i≠j} is the set of edges. Let ***C***={$C_1, C_2, \ldots, C_m$} be the set of all communities in **G**, then **C** satisfy the following conditions [Equation 3]:

$$\begin{gathered} \boldsymbol{C}_i \subset \boldsymbol{V},\ i=1,\ 2,\ \ldots,\ m \\ \boldsymbol{C}_i \neq \varnothing,\ i=1,\ 2,\ \ldots,\ m \\ \boldsymbol{C}_i \neq \boldsymbol{C}_j, \forall i, j \in \{1,2,\ldots,m\} \text{ and } i \neq j \\ \bigcup_{i=1}^{m} \boldsymbol{C}_i = \boldsymbol{V} \end{gathered} \tag{3}$$

It should be noted is that each community is a true subset of ***V***, which means it is meaningless to find a community including all nodes. The joint set of all communities is equivalent to ***V***. Clearly, if every pair of communities satisfy $\boldsymbol{C}_i \cap \boldsymbol{C}_j = \varnothing,\ \forall i, j \in \{1, 2, ..., m\}$ and $i \neq j$ then ***C*** is a set of separated communities.

An indirect encoding method based on permutation was designed. In this method, each individual denotes a type of division of communities, and consists of two components, namely a permutation component and a community component. An individual, , consists of two components. The first component is a permutation of all nodes in ***V***, which is labeled as $\mathcal{A}$<p>, that is calculated in [Equation 4]

$$\mathcal{A}\langle \boldsymbol{P} \rangle = \{v_{\pi_1}, v_{\pi_2}, ..., v_{\pi_n}\} \quad \textbf{(4)}$$

Where (π_1, π_2, …, π_n) is a permutation of (1, 2, .. , *n*). The second component is the set of communities derived from $\mathcal{A}$<p>, which is labeled as $\mathcal{A}$<C>.

Appropriately defined evolutionary operators can be performed on $\mathcal{A}$<p>. However, $\mathcal{A}$<C> is the result, and objective functions can evaluate $\mathcal{A}$<p>directly. Therefore, there is a need to design a decoding method to transform $\mathcal{A}$<p> to $\mathcal{A}$<C>.

It was stated that three core objective functions are designed to control the quality of communities, the separated communities, and the overlapping communities, respectively. These are labeled as f_{quality} (.) , $f_{\text{separated}}$ (.) , and $f_{\text{overlapping}}$ (.) . They have different effects on the evolutionary process, to find both separated and overlapping communities at the same time. The details of these three functions are given in (Equation 5) [34]

$$\begin{cases} \text{Maximize} \quad f_{\text{quality}}(\mathcal{A}) \\ \text{Maximize} \quad f_{\text{separated}}(\mathcal{A}) \\ \text{Maximize} \quad f_{\text{overlapping}}(\mathcal{A}) \end{cases} \quad \textbf{(5)}$$

For community detection problems, the most crucial objective is to detect high quality communities. Thus, the first objective is based on the community fitness. So, after transforming $\mathcal{A}$<p> to $\mathcal{A}$<C>, $\mathcal{A}$<C>={C_1, C_2, …, C_m}, the function to evaluate the quality of the set of communities obtained is defined in (Equation 6) [34].

$$f_{\text{quality}}(\mathcal{A}) = \frac{\sum_{i=1}^{m} f(\boldsymbol{C}_i)}{m}$$ **(6)**

And $f(\boldsymbol{C}_i) = \frac{k_{in}^{\boldsymbol{C}_i}}{\left(k_{in}^{\boldsymbol{C}_i} + k_{out}^{\boldsymbol{C}_i}\right)^{\alpha}}$, where $k_{in}^{C_i}$ and $k_{out}^{C_i}$ denote the total internal and external degrees of the nodes of Ci, and α is a positive real valued parameter, controlling the size of the communities.

Clearly, the higher the value of $f_{quality}$ is, the better $\mathcal{A}$<C> is, where the highest value of $f_{quality}$ is One. However, when detecting separated communities, the less the number of nodes belonging to more than one community results in more separated the communities. Thus, the objective function to evaluate the extent that each community separates to each other is defined in (Equation 7) [34].

$$\boldsymbol{V}_{\text{overlapping}} = \left\{ \begin{array}{l} v \mid v \in \boldsymbol{V} \text{ and} \\ \left| \left\{ \boldsymbol{C}' \mid \boldsymbol{C}' \in \mathcal{A}\langle \boldsymbol{C} \rangle \text{ and } v \in \mathcal{A}\langle \boldsymbol{C} \rangle \right\} \right| > 1 \end{array} \right\}$$ **(7)**

$$f_{\text{separated}}(\mathcal{A}) = -\left|\boldsymbol{V}_{\text{overlapping}}\right|$$ **(8)**

$f_{separated}$ in [Equation 8], is straightforward, as it counts the number of nodes belonging to more than one community and negate the value. Obviously, the higher the value of $f_{separated}$ indicates more separated communities. When detecting overlapping communities within a network, the nodes that belong to multiple communities with similar strength are eligible to join multiple communities. And this only occurs when the number of edges connecting a node to a certain community is nearly equal to the number of edges connecting this node to another community. As a result, it is unattainable to just maximize the number of nodes that belong to more than one community, but it is required to evaluate the strength of these connections and the number of the nodes. Consequently, the objective function to evaluate the extent that communities overlap each other is defined in (Equation 9) [34].

$$f_{\text{overlapping}}(\mathcal{A}) = \sum_{v \in \boldsymbol{V}_{\text{overlapping}}} \min_{c \in \{\boldsymbol{C}' \mid \boldsymbol{C}' \in \mathcal{A}\langle \boldsymbol{C} \rangle \text{ and } v \in \mathcal{A}\langle \boldsymbol{C} \rangle\}} \left\{ \frac{k_v^c}{k_v} \right\}$$ **(9)**

Where k_v^c denotes the number of edges connect node v and community c, and k_v denotes v's degree. To run Equation 9, the fraction of edges a node connects to each community to the node's degree first need to be calculated out, and then the minimum fraction of each node belong to multiple communities is aggregated.

Homogeneity

Homogeneity was first introduced by Wu and Pan [112]. This objective function works based on Shannon information entropy theory in which the entropy of a set, measures the average Shannon information content of the set. A set with high disorder rate has higher Shannon information. This leads to high entropy, indicating that a given set has high entropy and hence low homogeneity rate [82]. Thus, the entropy-based criterion can be used to measure how homogeneous the elements of a set or category are.

A network with attributes is named as attributed network which is defined as a 4-tuple $\mathcal{G}=(\mathcal{V}, \mathcal{E}, \mathcal{B}, \mathcal{H})$. In which $\mathcal{V}=\{v_1, v_2, \ldots, v_n\}$ is a set of n nodes, $\mathcal{E}= \{ (v_i, v_j)| \ v_i, v_j \in \mathcal{V}, v_i \neq v_j \}$ denote the edge set, $\mathcal{B} = \{b_1, b_2, .., b_t\}$ is a set of categorial attributes and $\mathcal{H} = \{h_1, h_2, .., h_t\}$ is an attribute function set. And t is the number of attributes in the network. Every attribute b_i has a domain $dom(b_i)=$ $\{b_i^1, b_i^2, \ldots, b_i^{d_i}\}$ where d_i is the size of the domain of b_i. Each attribute function $h_i: \mathcal{V} \rightarrow dom(b_i)$ assigns each node in $\mathcal{V}$ with an attribute value on b_i. Thus each node v_i is associated with an attribute vector $\mathcal{H}(v_i)$ where the jth element is given by the function $h_j(v_i)$.

Wu and Pan [112] stated that the homogeneity of a categorization X on attribute b_j is $H_{b_j}(X)$ can be defined in (Equation 10).

$$H_{b_j}(X) = \ln(d_j + 1) - PCE_{b_j}(X). \qquad \textbf{(10)}$$

Given a homogeneous category $\mathcal{G}_1$ with size n_1 whose nodes have the same attribute value on attribute b_j, its degree of homogeneity with respect to attribute b_j is defined as $D_j(\mathcal{G}_l) = \frac{n_l}{n_l + 1} \ln \frac{n_l}{n_l + 1}.$ where n_1 denotes the number of nodes in category $\mathcal{G}_1$ and n_{lj}^q is the number of nodes with attribute value b_j^q in category $\mathcal{G}_1$.

$$PCE_{b_j}(X) = \sum_{\mathcal{G}_l \in X} \frac{n_l}{n} PE_{b_j}(\mathcal{G}_l). \qquad \textbf{(11)}$$

Where $PE_{b_j}(\mathcal{G}_l) = -\sum_{q=1}^{d_j} pp_{lj}^q \ln pp_{lj}^q,$ and $pp_{lj}^q = \frac{n_{lj}^q}{n_l + 1}$.

The value of homogeneity is positive and performs a high value if the attribute categorization objective is attained. The attributes of objects in real-world networks are numerous and diversified, but not all attributes are suitable for significative classification. When multiple attributes are selected, a weighted homogeneity *H(X)* is defined to handle them [112].

Weighted homogeneity *H(X)* grants the control of the significance of different attributes in the classification. Assuming that *t* attributes are chosen for categorization, the weighted homogeneity is defined in (Equation 12) [112].

$$H(X) = \sum_{j=1}^{t} \omega_j H_{b_j}(X), \qquad \textbf{(12)}$$

Where ω_j is the weight of $H_{b_j}(X)$ which measures the importance of attribute b_j in categorization.

However, Moayedikia [82] claims that Modularity and homogeneity are conflicting, which means improving one of them leads to degradation of another.

Rand Index

This measure was proposed by Rand [113], it focuses on pairwise agreement, for each possible pair of elements in the considered set, the Rand index evaluates how similarly the two partitions treat them.

Two partitions of the same set *S* can be denoted by X={x_1, .., x_I} and Y={y_1, .., y_J, where x_i and y_j are the parts (1≤ i ≤ I) and (1≤ j ≤ J). And to denote the cardinalities, n=|S| is used for the total number of elements in the partitioned set, and n_{ij}=| $x_i \cap y_j$| for the intersections of two parts. And n_{i+}=| x_i| and n_{+j}=| y_j| as the part sizes.

Let a (respectively d) be the number of pairs in a community (respectively to different parts) in both partitions. And b (respectively c) be the number of pairs in which nodes belong to the same part in the first (respectively second) community, while they belong to distinct parts in the second (respectively first) one. So, a can be generated by counting the number of pairs belonging to part intersections $x_i \cap y_j$:

$$a = \sum_{ij} \binom{n_{ij}}{2} \qquad \textbf{(13)}$$

Furthermore, b and c match to pairs of different intersections elements. In b, the number of pairs in part x_i are considered, if they were not counted in a. So b and c can be calculated using (Equation 14), (Equation 15) respectively.

$$b = \sum_{i} \binom{n_{i+}}{2} - \sum_{ij} \binom{n_{ij}}{2} \qquad \textbf{(14)}$$

$$c = \sum_{j} \binom{n_{+j}}{2} - \sum_{ij} \binom{n_{ij}}{2} \qquad \textbf{(15)}$$

And d is calculated by subtracting a, b, and c by the total pairs number.

$$d = \binom{n}{2} + \sum_{ij} \binom{n_{ij}}{2} - \sum_{i} \binom{n_{i+}}{2} - \sum_{j} \binom{n_{+j}}{2} \qquad \textbf{(16)}$$

And finally, the Rand index (RI) is generated by processing the proportion of pairs on which both partitions agree:

$$RI(X,Y) = \frac{a+d}{a+b+c+d} \qquad \textbf{(17)}$$

The maximum value of *RI* is 1, and it indicated an ideal resemble of communities, while the minimum value is 0.

Adjusted Rand Index

The Rand Index was enhanced and proposed later under the name of Adjusted Rand Index or ARI [114].

It should be noted that the two partitions of the same set *S* can be denoted by X={x_1, .., x_I} and Y={y_1, .., y_J, where x_i and y_j are the parts (1≤ i ≤ I) and (1≤ j ≤ J). And to denote the cardinalities, n=|S| is used for the total number of elements in the partitioned set, and n_{ij}=| $x_i \cap y_j$| for the intersections of two parts. And n_{i+}=| x_i| and n_{+j}=| y_j| as the part sizes.

Hubert and Arabie proposed a model that produced arbitrary partitions with the constraint of having fixed number of parts (*I* and *J*) and part sizes (n_{i+} and n_{+j}). And based on that, the expected value for the number of pairs in a part intersection $x_i \cap y_j$ can be calculated in (Equation 18).

$$E\left(\binom{n_{ij}}{2}\right)=\binom{n_{i+}}{2}\binom{n_{+j}}{2}\Big/\binom{n}{2} \qquad \textbf{(18)}$$

And therefore, the proposed ARI measure can be calculated in (Equation 19)

$$ARI(X,Y)=\frac{\sum_{ij}\binom{n_{ij}}{2}-\sum_{i}\binom{n_{i+}}{2}\sum_{j}\binom{n_{+j}}{2}\Big/\binom{n}{2}}{\frac{1}{2}\left(\sum_{i}\binom{n_{i+}}{2}+\sum_{j}\binom{n_{+j}}{2}\right)-\sum_{i}\binom{n_{i+}}{2}\sum_{j}\binom{n_{+j}}{2}\Big/\binom{n}{2}} \qquad \textbf{(19)}$$

The *ARI* is symmetrical, and the ideal value is 1 indicating that partitions are identical, while values less than or equal to 0 indicates a low accuracy.

Variation of Information

The variation of information introduced by Meila is a dissimilarity measure, it compares two partitions and indicates whether or not they are different from one another [115].

In a community C_k that contains n_k nodes in network *N*. And $f_k = n_k / N$ is the fraction of nodes that belong to community C_k. The amount of information contained in a partition P can then be defined by its Shannon entropy.

And the Variation of Information between partition $\mathcal{P}$ and $\mathcal{P}'$ will indicate the amount of unshared information by the two partitions, and can be expressed by marginal ($H(\mathcal{P}), H(\mathcal{P}')$) and joint ($H(\mathcal{P}, \mathcal{P}')$) entropies

$$\mathrm{VI}(\mathcal{P},\mathcal{P}')=2\mathrm{H}(\mathcal{P},\mathcal{P}')-\mathrm{H}(\mathcal{P})-\mathrm{H}(\mathcal{P}') \qquad \textbf{(20)}$$

Where $\mathrm{H}(\mathcal{P})=-\sum_{k} f_k \log f_k.$ **(21)**

And ($H(\mathcal{P}, \mathcal{P}')$) can be defined as

$$\mathrm{H}(\mathcal{P},\mathcal{P}') = -\sum_{k}\sum_{k'} f_{k,k'} \log f_{k,k'}\,. \qquad \textbf{(22)}$$

This measure is a true metric distance, symmetric, non-negative, and satisfies the triangle inequality. It is a number between 0 and 1, and is equal to 0 only when the partitions are identical. Which means that the closer the value is to 0, the better.

Split-Join Distance

Split-Join is used to measure the privacy level after anonymization in the social network in the community detection [116]. The value of this measure is meant to be as low as possible, to indicate an ideal privacy level. It is basically the sum of the projection distance between partitions A and B, Split-Join can be calculated

$$\rho_A(B) = \sum_{a\in A} max_{b\in B}|a \cap b| \qquad \textbf{(23)}$$

Where |*a*∩*b*| denotes the number of common members between any subset a ∈ A and b ∈ B.

F-Score

F-Score of a network is based common class membership of object pairs in a true community structure of the network C and the com- munity structure achieved by the algorithm network C'. Let T denote the set of object-pairs that belong to the same class in C and S denote the set of object- pairs that are assigned to the same cluster in C' [69].

In a certain network, if *T* is the set of object-pairs in the same class in and *S* is the set of object pairs that are assigned to the same group in the network. The F-score can be calculated in (Equation 24) [69].

$$\text{Precision} = \frac{|S \cap T|}{|S|}; \ \text{recall} = \frac{|S \cap T|}{|T|}$$
$$F - \text{score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

(24)

The value of F-score falls in the range from 0 to 1. The maximum values of the F-score, indicates that the partition is closer to the ground truth.

Computational Cost

Computational cost of an algorithm includes the time complexity, and convergence rate. In this field of study, a few researchers have measured the run time of their algorithm and compared it to other existing algorithms to find an effective method. This is probably because the efficiency can be sacrificed as long as the method is generating more accurate results.

A set of Objectives

A research study was proposed to focus on the correlations (i.e., positively correlated, independent, or negatively correlated) of 11 objective functions to determine the effect of optimization objectives on the performance of multi-objective community detection [117]. These are listed as follow:

- Conductance (Q1) measures the fraction of total edge volume that points outside the cluster.

- Expansion (Q2) measures the number of edges per node that point outside the cluster.

- Cut Ratio (Q3) is the fraction of all possible edges leaving the cluster.

- Normalized Cut (Q4) is the normalized fraction of edges leaving the cluster.

- Maximum-ODF (Out Degree Fraction) (Q5) is the maximum fraction of edges of a node pointing outside the cluster.

- Average-ODF (Q6) is the average fraction nodes' edges pointing outside the cluster.

- Flake-ODF (Q7) is the fraction of nodes in S that have fewer edges pointing inside than to the outside of the cluster.

- Q (Q8) measures the number of within community edges, relative to a null model of a random graph with the same degree distribution.

- Description Length (Q9) is the number of edges between the community i and j. The objective regards the community as an optimal compression of network's topology.

- Community Score (Q10) measures the density of a sub-matrices based on volume and row/column means.

- Internal Density (Q11) is the internal edge density of the cluster

It was observed that the definitions of some objectives are correlated. For example, the first four objectives from the graph theory community are called the cut- based objectives, and the last three objectives are called degree-based objectives. So, the Pearson correlation coefficients were applied to characterize the relations.

In fact, evaluation metrics are based on the functional ground-truth community structure while quality metrics describe topological properties linked to cohesiveness. It can be observed that the most used metrics are Modularity and NMI. While other measures (such as Homogeneity and F-Score) did not receive much attention. On the other hand, a number of research studies have proposed adjusted versions of the evaluation metrics (such as NMI and ARI) [118]. Whereas homogeneity, as it was introduced in 2016, was not considered in many studies, and the measure proposed by Wu and Pan [112] does not consider the network structure, as real-world datasets might have some aspects that need to be considered when measuring the homogeneity.

It should also be noted that several metrics were proposed to measure the accuracy on detected partitions, such as NMI, RI, ARI, VI, and F-Score. And different research studies have used different accuracy measures in order to compare the detected communities by the ground-truth. However, NMI seems to be the most frequently used metric among all.

Maximizing all the evaluation metrics at once might not be practical, therefore, different research studies were conducted to optimize a measure while maintaining the others. Table 2-11 below summarizes the evaluation parameters used in several research papers. It is noticeable that the most common parameters are Modularity and NMI. While computational cost has not received much attention compared to other metrics, as the detection of accurate and high-quality community structure is more important aspects.

The proposed ABLP method reveals disjoint communities in social networks, which means that overlapping metrics are not examined. The main concern for this method is to detect homogeneous communities while maintaining a high Modularity measure.

Therefore, Modularity is used to measure the quality and strength of the division and Homogeneity to evaluate the similarity in communities. And to measure the accuracy when compared to ground-truth structures NMI is considered and supported with other accuracy measures (RI, ARI, VI). In addition to Split-Join to measure the privacy level.

Table 2-11: Evaluation Metrics used for community detection problem, where NMI is Normalized Mutual Information, RI is Rand Index, ARI is Adjusted Rand Index, VI is Variation of Information

Source	Year	Evaluation Metrics									
		NMI	Modularity	Cost	Overlapping	RI	ARI	VI	Homogeneit y	F1-Score	Split-Join
[119]	2021	✓	✓				✓				
[120]	2021	✓	✓			✓	✓	✓			✓
[86]	2021	✓		✓	✓					✓	
[95]	2021	✓			✓				✓		
[62]	2021	✓			✓				✓		
[79]	2021	✓	✓	✓							
[87]	2021	✓	✓	✓							
[18]	2020	✓	✓					✓			
[80]	2020	✓	✓							✓	
[65]	2020	✓	✓	✓							
[66]	2019	✓	✓								
[121]	2019	✓	✓	✓		✓	✓	✓			✓
[60]	2019	✓	✓								
[59]	2019	✓		✓							
[60]	2019	✓	✓	✓	✓						
[59]	2019		✓	✓	✓						
[77]	2018			✓							
[46]	2018	✓	✓								
[97]	2018	✓							✓		
[68]	2018	✓	✓								
[77]	2018			✓							
[78]	2018	✓	✓								
[58]	2018	✓	✓	✓	✓						
[82]	2018		✓	✓					✓		
[122]	2018	✓									✓
[67]	2018	✓	✓		✓						
[89]	2018	✓	✓	✓							
[90]	2018		✓								
[108]	2018	✓	✓								
[84]	2018	✓	✓	✓							
[46]	2018	✓	✓								
[69]	2017	✓	✓	✓						✓	
[123]	2017	✓				✓	✓	✓			
[92]	2017	✓	✓								
[124]	2017	✓	✓								
[91]	2017	✓	✓								
[45]	2017	✓	✓								
[70]	2017		✓								
[71]	2016		✓								
[85]	2016	✓	✓								
[4]	2015	✓									
[72]	2015	✓	✓		✓						
[125]	2015	✓	✓				✓				
[53]	2014	✓	✓								
[74]	2013	✓	✓	✓							
[75]	2013	✓	✓	✓							
[126]	2012	✓			✓						
[127]	2011			✓				✓			
ABLP *	2022	✓	✓			✓	✓	✓	✓		✓

*ABLP is the (Attribute Based Label Propagation Algorithm) Proposed in chapter 4

The homogeneity degree was proposed based on Shannon information entropy theory, and it does not consider the network structure, as real-world datasets might have some aspects that need to be considered. Also, if an algorithm detects a number of communities, with similar attribute nodes in each community, it will score a high homogeneity measure, regardless of the efficiency of the results. Therefore, a homogeneity measure that can be flexible and customized based on the dataset and the specified attribute values and evaluates the homogeneity degree in each community is needed. And since it was stated that Modularity and homogeneity are conflicting[82], there is a clear research gap as this area needs to be studied further, and methods that maximize both homogeneity and Modularity are needed to make sure that the community detection results are accurate and homogeneous. The evaluation measures used also consider one perspective of assessing the results, which is why there are various measures and techniques. This can be resolved by examining a penalized measure that takes more than one aspect, and therefore provides a more comprehensive evaluation.

2.2.3 Community Detection Applications

Community detection problem was studied from different disciplines, various types of networks (datasets) were used as community detection datasets. The type of dataset mainly depends on the purpose of community detection, the practical applications of community detection were discussed and categorized by Karatas and Sahin [128]. The categories are listed below, and recent research studies are updated and discussed for each category:

Criminology:

Community detection was used to identify criminal user groups, which are formed by real accounts, or bot accounts. This resulted in supporting or diffusing criminal ideas or terrorism-like activities. Karatas and Sahin also conducted a research study to reveal the potential hazards of malicious social bots and review the detection techniques within a methodological categorization [129]. A recent study also proposed a system to detect bank fraud using a community detection algorithm that identifies the patterns that can lead to fraud occurrences [130].

Public Health:

Community detection was applied on medical data (e.g., genomic, tissue phenotyping), to discover dynamics of certain groups susceptible to an epidemic disease. It was used to study a data-driven clustering approach by clustering subjects from twelve cancer types using modularity maximization-based community detection technique [131].

Another study proposed a semi-supervised cellular community detection algorithm for tissue phenotyping based on cell detection and classification, and clustering of image patches into biologically meaningful communities [132]. As issue phenotyping in a whole-slide images can aid with understanding the contents and the tumor microenvironment associated with cancer subtypes in terms of survival and clinical outcomes.

Politics:

Community detection can be used for observation of influences of political ideologies or individual politicians on some social group. It was used to detect bots that try to create a fake impression on real grassroots for political reasons [129]. And since Twitter has become one of the main stages of political activity both among politicians and partisan crowds, a study examined Twitter content information including word, hashtag, and domain, it showed that that user content and endorsement filtered connectivity information are integral detecting politically motivated users into pure political communities [133].

Customer Segmentation, Smart Advertising and Targeted Marketing:

Detecting customers/ shoppers' groups can provide help for marketing companies. A research study used community detection greedy algorithm to detect influential people based on their greedy index, and then improved the accuracy due to users' index similarity [134].

Recommendation Systems:

Detecting people with similar preferences will help the recommender system in proposing something that users will most likely prefer. A Louvain method to partition customers into clusters based on their purchases' similarity was proposed, it generated interpretable clustering results with distinct product purchase patterns which led to higher response rates in the recommendation of products to customers in the same cluster [135].

Network Summarization and Privacy:

Community detection can provide privacy to the network when sharing generalized properties with other parties. It was used to reidentify multiple addresses that belong to same user in weak signal bitcoin network [136].

Link Prediction:

It is used to detect fake, missing, and future links between the network's nodes. A generative model for multilayer networks that extends and generalizes the mixed membership stochastic block model was proposed [137], it assumes that the layers share a common community structure but allows links in different layers to be correlated with the community memberships in different ways.

Social Network Analysis:

It helps in understanding the networks' behavior. Social networks as discussed earlier in this section can include online networks such as Facebook, Linked-in, Twitter etc. A survey was conducted to constitute the concept of community and the problem of community detection in the social media area and categorize the existing community detection algorithms based on their methodological principles [15]. It also proposes five strategies for scaling community detection to real-world networks of huge scales, which are: sampling techniques, local graph processing, iterative schemes, multi-level approaches and parallelization. And then the techniques were implemented to social media applications, such as topic detection, tag disambiguation, user profiling construction, photo clustering and event detection.

By reviewing the most relevant and recent research papers, more applications and studies that used community detection techniques to solve their problems were found. And hence, several specific categorizations can be added, which are listed and discussed below:

Psychology:

To overcome the social tragedy of the increasing psychological pressure, research studies were conducted as group psychology analysis. Psychologists tend to study groups as most of human activities such as working, learning, playing happen in groups. Instead of the abstract data presentation, a recent study has used the Reinforcement Learning technique to visualize the group psychology analysis [138].

Citations Analysis:

Citation analysis is a prevalent research field, it is used to rank the authors and the publication venues of research papers. As the number of publications is rapidly increasing, the users are not able to locate pertinent studies. Community detection was employed to identify the latent groups of citation networks, and then recommend the studies based on the groups' relationships. A recent study has discussed the community detection accuracy and the impact of improving direct citations, with regards to publication–publication relatedness measurement, by indirect citation relations (bibliographic coupling, co-citation, and extended direct citations) and text citations [139]. The results indicate that the co-citation performed poorly, whereas the extended direct citations achieved the best performance. Another research study also investigates the community detection problem in bibliometrics, it uses overlapping community detection techniques to reveal the groups of authors, papers, and venues [140]. This area was also studied by Gupta et al., that proposed a community detection method for evolution diagnosis of Bibliographic networks [141].

Music:

An interesting application of community detection is the musical rhythmic pattern extraction, a musical piece is generally formed by one or more predefined rhythmic patterns and such patterns are composed of rhythmic cells (RCs), which are groups of rhythmic figures derived from nth division of a larger rhythmic figure [142]. This research study proposed a method that can extract any type of rhythm pattern, both monophonic and polyphonic, represented by symbolic data. And another research study proposed a hierarchical analysis of music structure that is based on graph theory and multi-resolution community detection [143].

Business and Enterprise:

The idea is that employees within the organization can be distributed on several groups or communities. In the enterprise context, valuable information related to the employees can be obtained in offline company internal sources and online enterprise social networks (ESNs). A study proposed a method to reveal the employees' social communities, and the problem is formally called the "Enterprise Community Detection"

(ECD) problem [144]. On the other hand, a framework was proposed a method that performs co-clustering on enterprise social networks for effective communities' detection and recognition [145]. The model integrates the network's topology structure and rich content information which covers the interactions and correlations between employees.

In general, various algorithms were used to reveal the detected communities in different applications and networks, however, most of the algorithms were based on modularity maximization, as shown in Table 2-12. Which indicates that evaluating the communities' structure is an important metric regardless of the network type or the purpose of detection. While some algorithms were based on centrality measures such as betweenness and PageRank. On the other hand, some applications employed community detection algorithms based on node importance, or the top actors' interactions, where the communities are structured based on the association of nodes/actors. Nonnegative Matrix Factorization and simple clustering techniques was also used in some fields.

Table 2-12: Community Detection methods based on their applications

Type of Detected Communities (Application)	**Community Detection Methods**				
	Centrality Measures	**Top actors' interactions**	**Modularity Maximization**	**Nonnegative Matrix Factorization**	**Clustering**
Criminology	[146] [130]	[146]			
Public Health		[132]	[131]		
Politics				[133]	
Advertising and Marketing					[134]
Recommendation Systems			[135]		
Network Summarization and Privacy			[136]		
Link Prediction			[137]		
Social Network Analysis			[15]		
Psychology			[138]		
Citations Analysis			[139]		
Music			[143]		
Business and Enterprise				[144]	

To understand the characteristics of the detected communities based on the application they were used for, Table 2-13 illustrates the of detected communities and their applications. It can be observed that social communities play dominates the community detection problem, as they were utilized to detected social communities of strongly connected/ related people in several fields such as criminology, politics, advertising and marketing, link predication, social network analysis, psychology, and business and enterprise. Whereas biological networks were employed in the public health types of research, and communities based on hierarchical segmentation were detected in musical tracks.

Table 2-13: Characteristics of the detected communities based on their application

Type of Detected Communities (Application)	Source/ Year	Characteristics of detected communities			
		Social	biological	Based on user transactions/ relationships	hierarchical segmentation
Criminology	[146] 2014	✓			
	[130] 2020	✓			
Public Health	[131] 2017		✓		
	[132] 2020		✓		
Politics	[133] 2016	✓			
Advertising and Marketing	[134] 2021	✓			
Recommendation Systems	[135] 2021			✓	
Network Summarization and Privacy	[136] 2017			✓	
Link Prediction	[137] 2018	✓	✓		
Social Network Analysis	[15] 2012	✓			
Psychology	[138] 2021	✓			
Citations Analysis	[139] 2020			✓	
Music	[143] 2020				✓
Business and Enterprise	[144]	✓			

It is noteworthy that community detection plays an important role in numerous fields and can be utilized to solve many substantial problems. And although there are many purposes for community detection, and various research studies were conducted, there isn't a study that examines the spread of a certain disease in a social network. The main concept in finding communities in social network seems to be used for marketing, recommender systems, or preferences grouping. And with the emerge of COVID-19, the world has employed many techniques to identify the problem, provide solutions and limit the damages. It is considered in this thesis that the community detection problem can be used to address the pandemic of COVID-19.

2.2.4 Datasets for Community Detection in Social Networks

To evaluate the performance and accuracy verification of a certain algorithm, it is implemented on several datasets. The datasets are basically networks represented by graphs, and the datasets used can be synthetic, or real words networks.

Synthetic networks are able to obtain real community structure through the setting of tunable parameters [80]. One of the most popular synthetic networks is the classic network proposed by Girvan-Newman [148]. It is a large set of artificial, computer-generated graphs, each graph was constructed with 128 vertices, the graph is split into four communities, each community contains 32 vertices. Edges were set between pairs of vertices randomly and autonomously, with a certain probability. This generates graphs that have known community structure, but which are particularly random in other respects.

Another commonly used synthetic network is the Lancichinetti–Fortunato–Radicchi (LFR) benchmark network [149], it accounts for the heterogeneity in the distributions of node degrees and of community sizes. The network was constructed by several tunable parameters which has similar identical attributes as real networks, and different types of networks can be generated by changing the values of related exponents [80]. The LFR benchmark network assumes that both the degrees of vertices and the sizes of communities obey exponential distribution [149].

Real world datasets benchmark is another way to measure the performance of the community detection algorithms, some networks have a known community divisions while others do not. However, different scales datasets from various domains can be adopted for this purpose, these include social networks (e.g., Zachry’s Karate Club [150], American College Football [148]), and biological networks (e.g., Yeast [151]).

Table 2-14 summarizes the datasets used in a several community detection research papers over the past years, datasets are divided into real-world and synthetic datasets.

Table 2-14: Datasets used in Community Detection

Real-world Networks		
Source	[86] 2021 [87] 2021 [65] 2020 [59] 2019 [58] 2018 [68] 2018 [89] 2018 [90] 2018 [84] 2018 [69] 2017 [92] 2017 [91] 2017 [45] 2017 [44] 2016 [109] 2016 [4] 2015 [53] 2014 [126] 2012	[79] 2021 [80] 2020 [66] 2019 [60] 2019 [78] 2018 [46] 2018 [82] 2018 [67] 2018 [108] 2018 [124] 2017 [54] 2017 [70] 2017 [85] 2016 [71] 2016 [72] 2015 [73] 2014 [75] 2013
Synthetic Networks		
Source	[86] 2021 [87] 2021 [60] 2019 [59] 2019 [46] 2018 [68] 2018 [67] 2018 [84] 2018 [124] 2017 [54] 2017 [70] 2017 [4] 2015 [75] 2013	[79] 2021 [80] 2020 [66] 2019 [58] 2018 [78] 2018 [89] 2018 [108] 2018 [69] 2017 [92] 2017 [91] 2017 [109] 2016 [72] 2015 [126] 2012

It can be observed that there is a wide range of real-world networks, so the most used networks are discussed below:

Zachry's Karate Club [150], this network has a known community structure, the nodes represent the members of a karate club, observed over three years. A disagreement occurred between the club president and the instructor of the karate club, which resulted in a split in the team. Hence, the club was divided into two separate groups due to this conflict, so one group sided with the instructor and the other the club president. The club is divided into two groups which are considered as communities (ground-truth is illustrated in Figure 2-3 communities are distinguished by circles and squares). This network consists of 34 members, node 1 represents instructor and node 33 represents instructor administrator. The edges represent the relationships of those club members who interacted with each other.

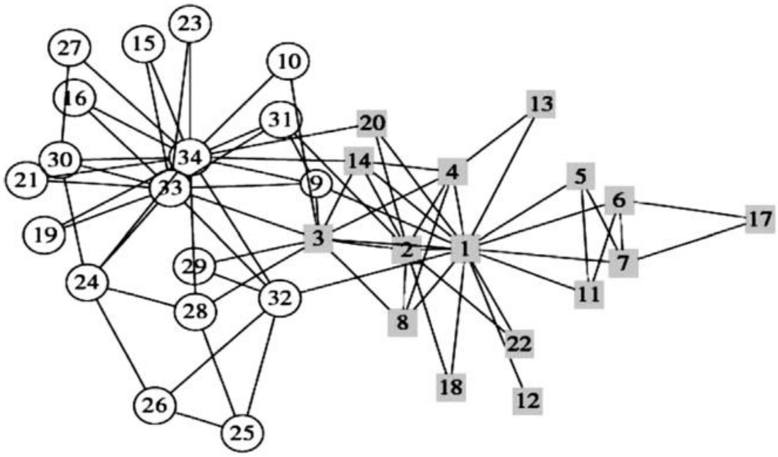


Figure 2-3: Zachry's Karate Club Network [78]

American college football [148], is an attributed network is constructed from a university football matches in USA. Nodes represent teams and edges represent the matches played between the teams. The network includes 12 teams considered as communities which consists of 115 nodes and 616 edges (ground-truth is illustrated in Figure 2-4 each team is in a different color). The nodes in this network belong to a conference, which is considered as an attribute, there are 11 conferences.

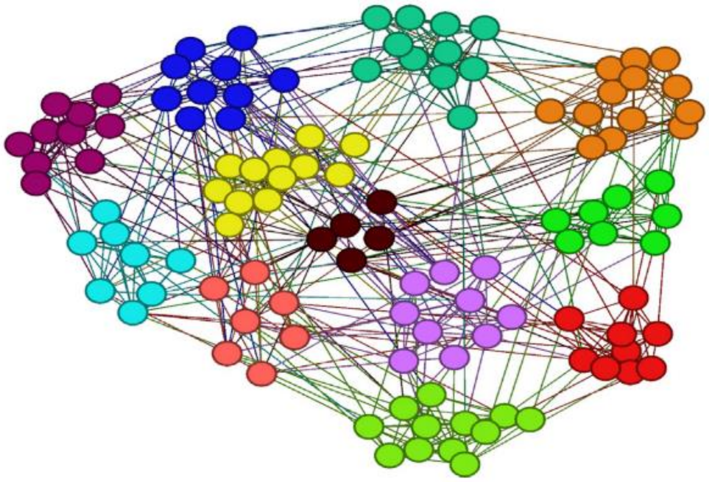


Figure 2-4: American College Football Network [78]

American Political Books [110], it is an attributed social network of US politics books. This network is of co-purchase relationships of online booksellers. The nodes of the network represent books about United States politics, while the edges between two books occur when these books were purchased together by a shopper. There are three different categories of books which are considered as their attributes: liberal; neutral; and conservative. (ground-truth is illustrated in Figure 2-5, Triangles: neutral books, dots: conservative books, and squares: liberal)

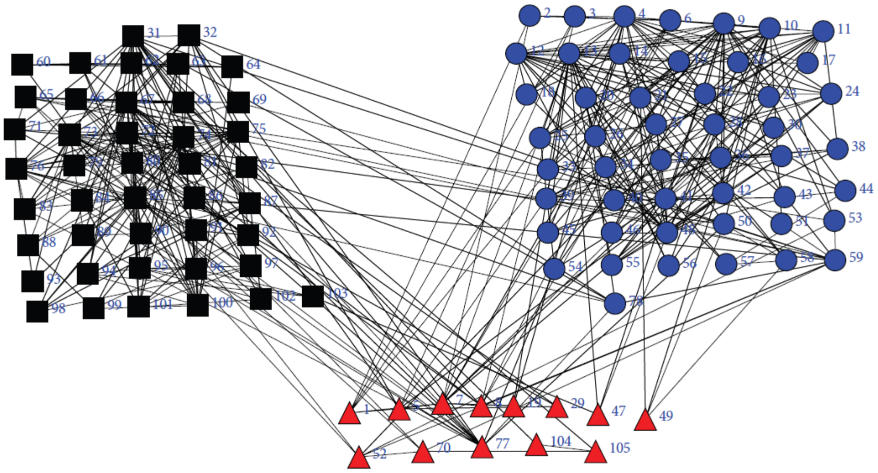


Figure 2-5: Political Books Networks [152]

These social networks datasets were extensively used in the literature to test and benchmark various community detection algorithms, which helped in understanding the networks' structures and evaluate the results.

2.2.4.1 COVID-19 Constructed Datasets

Since the COVID-19 pandemic, different types of datasets were created, a comprehensive survey was conducted to review COVID-19 open-source datasets [153], this includes medical datasets such as medical images (CT scans and X-rays), textual datasets (such as tweets, scholarly articles, mobility, and non-pharmaceutical interventions (NPI). For instance, a social dataset to analyze human emotions and worries regarding COVID-19 was created, the goal was to understand emotional responses on a wide scale [154]. The study involved 2500 participants from the UK reporting their emotions during two days over the Lock-down in ICU and created a dataset of 5000 texts (2500 short and 2500 long texts). The participants were also required to report their feelings about COVID-19 situations using a defined scale, in addition to a scale of how much anger, anxiety, desire, disgust, fear, happiness, relaxation, and sadness they felt. In this context, several datasets were also created based on Twitter tweets during the pandemic: [155], [156], [157]. For example a research study considered all tweets that contained "5Gcoronavirus" keyword or the "#5GCoronavirus" hashtag, or were replied to or mentioned in these tweets in the period March to April 2020 [14]. It was used to analyze the 5G conspiracy theory in the context of COVID-19 on Twitter offering practical guidance to health authorities in how, in the context of a pandemic, rumors may be combated in the future. Another dataset was formed based on Instagram hashtags related to COVID-19 #coronavirus, #covid19, #corona, etc.) [158]

However, none of these datasets were networks that can be used in the community detection. Nonetheless, with the emerge of Coronavirus, comes a need of tracing infected persons, and break the chain of the virus. As defined by the World Health Organization: "Contact tracing is used to identify and provide supported quarantine to individuals who have been in contact with people who are infected with SARS-CoV-2 and can be used to find a source of infection by identifying settings or events where infection may have occurred, allowing for targeted public health and social measures." [159]

Therefore, targeting the network, and dividing it into communities will facilitate those purposes. Additionally, instead of focusing on grouping people in communities, there is a necessity of breaking these groups to stop the spread of the virus, which can only be done by investigating a network of patients. As WHO also stated, "along with robust testing, isolation and care of cases – is a key strategy for interrupting chains of transmission of SARS-CoV-2 and reducing mortality associated with COVID-19". Indeed, a dataset that presents a network of virus-infected-persons does not exist.

2.2.5 The Process of Community Detection

To recapitulate the literature review, the process of Community Detection can be divided in three parts; design and implementation of algorithm, experiments on different types of datasets and evaluation metrics to measure the performance of algorithm. This can be illustrated in Figure 2-6.

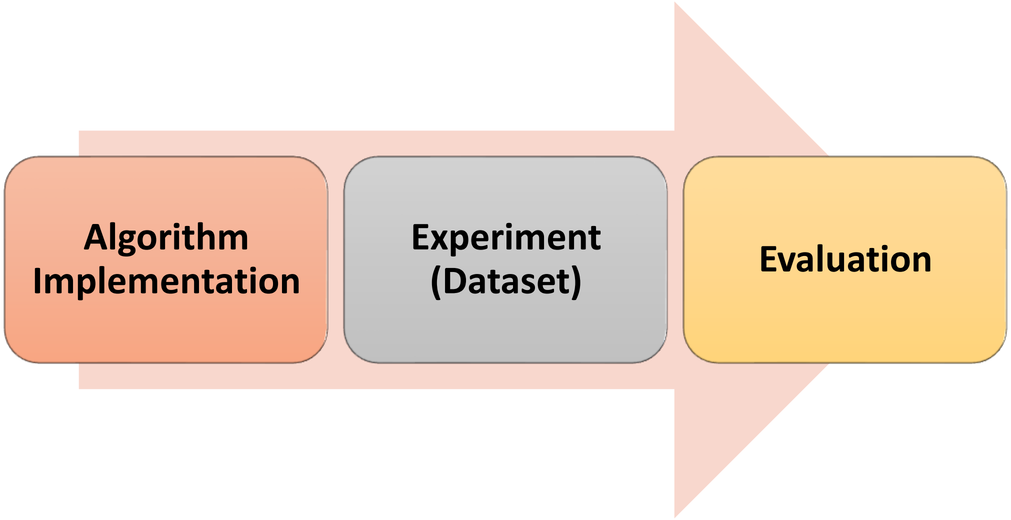


Figure 2-6: The Process of Community Detection

In this thesis, the three main areas of Community Detection process were investigated and contributed to, as shown in Figure 2-7, an algorithm, a dataset and an evaluation measure are proposed to fulfill the gap found in the literature.

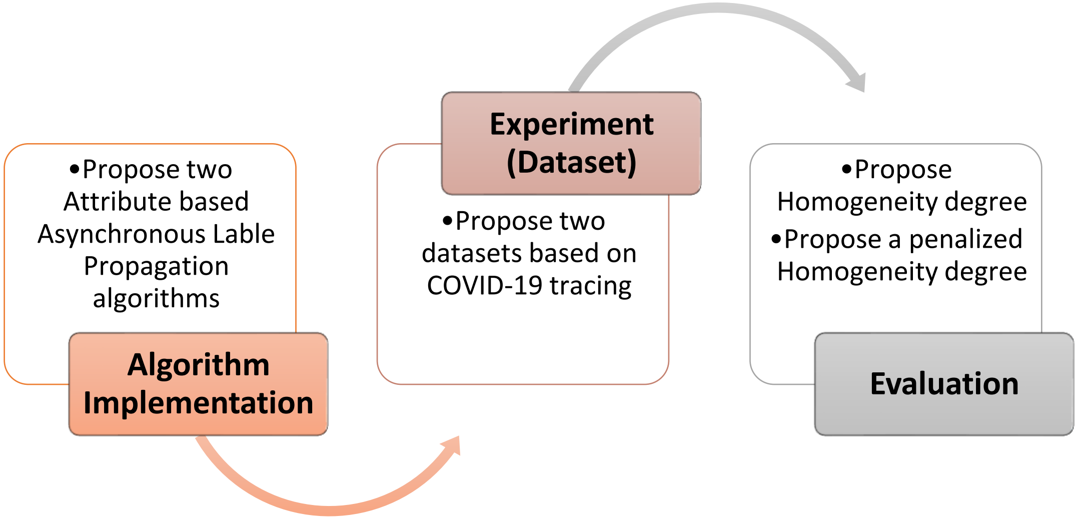

•Propose two Attribute based Asynchronous Lable Propagation algorithms
Algorithm Implementation
Experiment (Dataset)
•Propose two datasets based on COVID-19 tracing
•Propose Homogeneity degree
•Propose a penalized Homogeneity degree
Evaluation

Figure 2-7: Contribution to the Community Detection problem

The community detection problem has different angles that need to be covered, this includes the algorithm design, and then implementing it on network-based datasets, and finally examine the results of detected communities' performance through experiments of evaluation metrics.

Moreover, as it can be noticed from this chapter, that the evaluation techniques vary from one aspect to another, so a method is good at one criterion but might not be on the other, -as it was not designed or tested in that regards-; which makes it hard to decide to whether the proposed method satisfies the requirements of the user or not. In addition, algorithms return a set of solutions; the decision makers have to choose the optimal one, how to decide is basically the key issue to improve the algorithm performance [160].

So, these three essential components of the community detection problem can vary based on the main reason behind the detection, type of network and the nature of detected communities. In general, The main idea of these methods is the same, i.e., they model the community detection as either a single objective or a multi-objective optimization problem and then design optimization metaheuristics to solve it [52].

2.3 Summary

Based on the literature review conducted in this chapter, several research gaps were identified. First, homogeneity evaluation measure in attributed networks needs to be studied and optimized, as the number of research studies conducted in this area is comparatively low, and only one measure of Homogeneity was proposed, which might not be ideal in real-world datasets that require some sort of adjustments.

Moreover, it is noticeable that homogeneity was not studied as a constraint in the Community Detection, even though several constraints were employed in this context. As a result, an evaluation method that considers both Modularity and homogeneity does not exist.

In addition, with the emerge of COVID-19, there is an imperious need of studying and understanding the distribution of the virus, in order to limit spread of the virus. The problem of COVID-19 was never formed by a social network and studied as a part of the Community Detection problem.

3 COVID 19 Datasets

3.1 Overview

An epochal type of data is being excessively generated since December 2019, when a novel virus named COVID-19 emerged, and while the first case was discovered in Wuhan, China, it did not take long for the disease to travel across the globe and infect every continent (except Antarctica) [153], causing widespread infections and deaths due to its contagious characteristics and no medically proven treatment [161]. The COVID-19 pandemic has been termed as the most consequential global crisis since the World Wars. Because of the rapid prevalence of this virus, health organizations all over the world tend to track and store all data related to this pandemic. This includes the contact tracing, number of cases, number of deaths, etc. The availability of rich textual data from various online sources can be used to understand the growth, nature and spread of COVID-19 [162].

Different types of datasets were created since the emerge of COVID-19, this includes medical datasets such as medical images (CT scans and X-rays), textual datasets (such as tweets, scholarly articles, mobility, and non-pharmaceutical interventions.

However, according to World Health Organization (WHO) [159], contact tracing is the process of identifying, assessing, and managing people who have been exposed to a disease to prevent onward transmission. The contact tracing data were not employed to create a network, even though, when systematically applied, contact tracing will break the chains of transmission of an infectious disease and is thus an essential public health tool for controlling infectious disease outbreaks. When contact tracing data is compiled it can be represented by a network, and hence structured into a graph, which can be analyzed using graph mining techniques. In this section, the main concern is to use the contact tracing data to create a new network dataset, which is motivated by the open-source efforts by the health organizations.

3.2 Proposed Datasets

As any network can be outlined in a graph, and the graph is composed of a set of nodes which can be individuals or entities, and edges that represent the connections and interactions between the nodes [3]. This research creates and studies COVID19 datasets, to examine how the virus is spreading among nodes. According to CDC (Centers for Disease Control and Prevention, US), the use of digital contact tracing tools may help with certain case investigation and contact tracing activities but will not replace the need for a large public health workforce. As the complete clinical picture of COVID-19 is not fully known, the virus can be spread before symptoms occur or when no symptoms are present. Therefore, case investigation and contact tracing activities need to be studied. This type of analysis helps us understand the transmission of the virus, the distribution of countries based on their infection. It might also help predict where the virus might strike next, or how a mutated virus will behave, and therefore take the precautions needed.

3.2.1 Proposed COVID-19 Dataset: World Countries

Data science, defined broadly, will play a central role in the global response to the COVID-19 pandemic[163]. All the scientific efforts necessitate that the data brought to service for the analysis should be open source to promote the extension, validation, and collaboration of the work in the fight against the global pandemic. The open-source tracing data is utilized to create a novel network dataset, that is concerned with the spread of COVID19.

The dataset proposed here is based on the spread of virus between countries. An open-source contact tracing data are used to follow the spread of virus from January to March 2020, between the countries worldwide, which started in China and expanded to other countries. The data originally contained the following:

- Reporting date: the date on which the case was officially reported.
- Summary: this includes (if applicable) the location of the patient, the hospital, the duration of hospital stay, health condition, travelling history, nationality, how this patient got infected. However, some were left blank, so they were ignored.
- Location (town, state)
- Country
- Gender
- Age
- Symptom onset (date)
- Symptoms
- Hospital visit date
- Whether or not this patient visited Wuhan
- Source

The cases considered in this dataset are the ones that had a travelling history, so a relationship between the visited country and their hometown is created. Only the frequent countries were examined, so out of 3397 cases, 36 countries are considered. To understand the procedure of converting the data into a network, it is illustrated in Figure 3-1.

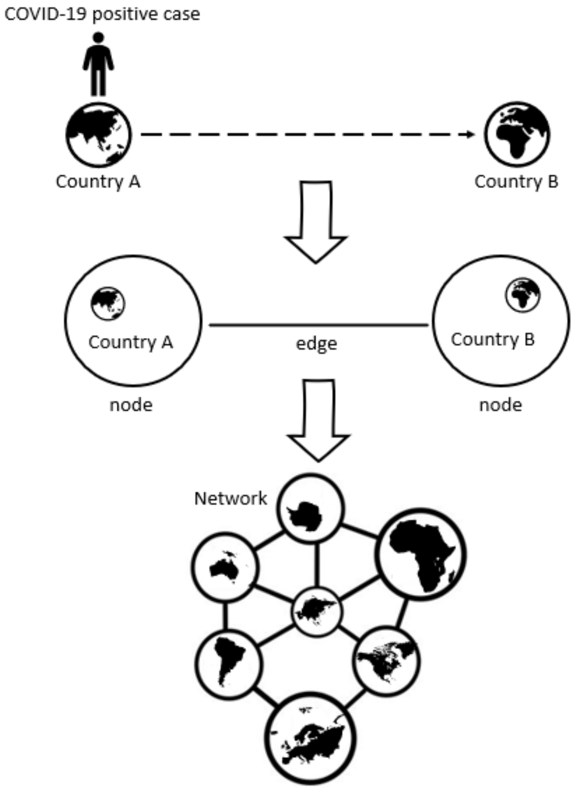
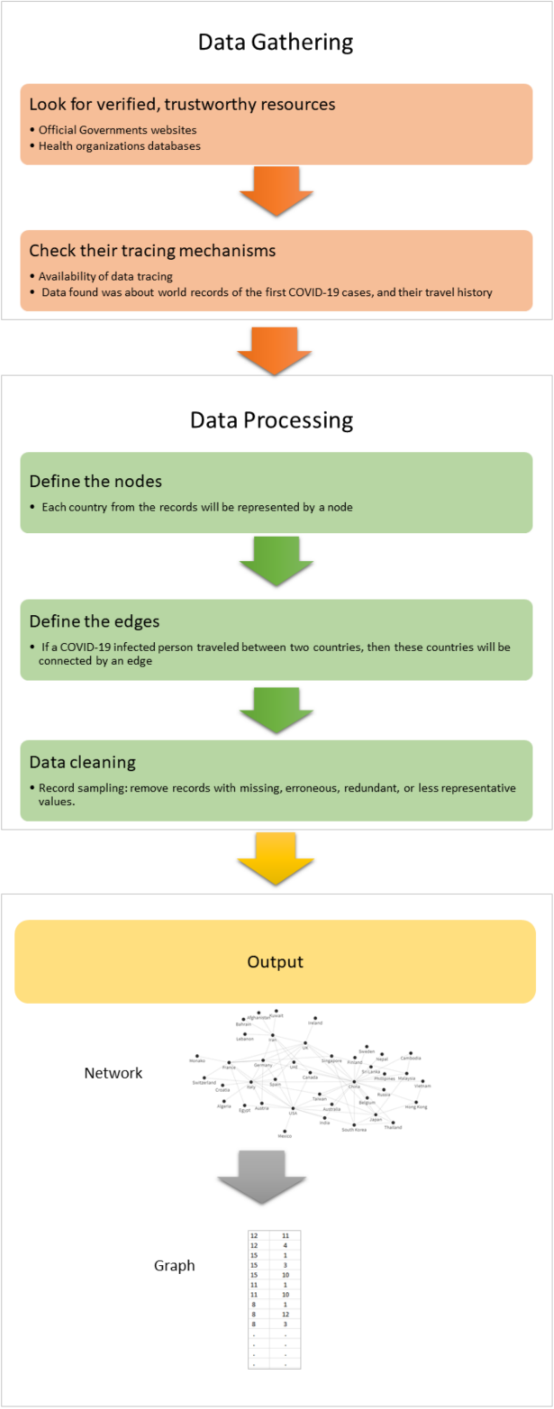


Figure 3-1: Procedure of converting the data into a network

The process of creating this dataset is illustrated in Figure 3-2, it started by gathering data from reliable sources, and then data cleaning to remove any redundancy, faulty or erroneous data. Finally, the network is formed by representing each country with node, and an edge is used when a country has a contact infected person from another country.


Data Gathering
Look for verified, trustworthy resources
• Official Governments websites
• Health organizations databases
Check their tracing mechanisms
• Availability of data tracing
• Data found was about world records of the first COVID-19 cases, and their travel history
Data Processing
Define the nodes
• Each country from the records will be represented by a node
Define the edges
• If a COVID-19 infected person traveled between two countries, then these countries will be connected by an edge
Data cleaning
• Record sampling: remove records with missing, erroneous, redundant, or less representative values.
Output
Network
Afghanistan
Kuwait
Bahrain
Ireland
Lebanon
Iran
UK
Sweden
Nepal
Cambodia
Monako
France
Germany
UAE
Singapore
Finland
Sri Lanka
Philippines
Malaysia
Canada
Switzerland
Italy
Spain
China
Vietnam
Croatia
Taiwan
Russia
Belgium
Hong Kong
Algeria
Egypt
Austria
USA
Australia
Japan
India
South Korea
Thailand
Mexico
Graph

12	11
12	4
15	1
15	3
15	10
11	1
11	10
8	1
8	12
8	3
.	.
.	.
.	.
.	.

Figure 3-2: Proposed COVID-19 dataset: World Countries Process

The dataset is non-attributed, it consists of 36 nodes (countries), and 75 edges (Figure 3-3). It should be noted that the data obtained to design this network does not include any personal information, it was publicly published and does not require any ethical approval.

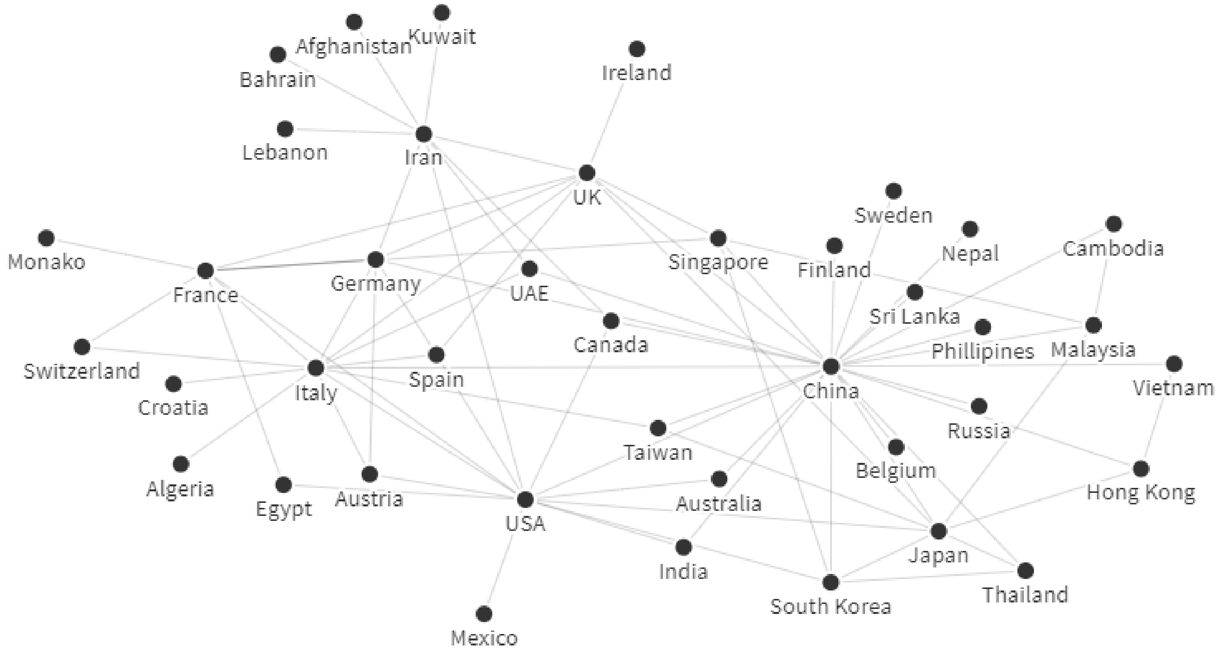


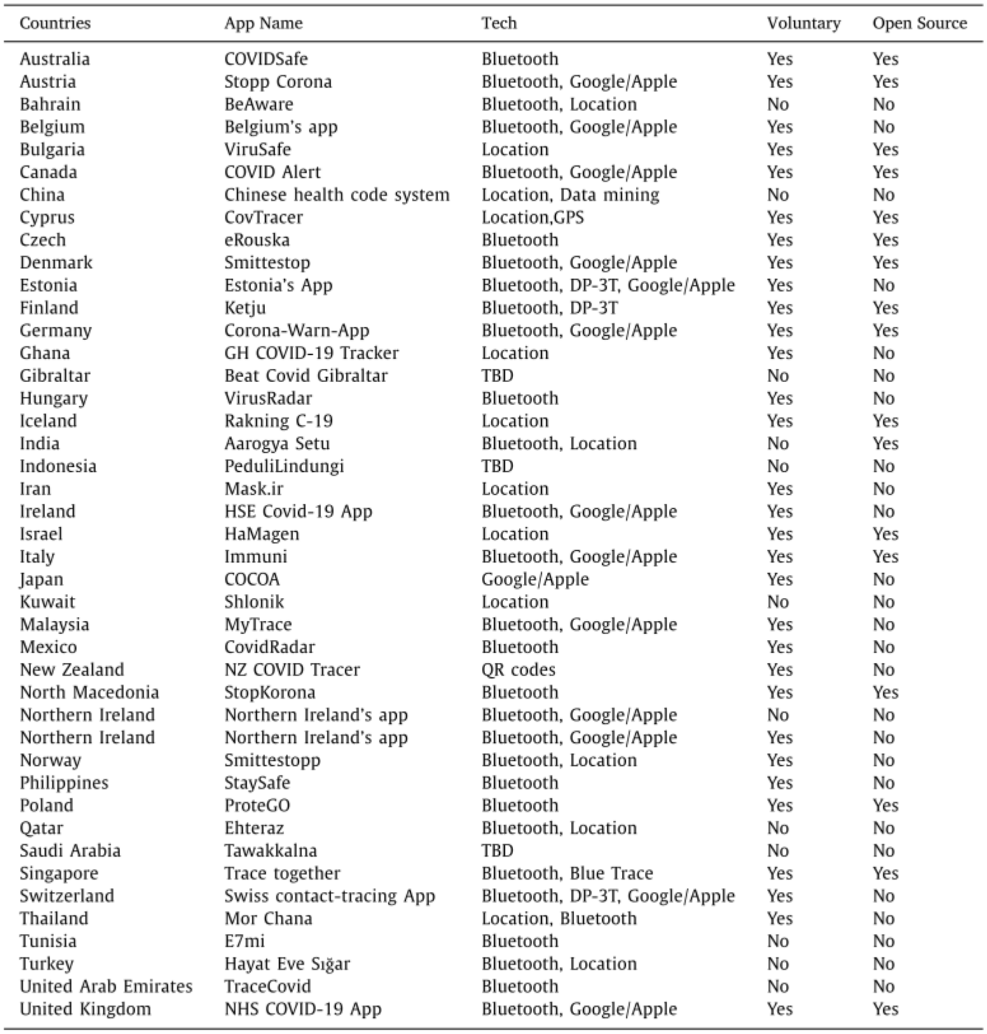
Figure 3-3: Proposed COVID-19 Dataset: World Countries

Practical limitation to creating this network is the missing data, however, this is comprehensible as the data collection was during a global pandemic and the cause of infection is not necessarily recognized.

3.2.2 Proposed COVID-19 Dataset: Contact Tracing in the Kingdom of Bahrain

With the rapid spread of COVID-19, different precautions were taken by governments of countries over the world, one is that many applications were developed to assist in the contact tracing process. And with the aim of upgrading the effectiveness of contact tracing, countries are exploiting advancements in mobile technology and Internet of Things (IoT) to assist the conventional manual process to trace individuals who were in contact with a positive COVID-19 case [162]. The majority of countries implemented digital contact tracing solutions to support the manual procedure. The execution of these applications used GPS or Bluetooth technologies. The usage of these applications was mandatory in some countries, and facultative in others. Table 3-1 enumerates the different mobile applications used by diverse countries [162].

Table 3-1: List of countries using different apps and their adopted technologies [162]



Countries	App Name	Tech	Voluntary	Open Source
Australia	COVIDSafe	Bluetooth	Yes	Yes
Austria	Stopp Corona	Bluetooth, Google/Apple	Yes	Yes
Bahrain	BeAware	Bluetooth, Location	No	No
Belgium	Belgium's app	Bluetooth, Google/Apple	Yes	No
Bulgaria	ViruSafe	Location	Yes	Yes
Canada	COVID Alert	Bluetooth, Google/Apple	Yes	Yes
China	Chinese health code system	Location, Data mining	No	No
Cyprus	CovTracer	Location,GPS	Yes	Yes
Czech	eRouska	Bluetooth	Yes	Yes
Denmark	Smittestop	Bluetooth, Google/Apple	Yes	Yes
Estonia	Estonia's App	Bluetooth, DP-3T, Google/Apple	Yes	No
Finland	Ketju	Bluetooth, DP-3T	Yes	Yes
Germany	Corona-Warn-App	Bluetooth, Google/Apple	Yes	Yes
Ghana	GH COVID-19 Tracker	Location	Yes	No
Gibraltar	Beat Covid Gibraltar	TBD	No	No
Hungary	VirusRadar	Bluetooth	Yes	No
Iceland	Rakning C-19	Location	Yes	Yes
India	Aarogya Setu	Bluetooth, Location	No	Yes
Indonesia	PeduliLindungi	TBD	No	No
Iran	Mask.ir	Location	Yes	No
Ireland	HSE Covid-19 App	Bluetooth, Google/Apple	Yes	No
Israel	HaMagen	Location	Yes	Yes
Italy	Immuni	Bluetooth, Google/Apple	Yes	Yes
Japan	COCOA	Google/Apple	Yes	No
Kuwait	Shlonik	Location	No	No
Malaysia	MyTrace	Bluetooth, Google/Apple	Yes	No
Mexico	CovidRadar	Bluetooth	Yes	No
New Zealand	NZ COVID Tracer	QR codes	Yes	No
North Macedonia	StopKorona	Bluetooth	Yes	Yes
Northern Ireland	Northern Ireland's app	Bluetooth, Google/Apple	No	No
Northern Ireland	Northern Ireland's app	Bluetooth, Google/Apple	Yes	No
Norway	Smittestopp	Bluetooth, Location	Yes	No
Philippines	StaySafe	Bluetooth	Yes	No
Poland	ProteGO	Bluetooth	Yes	Yes
Qatar	Ehteraz	Bluetooth, Location	No	No
Saudi Arabia	Tawakkalna	TBD	No	No
Singapore	Trace together	Bluetooth, Blue Trace	Yes	Yes
Switzerland	Swiss contact-tracing App	Bluetooth, DP-3T, Google/Apple	Yes	No
Thailand	Mor Chana	Location, Bluetooth	Yes	No
Tunisia	E7mi	Bluetooth	No	No
Turkey	Hayat Eve Sığar	Bluetooth, Location	No	No
United Arab Emirates	TraceCovid	Bluetooth	No	No
United Kingdom	NHS COVID-19 App	Bluetooth, Google/Apple	Yes	Yes

In the Kingdom of Bahrain, the Information and eGovernment Authority (iGA), in collaboration with the National Taskforce for Combatting the Coronavirus COVID-19 introduced a mobile application called “BeAware”[164]. It provided citizens with all the services they need to help prevent the spread of the virus, including test appointments, results, PCR certificates, announcement, etc. In addition to the contact tracing feature, which uses the users’ geographical location and involves the Google/Apple Exposure Notification (GAEN) system which uses the phone’s Bluetooth signal to check if they were near any infected person and alerts them to get tested once they have been in contact with a positive case typically, within 2 meters of one another for more than 15 minutes. And there were cases where citizens got notified and tested positive, even though they were not aware that they were in contact with other positive cases, so the application helped an early detection of the virus and prevented a wider spread.

It is explicit that the contact tracing feature has been a great assistance to the early detection of the virus, and if the data gathered from the contact tracing applications was utilized, extensive analysis and fruitful information can be carried out. One of the employments is to form the contacts into a network and then divide it into communities, to maximally analyze the behavior of the positive cases and supposedly the virus expansion.

While community detection in social network is originally interested in finding the people with similar taste or preferences, the motif behind this dataset is to understand the properties of communities or closer nodes, and therefor break the chain of spreading the virus among the nodes. There is a requirement to understand the disease spread patterns and its routes among neighboring individuals for the timely implementation of corrective measures at the required placement [162]. To aid the analysis purposes, the dataset needs to be designed with attributes, to find and study the characteristic of each community.

So, another dataset developed and proposed in this section is the contact tracing in the Kingdom of Bahrain. The data used to form the dataset was available on Bahrain's Ministry of Health website [165], and was publicly available, it contained the contact tracing of citizens who were infected by the COVID-19 virus. The cases used in this dataset are the ones covered in the period 01/April/2020 to 10/May/2020, this contains 2972 cases.

The details included:

- Case number
- Age
- Nationality
- Gender
- Travel history: if this person tested positive when arriving from another country
- The other case number contacted which caused the infection.

The procedure of transforming the contact tracing data into a social network dataset that can be used for community detection problem can be visualized in Figure 3-4.

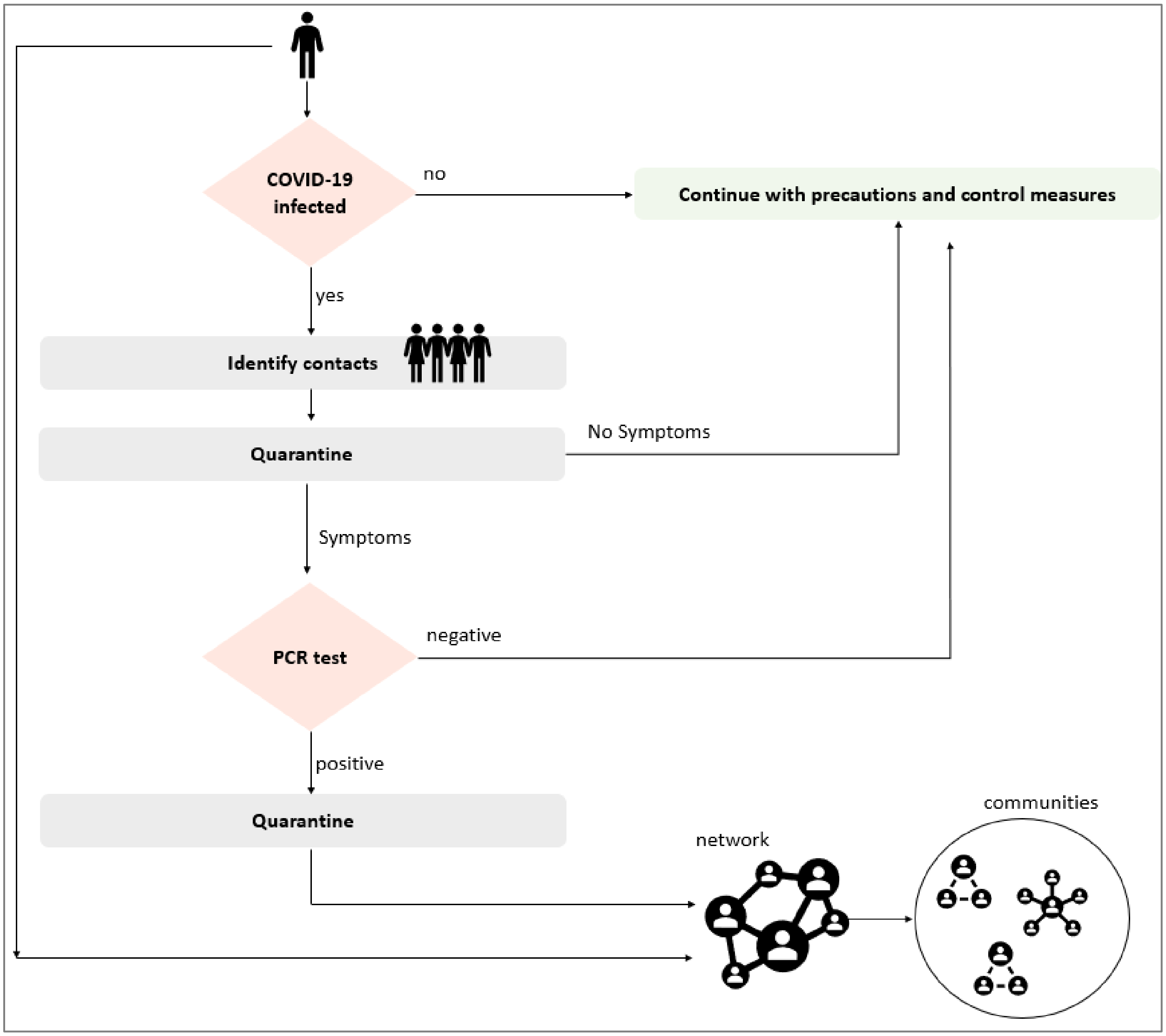
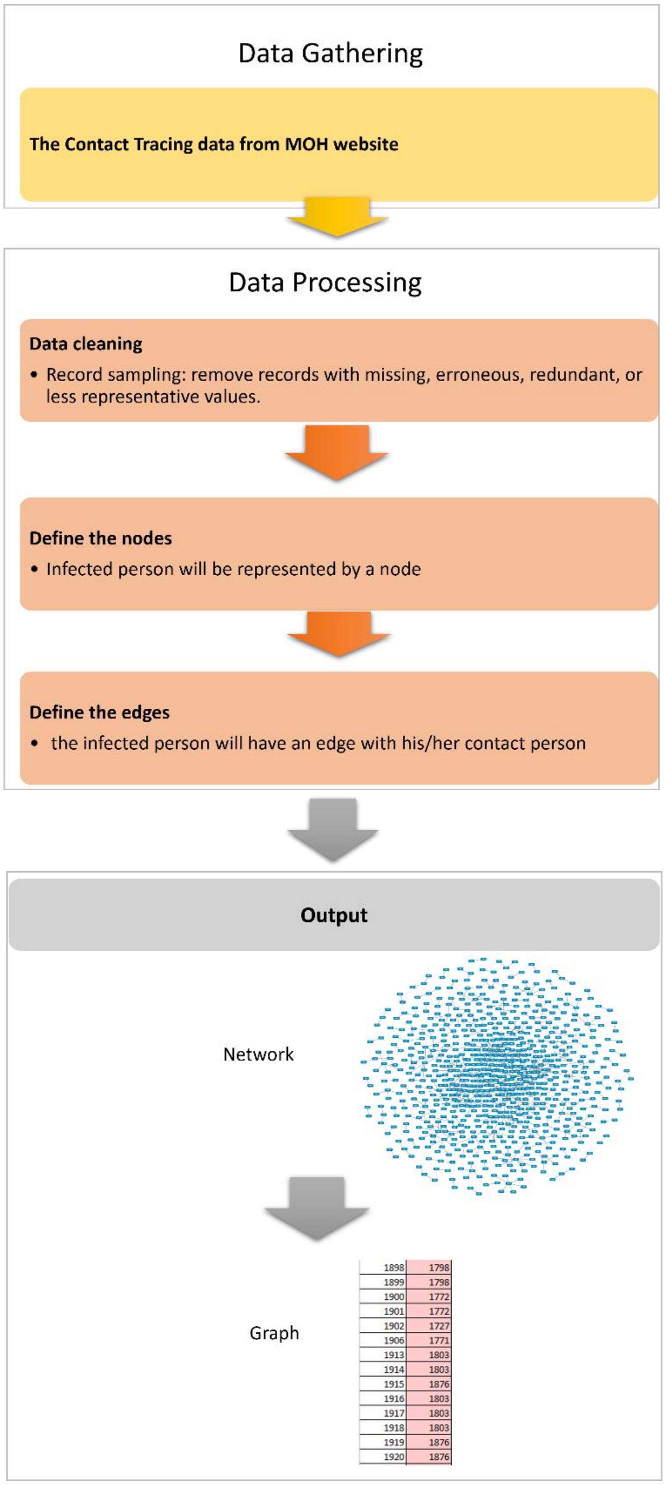


Figure 3-4: Procedure of transforming the contact tracing data into dataset

Some interesting facts were observed from this data, as the number of male infected persons was 2741, while the number of females was 225, the remaining 7 cases' gender were missing, and the average age of the cases was 35 years old. Moreover, 563 of the cases were reported after developing symptoms, but without knowing the source of infection.

In addition, as a part of the comprehensive national response to the global spread of the virus, the Ministry of Health in the Kingdom of Bahrain offered a random Polymerase Chain Reaction (PCR) test to help in detection of COVID-19 at its early stages. These tests were conducted at the drive-through testing facility, or through the mobile testing units which are transportable units (buses, which have been equipped with examination units according to medical standards) organize daily random COVID-19 testing for citizens and residents across the Kingdom, or in any announced random location. According to the data examined in the specified period, 1159 cases were detected through these random tests.

Although the number of cases in the covered period was almost 3k, only 753 cases are considered, which are the ones that were infected by other cases. As the idea is to trace the citizens who got infected by other citizens, and hence form a network of infected persons, then detect the communities and examine the cases. The process of creating this dataset can be illustrated in Figure 3-5.



Data Gathering

The Contact Tracing data from MOH website

Data Processing

Data cleaning

- Record sampling: remove records with missing, erroneous, redundant, or less representative values.

Define the nodes

- Infected person will be represented by a node

Define the edges

- the infected person will have an edge with his/her contact person

Output

1898	1798
1899	1798
1900	1772
1901	1772
1902	1727
1906	1771
1913	1803
1914	1803
1915	1876
1916	1803
1917	1803
1918	1803
1919	1876
1920	1876

Figure 3-5: COVID-19 dataset: Contact Tracing in the Kingdom of Bahrain Process

Since the data was publicly available on the website and it does not contain any personal information which makes it impossible to recognize any of the cases, it did not require any ethical approval. Figure 3-6 presents the network formed by the contact tracing data.

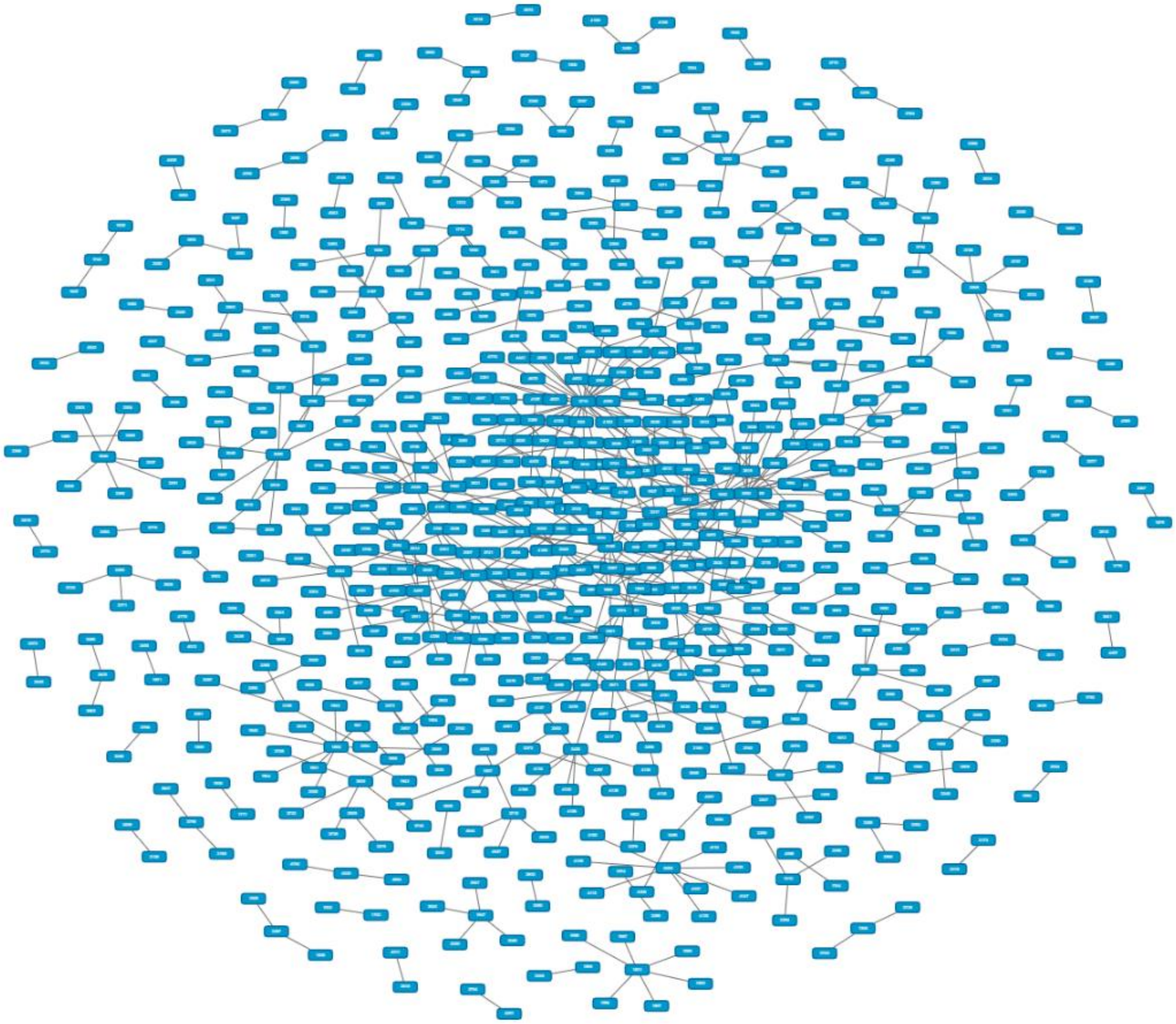


Figure 3-6: Proposed COVID-19 Dataset: Contact Tracing in the Kingdom of Bahrain

The dataset is an attributed network, it consists of 753 cases represented by nodes and 589 relationships between the cases (contacted persons) represented by edges. The attributes are age, nationality, and gender. Travel history was not considered as an attribute in this dataset because the main concern was to link the infected citizens with each other rather than with other countries.

As discussed earlier, other cases were ignored as the source of getting the virus was unknown as they were tested as part of a campaign to obtain random samples from the community or tested positive after developing symptoms without clear idea of the contacted persons.

3.3 Datasets Validity

The main idea behind building these networks is to understand the social structures, their intricacies, and therefore understand the behavior of COVID-19 virus and its transition. To build a social network, two factors are first defined which are actors and relationships, in the first dataset the actors are the countries, and the relationships are the visiting citizens between them, whereas in the second dataset, the actors are the COVID-19 infected persons, and the relationships are their contacts who also caught the virus. By defining these factors, the nodes and edges are built, and the networks are formed. As the direction did not really matter to the networks and weight is undefined, the networks are undirected, and unweighted.

As the datasets used in the community detection problem are basically graphs of networks, in both proposed datasets, the process is based on transforming the tracing data into a network, and thus into a graph consisted of nodes and edges. Which makes the datasets valid and functional to be used in community detection in social networks algorithms. The datasets are created based on real data, and from trustworthy resources, as a result, the network formed based on this information is more likely to be accurate.

In fact, both datasets can also be used in different fields; for example, the proposed network in section 3.2.1 can be used in economics, and the proposed network defined in section 3.2.2 can be used in the medical field as it contains the patients’ information as well.

3.4 Datasets Availability

Both datasets are available on IEEE DataPort. To achieve research reproducibility and allow the collaboration of research data with others and enhance the visibility of the proposed datasets.

DOI: 10.21227/8zxr-vh55

3.5 Summary

According to the literature review, there does not exist any networks datasets that trace the spread of COVID-19, even though the contact tracing mechanization was used in almost all countries around the world since the start of this pandemic. As a result, two contact tracing COVID-19 datasets are built and proposed in this chapter.

And as this type of datasets was not studied, implementation of different algorithms needs to be done. A comparative study of algorithms' performance on the first dataset defined in section 3.2.1 is conducted in Chapter 5.

And since the second dataset defined in section 3.2.2 is an attributed network, it is used to study the homogeneity of communities as presented in Chapter 4.

However, finding and collecting this type of data is challenging, as it was not publicly published, and even though the contact tracing details used for this dataset were on the MOH website [165], they were then removed from the on mid-2020, as it was traced by groups rather than individuals. And as for other countries, only final statistics are now announced, without any details of the cases. Undoubtedly, this data is sensitive and might need an ethical approval, but despite of that, the proposed type of analysis does not require any personal information. This is a limitation that might be shared and considered.

4 Attributed-Based Label Propagation Method for Balanced Modularity and Homogeneity Community Detection

4.1 Overview

Extensive research was done to detect communities within networks, detected communities are densely connected nodes that are strongly connected to each other in or the subnetwork (community) than to the rest of the network[166]. In social networks, a community can be defined as a group of nodes or persons that are similar to each other and dissimilar from the rest of the group [29]. This indicates that the group of nodes in one community will most likely share the same characteristics or interests. Whereas in attributed networks, the nodes in a community will most likely share the same attributes' values.

To assess the output of generated communities, different number of measures are being used, including Modularity measure which indicates the quality of the generated partitions or communities, or NMI value which denotes the accuracy of results compared to the real communities. However, the integration of different types of constrains or external information on community composition was rarely investigated [167], and homogeneity as constraint still remains uncharted. In consequence, the detected communities might contain irrelevant nodes in one cluster even-though the communities scored a good fitness score in other measures such as Modularity and NMI.

To overcome this, a homogeneity measure can be integrated with Modularity, to consolidate the evaluation process. So, a method that maximizes both Modularity and homogeneity is proposed, with Modularity and homogeneity as objective functions. On the other hand, as constrained community detection shows robust performance on noisy data since it uses background knowledge[168] and the restriction of the type considered here has, to our knowledge, remained unstudied, Modularity with homogeneity as a constraint is also tested to adjust the detection of homogenous communities.

To this end, this chapter is organized as follows. Section 4.2 states some preliminaries, and section 4.3 proposes a community detection method, and section 4.4 proposes a homogeneity measure with Penalty regulation. Experiments on social networks from the literature are carried out in section 4.5, in addition to a proposal of a novel dataset of COVID-19 contact tracing dataset in the Kingdom of Bahrain, to help in identifying the infected persons clustered in communities. And then homogeneity is treated as an objective function, and as a constraint in sections 4.5.3 and 4.5.4.

4.2 Preliminaries

The detected communities are evaluated using a number of evaluation measures such as ***Modularity*** [36], which measures the fraction of the edges in the network that connect vertices of the same type (i.e., within-community edges) minus the expected value of the same quantity in a network with the same community divisions but random connections between the vertices. Modularity has been used to compare the quality of the partitions obtained by different methods, but also as an objective function to optimize [169].

Homogeneity was also used as an objective function [112], a measure was proposed based on Shannon information entropy theory in which the entropy of a set, measures the average Shannon information content of the set. Unfortunately, the modularity values produced in this research were significantly lower than others, which negatively affects the results as Modularity quantifies the quality of the detected communities in the network.

Moayedikiaa [82] used the proposed homogeneity in [112] as an objective function by developing an attributed community detection algorithm wrapped by Harmony Search that relies on nodes' importance to form communities. Yet this algorithm performed a long execution time, and it also suffered from entrapment in local optima. Another research proposed a method for community detection based on a higher-order feature termed Attribute homogenous Motif [97], which integrates both node attributes and higher-order structure of the network. However, the modularity was neglected in this research.

On the other hand, some researchers validated their community detection method by looking at the linguistic homogeneity of communities [170], they performed a language analysis on the detected communities. And the homogeneity of a community is characterized by the percentage of those speaking the dominant language in that community, to reveal the monolingual communities.

The two objective functions (Modularity and homogeneity) are conflicting, which means that improving one of them leads to degradation of another [82]. And Modularity has proven its effectiveness in evaluating community detection problem, many algorithms are based on modularity maximization [27],[171]. Hence comes the idea of testing the homogeneity as a constraint, in addition to testing it as an objective function. Constrained algorithms are effective in dealing with combined optimization problems, due to its wide representation scope and generally applicable solving methods[172]. The optimized solution is obtained by solving a Constraint Satisfaction Problem in which the objective function is considered as a constraint that forces it to be equal to a new value [173].

Barber and Clarck [107] also stated that a well-established technique for excluding undesirable results is to adjust the objective function by adding a constraint term that penalized those undesirable results. This was based on what Pillo proclaimed that “the exact penalty methods for the solution of constrained optimization problems are based on the construction of a function whose unconstrained minimizing points are also solution of the constrained problem”[174]. And that the approaches based on exact penalty or exact augmented Lagrangian functions have the features of a built-in preference for minimum points rather than saddle points of the original problem, in addition to the amended convergence advantage in the existence of a barrier term.

In addition, there are several accuracy measures that are used to compare the detected communities to the ground truth of the dataset. The most used measure is ***Normalized Mutual Information*** (NMI) [175] which surveys the quality of a cluster correctness. It is a preferred approach for verifying the correctness of algorithm-identified community structures when a network has a ground-truth community partition for calculating similarities between actual and identified partitions. Similarly, Rand Index [113], focuses on pairwise agreement, for each possible pair of elements in the considered set, it evaluates how similarly the two partitions treat them. The Rand Index was enhanced and proposed later under the name of Adjusted Rand Index or ARI [114]. It is a model that produced arbitrary partitions with the constraint of having fixed number of parts. The ARI is symmetrical, and the ideal value is 1 indicating that partitions are identical, while values less than or equal to 0 indicates a low accuracy. And finally, The variation of information is a dissimilarity measure, it compares two partitions and indicates whether or not they are different from one another [115].

On the other hand, to evaluate the privacy in community detection, Split-Join is used to measure the privacy level after anonymization in the social network [116]. The value of this measure is meant to be as low as possible, to indicate an ideal privacy level.

The evaluation measures used can assess one criterion only, so different measures are used to evaluate different aspects of the result. As one method might generate results that perform well in one evaluation measure while fail to achieve a fair result in another one. Thus, an evaluation technique that takes this issue into account needs to be studied.

In addition, most of the current community detection methods consider the structural information of networks, but disregard the fruitful information of the nodes, and this results in the failure of detecting semantically meaningful communities [97]. However, homogeneity was never studied as a constraint, and was always treated as an objective function. In this research a new measure of homogeneity is proposed and used as an objective function once and as a constraint once again. Therefore, scientific contributions described in this chapter are:

1. Develop improved Label Propagation algorithms (**Attribute-Based Label Propagation)** that considers the nodes' attributes to achieve a fair homogeneity value, while maintaining high Modularity and Accuracy measures (discussed in 2.2.2: NMI, ARI, RI, VI, and Split- Join distance).
2. Formulate and propose an adaptive Penalized Homogeneity measure, with penalty and weight modulation, that can be utilized in consonance with the user's requirements.
3. Based on the literature review, a research gap of employing homogeneity as a constraint in the Community Detection problem is identified, and accordingly, homogeneity as a constraint in Modularity based methods is investigated.

4.3 Proposed Methods

In this section, a new algorithm called **Attribute-Based Label Propagation (ABLP)** based on attributes' regulation is proposed. Results have been tested with homogeneity and Modularity as objective functions and as Modularity constrained with homogeneity. The proposed method is an Attribute-Based Label Propagation Algorithm which is a Modularity maximization based on Label Propagation algorithm with regards to homogeneity. The main concept is in each run, the Modularity measure is calculated, and the maximum value is considered, and at the same time, examining the nodes' attributes to assign similar nodes in a single community. Both methods do not require the number of communities to be set in advance.

Label Propagation Algorithm (LPA) is considered as one the effective algorithms amongst the existing algorithms used for community detection because of its time efficiency [102]. It was first introduced in 2007, by Nandini Raghavan, Réka Albert, and Soundar Kumara [29], it uses the network structure alone as its guide and requires neither optimization of a predefined objective function nor prior information about the communities. The simplicity and near linear complexity of the LPA makes it feasible to detect communities in complex large networks. The main notion is that each node makes its own decision regarding the community to which it belongs based on the communities of its immediate neighbors [29].

However, there are some explicit drawbacks in this algorithm that affect its performance. The randomness that is induced in its update sequences and tie breaking processes cause the LPA to return multiple detections, thus making it a non-deterministic detection algorithm.

So, the concept of ABLP is enhance the Label Propagation Algorithm by considering the nodes' attributes to embrace the randomness of the LPA. And since this algorithm is probabilistic, it generates a different result in every run, the run that produces the maximum Modularity and the proposed measure of Penalized Homogeneity degree is considered.

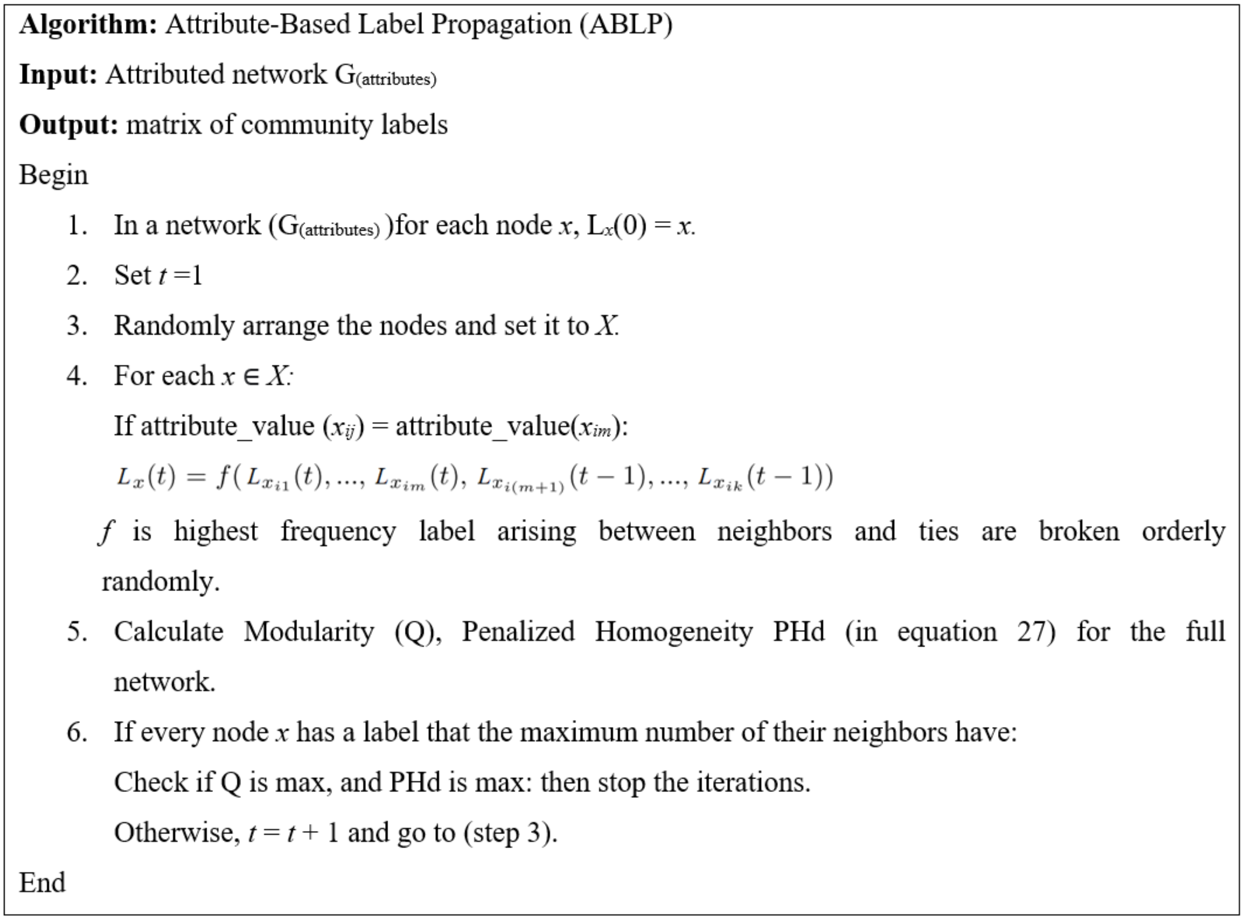
The proposed ABLP algorithm is meant to be executed in an attributed network $G_{(attributes)}$, that contains nodes (denoted by *x*) and edges (denoted by e). All nodes are initialized with a unique label ($L_x(t)$) which is the label of node *x* in time *t*, and then these labels are propagated through the network. As the propagation is iterated, in each iteration the nodes amend their labels based on their neighbors, in which nodes will choose the label that the maximum of their neighbors that share the same attribute value belongs to. And with ties broken orderly yet haphazardly, the labels of nodes could be modified in each iteration even if the labels of their neighbors do not change. Typically, the iterations are performed until no node modifies its label. Groups of densely connected nodes are be formed based on these labels and start to broaden through the network, and finally nodes that have the same labels will form a community.

In an attributed network with *n* nodes and *m* edges: $G_{(attributes)}(n, m)$. Let L_1 ,L_2 ,.. ,L_p be the detected communities. In most algorithms, the communities found satisfy the following constraints: $L_i \cap L_j = \emptyset$ for $i \neq j$. In the network, each node is denoted by *x*, and $L_x(t)$ is the node *x* at time *t*. Asynchronous updating is used where $L_x(t) = f(L_{x_{i1}}(t), ..., L_{x_{im}}(t), L_{x_{i(m+1)}}(t-1), ..., L_{x_{ik}}(t-1))$.

And x_{i1},.. ,$\underline{x}_{im}$ are neighbors of *x* that have already been updated in the current iteration while $x_{i(m+1)}$,.. ,$\underline{x}_{ik}$ are neighbors that are not yet updated in the current iteration. The order in which all the *n* nodes in the network are updated at each iteration is chosen randomly. Although there are *n* different labels at the beginning of the algorithm, the number of labels reduces over iterations, resulting in only as many unique labels as there are communities. The iterative process is performed until every node in the network has a label to which the maximum number of its neighbors belongs.

4.3.1 ABLP Algorithm

The proposed algorithm can be described as follow:



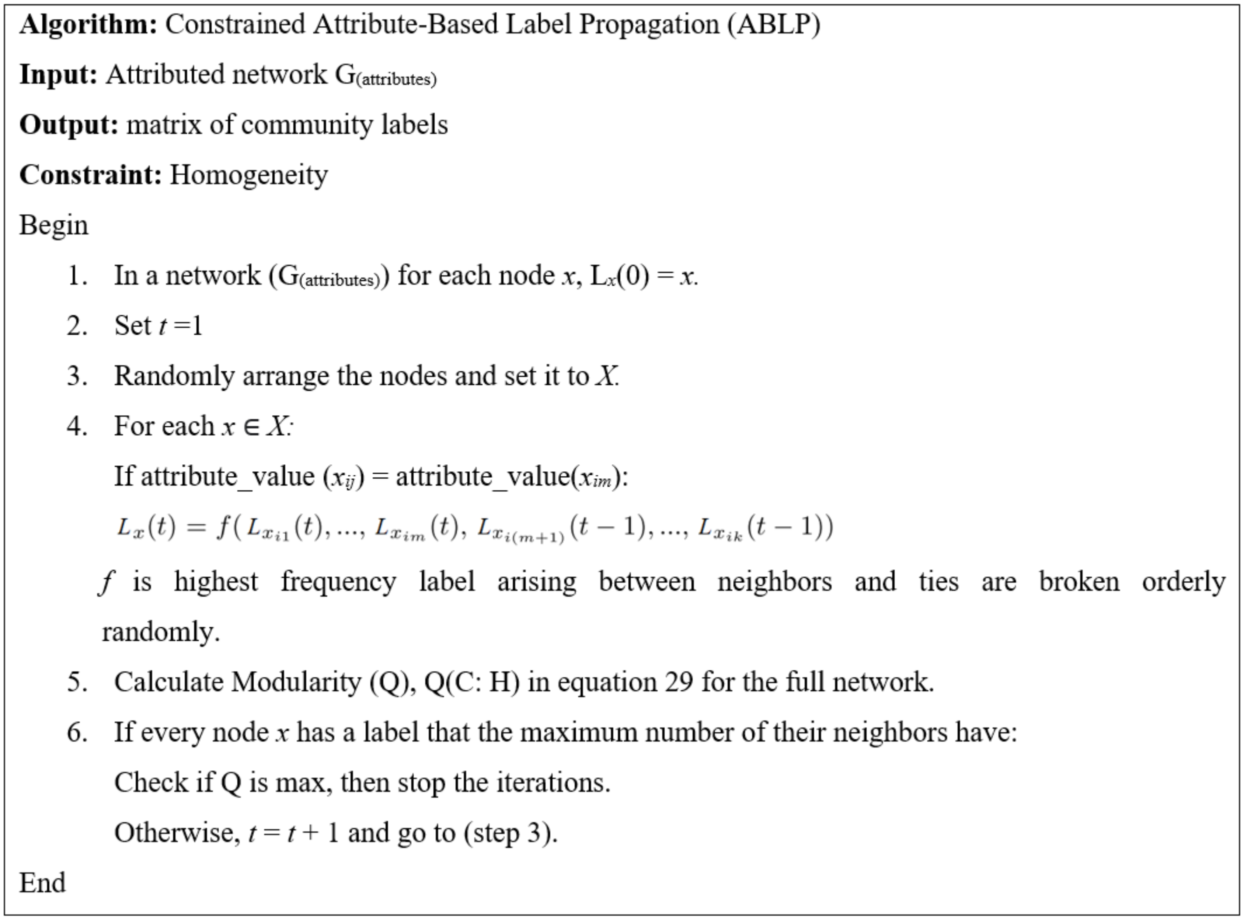
```
**Algorithm:** Attribute-Based Label Propagation (ABLP)
**Input:** Attributed network $G_{(attributes)}$
**Output:** matrix of community labels
Begin
    1. In a network ($G_{(attributes)}$ )for each node $x$, $L_x(0) = x$.
    2. Set $t$ =1
    3. Randomly arrange the nodes and set it to $X$.
    4. For each $x \in X$:
       If attribute_value ($x_{ij}$) = attribute_value($x_{im}$):
       $L_x(t) = f(L_{x_{i1}}(t), ..., L_{x_{im}}(t), L_{x_{i(m+1)}}(t-1), ..., L_{x_{ik}}(t-1))$
      $f$ is highest frequency label arising between neighbors and ties are broken orderly randomly.
    5. Calculate Modularity (Q), Penalized Homogeneity PHd (in equation 27) for the full network.
    6. If every node $x$ has a label that the maximum number of their neighbors have:
       Check if Q is max, and PHd is max: then stop the iterations.
       Otherwise, $t = t + 1$ and go to (step 3).
End
```

Figure 4-1: ABLP Algorithm

In this way, the proposed ABLP algorithm assigns similar labels to the node and its neighbors while checking the attribute value. Hence, the algorithm tends to maximize Modularity and homogeneity at the same time, which will also result in maintaining a high accuracy value (NMI, RI, ARI, VI). As the Ground Truth table for attributed networks mostly divides the nodes according to their attributes' values.

4.3.2 Constrained ABLP Algorithm

As observed in the literature review, a gap of utilizing homogeneity as a constraint is found, and accordingly, the proposed method has been modified to include this constraint. The same concept of the proposed ABLP is followed, with regards of homogeneity as a constraint, which penalizes the Modularity measure by minimizing it based on the achievement of the homogeneity value. So ideally, if the homogeneity degree is high, the modularity measure should remain at its best. However, if the homogeneity degree is low, the Modularity value should be punished and reduced.



```
Algorithm: Constrained Attribute-Based Label Propagation (ABLP)
Input: Attributed network $G_{(attributes)}$
Output: matrix of community labels
Constraint: Homogeneity
Begin
    1. In a network ($G_{(attributes)}$) for each node $x$, $L_x(0) = x$.
    2. Set $t$ =1
    3. Randomly arrange the nodes and set it to $X$.
    4. For each $x \in X$:
       If attribute_value ($x_{ij}$) = attribute_value($x_{im}$):
       $L_x(t) = f(L_{x_{i1}}(t), \ldots, L_{x_{im}}(t), L_{x_{i(m+1)}}(t-1), \ldots, L_{x_{ik}}(t-1))$
      $f$ is highest frequency label arising between neighbors and ties are broken orderly randomly.
    5. Calculate Modularity (Q), Q(C: H) in equation 29 for the full network.
    6. If every node $x$ has a label that the maximum number of their neighbors have:
       Check if Q is max, then stop the iterations.
       Otherwise, $t = t + 1$ and go to (step 3).
End
```

Figure 4-2: Constrained ABLP

The Constrained Attribute-Based Label Propagation algorithm is a highest-modularity, homogeneity constraint-satisfying solution for the community detection problem in attributed networks. The algorithm considers the run that generates the maximum constrained Modularity and proposed measure of Penalized Homogeneity degrees is considered.

4.3.3 Time Complexity

The proposed methods do not require prior knowledge of the number of communities to be detected. The time complexity of these methods depends on the number of nodes in the network, as it determines the time for iteratively updating the nodes' labels. Therefore, the complexity is linear O(*n*), where *n* is the number of nodes in the network.

4.4 Proposed Evaluation Measure

4.4.1 Homogeneity Degree

Homogeneity in community detection was first proposed by [112], it was defined based on Shannon information entropy theory, the entropy of a set measures the average Shannon information content of it. It is an entropy-based criterion; a homogeneous set of elements has a low entropy. This homogeneity measure is weighted; it is based on the importance of different attributes in the network. This measure assumes that each community contains at least node of each attribute value, and that each attribute is assigned to a weight. This homogeneity measure considers the proportion of the number of nodes with a certain attribute in a community to the total number of nodes in a community. So, if an algorithm detects a number of communities, with similar attribute nodes in each community, it will score a high homogeneity measure, regardless of the efficiency of the results. In other words, if an algorithm detects 10 different communities, and all have the same number of nodes with same attribute, it will result in a high homogeneity score. Which means that it does not consider the number of communities detected, in which if a network consists of 5 nodes and the algorithm detected 5 communities with 1 node in each, it will still give a good homogeneity value.

On the other hand, this measure does not consider the network structure, as real-world datasets might have some aspects that need to be considered. As discussed later, a real network is proposed, and the ratio between the number of nodes in each community to the total number of nodes in the network should not matter, therefore it should not be considered in the calculations.

As homogeneity was used as an objective function to measure the homogeneity of the detected communities in the network as one unit, here is the proposal of a new of homogeneity measure that evaluates the homogeneity degree in each community, based on specified attribute values.

The formula calculates the number of nodes with the specified attribute divided by the total number of nodes in the cluster. It reflects the standard deviation; however, standard deviation finds how concentrated the data is around the mean, in our case, the mean is be ignored, μ=0;

The closer the value is to 1, the more homogeneous the cluster is. This can be calculated in Hd

$$\text{Hd} = \sum_{i=1}^{k} \left(\frac{n_{att}}{n}\right)^2 \qquad \textbf{(25)}$$

Where where *k* is the number of communities, *att* is the number of attributes in the network, n_{att} is the number of nodes with each attribute in a community, and *n* is the total number of nodes in the community. The square value is calculated as it adds more weighting to the differences which makes the value more significant.

4.4.2 Penalized Homogeneity Degree

It should be noted that the Homogeneity degree (Hd) measure proposed in section 5.1 does not consider the number of communities and the number of nodes in each community compared to the total number of nodes in the network. To add more flexibility and user-preference to the proposed measure, a penalty is given, to ensure that nodes among all detected communities are homogeneous, and that distribution is fair.

To add more restrictions to the homogeneity degree, (P) is considered, which is a penalty that takes the number of nodes for each attribute in the community compared to the total number of nodes with this attribute in the network.

$$\mathrm{P} = 1 - \left(\frac{n_{att(\max)}}{\mathrm{N}_{att(\max)}}\right) \qquad \textbf{(26)}$$

Where n_{att} is the number of nodes with each attribute in a community, and N_{att} is the number of nodes with this attribute in the network, for the attribute that owns the maximum number of nodes in each community.

PHd= Hd - P **(27)**

Where PHd measures the Penalized Homogeneity degree. This allows the user to apply an impartial penalty for algorithms that detect a large number of communities that contain a small number of nodes with a certain attribute.

In some cases, the network might contain multiple attributes and more than one attribute needs to be considered in the community detection process. To calculate homogeneity in such communities, it is possible to set a weight for the attribute, based on the user's preference and how important each attribute is. Multi-Attribute Weighted Penalized Homogeneity degree MAWPHd is an optional measure that can be calculated to assign weight for different attributes in the network.

MAWPHd= $\sum_{\boldsymbol{i=1}}^{\boldsymbol{z}} \boldsymbol{w} * \mathbf{PHd}$ **(28)**

Where *z* is the number of attributes to be considered, and *w* is the weight assigned to each attribute. So, if more than one attributes were considered, a weight might be set to each attribute, which might be equal for all attributes of varies based on the attribute's importance.

On the other hand, to calculate Modularity constrained by Homogeneity, the PHd is subtracted from 1 to minimize the penalty of constraint. Because the higher the homogeneity value, the less punishment is applied on the Modularity.

Q(C: H) = |Q-1- PHd | **(29)**

Where Q(C: PHd) calculates the Modularity with Penalized Homogeneity as Constraint, *Q* represents Modularity, H is the Homogeneity (can be Hd or PHd, based on the experiment, dataset or research requirements).

The proposed measures of Penalized Homogeneity degree (PHd) and the Multi-Attribute Weighted Penalized Homogeneity degree (MAWPHd) allows a more flexible mensuration of homogeneity on different types of attributed networks based on the user-defined requirements.

4.5 Experiments

The proposed algorithm has been implemented on the following datasets: Political Books [110] and American Football [148], in addition to a proposed dataset of COVID-19 contact tracing. The results have been compared with a number of existing algorithms. The results of implementing the proposed algorithms have been compared in term of Modularity, NMI, and the proposed measures of homogeneity. Furthermore, homogeneity has been also considered as a constraint for Modularity, and applied on the datasets, all results are discussed and analyzed.

The experiments are conducted on one computer server which is equipped with core i7 CPU, and 32G memory. Other software environments include python 3.6, pycharm 2019.3.1, and RStudio 1.4.1717. In order to overcome the effect of randomness, algorithms are implemented and run 100 times, and the average results are considered.

4.5.1 Datasets used in the experiments

The datasets used for the experiments are attributed social networks from the literature, in addition to a proposed real-world dataset based on the contact tracing of COVID-19 infected persons in the Kingdom of Bahrain.

Political Books (PolBooks) network social network consists of nodes representing books about US politics. Edges represent frequent co-purchasing of books by the same buyers. Books were labeled by Newman [110] with an attribute describing their political alignments, i.e., liberal, neutral, and conservative. This undirected network consists of 105 nodes, and 441 edges. In this dataset, 49 of the books are conservative, 43 are liberal and the remaining 13 are neutral.

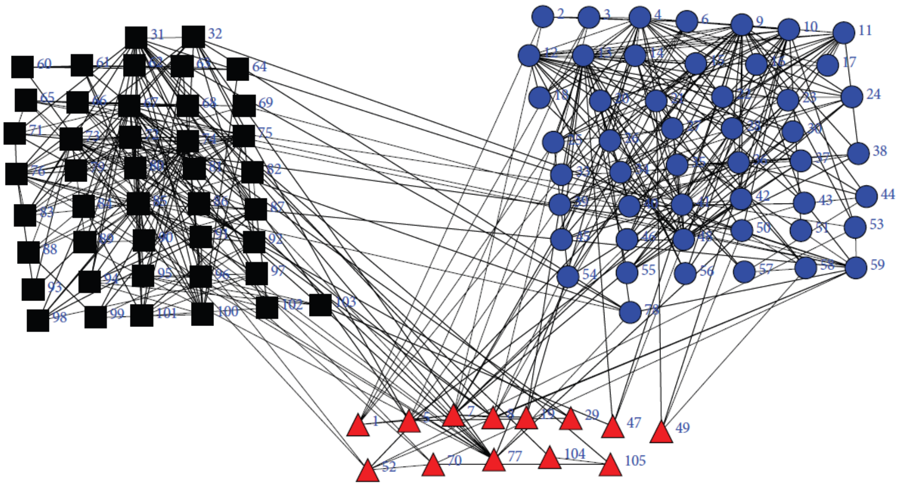


Figure 4-3: The ground truth partition of political books dataset (Triangles: neutral books, dots: conservative books, and squares: liberal) [152]

American College Football network, which is a network of American football games between Division IA colleges during regular season fall in 2000 [148]. It consists of 115 teams represented by nodes and 613 games represented by edges, divided into 12 conferences (attributes).

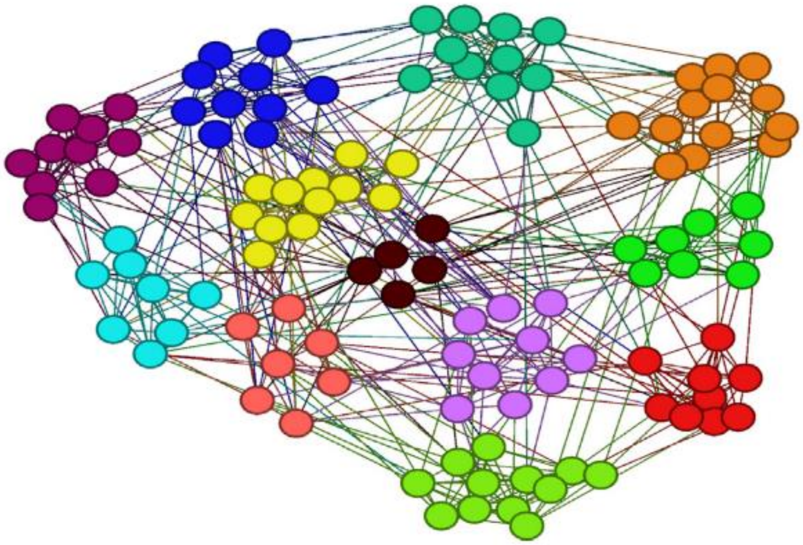


Figure 4-4: The ground truth partition of American College Football Network (each team is represented in a different color) [78]

Proposed dataset: COVID-19 contact tracing

The data used to form the dataset was available on Bahrain's Ministry of Health website[165], and was publicly available, it contained the contact tracing of citizens who were infected by the COVID-19 virus, the details include the case number, age, nationality, gender, travel history (if any), and the other case number contacted which caused the infection. The contact tracing details was then removed from the website on mid-2020, as it was traced by groups rather than individuals. Since the data was publicly available on the website and it does not contain any personal information, it did not require any ethical approval.

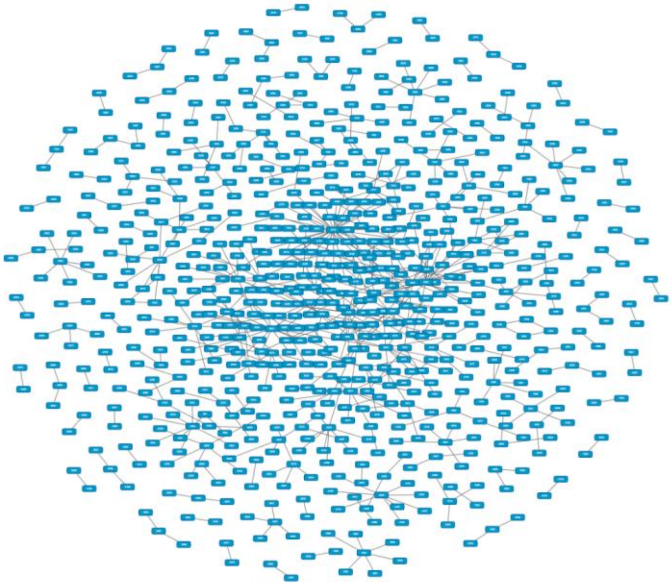


Figure 4-5: Proposed COVID-19 dataset: contact tracing in the Kingdom of Bahrain

This attributed dataset consists of 753 cases represented by nodes and 589 relationships between the cases (contacted persons) represented by edges. Each case has three attributes: The attributes are age, nationality, and gender. In this experiment, the nationality attribute was considered to study the ethnicity of communities, as citizens with the same nationality tend to live together and interact with one another. Considering this attribute is to help us understand the interconnection between different nationalities in the Kingdom of Bahrain.

Other cases were ignored as the source of getting the virus was unknown as they were tested as part of a campaign to obtain random samples from the community or tested positive after developing symptoms without clear idea of the contacted persons.

As this is a real-world network constructed from contact tracing data, there is no ground-truth table for this dataset.

4.5.2 Benchmarking Algorithms

The ABLP algorithm proposed in this chapter, along with several existing algorithms have been implemented. The algorithms used for the comparison are:

Asynchronous Label Propagation LPA [29]: this algorithm initializes each node with a unique label, it repeatedly sets the label of a node to be the label that appears most frequently among that nodes' neighbors. The algorithm halts when each node has the label that appears most frequently among its neighbors. The algorithm is asynchronous because each node is updated without waiting for updates on the remaining nodes.

Graph Embedding with Self Clustering (GEMSEC) [176]: it places nodes in an abstract feature space where the vertex features minimize the negative log-likelihood of preserving sampled vertex neighborhoods, and it incorporates known social network properties through a machine learning regularization.

An Edge Enhancement Approach for Motif-aware (EdMot) [177] this method partitions the top K largest connected components in the hypergraph into modules. And then, uses an edge enhancement approach for enhancing the connectivity structure of the original network. Which is achieved by constructing a new edge set to derive a clique from each module. Based on the new edge set, the original connectivity structure of the input network is enhanced to generate a rewired network, whereby the motif-based higher-order structure is leveraged, and the hypergraph fragmentation issue is well addressed. After the edge enhancement, the rewired network is partitioned to obtain the higher-order community structure.

Deep Autoencoder-like Nonnegative Matrix Factorization (DANMF) [178] Similar to deep autoencoder, DANMF consists of an encoder component and a decoder component. This architecture empowers DANMF to learn the hierarchical mappings between the original network and the final community assignment with implicit low-to-high level hidden attributes of the original network learnt in the intermediate layers.

ABLP (Proposed) The main purpose of proposing the algorithm is to maximize homogeneity, while maintaining high Modularity, accuracy values (NMI, RI, ARI, VI) and privacy Split-Join measure. So, the Homogeneity degree is be calculated and compared as an objective function. And then the results are tested again with consideration of homogeneity as a constraint.

4.5.3 Homogeneity as an objective function in ABLP

The proposed algorithm ABLP, along with the chosen benchmarking algorithms are executed on the three discussed datasets : Political Books (Polbooks) and the considered attribute is book's type (which can either be conservative, neutral or liberal), American College Football and the considered attribute is conference (there are 12 conferences), and proposed COVID-19 Contact Tracing in the Kingdom of Bahrain with nationality as the considered attribute (which can be Bahraini , Indian, Bangladeshi, Pakistani, Nepali, Ugandan, Filipino, Yemeni, Kenyan, Egyptian, Ghanian, Ethiopian, or Indonesian).

Communities obtained in Polbooks datasets are visualized in (Figure 4-6 to Figure 4-10), different colors identify different communities.

The proposed ABLP algorithm detected 4 communities, 2 of them appear to be the main communities that contain a large number of nodes, whereas the other 2 contain less numbers of nodes (As illustrated in Figure 4-6). ABLP algorithm detected a community where 87% of its nodes were conservative, another community was 86% liberal, while the remaining two were 50% conservative, and 37% neutral. ABLP is the only method that detected a homogeneous neutral community; in which there exists a community where the majority of its nodes type is neutral, as other algorithms distributed the neutral nodes among all other communities.

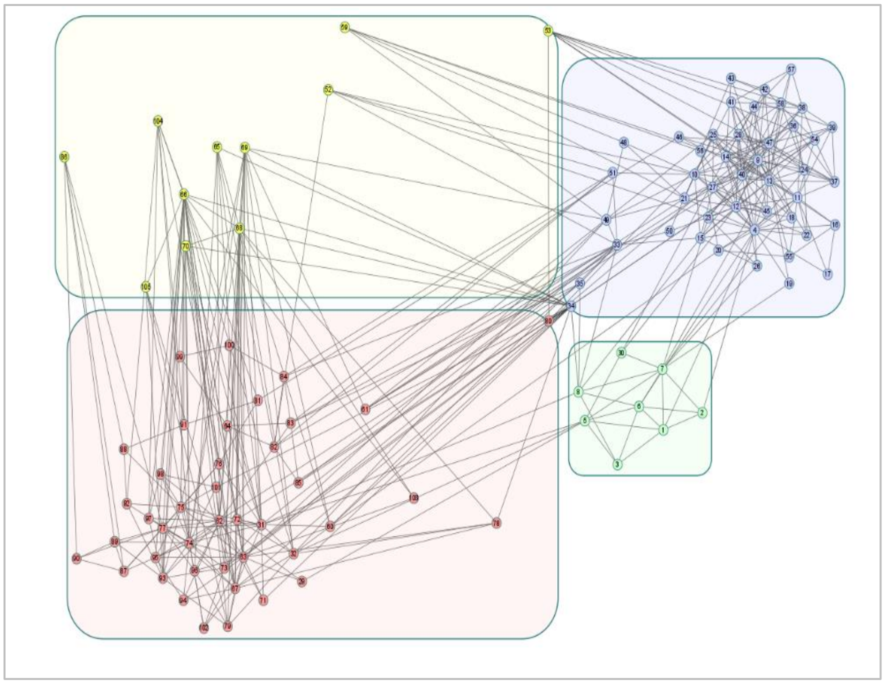


Figure 4-6: Detected 4 Communities in Polbooks dataset by the proposed ABLP algorithm

LPA (Figure 4-7) and Edmot (Figure 4-8) algorithms, both generated 7 communities. While the number of nodes in each community is different, LPA detected 4 communities with a small number of nodes (4 or less nodes), and two main communities that contained the majority of nodes. Edmot appear to detect communities with fairly similar number of nodes, with one community that contains 3 nodes only, which is a small number compared to the total numbers in the network (105).

Communities detected by LPA can be classified as follow: one was 77% conservative, 65% liberal, 62% conservative, 36% conservative and 37% liberal. On the other hand, Edmot detected a community that contained two conservative communities with percentage of 87% and 67%, and 4 liberal communities with percentage 75%, 57%, 54%, and 45%, while one community contained an equal number of all types which was 33.3% homogeneous for all types.

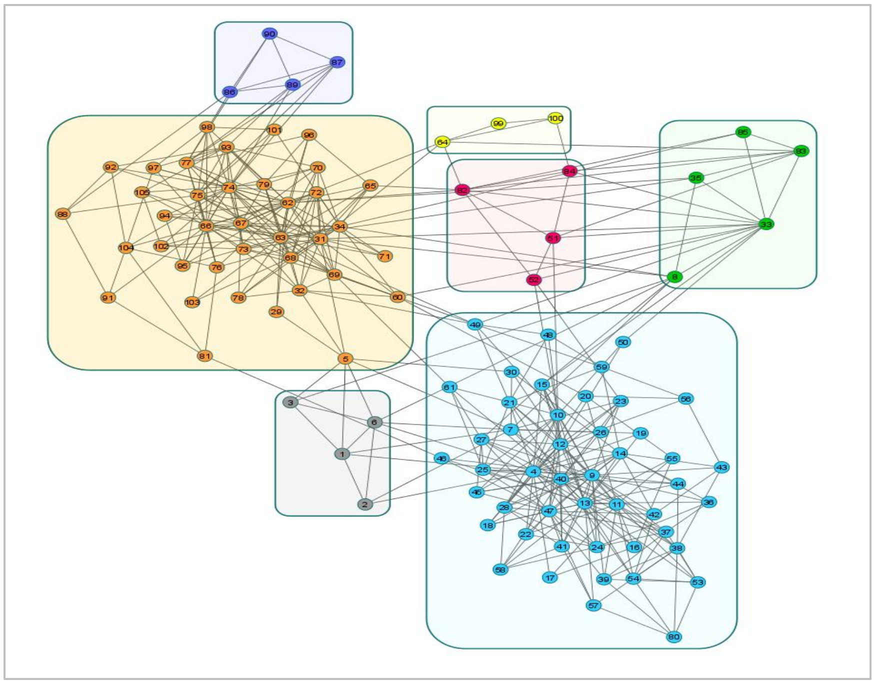


Figure 4-7: Detected 7 Communities in Polbooks dataset by LPA

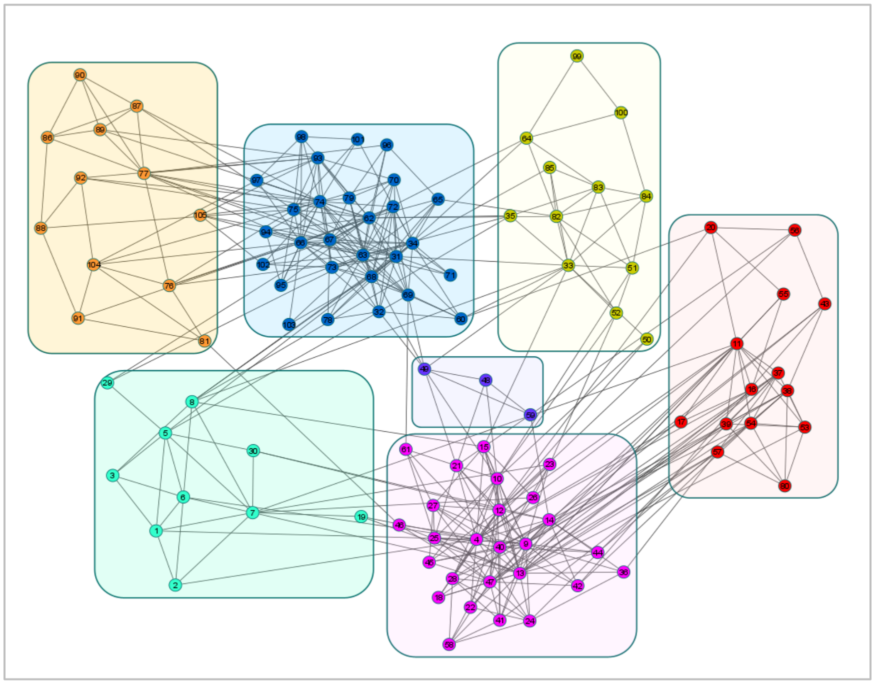


Figure 4-8: Detected 7 Communities in Polbooks dataset by Edmot algorithm

GEMSEC algorithm (as illustrated in Figure 4-9) generated 9 communities with a mixed number of nodes in each community, including one dominant community that contained the majority of nodes, and this distribution affects the Modularity and homogeneity results. This community contained 42% conservative books, while the small communities were 72%, 55% and 37% conservative, 68%, 54% and 55% liberal. The remaining two communities were 100% homogeneous as they contained 2 liberal nodes and 5 conservative nodes respectively.

Finally, DANMF detected 8 communities with a commensurate number of nodes in each community (Figure 4-10). The detected communities were conservative with a percentage of 70%, 62% and 54%, whereas the liberal communities were 74%, 68% 55%, 54%, 46% homogeneous. On the other hand, neutral books were distributed over the communities as there wasn't any homogeneous neutral communities.

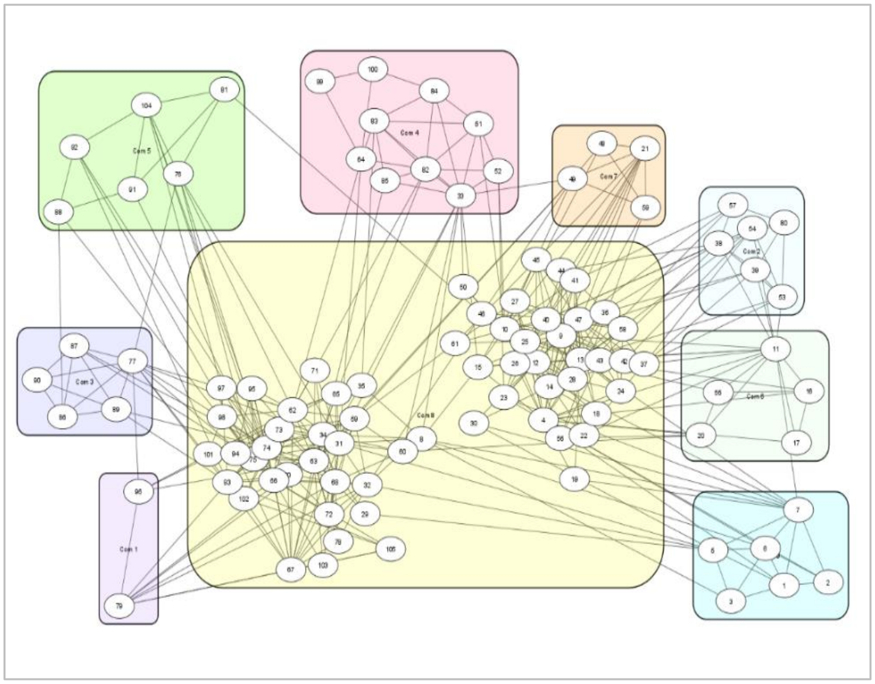


Figure 4-9: Detected 9 Communities in Polbooks dataset by GEMSEC algorithm

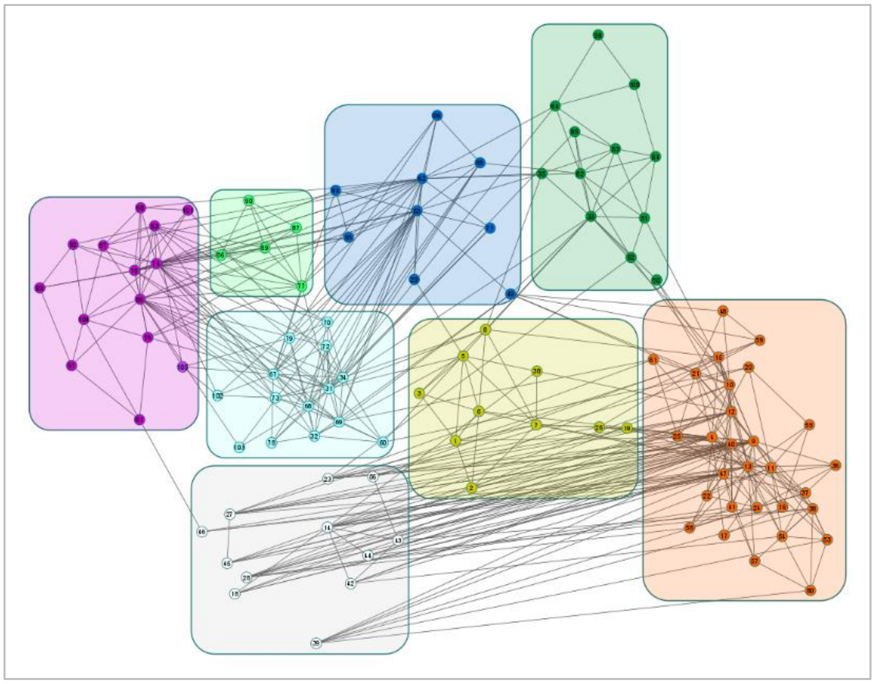


Figure 4-10: Detected 8 Communities in Polbooks dataset by DANMF algorithm

The results of Modularity (Q), and proposed measures on the three datasets are shown in
(Table 4-1 to

Table 4-4). Where Hd states the proposed Homogeneity degree measure in detected communities (Equation 25), P is the proposed penalty measure (Equation 26) and PHd is the proposed Penalized Homogeneity degree values (Equation 27).

It is observed that considering the nodes' attributes values result in more homogeneous communities. Nodes with similar attributes are beyond any doubt share the same value, however, they may not necessarily be neighbors or share direct ties. So, paying more attention to the node's values helps detect denser communities in terms of interests or preferences.

Table 4-1: Results on Books Dataset, where Q is Modularity, Hd is proposed Homogeneity degree, P is Penalty and PHd is the proposed Penalized Homogeneity degree

Algorithm	**Number of detected communities**	**Q**	**Proposed Hd**	**Proposed P**	**Proposed PHd**
Proposed ABLP (2021)	**4**	**0.527**	0.652	**0.515**	**0.137**
LPA [29] (2007)	7	0.481	**0.683**	0.736	-0.053
EdMot [177] (2019)	7	0.509	0.649	0.748	-0.099
GEMSEC [176] (2019)	9	0.336	0.650	0.852	-0.202
DANMF [178] (2018)	8	0.492	0.604	0.784	-0.180

The values obtained from the above-discussed evaluation measures achieved by different com- munity detection algorithms for Polbooks dataset are visualized in Figure 4-11 to Figure 4-13. Figure 4-11 shows the number of communities detected in each method, in comparison with the true number of communities in the network, which is 3.

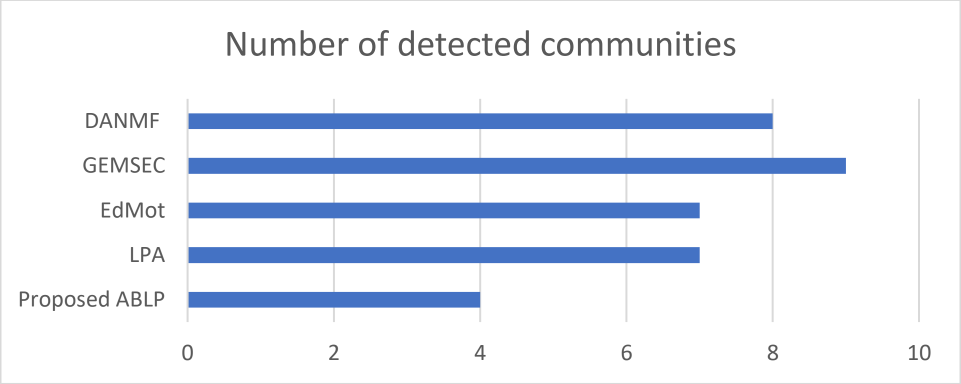


Figure 4-11: Number of detected communities in Polbooks dataset.
Where the number of communities in ground truth table= 3

The proposed ABLP has scored the closest number of detected communities to the ground truth communities. The proposed Homogeneity degree and Penalty of Homogeneity degree are visualized in Figure 4-12. High Homogeneity degree values indicate homogeneous results; however, high penalty measures will affect these results. In other words, Homogeneity is best achieved when the Homogeneity degree is high, the penalty measure should remain as low as possible, to avoid affected the total proposed measure of Penalized Homogeneity degree.

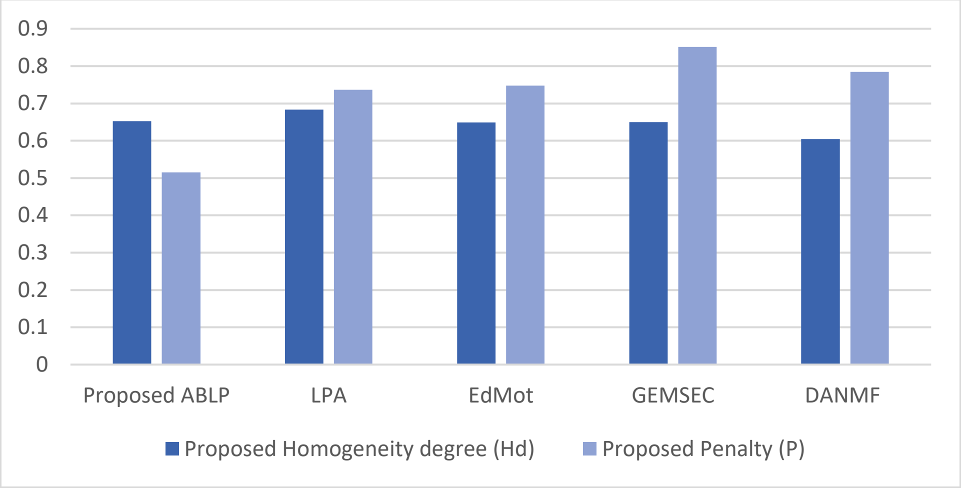


Figure 4-12: Proposed Homogeneity degree and proposed Penalty on Polbooks dataset. Where the maximum values of Homogeneity and minimum values of penalty indicate best homogeneous results.

The final values of Penalized Homogeneity degrees and Modularity measure are illustrated in Figure 4-13. Highest values of both measures are achieved by the proposed measure.

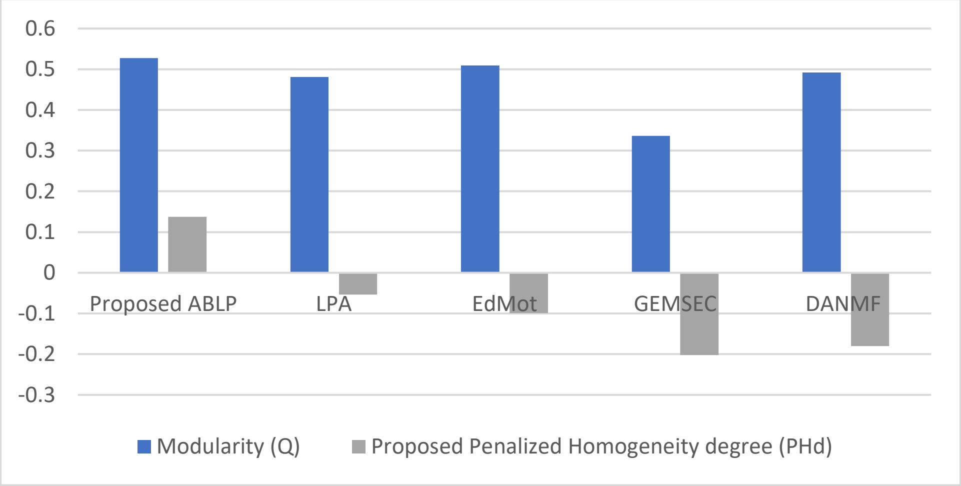


Figure 4-13:Proposed Penalized Homogeneity degree and Modularity on Polbooks dataset, where highest values of both measures denotes the best results in Modularity and Penalized Homogeneity.

It can be stated that the highest modularity value is achieved in Books dataset by the proposed ABLP (Attribute-Based Label Propagation) algorithm with a value of 0.527, followed by EdMot. Whereas the lowest value is given by GEMSEC with the value of 0.3362.

As for the Homogeneity degree (before applying the penalty), LPA achieved a high rate, however its penalty is high because it detected two small communities with node sizes 4 and 3, and all nodes in both communities had the same attribute value. This resulted in a very low Penalized Homogeneity degree. GEMSEC also had an elevated penalty value for the same reason. This gives rise to the proposed ABLP algorithm achieving the highest assessment value among all other algorithms.

Furthermore, the experiments carried out on the American College Football are visualized in (Figure 4-14 to Figure 4-18). The proposed algorithm as illustrated in Figure 4-14, has generated 9 communities, with a convergent number of nodes in each community.

The following percentages represent how homogeneous the detected communities by ABLP are; in which the nodes in one community share the same conference (attribute) value: 82%, 80%, 78%, 75%, 68%, 63%, 53%, 42% and 41%. The communities contained nodes from no more than four different conferences, which resulted in higher homogeneity value.

While LPA (Figure 4-15) seem to have the same number of communities, the distribution of nodes is slightly distinct. Four of the detected communities by LPA were more than 60% homogeneous (the nodes belong to the same conference), two communities were 58% and 59%, while the remaining three were 42%, 39% and 32% homogeneous.

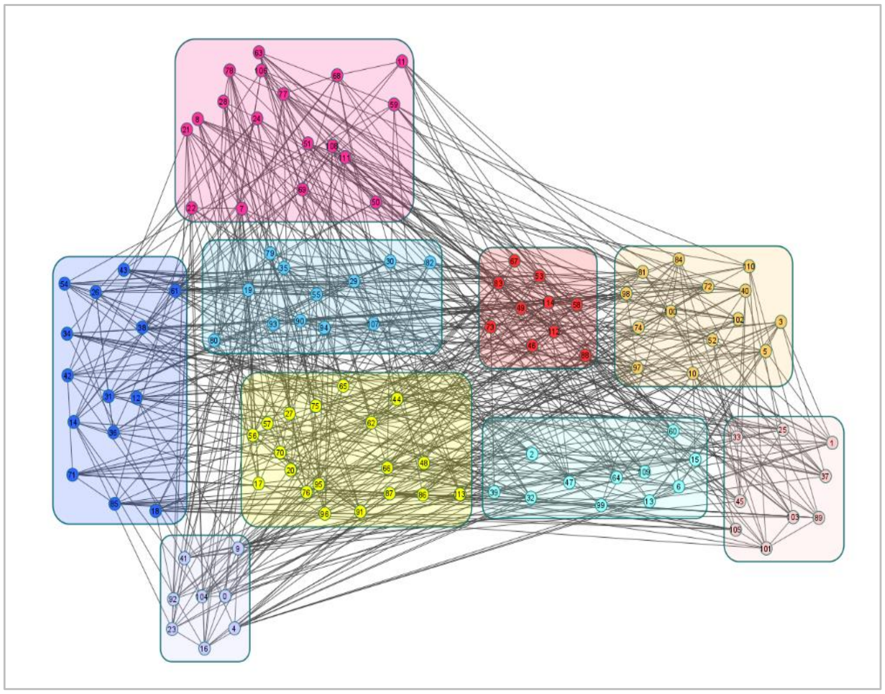


Figure 4-14: Detected 9 Communities in Football dataset by the proposed ABLP algorithm

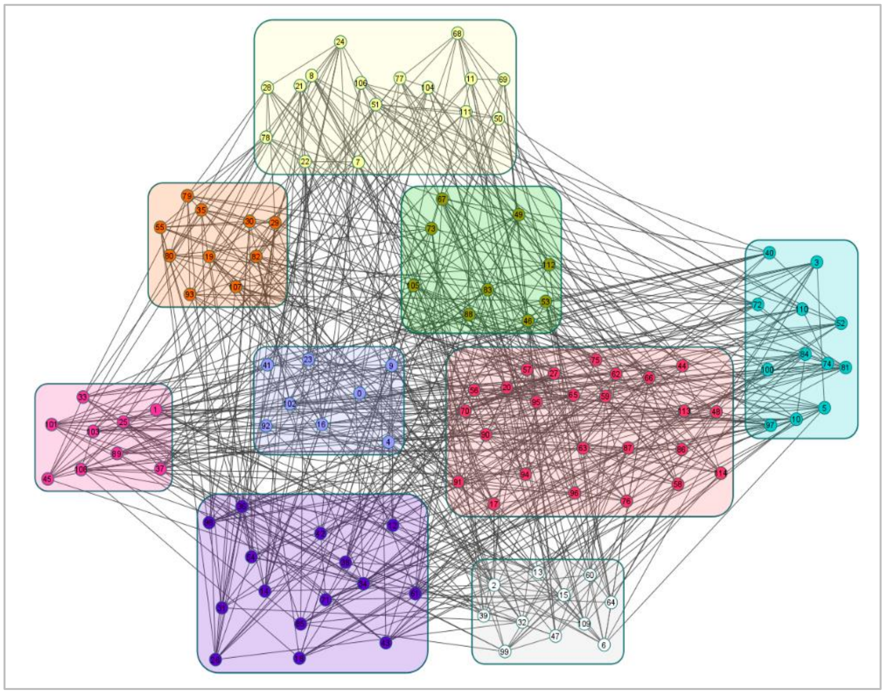


Figure 4-15: Detected 9 Communities in Football dataset by LPA

On the other hand, Edmot (Figure 4-16) has generated 10 communities, with similar number of nodes but quite different distribution of nodes than the proposed ABLP and LPA.

The communities detected by Edmot were not highly homogeneous in terms of the conference value; all communities were between 12% to 25% homogeneous. Which means that the communities contained nodes from more than 5 conferences.

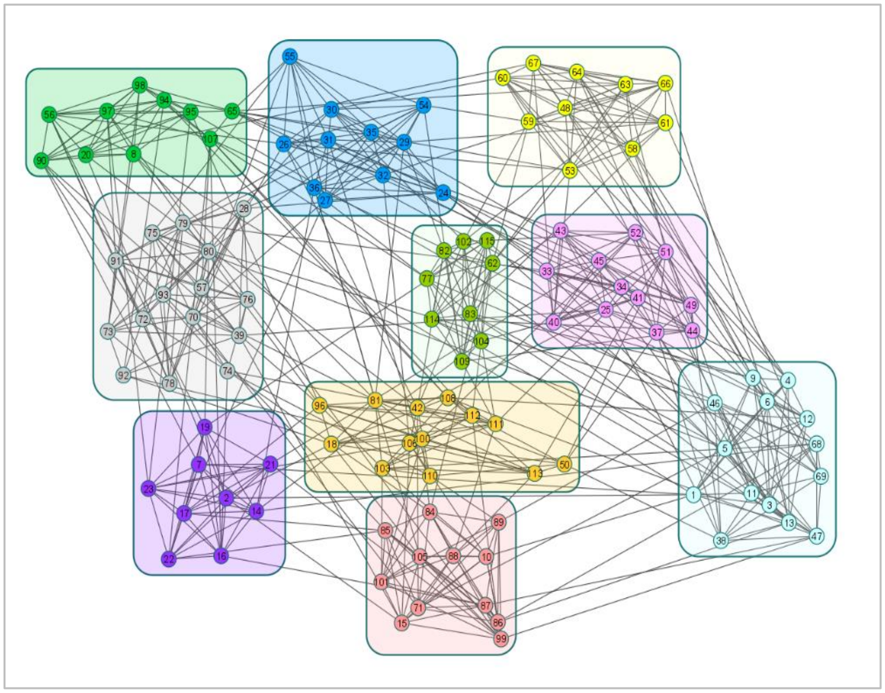


Figure 4-16: Detected 10 Communities in Football dataset by Edmot algorithm

And following the same pattern done in Polbooks dataset, GEMSEC algorithm (as visualized in Figure 4-17) has once again produced one large community consisting of the majority of the nodes, even though the algorithm has managed to detect 10 distinct communities. But the number of nodes in these 9 remaining communities is comparatively low.

The dominant community was 11% homogeneous in which nodes in this community belonged to different conferences, while the other communities ranged between 15% to 27%.

Alternatively, DANMF algorithm (visualized in Figure 4-18) has detected 8 communities from the American College Football network, with relatively close number of nodes in each community.

Communities detected by DANMF also ranged from 14% to 24% homogeneous, where the number of conferences in each community was not less than 6.

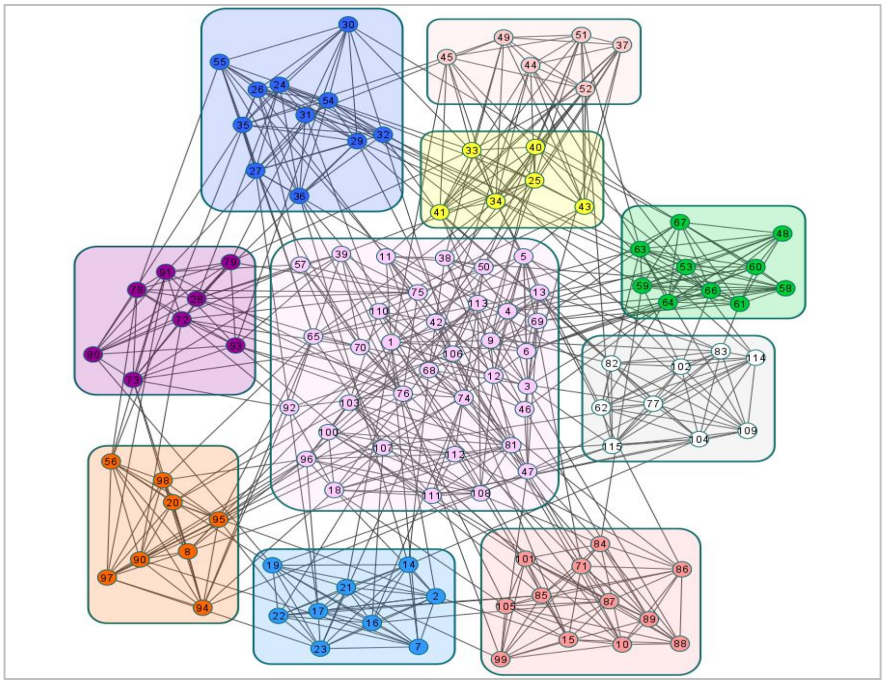


Figure 4-17: Detected 10 Communities in Football dataset by GEMSEC algorithm

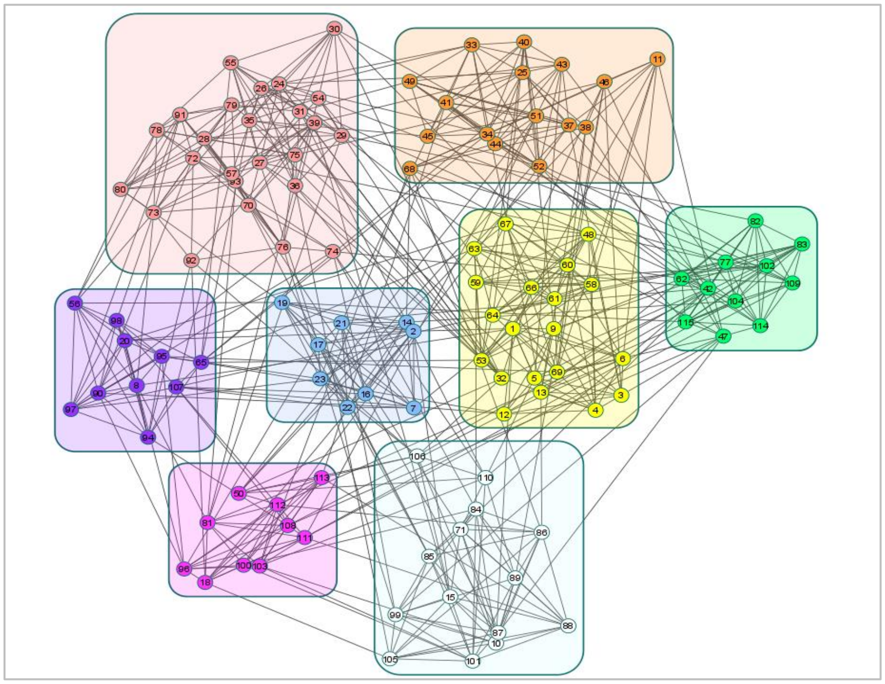


Figure 4-18: Detected 8 Communities in Football dataset by DANMF algorithm

Similarly, Modularity measure, Proposed Homogeneity degree, proposed Penalty measure and the proposed Penalized Homogeneity degree are calculated and compared in Table 4-2.

Table 4-2: Results on Football Dataset, where Q is Modularity, Hd is proposed Homogeneity degree, P is Penalty and PHd is the proposed Penalized Homogeneity degree

Algorithm	**Number of detected communities**	**Q**	**Proposed Hd**	**Proposed P**	**Proposed PHd**
Proposed ABLP (2021)	**9**	**0.660**	**0.650**	**0.108**	**0.543**
LPA [29] (2007)	9	0.610	0.547	0.181	0.365
EdMot [177] (2019)	10	0.651	0.179	0.670	-0.491
GEMSEC [176] (2019)	10	0.561	0.197	0.707	-0.510
DANMF [178] (2018)	8	0.621	0.157	0.626	-0.469

The evaluation measures by discusses algorithms performed on the American College Football are also visualized in Figure 4-19 to Figure 4-21.

The number of detected communities in American College Football dataset are shown in Figure 4-19. None of the algorithms has detected the true number of communities (which is 12), although, GEMSEC and Edmot generated 10 communities, which is the closet, ABLP and LPA have generated 9 communities only. However, the number of communities does not always indicate a good division of communities; other evaluation measures are more reliable as they reveal the structure and accuracy of detected communities.

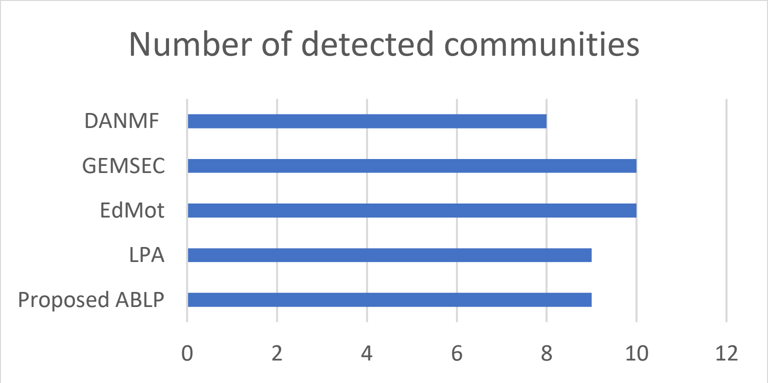


Figure 4-19: Number of detected communities in Football dataset.
Where the number of communities in ground truth table= 12

While GEMSEC and Edmot have detected the highest number of communities, the Homogeneity and Modularity values are considerably low. (As show in Figure 4-20 and Figure 4-21)

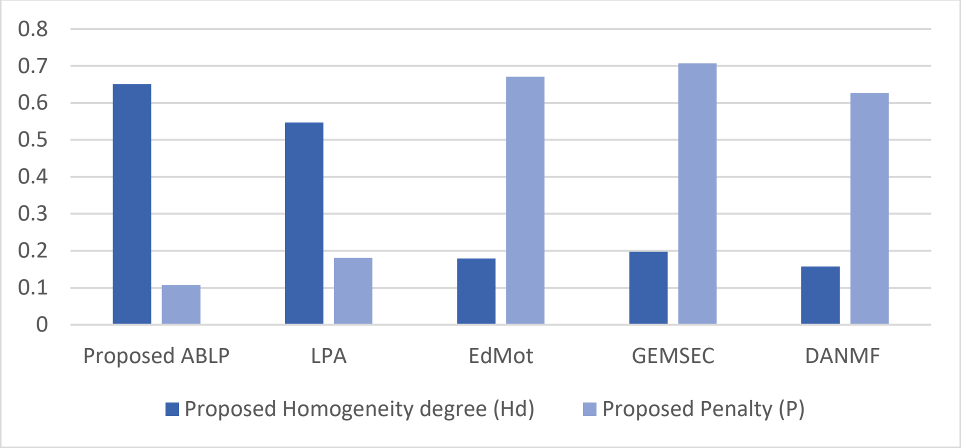


Figure 4-20: Proposed Homogeneity degree and proposed Penalty on Football dataset, Where the maximum values of Homogeneity and minimum values of penalty indicate best homogeneous results.

The proposed algorithm has obtained the highest Homogeneity degree, while maintaining the lowest penalty value. On the other hand, DANMF has achieved the lowest homogeneity, and GEMSEC has performed the highest penalty.

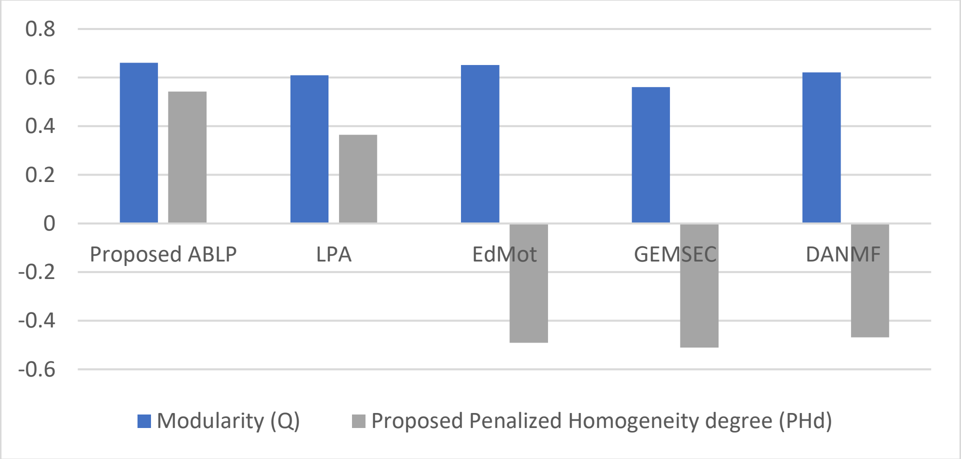


Figure 4-21:Proposed Penalized Homogeneity degree and Modularity on Football dataset, where highest values of both measures denotes the best results in Modularity and Penalized Homogeneity.

The Modularity measure values of American College Football dataset are likely close by the experimented algorithm. However, Homogeneity measures are significantly low as the communities detected included nodes from diversified conference values.

For a higher Homogeneity value, the community should contain nodes with the least number of attribute values possible. To better understand what happened, the average number of attribute values in a community can be calculated, and obviously, the closer the value to 1, the better.

In American College football dataset, the number of attribute values is 12, which can be considered high to some extent compared to Books dataset which consisted of 3 attribute values. It is observed that when a community consists of nodes with more than 3 different attribute values, the Homogeneity value is relatively low. To prove this, a measure of Average Attribute value (AAv) in a community is proposed and calculated, as seen in Table 4-3. The AAv counts the average number of attributes in each community, to check if the community contains nodes with similar attributes (which results in increasing the Homogeneity degree).

It can be clearly perceived that higher Average Attribute value result in higher penalty and thus a lower PHd value. This draws a conclusion, that having multiple attribute values in one community results in a non-homogeneous environment

Table 4-3: The average number of attribute value (AAv), where the maximum value of PHd, and the values closer to one in AAv and P indicate the most homogeneous communities

Algorithm	Proposed P	Proposed PHd	AAv
	min	max	min
Proposed ABLP (2021)	**0.108**	**0.543**	**2.78**
LPA [29] (2007)	0.181	0.365	5.44
EdMot [177] (2019)	0.670	-0.491	7.2
GEMSEC [176] (2019)	0.707	-0.510	6.6
DANMF [178] (2018)	0.626	-0.469	9.625

For the proposed dataset, since it is a real-world contact tracing network, and the number of edges is less than the number of nodes, so the penalty is not considered as the nodes did not have enough connections with one another.

Table 4-4: Results on Proposed COVID-19 Contact Tracing in the Kingdom of Bahrain dataset, where Q is Modularity, Hd is proposed Homogeneity degree, P is Penalty and PHd is the proposed Penalized Homogeneity Degree

Algorithm	Number of detected communities	Q	Proposed Hd
Proposed ABLP (2021)	191	0.983	**0.801**
LPA [29] (2007)	204	0.938	0.722
EdMot [177] (2019)	183	**0.986**	0.482
GEMSEC [176] (2019)	8	0.327	0.337
DANMF [178] (2018)	10	0.394	0.306

As the number of detected communities in the proposed COVID-19 Contact Tracing in the Kingdom of Bahrain dataset was high, the characteristics of the large communities (contain more than 7 nodes) only will be discussed. As the nationality attribute was investigated in this dataset, the percentage of the nationality with the majority of nodes in one community is considered. It was observed that most nodes in the detected communities were either Bahraini, Indian, or Bangladeshi.

For example, detected communities by ABLP contained Bahraini patients with the following proportions: 100%, 95%, 75%, 71% and 38%. While Indian communities were 100%, 62% 60%, 55%, 48%, 46%, 44%, 42%, 38%, 24%, with an average of 52%. Moreover, Bangladeshi communities were 55%, 32%, and 31% homogenous.

Other algorithms also detected homogeneously communities based on nodes' nationality where most of the communities were Bahraini, Indian, and Bangladeshi, which means that these three nationalities were governing the virus distribution among citizens.

The results obtained on the proposed dataset of COVID-19 contact tracing in the Kingdom of Bahrain can be visualized in Figure 4-22.

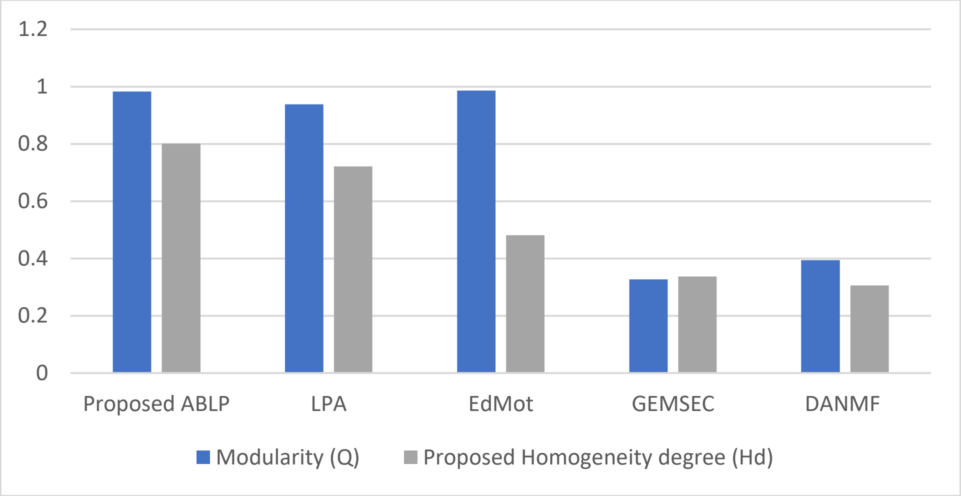


Figure 4-22: Proposed Homogeneity degree and Modularity on proposed dataset, where highest values of both measures denotes the best results in Modularity and Penalized Homogeneity.

As detected from these results, the highest modularity value is achieved by Edmot, followed by the proposed ABLP algorithm. However, the Homogeneity value is the highest in the proposed measure ABLP and the Label Propagation Algorithm. It is noticeable that Edmot achieved as high modularity value whilst scored a comparatively low Homogeneity measure. As for GEMSEC and DANMF, both algorithms detected a low number of communities with high number of nodes in one community, then divided the rest of the nodes on the remaining communities. This manifestly resulted in a low modularity value as well as a low Homogeneity measure.

4.5.3.1 Evaluating the accuracy and privacy of the detected communities' results

The accuracy of generated results is calculated by several measures: Variation of Information (VI) where the closest values to zero, the better. While for Normalized Mutual Information (NMI), Rand Index (RI), and Adjusted Rand Index (ARI), when the value is near to 1 it denotes that both the clustering results are more similar and the distance between them is zero. Whereas the privacy measure is captured by Split-Join, and values closer to 0 indicate highest privacy.

These measures compare the communities discovered by the algorithms to the ones given by the ground-truth, which means that they can only be applied on datasets provided with a ground-truth community structure.

In these experiments, the American College Football and Political books datasets are used, as the proposed dataset COVID-19 contact tracing does not have a ground-truth structure. And the algorithms are the proposed ABLP, LPA, Edmot, GEMSEC, DANMF.

Table 4-5: Accuracy measures on Polbooks dataset (Where NMI is Normalized Mutual Information, RI is Rand Index, VI is Variation of Information, ARI is Adjusted Rand Index**)**

Methods	**NMI**	**RI**	**ARI**	**VI**
	Maximum value indicates maximum accuracy		Minimum value is best	
Proposed ABLP (2021)	**0.554**	**0.843**	**0.665**	**0.956**
LPA [29] (2007)	0.371	0.746	0.454	1.463
EdMot [177] (2019)	0.498	0.695	0.296	1.770
GEMSEC [176] (2019)	0.179	0.559	0.057	2.078
DANMF [178] (2018)	0.333	0.672	0.231	1.966

The results carried out in the accuracy evaluation measures on the Political Books dataset are visualized in Figure 4-23. Where the highest values indicate a better accuracy of community detection. However, it was noted that the values of RI are high compared to other measures.

As NMI compares the results with the ground truth table of the dataset, and the ground truth of PolBooks dataset divides the nodes in terms of their political alignments which is the attribute value considered in the experiment, the proposed algorithm ABLP managed to score the highest NMI value, which is 0.554, followed by EdMot again. And the lowest score is 0.179 which is achieved by GEMSEC. Similarly, the highest Rand and ARI values are achieved by proposed algorithm with values 0.84 and 0.66 respectively. While the lowest values are generated by GEMSEC algorithm. Edmot has scored a relatively high NMI, whilst LPA scored a higher Rand and ARI values.

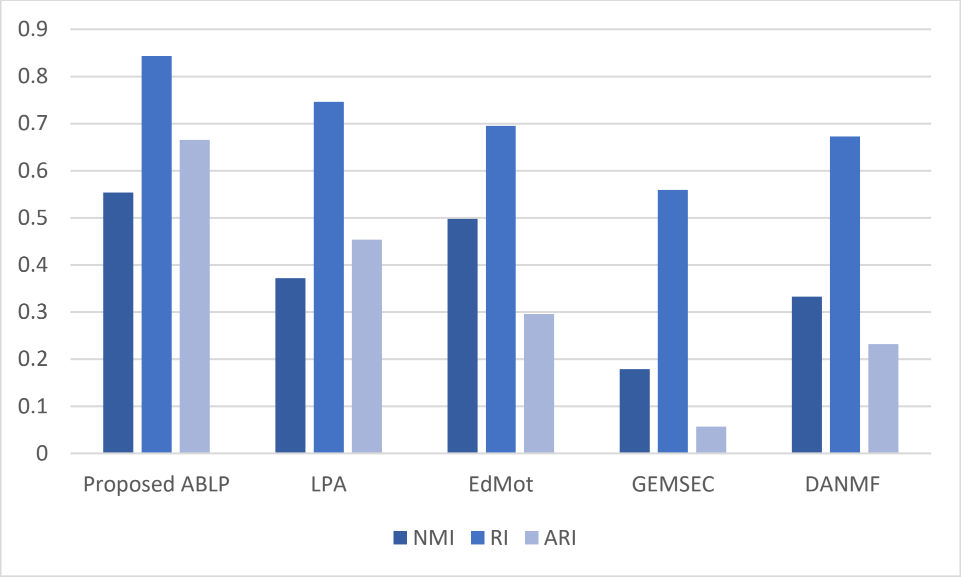


Figure 4-23: Accuracy Evaluation measures on Polbooks dataset, where maximum values of NMI, RI and ARI denote the best accuracy level.

On the other hand, the proposed algorithm scored the lowest Variation of Information value which makes it the optimal result among other algorithms, as the ideal value of VI should be 0, which states that there is no variation of information across clusters. This is illustrated in Figure 4-24.

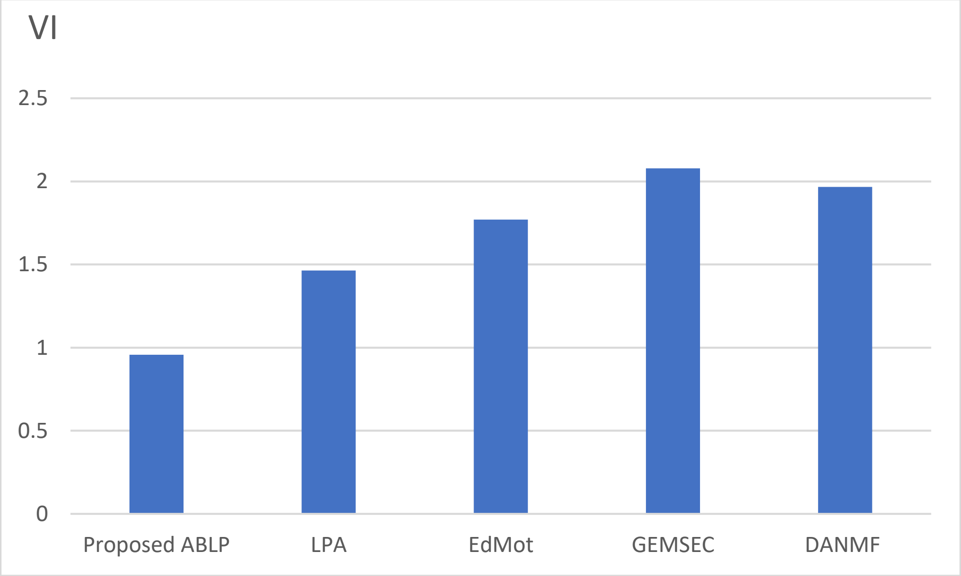


Figure 4-24: Variation of Information on Polbooks dataset,
where minimum value of VI denotes the shortest and hence the best shared information distance.

The same accuracy evaluation measures are computed on American College Football dataset, and the results are carried out in Table 4-6.

Table 4-6: Accuracy measures on Football dataset (Where NMI is Normalized Mutual Information, RI is Rand Index, VI is Variation of Information, ARI is Adjusted Rand Index**)**

Method	NMI	RI	ARI	VI
	Maximum value indicates maximum accuracy		Minimum value is best	
Proposed ABLP (2021)	**0.778**	**0.934**	**0.622**	**1.023**
LPA [29] (2007)	0.628	0.911	0.477	1.723
EdMot [177] (2019)	0.228	0.840	0.0006	3.662
GEMSEC [176] (2019)	0.234	0.799	0.004	3.502
DANMF [178] (2018)	0.173	0.806	-0.007	3.696

And similarly, the results of the accuracy evaluation performed on American College Football dataset are illustrated in Figure 4-25. And the values of RI are high compared to other measures.

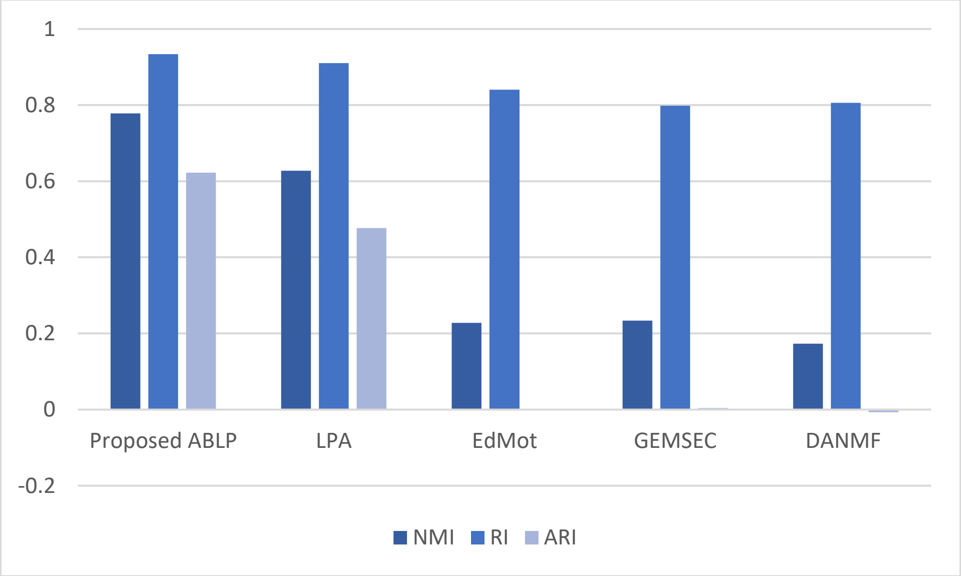


Figure 4-25: Accuracy Evaluation measures on Football dataset, where maximum values of NMI, RI and ARI denote the best accuracy level.

As observed, the highest NMI, Rand and ARI values are produced by the proposed algorithm with values of 0.778, 0.934 and 0.622 respectively. However, DANMF has achieved the lowest NMI value, which is 0.173, and the lowest ARI value which is the only negative value between all. And the lowest Rand value is scored by GEMSEC.

In addition, the lowest VI among all algorithms is achieved by the proposed algorithm with a value of 1.023, followed by LPA with a value of 1.723. While the highest value of 3.696 is obtained by DANMF and followed by Edmot and GEMSEC with values of 3.662 and 3.502 respectively, as visualized in Figure 4-26.

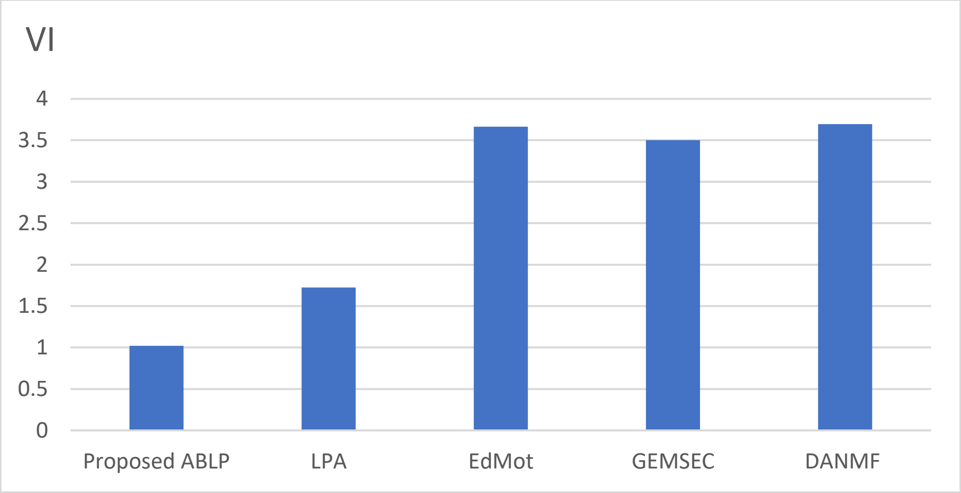


Figure 4-26: Variation of Information on Football dataset
where minimum value of VI denotes the shortest and hence the best shared information distance.

It is known that the accuracy measures are calculated using the ground truth table of the network, and in the ground truth the nodes are basically divided according to their original attribute value. This consequence in a comparatively low NMI, Rand and ARI values in the algorithms that generated communities with high average number of attribute values (Table 4-3).

And to rank the privacy measure, Split-Join is calculated (Table 4-7 and Table 4-8), where lower values indicate maximum privacy as they indicate optimal parameter for privacy preserving in the graph.

Table 4-7: Split-Join measure on Polbooks dataset

Method	Split-Join
Proposed ABLP (2021)	**37.0**
LPA [29] (2007)	48.0
EdMot [177] (2019)	69.0
GEMSEC [176] (2019)	85.0
DANMF [178] (2018)	83.0

It is observed that the lowest Split-Join, that indicates the better-preserved privacy is attained by the proposed algorithm, followed by LPA. On the contrary, the highest Split-Join is achieved by GEMSEC and DANMF with a quite close value.

Table 4-8: Split-Join measure on Football dataset

Method	Split-Join
Proposed ABLP (2021)	**44.0**
LPA [29] (2007)	68.0
EdMot [177] (2019)	174.0
GEMSEC [176] (2019)	162.0
DANMF [178] (2018)	172.0

In American College football, the results follow the same methodology, where the values of Split-Join values arranged in a descending order are the proposed dataset, LPA, GEMSEC, Edmot, and DANMF.

The Split-Join distance value of both datasets can be illustrated in Figure 4-27.

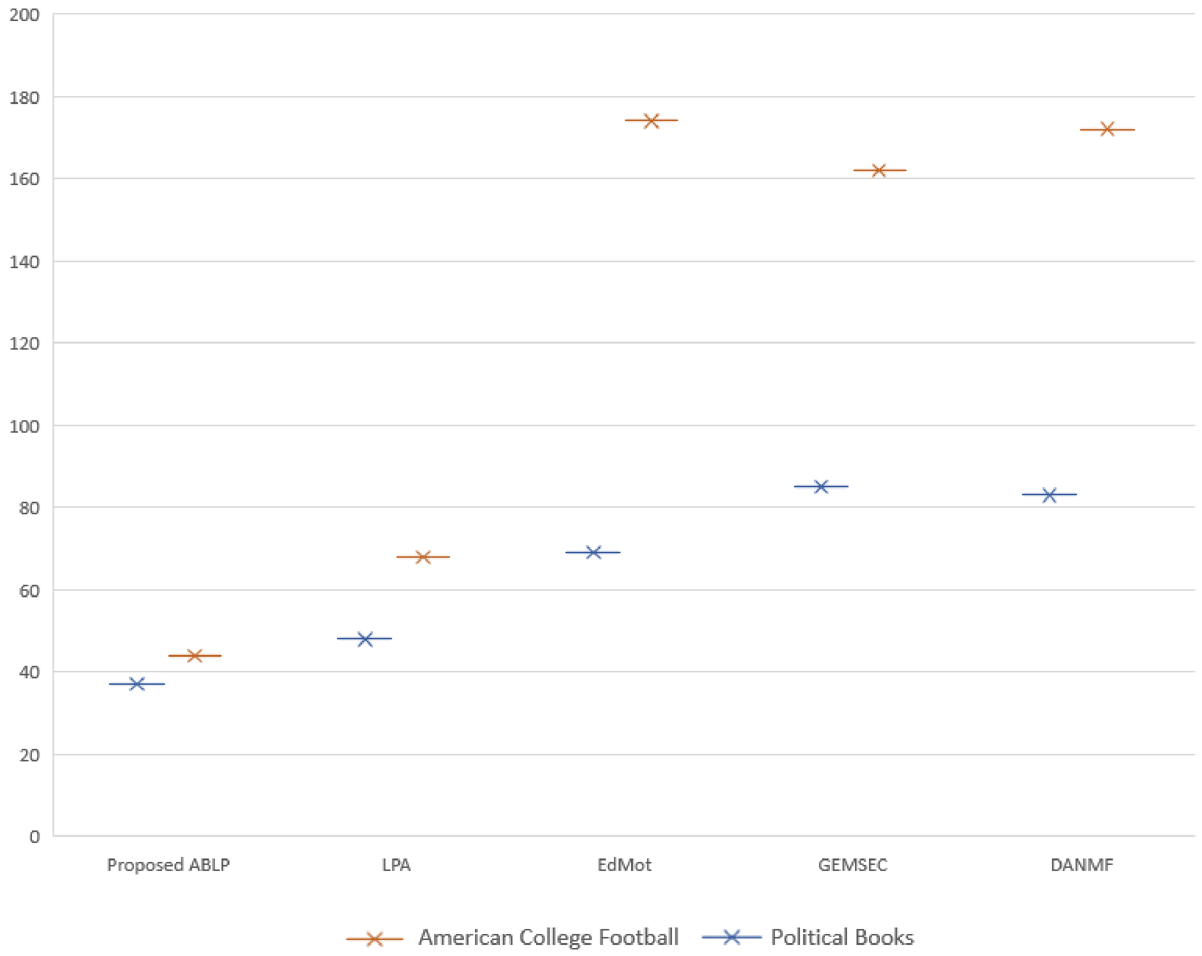


Figure 4-27: Split-Join distance value on Polbooks and Football datasets, where lowest values indicate an ideal privacy level

By calculating the accuracy and privacy measures, it can be clearly noted that the best results in this experiment are achieved by the proposed algorithm ABLP. Which denotes that the proposed algorithm can detect communities that are close to the ground truth. And since the ground truth communities are basically divided based on attributes' values, the detected communities by the proposed algorithm are also homogeneous.

4.5.4 Homogeneity as a constraint in ABLP

Community detection problem was proposed as a constrained approach to exploit the existing side information of the network [99]. Adding constraints helps in generating more efficient and actionable results, and help develop data mining techniques that can handle complex and domain-specified constraints [100]. And as demonstrated in section 0, most of the current community detection methods consider the structural information of networks, but disregard the fruitful information of the nodes, and this results in the failure of detecting semantically meaningful communities [97]. And based on the literature review, Table 2-9 illustrated that homogeneity was never studied as a constraint, and was always treated as an objective function, and based on this research gap, homogeneity is studied as a constraint.

In this experiment, the homogeneity is treated as a constraint, which minimizes the Modularity value based on the achievement of the homogeneity value. When the homogeneity value is high, modularity measure should remain at its best. On the contrary, when the value of homogeneity is low, the Modularity value should be punished and reduced. This is tested with the same experiments, as seen in Table 4-9. Where Q(C: H) is the value of Modularity constrained with homogeneity (Equation 29). For Books and Football datasets, PHd homogeneity value is considered since a penalty is applied.

Table 4-9: Homogeneity as constraint in Books dataset, where Q is Modularity, PHd is the proposed Penalized Homogeneity degree, Q(C: H) is Modularity Constrained with Homogeneity

Algorithm	Q	PHd	Q(C: H)
Proposed ABLP (2021)	**0.527**	**0.137**	**0.610**
LPA [29] (2007)	0.480	-0.053	0.467
EdMot [177] (2019)	0.509	-0.099	0.392
GEMSEC [176] (2019)	0.336	-0.209	0.455
DANMF [178] (2018)	0.492	-0.180	0.328

The same experiments were conducted on the American Football College dataset, and results of evaluation measures are illustrated in Table 4-10.

Table 4-10: Homogeneity as constraint in Football dataset, where Q is Modularity, PHd is the proposed Penalized Homogeneity degree, Q(C: H) is Modularity Constrained with Homogeneity

Algorithm	Q	PHd	Q(C: H)
Proposed ABLP (2021)	**0.660**	**0.543**	**0.882**
LPA [29] (2007)	0.610	0.365	0.755
EdMot [177] (2019)	0.651	-0.456	0.107
GEMSEC [176] (2019)	0.561	-0.510	0.071
DANMF [178] (2018)	0.621	-0.469	0.090

And as the proposed COVID-19 dataset did not need the penalty measure, the value of constrained Homogeneity is Hd.

Table 4-11: Homogeneity as constraint in proposed dataset, where Q is Modularity, PHd is the proposed Penalized Homogeneity degree, Q(C: H) is Modularity Constrained with Homogeneity

Algorithm	Q	Proposed Hd	Q(C: H)
Proposed ABLP (2021)	0.983	**0.801**	0.788
LPA [29] (2007)	0.938	0.722	0.660
EdMot [177] (2019)	**0.986**	0.482	0.468
GEMSEC [176] (2019)	0.327	0.337	0.336
DANMF [178] (2018)	0.394	0.306	0.299

Testing the homogeneity as a constraint is to help in evaluating the results in terms of Modularity and homogeneity at the same time. For example, even though Edmot algorithm has generated the highest Modularity value, the homogeneity value was lower than proposed ABLP and LPA. So, using the constrained Homogeneity will help compare these methods.

Here is it assumed that both measures have the same importance or weight in the results. However, a weight can be assigned to the measures based on how important each measure is. This facilitates in the evaluation process based on the defined user requirements, which are aligned with the dataset itself. So, if the user is interested more in the homogeneity than Modularity, a ratio of 70/30 can be applied, where homogeneity is responsible for 70% of the measure and the Modularity is for the other 30%. This can be calculated as |1- (0.3 * Q – 0.7 *H). In other words, this way can be personalized according to the nature of the dataset and the expected detected communities.

Both proposed methods have achieved remarkable results in terms of Modularity, Normalized Mutual Information, and Homogeneity degree. However, while the performance of the proposed ABLP algorithm is better in terms of time efficiency, the proposed constrained ABLP can be easily adjusted and personalized based on the user’s requirements. It is always required to check the main objectives of the community detection problem and hence decide on the method. So, if Homogeneity and Modularity are equally important then the Constrained ABLP is a better choice as it provides two different perspectives through one measure. However, ABLP with homogeneity as an objective function provides separate measures to be looked into in order to evaluate the results of detected communities.

4.6 Summary

Community detection in attributed networks can be evaluated in many aspects. The mostly used evaluation measures such as Modularity and NMI, cannot address the evaluation of homogeneity. Hence, Attribute-Based Label Propagation ABLP algorithm, that considers the attribute values of nodes while maintaining a high Modularity, homogeneity and accuracy measures is proposed. It was studied with homogeneity as an objective function, and another time as a constraint.

And to support evaluating the proposed algorithm, an adaptable homogeneity measure is also proposed. This measure assesses the homogeneity in an attributed network and can be penalized based on the type of the dataset. Experiments on existing social networks are conducted as well as on the newly proposed COVID-19 dataset which is based on the contact tracing of the virus infected persons in the Kingdom of Bahrain. The algorithm appears to have good results in terms of the discussed evaluation measures.

5 Algorithms Performance Evaluation on Proposed COVID-19 World Countries Dataset

Comparative studies are performed on a number of methods, and on different datasets to examine and reconnoiter their implementation and performance. The objective of this type of research is to exhibit paradigm suggestions according to graph structure. In the field of community detection, multiple research studies aimed to proceed a comparative study to study different datasets, and to evaluate the distinct algorithms [179], [120].

In this chapter, several algorithms such as Girvan Newman, Greedy, Louvain, Clustering, and Label Propagation are implemented on the proposed COVID-19 dataset: World Countries, and the results are evaluated using the modularity measure. It should be noted that the proposed methods in 4.3 were not used in this experiment as they are attribute-based, and this dataset has no attributes.

5.1 Algorithms used in this experiment

A number of the most used algorithms to solve the community detection problems, that are implemented and evaluated on two different datasets.

Louvain Best partition[26] a hierarchical clustering algorithm, that recursively merges communities into a single node and executes the modularity clustering on the condensed graphs. The running time on a network with *n* vertices is linear O($n\log^2 n$).

Greedy[25] is a heuristic method that finds communities in graph using Clauset-Newman-Moore greedy modularity maximization. This method begins with each node in its own community and joins the pair of communities that most increases modularity until no such pair exists. The running time on a network with *n* vertices is linear O($n\log^2 n$).

Newman 2008[27] clusters the network into several modules using modularity maximization by spectral methods. Supports directed and undirected networks. Edge weights are ignored.

Spectral Clustering[28] is a type of unsupervised community discovery algorithm. The number of communities k should be given in advance. Time complexity for a network with *n* vertices is O(n^2).

Kernighan–Lin[24] is a heuristic method that partitions a network into two sets by iteratively swapping pairs of nodes to reduce the edge cut between the two sets. An upper bound is justified to the execution time of O(n^2log(C)), where N is the number of nodes, and C the number of communities in the network.

Asynchronous label propagation[29] initializes each node with a unique label, it repeatedly sets the label of a node to be the label that appears most frequently among that nodes' neighbors. The algorithm is asynchronous because each node is updated without waiting for updates on the remaining nodes. The time complexity of this method is near linear, which is O(n+m) in a network of *n* nodes, and *m* edges.

5.2 Experiments

The aim of this experiment is to compare the different community detection algorithms on two datasets. The first dataset is (Zachry's Karate Club) [150], it was used in the literature and its structure is somewhat close to the proposed one. It is a social network of friendships between 34 members of a karate club at a US university in the 1970s. And the second one is the dataset proposed in 3.2.1 (World Countries), it consists of 36 nodes and 75 edges. Upon detecting the communities within these two networks, the results are tested in terms of number of communities detected (K) and modularity value; the latter is calculated to measure the quality of a disjoint partition of a network.

The experiments are conducted on one computer server which is equipped with core i7 CPU, and 32G memory. Other software environments include python 3.6 and pycharm 2019.3.1

Experiment A: Investigation of literature dataset (Zachry's Karate Club dataset)

This section presents the simulation results for the Zachry's Karate club dataset (as an example of commonly used datasets) tested on community detection algorithms; results are presented in Table 5-1.

The algorithms were already tested in previous research papers, the main idea of this experiment is to compare these existing examined results with the results of the proposed dataset generates. It is remarkable that the highest modularity measure is achieved by Girvan Newman algorithm, followed by spectral clustering when the number of detected communities is set to 4. Furthermore, the lowest modularity measure is achieved by Kernighan–Lin algorithm, although most algorithms gave convergent values.

Table 5-1: Zachry's Karate Club dataset results

Algorithm	Number of Communities	Modularity
Louvain	4	0.415
Greedy	3	0.381
Girvan Newman 2008	4	0.419
Spectral Clustering **Pre-defined number of communities**	K=3	0.399
	K=4	0.411
	K=5	0.390
Kernighan–Lin	2	0.372
Asynchronous label propagation	5	0.394

5.2.1 Experiment B: Investigation of proposed dataset (COVID 19 World Countries)

On the other hand, Table 5-2 sums up the studies regarding testing the algorithms explained in the literature on the World Countries designed dataset.

As can be seen in Table 5-2, the modularity values for the proposed dataset are likewise close, except for Kernighan–Lin which seems to give the lowest value, it also gave the lowest with karate dataset. While the highest modularity value is achieved by Louvain, which is not the case in experiment A.

Louvain algorithm randomly order all nodes in the network, then, one by one, it will remove and insert each node in a different community until no significant increase in modularity is verified [170]. The modularity measure achieved by the proposed dataset, are comparatively close to the measure produced in the Zachry's Karate dataset. For spectral clustering algorithm, the number of communities should be set in advance, so the common numbers observed from other algorithms (3, 4 and 5) were selected and applied, and the modularity gave its best result when the number of communities is set to 3. This algorithm also achieved promising results on both datasets.

Table 5-2: Proposed COVID-19 dataset: World Countries results

Algorithm	Number of Communities	Modularity
Louvain	4	0.344
Greedy	3	0.333
Girvan Newman 2008	4	0.332
Spectral Clustering **Pre-defined number of communities**	K=3	0.333
	K=4	0.315
	K=5	0.323
Kernighan–Lin	2	0.253
Asynchronous Label Propagation	6	0.305

In general, Newman achieved the highest modularity value in the first experiment, while Louvain achieved better in the second experiment. Both methods have generated close modularity values in both experiments. As the modularity measures the strength and structure of the detected communities, both methods can be considered effective, and they don't require a prior knowledge of the number of nodes unlike the spectral clustering which may not be practical if this number is unknown.

5.3 Results and Discussion

This chapter proposes a novel dataset based on the coronavirus contact tracing within the world countries. Since it has become a global concern, the records are utilized and transformed into a network, where each country is represented by a node, and their relationships modelled by edges. Analysis of diverse algorithm is carried out on the newly generated dataset, for comparison purpose in terms of numbers of detected communities and modularity performance measure. Louvain algorithm tends to score the highest modularity value. The average number of detected communities is 4, and the average modularity value is 0.317. Analysis of communities generated by algorithms have demonstrated the common features of communities identified in COVID 19 dataset that are not always generated by geographical locations.

In general, it is intelligible that the countries are distributed geographically, even though the dataset is non-attributed, and the community detection algorithms are not aware for the geographic location of the countries.

Some countries are more difficult to be classified and separated, as an example, the community of Europian countries ”Italy, Spain, Germany, France” is consistent in all algorithms. As well as Asian countries “Singapore, Japan, Malaysia, Cambodia, Thailand, Vietnam, Sri Lanka, Hong Kong, Philippines, were always grouped in one community with Russia and Belgium which is a European country.

It is also explicit that some countries belonged to the same clusters among all algorithms, such as [France, Italy, Germany, Spain, Austria, Algeria, Switzerland, Croatia, UK]. This indicates that the connections between these countries is strong, and there were several interactions between them. As nodes that belong to the same community have tighter connections with one another than with the rest of the network, which is why they are classified in the same community by different algorithms.

On the other hand, as can be seen in the visual representation (Figure 5-1 to Figure 5-8), although some algorithms generate the same number of communities, the nodes formation in each community is distinct.

For example, Louvain in Figure 5-1, classifies “Canada” in a community with [“Iran”, “UAE”, “Lebanon”, “Kuwait”, “Bahrain”, “Afghanistan”] whereas in Girvan Newman (Figure 5-3), “Canada” does not belong to this community even though the other six countries do. And in Spectral clustering when k=4 (Figure 5-5), the community that contains the majority of these countries did not include “Canada” neither “UAE”. It can be observed that the characteristics of the detected communities is not always based on the geographical location, but on the type of connections between the different countries, which is basically the travelling history of their citizens.

Another interesting finding is that ”China” and ”Iran” are acting as core nodes in the network, they both consistently have the focus in the subnetwork, except in Kernighan-Lin (Figure 5-7), as there are only two communities. This denotes that these countries have tighter connections with the other countries and they supposedly spread the virus to the rest of the countries, as the virus originally started in China.

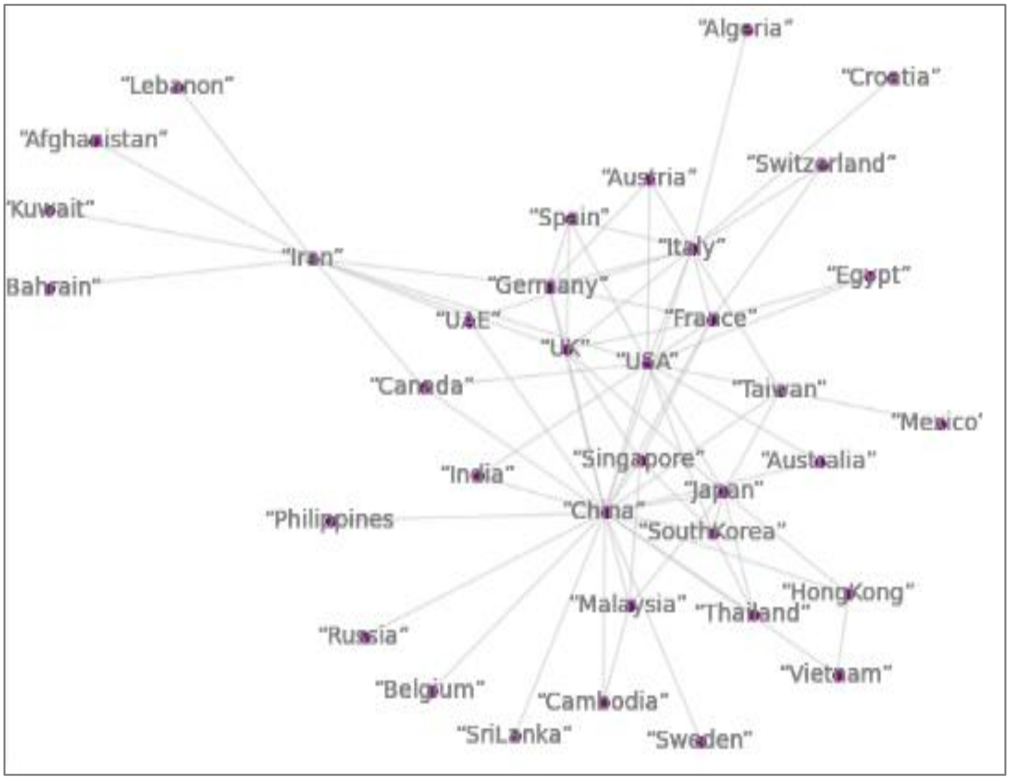


Figure 5-1: Louvain Algorithm on Proposed COVID-19 dataset: World Countries

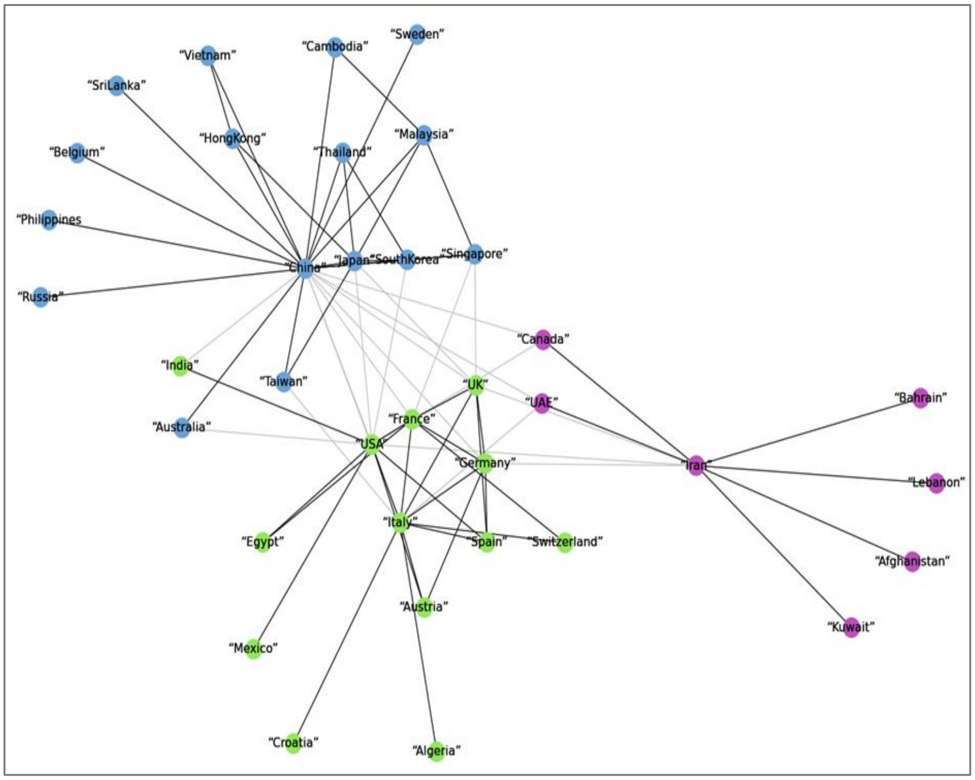


Figure 5-2: Greedy Algorithm on Proposed COVID-19 dataset: World Countries

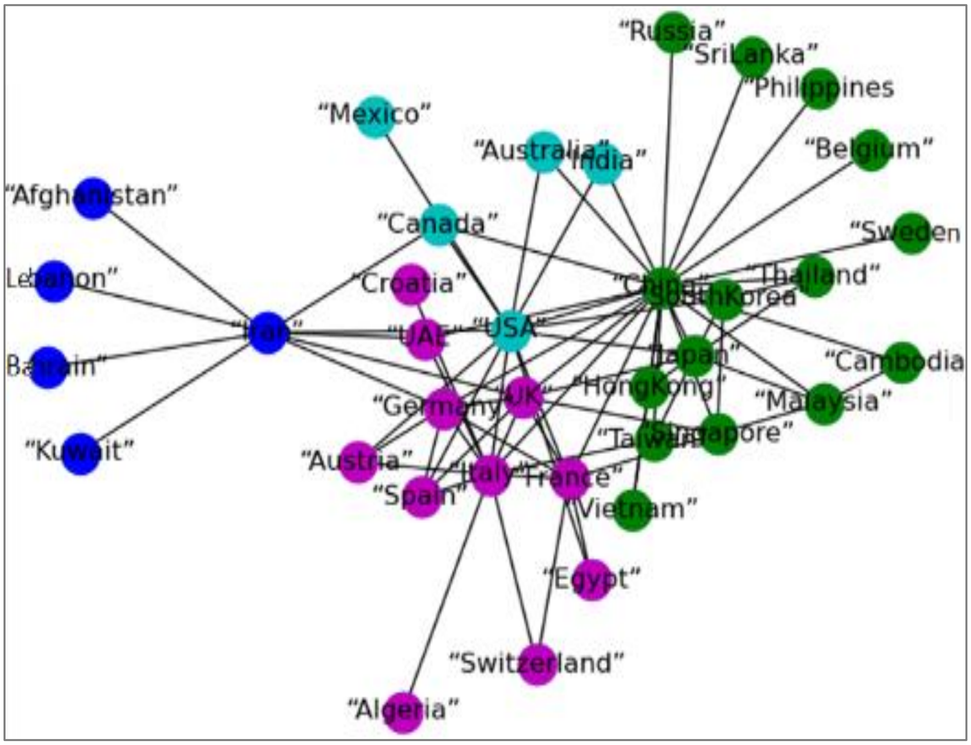


Figure 5-3: GN Algorithm on Proposed COVID-19 dataset: World Countries

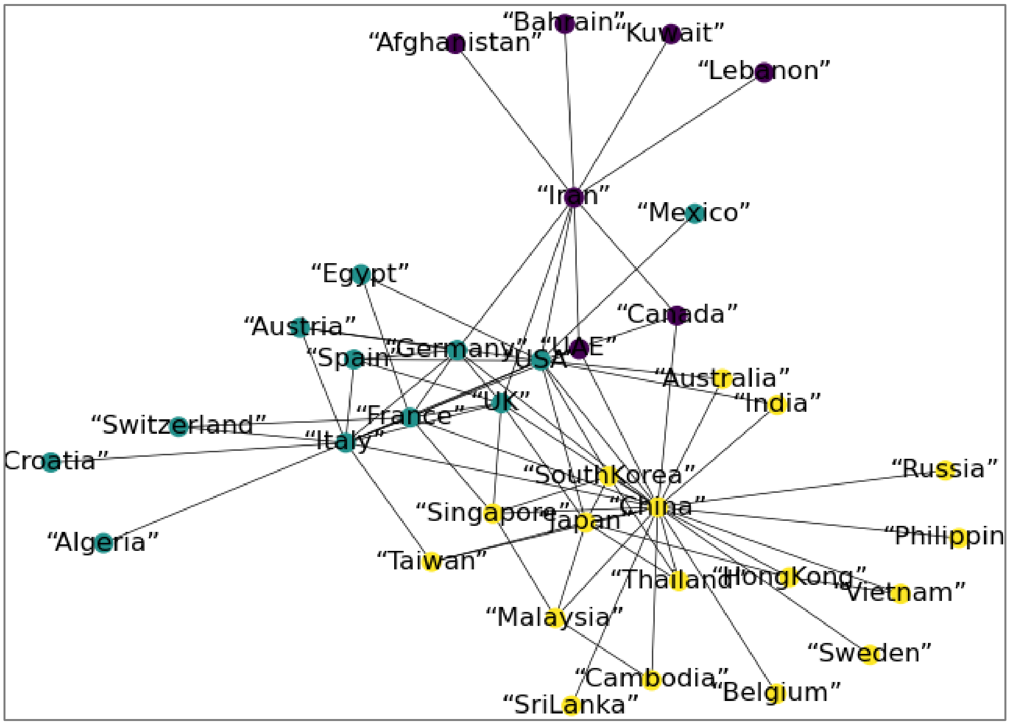


Figure 5-4: Spectral Clustering Algorithm k=3 on Proposed COVID-19 dataset: World Countries

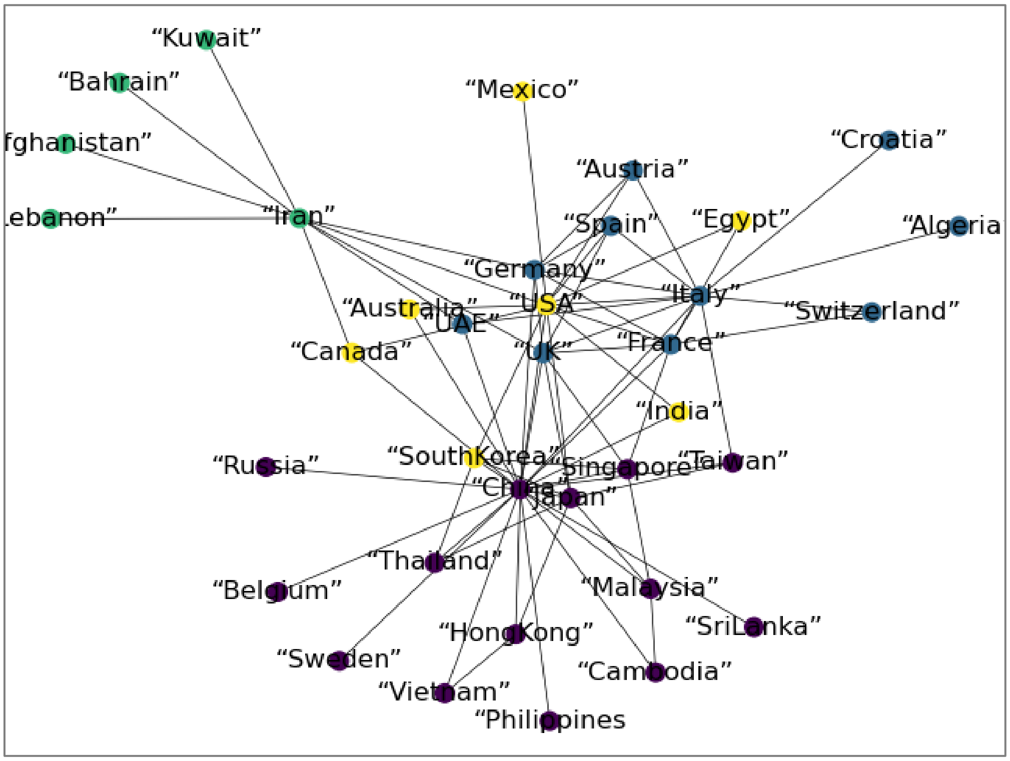


Figure 5-5: Spectral Clustering Algorithm k=4 on Proposed COVID-19 dataset: World Countries

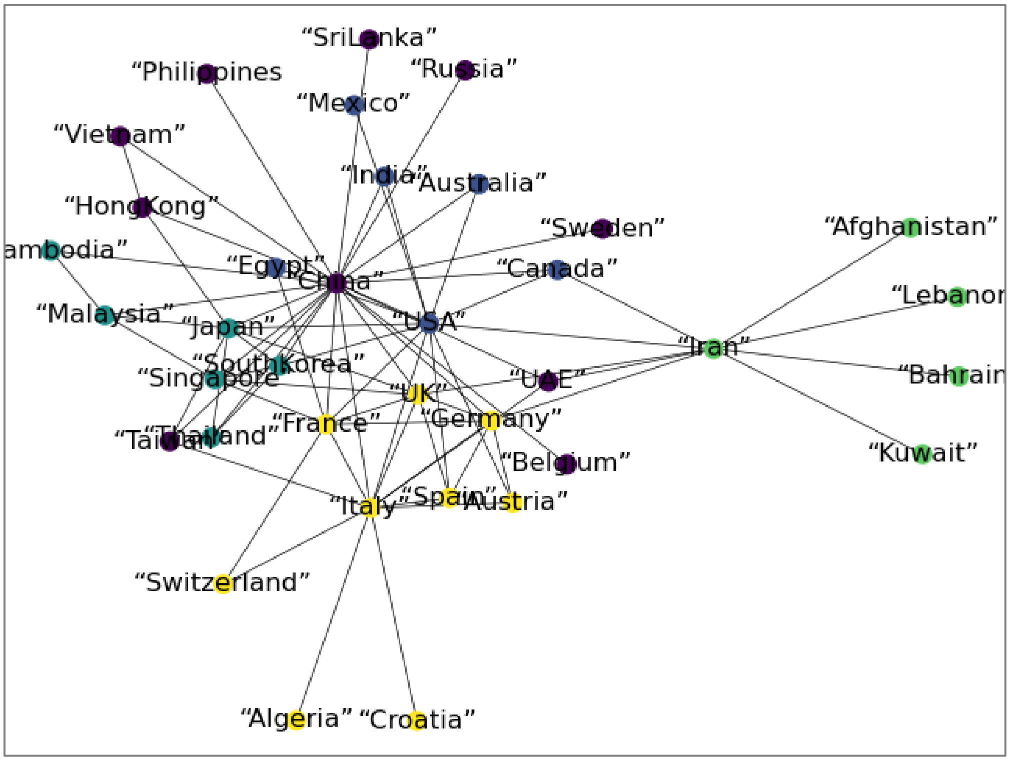


Figure 5-6: Spectral Clustering Algorithm k=5 on Proposed COVID-19 dataset: World Countries

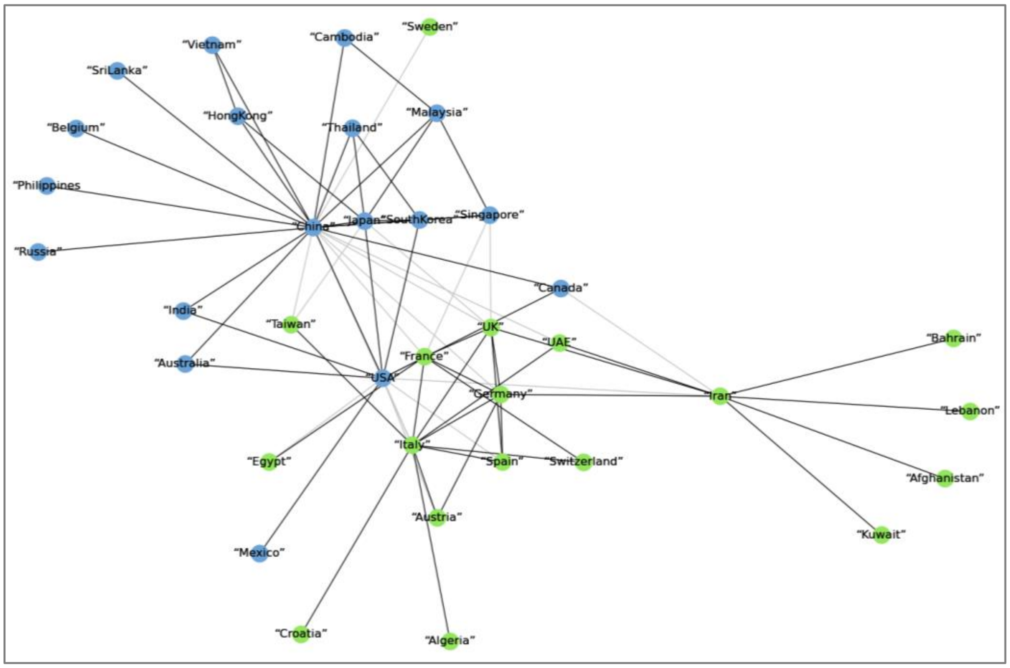


Figure 5-7: Kernighan-Lin Algorithm on Proposed COVID-19 dataset: World Countries

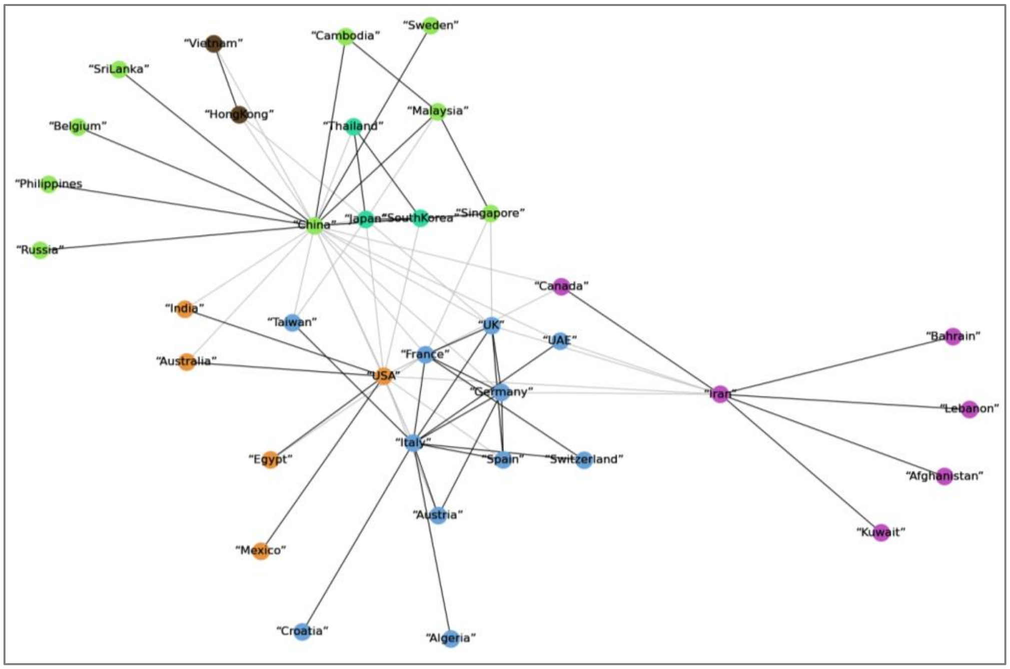


Figure 5-8: ALP Algorithm on Proposed COVID-19 dataset: World Countries

In fact, the exact community is generated by Louvain, GN, and Spectral Clustering, unless Louvain’s also contained Egypt, and Spectral Clustering contained UAE. Another common community includes [China, Singapore, Malaysia, Cambodia, SriLanka, Russia, Philippines, Sweden, and Belgium]. Furthermore, Label Propagation algorithm would have produced only 4 communities and its’ results would have been very close to Louvain’s if it had merged the East Asian countries in one community instead of three.

5.4 Summary

Analysis of diverse algorithm is carried out on the newly generated dataset, for comparison purpose in terms of numbers of detected communities and modularity performance measure. Louvain algorithm tends to score the highest modularity value. The average number of detected communities is 4, and the average modularity value is 0.317. Analysis of communities generated by algorithms have demonstrated the common features of communities identified in COVID 19 dataset that are not always generated by geographical locations.

The main strength among all algorithms is the consistency in countries' allocation, regardless the number of detected communities. The consistency of the results demonstrate that the proposed dataset is created with balanced number of communities, and the results are comparable with the literature dataset Zachry's Karate Club.

6 Conclusion and Future Work

As a tool of network analysis, Community Detection is an emerging field nowadays, as it reveals structures and functional characteristics of the network and discovers the identification of latent communities. The detected communities outline the graphs into strongly connected groups of nodes, in which the nodes that belong to each group or community are substantially connected to each other than to the rest of the network.

In this research a critical literature review of Community Detection problem is provided to explore the scope of community detection solutions and introduce prominent research studies in this decade. The fundamental concepts of this problem are covered, including the different algorithms used for community detection, for example clustering, fuzzy, node-importance, genetic and swarm intelligence algorithms. Different evaluation metrics used to examine various algorithms were also reviewed. In addition, the community detection applications and uses are categorized and explained, and real-world as well as synthetic datasets are discussed.

Based on the research gap, this thesis proposes an Attribute Based Label Propagation algorithm that maximizes Modularity and homogeneity at the same time. The proposed method is experimented with homogeneity as an objective function once, and as a constraint another time. Additionally, a new Penalized Homogeneity degree is proposed, to be personalized and used on real-world networks. In addition to a Multi-Attribute Weighted Penalized Homogeneity degree (MAWPHd) which allows a more flexible mensuration of Homogeneity on different types of attributed networks based on the user-defined requirements.

Both proposed algorithms and the proposed Penalized Homogeneity degree are experimented and benchmarked with other algorithms. Results of proposed algorithms are comparably superior in terms of Modularity, and the proposed Penalized Homogeneity degree, in addition to Rand Index, Adjusted Rand Index, Normalized Mutual Information, Variation of Information, and Split-Join distance.

On the other hand, COVID-19 tracing data are employed to build two different networks that can be used in the community detection problem. The first dataset is based on the virus transition between the world countries, and the second dataset is an attributed network based on the virus transition among the contact tracing in the Kingdom of Bahrain. Networks that are utilized in tracking COVID-19 virus transmission were not formed before and were never studied as a part of the community detection problem.

The limitation of this work was the difficulty in accessing the COVID-19 tracing data, while most countries already have tracing applications, this data is not publicly published, which limited the number of datasets created in this matter, and therefore the comparison between different countries contact tracing was not carried out. As this research is limited to community detection problem, the implementation and employment of the proposed datasets were conducted accordingly. However, these datasets can help researchers and policy makers in studying different societal issues related to the pandemic. So as future work, different types of network analysis can be performed on these datasets, as well as employing them in the medical field, economics, in addition to social sciences.

As future work, the proposed algorithms can be extended to contain other types of networks like directed and weighted networks. Because considering the direction and weight of relationship, in addition to the nodes' the attributes can reveal some interesting information and therefore assist in the community detection process. In addition, as observed from the literature review, community detection needs more attention in some fields as this concept is quite new, such as in enterprise, psychology, and folksonomy. Mental health has become a serious concern in peoples' work loaded lives. Which is why this area needs to be studied further to help in understanding human behavior and mental processes.

Bibliography

[1] "The World Economic Forum," *Https://www.weforum.org/*, 2021. .

[2] S. U. Rehman, A. U. Khan, and S. Fong, "Graph mining: A survey of graph mining techniques," *7th Int. Conf. Digit. Inf. Manag. ICDIM 2012*, no. Icdim, pp. 88–92, 2012, doi: 10.1109/ICDIM.2012.6360146.

[3] P. Bedi and C. Sharma, "Community detection in social networks," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 6, no. 3, pp. 115–135, 2016, doi: 10.1002/widm.1178.

[4] P. M. Zadeh and Z. Kobti, "A Multi-Population Cultural Algorithm for Community Detection in Social Networks," *Procedia - Procedia Comput. Sci.*, vol. 52, pp. 342–349, 2015, doi: 10.1016/j.procs.2015.05.105.

[5] L. Tang and H. Liu, "Graph Mining Applications to Social Network Analysis," *Manag. Min. Graph Data*, vol. Springer, pp. 487–513, 2010, doi: 10.1007/978-1-4419-6045-0_16.

[6] S. Mishra, R. Borboruah, B. Choudhury, and S. Rakshit, "Modeling of Social Network using Graph Theoretical Approach," *Int. J. Comput. Appl.*, pp. 34–37, 2014.

[7] J. Leppink and P. Pérez-fuster, "Social Networks as an Approach to Systematic review," *Heal. Prof. Educ.*, vol. 5, no. 3, pp. 218–224, 2019, doi: 10.1016/j.hpe.2018.09.002.

[8] R. Vahidzadeh, G. Bertanza, S. Sbaffoni, and M. Vaccari, "Regional industrial symbiosis : A review based on social network analysis," *J. Clean. Prod.*, vol. 280, p. 124054, 2021, doi: 10.1016/j.jclepro.2020.124054.

[9] S. Banerjee, M. Jenamani, and D. Kumar, "A survey on influence maximization in a social network," *Knowl. Inf. Syst.*, vol. 62, no. 9, pp. 3417–3455, 2020, doi: 10.1007/s10115-020-01461-4.

[10] W. Ahmed, K. Antonio, B. Baesens, T. Donas, and T. Reynkens, "Social network analytics for supervised fraud detection in insurance," *Risk Anal.*, 2020.

[11] M. Aivazoglou, A. O. Roussos, D. Margaris, C. Vassilakis, and S. Ioannidis, "A fine-grained social network recommender system," *Soc. Netw. Anal. Min.*, pp. 1–18, 2020, doi: 10.1007/s13278-019-0621-7.

[12] M. Valeri and R. Baggio, "Social network analysis : organizational implications in tourism management," *Int. J. Organ. Anal.*, pp. 342–353, 2021, doi: 10.1108/IJOA-12-2019-1971.

[13] M. Hung *et al.*, “Social network analysis of COVID-19 sentiments: Application of artificial intelligence,” *J. Med. Internet Res.*, vol. 22, no. 8, pp. 1–13, 2020, doi: 10.2196/22590.

[14] W. Ahmed, J. Vidal-Alaball, J. Downing, and F. L. Seguí, “COVID-19 and the 5G conspiracy theory: Social network analysis of twitter data,” *J. Med. Internet Res.*, vol. 22, no. 5, pp. 1–9, 2020, doi: 10.2196/19458.

[15] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, “Community detection in social media performance and application considerations,” *Data Min. Knowl. Discov.*, vol. 24, no. 3, pp. 515–554, 2012, doi: 10.1007/s10618-011-0224-z.

[16] J. Moosa, W. Awad, and T. Kalganova, “Intelligent Community Detection : Comparative Study (COVID19 Dataset),” in *EAMMIS 2021: Artificial Intelligence Systems and the Internet of Things in the Digital Era*, 2021, vol. 239, pp. 189–196.

[17] J. Moosa, W. S. Awad, and T. Kalganova, “Intelligent Community Detection: Review,” *SSRN Electron. J.*, pp. 1–7, 2020, doi: 10.2139/ssrn.3659107.

[18] H.-J. Li, L. Wang, Y. Zhang, and M. Perc, “Optimization of identifiability for efficient community detection,” *New J. Phys.*, vol. 22, no. 6, 2020, doi: 10.1088/1367-2630/ab8e5e.

[19] F. Liu *et al.*, “Deep learning for community detection: Progress, challenges and opportunities,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2021-Janua, pp. 4981–4987, 2020, doi: 10.24963/ijcai.2020/693.

[20] S. Fortunato, “Community detection in graphs,” *Phys. Rep.*, vol. 486, no. 3–5, pp. 75–174, 2010.

[21] H. Shen, X. Cheng, F. Guo, L. Gao, and X. Yong, “Detecting the overlapping and hierarchical community structure in complex networks,” *New J. Phys.*, vol. 11, no. 3, p. 033015, 2009, doi: 10.1088/1367-2630/11/3/033015.

[22] F. Karimi, S. Lotfi, and H. Izadkhah, “Multiplex community detection in complex networks using an evolutionary approach,” *Expert Syst. Appl.*, vol. 146, p. 113184, 2020, doi: 10.1016/j.eswa.2020.113184.

[23] N. Chen, B. Hu, and Y. Rui, “Dynamic Network Community Detection with Coherent Neighborhood Propinquity,” *IEEE Access*, vol. 8, no. November, pp. 27915–27926, 2020, doi: 10.1109/ACCESS.2020.2970483.

[24] S. Sobolevsky, R. Campari, A. Belyi, and C. Ratti, “A General Optimization Technique for High Quality Community Detection in Complex Networks,” *Phys. Rev. E*, vol. 90, no. 1, p. 012811, 2014.

[25] A. Clauset, M. E. J. Newman, and C. Moore, “Finding community structure in very large networks,” *Phys. Rev. E*, vol. 70, no. 6, p. 066111, 2004.

[26] H. Lu, M. Halappanavar, and A. Kalyanaraman, “Parallel heuristics for scalable community detection,” *Parallel Comput.*, vol. 47, pp. 19–37, 2015, doi: 10.1016/j.parco.2015.03.003.

[27] E. A. Leicht and M. E. J. Newman, “Community structure in directed networks,” *Phys. Rev. Lett.*, vol. 100, no. 11, pp. 1–4, 2008, doi: 10.1103/PhysRevLett.100.118703.

[28] U. Von Luxburg, “A Tutorial on Spectral Clustering,” *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[29] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 76, no. 3, pp. 1–11, 2007, doi: 10.1103/PhysRevE.76.036106.

[30] J. Bruna, “Community Detection with Graph Neural Networks,” *Stat*, vol. 1050, p. 27, 2017.

[31] W. Kernighan and S. Lin, “An Efficient Heuristic Procedure for Partitioning Graphs,” 1969.

[32] M. E. J. Newman, “Community detection and graph partitioning,” *EPL (Europhysics Lett.*, vol. 103, no. 2, 2013.

[33] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, and A. Baig, “Community detection in networks: A multidisciplinary review,” *J. Netw. Comput. Appl.*, vol. 108, no. September 2017, pp. 87–111, 2018, doi: 10.1016/j.jnca.2018.02.011.

[34] J. Liu, W. Zhong, H. A. Abbass, and D. G. Green, “Separated and overlapping community detection in complex networks using multiobjective Evolutionary Algorithms,” *IEEE Congr. Evol. Comput.*, pp. 1–7, 2010, doi: 10.1109/CEC.2010.5586522.

[35] M. E. J. Newman, “Spectral methods for community detection and graph partitioning,” *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 88, no. 4, 2013, doi: 10.1103/PhysRevE.88.042822.

[36] M. E. J. Newman and M. Girvan, “Finding and evaluating community structure in networks,” *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 69, no. 2 2, pp. 1–16, 2004, doi: 10.1103/PhysRevE.69.026113.

[37] A. Karaaslanli and S. Aviyente, “Constrained spectral clustering for dynamic community detection,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8474–8478.

[38] Y. Li, K. He, K. Kloster, D. Bindel, and J. Hopcroft, “Local Spectral Clustering for Overlapping Community Detection,” *ACM Trans. Knowl. Discov. Data*, vol. 12, no. 2, pp. 1–27, 2018, doi: 10.1145/3106370.

[39] M. Okuda, S. Satoh, Y. Sato, and Y. Kidawara, “Community Detection Using Restrained Random-Walk Similarity,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 89–103, 2021, doi: 10.1109/TPAMI.2019.2926033.

[40] Y. Xin, Z. Xie, and J. Yang, “An adaptive random walk sampling method on dynamic community detection,” *Expert Syst. Appl.*, vol. 58, pp. 10–19, 2016, doi: 10.1016/j.eswa.2016.03.033.

[41] C. Pizzuti and V. Pietro Bucci, “A Multi-objective Genetic Algorithm for Community Detection in Networks,” in *21st IEEE International Conference on Tools with Artificial Intelligence*, 2009, no. October, doi: 10.1109/ICTAI.2009.58.

[42] C. Pizzuti, “GA-Net : A Genetic Algorithm for Community Detection in Social Networks,” in *Parallel Problem Solving from Nature - PPSN X, 10th International Conference*, 2008, pp. 1081–1090, doi: 10.1007/978-3-540-87700-4.

[43] C. Pizzuti, “A Multiobjective Genetic Algorithm to Find Communities in Complex Networks,” *IEEE Trans. Evol. Comput.*, vol. 16, no. 3, pp. 418–430, 2012, doi: 10.1109/TEVC.2011.2161090.

[44] X. Meng, L. Dong, Y. Li, and W. W. Guo, “A genetic algorithm using K-path initialization for community detection in complex networks,” *Cluster Comput.*, vol. 20, no. 1, pp. 311–320, 2016, doi: 10.1007/s10586-016-0698-y.

[45] M. Guerrero, F. G. Montoya, R. Baños, A. Alcayde, and C. Gil, “Adaptive community detection in complex networks using genetic algorithms,” *Neurocomputing*, vol. 266, pp. 101–113, 2017, doi: 10.1016/j.neucom.2017.05.029.

[46] A. Said, R. Ayaz, O. Maqbool, and A. Daud, “CC-GA : A Clustering Coefficient based Genetic Algorithm for Detecting Communities in Social Networks,” *Appl. Soft Comput. J.*, vol. 63, pp. 59–70, 2018, doi: 10.1016/j.asoc.2017.11.014.

[47] M. Moradi, S. Parsa, and M. Rostami, “An Improved Multi-objective Genetic Algorithm for Revealing Community Structures of Complex,” *J. ofWaterway, Port, Coastal, Ocean Eng.*, vol. 127, no. February, pp. 45–52, 2020.

[48] A. Panizo-LLedot, G. Bello-Orgaz, and D. Camacho, “A Multi-Objective Genetic Algorithm for detecting dynamic communities using a local search driven immigrant’s scheme,” *Futur. Gener. Comput. Syst.*, vol. 110, no. November, pp. 960–975, 2019, doi: 10.1016/j.future.2019.10.041.

[49] G. Bello-orgaz, S. Salcedo-sanz, and D. Camacho, “A Multi-Objective Genetic Algorithm for overlapping community detection based on edge encoding,” *Inf. Sci. (Ny).*, vol. 462, pp. 290–314, 2018, doi: 10.1016/j.ins.2018.06.015.

[50] M. S. M. Sangeetha, “A weak clique based multi objective genetic algorithm for overlapping community detection in complex networks,” *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 6, pp. 6761–6771, 2021, doi: 10.1007/s12652-020-02301-7.

[51] M. Gong, L. Ma, Q. Zhang, and L. Jiao, “Community detection in networks by using multiobjective evolutionary algorithm with decomposition,” *Physica A*, vol. 391, no. 15, pp. 4050–4060, 2012, doi: 10.1016/j.physa.2012.03.021.

[52] Q. Cai, L. Ma, M. Gong, and D. Tian, “A survey on network community detection based on evolutionary computation,” *Int. J. Bio-Inspired Comput.*, vol. 8, no. 2, pp. 84–98, 2014, doi: 10.1504/IJBIC.2016.076329.

[53] A. I. Hafez, H. M. Zawbaa, A. E. Hassanien, and A. Aly, “Networks community detection using artificial bee colony swarm optimization,” in *The Fifth International Conference on Innovations in Bio-Inspired Computing and Applications IBICA*, 2014, pp. 229–239.

[54] A. Reihanian, M. Reza, and H. S. Aghdasi, “Community detection in social networks with node attributes based on multi- objective biogeography based optimization,” *Eng. Appl. Artif. Intell.*, vol. 62, pp. 51–67, 2017, doi: 10.1016/j.engappai.2017.03.007.

[55] P. G. Sun, “Community detection by fuzzy clustering,” *Physica A*, vol. 419, pp. 408–416, 2015, doi: 10.1016/j.physa.2014.10.009.

[56] H. Zhang, X. Chen, J. Li, and B. Zhou, “Fuzzy community detection via modularity guided membership-degree,” *Pattern Recognit. Lett.*, vol. 70, pp. 66–72, 2016.

[57] W. Luo, S. Member, Z. Yan, C. Bu, and D. Zhang, “Community Detection by Fuzzy Relations,” *IEEE Trans. Emerg. Top. Comput.*, vol. 8, no. 2, pp. 478–492, 2017, doi: 10.1109/TETC.2017.2751101.

[58] A. Biswas and B. Biswas, “FuzAg : Fuzzy Agglomerative Community Detection by Exploring the Notion of Self-Membership,” *IEEE Trans. FUZZY Syst.*, vol. 26, no. 5, pp. 2568–2577, 2018.

[59] M. Al-ayyoub, M. Al-andoli, and Y. Jararweh, “Improving fuzzy C-mean-based community detection in social networks using dynamic parallelism,” *Comput. Electr. Eng.*, vol. 74, pp. 533–546, 2019.

[60] Y. Tian, S. Yang, X. Zhang, and S. Member, “An Evolutionary Multiobjective Optimization Based Fuzzy Method for Overlapping Community Detection,” *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 11, 2019, doi: 10.1109/TFUZZ.2019.2945241.

[61] S. Yazdanparast, T. C. Havens, and M. Jamalabdollahi, “Soft Overlapping Community Detection in Large-Scale Networks via Fast Fuzzy Modularity Maximization,” *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 6, pp. 1533–1543, 2021, doi: 10.1109/TFUZZ.2020.2980502.

[62] M. Naderipour, M. Hossein, F. Zarandi, and S. Bastani, “Fuzzy community detection on the basis of similarities in structural / attribute in large - scale social networks,” *Artif. Intell. Rev.*, pp. 1–35, 2021, doi: 10.1007/s10462-021-09987-x.

[63] B. Cai, Y. Wang, L. Zeng, Y. Hu, and H. Li, “Edge classification based on Convolutional Neural Networks for community detection in complex network,” *Physica A*, vol. 556, p. 124826, 2020, doi: 10.1016/j.physa.2020.124826.

[64] P. Wang, B. Kong, C. Bao, Z. L.H, and C. . Wang, “Community Detection Based On Graph Neural Network,” in *IEEE 6th International Conference on Intelligent Computing and Signal Processing*, 2021, pp. 89–93.

[65] Z. Wang *et al.*, “Evolutionary Markov Dynamics for Network Community Detection,” *IEEE Trans. Knowl. Data Eng.*, 2020, doi: 10.1109/TKDE.2020.2997043.

[66] P. Ji, S. Zhang, and Z. Zhou, “A decomposition-based ant colony optimization algorithm for the multi-objective community detection,” *J. Ambient Intell. Humaniz. Comput.*, vol. 11,

no. 1, pp. 173–188, 2019, doi: 10.1007/s12652-019-01241-1.

[67] M. Ebrahimi, M. R. Shahmoradi, and Z. Heshmati, “A novel method for overlapping community detection using Multi-objective optimization,” *Physica A*, vol. 505, pp. 825–835, 2018, doi: 10.1016/j.physa.2018.03.033.

[68] S. Rahimi, A. Abdollahpouri, and P. Moradi, “A multi-objective particle swarm optimization algorithm for community detection in complex networks,” *Swarm Evol. Comput.*, vol. 39, no. February 2017, pp. 297–309, 2018, doi: 10.1016/j.swevo.2017.10.009.

[69] A. Verma and K. K. Bharadwaj, “Identifying community structure in a multi-relational network employing non-negative tensor factorization and GA k-means clustering,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 7, no. 1, p. e1196, 2017, doi: 10.1002/widm.1196.

[70] S. Liu and Z. Li, “A Modified Genetic Algorithm For Community Detection In Complex Networks,” in *International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*, 2017.

[71] Z. Ghaffaripour, A. Abdollahpouri, and P. Moradi, “A Multi-objective Genetic Algorithm for Community Detection in Weighted Networks,” in *Eighth International Conference on Information and Knowledge Technology (IKT)*, 2016, pp. 193–199.

[72] Y. Li, Y. Wang, J. Chen, L. Jiao, and R. Shang, “Overlapping community detection through an improved multi-objective quantum-behaved particle swarm optimization,” *J. Heuristics*, vol. 21, pp. 549–575, 2015, doi: 10.1007/s10732-015-9289-y.

[73] E. A. Hassan, A. I. Hafez, A. E. Hassanien, and A. A. Fahmy, “Community Detection Algorithm Based on Artificial Fish Swarm Optimization,” in *Intelligent Systems*, 2014, pp. 509–521.

[74] L. Yun, L. Gang, and L. Song-yang, “A genetic algorithm for community detection in complex networks,” *J. Cent. South Univ.*, vol. 20, no. 5, pp. 1269–1276, 2013, doi: 10.1007/s11771-013-1611-y.

[75] J. Li and Y. Song, “Community detection in complex networks using extended compact genetic algorithm,” *Soft Comput*, vol. 17, pp. 925–937, 2013, doi: 10.1007/s00500-012-0942-1.

[76] L. A. Zadeh, “The role of fuzzy logic in modeling, identification and control,” in *Advances*

in Fuzzy Systems — Applications and Theory, vol. 6, no. 3, 1994, pp. 191–203.

[77] M. Ghane, H. Vahdat, N. Majid, and A. Nezhad, “Detecting overlapping communities in LBSNs by fuzzy subtractive clustering,” *Soc. Netw. Anal. Min.*, vol. 8, no. 1, pp. 1–11, 2018, doi: 10.1007/s13278-018-0502-5.

[78] N. Binesh and M. Rezghi, “Fuzzy clustering in community detection based on nonnegative matrix factorization with two novel evaluation criteria,” *Appl. Soft Comput. J.*, vol. 69, pp. 689–703, 2018, doi: 10.1016/j.asoc.2016.12.019.

[79] H. Li, R. Zhang, Z. Zhao, and X. Liu, “LPA-MNI : An Improved Label Propagation Algorithm Based on Modularity and Node Importance for Community Detection,” *Entropy*, vol. 23, no. 5, p. 497, 2021.

[80] T. Ma, Q. Liu, J. Cao, Y. Tian, and A. Al-dhelaan, “LGIEM : Global and local node influence based community detection,” *Futur. Gener. Comput. Syst.*, vol. 105, pp. 533–546, 2020, doi: 10.1016/j.future.2019.12.022.

[81] M. Lu, Z. Zhang, Z. Qu, and Y. Kang, “LPANNI: Overlapping Community Detection Using Label Propagation in Large-Scale Complex Networks,” *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 9, pp. 1736–1749, 2018, doi: 10.1109/TKDE.2018.2866424.

[82] A. Moayedikia, “Multi-objective community detection algorithm with node importance analysis in attributed networks,” *Appl. Soft Comput. J.*, vol. 67, pp. 434–451, 2018, doi: 10.1016/j.asoc.2018.03.014.

[83] K. Berahmand, A. Bouyer, and M. Vasighi, “Community Detection in Complex Networks by Detecting and Expanding Core Nodes Through Extended Local Similarity of Nodes,” *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 4, pp. 1021–1033, 2018, doi: 10.1109/TCSS.2018.2879494.

[84] X. Zhang, K. Zhou, H. Pan, L. Zhang, X. Zeng, and Y. Jin, “A Network Reduction-Based Multiobjective Evolutionary Algorithm for Community Detection in Large-Scale Complex Networks,” *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 703–716, 2018.

[85] D. Chen, F. Zou, R. Lu, L. Yu, Z. Li, and J. Wang, “Multi-objective optimization of community detection using discrete teaching – learning-based optimization with decomposition,” *Inf. Sci. (Ny).*, vol. 369, pp. 402–418, 2016, doi: 10.1016/j.ins.2016.06.025.

[86] Y. Gao, X. Yu, and H. Zhang, “Overlapping community detection by constrained personalized PageRank,” *Expert Syst. Appl.*, vol. 173, no. February, p. 114682, 2021, doi: 10.1016/j.eswa.2021.114682.

[87] Y. Su, X. Zhang, and R. Cheng, “A Parallel multi-objective evolutionary algorithm for community detection in large-scale complex networks,” *Inf. Sci. (Ny).*, vol. 576, pp. 374–392, 2021, doi: 10.1016/j.ins.2021.06.089.

[88] H. Qing and J. Wang, “Consistency of spectral clustering for directed network community detection,” vol. arXiv prep, 2021.

[89] X. Zhu, Y. Ma, and Z. Liu, “A novel evolutionary algorithm on communities detection in signed networks,” *Physica A*, vol. 503, no. 71471106, pp. 938–946, 2018, doi: 10.1016/j.physa.2018.08.112.

[90] M. Yuanyuan and L. Xiyu, “Quantum inspired evolutionary algorithm for community detection in complex networks,” *Phys. Lett. A*, vol. 382, no. 34, pp. 2305–2312, 2018, doi: 10.1016/j.physleta.2018.05.044.

[91] S. Bilal and M. Abdelouahab, “Evolutionary algorithm and modularity for detecting communities in networks,” *Physica A*, vol. 473, pp. 89–96, 2017, doi: 10.1016/j.physa.2017.01.018.

[92] S. Gupta, S. Mittal, T. Gupta, I. Singhal, and B. Khatri, “Parallel quantum-inspired evolutionary algorithms for community detection in social networks,” *Appl. Soft Comput. J.*, vol. 61, pp. 331–353, 2017, doi: 10.1016/j.asoc.2017.07.035.

[93] S. Chobe and J. Zhan, “Advancing community detection using Keyword Attribute Search,” *J. Big Data*, vol. 6, no. 1, pp. 1–33, 2019, doi: 10.1186/s40537-019-0243-y.

[94] I. Falih, N. Grozavu, R. Kanawati, and Y. Bennani, “Community detection in Attributed Network,” in *Companion Proceedings of the The Web Conference*, 2018, pp. 1299–1306, doi: 10.1145/3184558.3191570.

[95] H. Lin, Y. Zhan, Z. Zhao, Y. Chen, and C. Dong, “Overlapping Community Detection Based on Attribute Augmented Graph,” *Entropy*, vol. 23, no. 6, p. 680, 2021.

[96] M. Contisciani, E. A. Power, and C. De Bacco, “Community detection with node attributes in multilayer networks,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–16, 2020, doi: 10.1038/s41598-020-72626-y.

[97] P. Li, L. Huang, C. Wang, D. Huang, and J. Lai, “Community Detection Using Attribute Homogenous Motif,” *IEEE Access*, vol. 6, pp. 47707–47716, 2018, doi: 10.1109/ACCESS.2018.2867549.

[98] Y. Asim, A. Majeed, R. Ghazal, and A. K. Malik, “Community Detection in Networks using Node Attributes and Modularity,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 1, pp. 382–388, 2017.

[99] M. Ganj, J. Bailey, and P. J. Stuckey, “Lagrangian Constrained Community Detection,” in *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018, pp. 2983–2990.

[100] M. Ganji, J. Bailey, and P. J. Stuckey, “A Declarative Approach to Constrained Community Detection,” in *International Conference on Principles and Practice of Constraint Programming*, 2017, pp. 477–494.

[101] E. Eaton and R. Mansbach, “A Spin-Glass Model for Semi-Supervised Community Detection,” in *the AAAI Conference on Artificial Intelligence*, 2012, pp. 900–906.

[102] J. H. Chin and K. Ratnavelu, “A semi-synchronous label propagation algorithm with constraints for community detection in complex networks,” *Nat. Publ. Gr.*, vol. 7, no. 1, pp. 1–12, 2017, doi: 10.1038/srep45836.

[103] J. H. Chin and K. Ratnavelu, “Detecting Community Structure by Using a Constrained Label Propagation Algorithm,” *PLoS One*, vol. 11, no. 5, p. e0155320, 2016, doi: 10.1371/journal.pone.0155320.

[104] S. Cafieri *et al.*, “Adding Cohesion Constraints to Models for Modularity Maximization in Networks,” *J. Complex Networks, Proc. Ind. Revolut. Bus. Manag. 11th Annu. PwR Dr. Symp. (PWRDS).*, vol. 3, no. 3, pp. 388–410, 2015.

[105] B. Thomas and S. Setzer, “Constrained fractional set programs and their application in local clustering and community detection,” in *International Conference on Machine Learning*, 2013, vol. PMLR, pp. 624–632.

[106] M. Ciglan and K. Nørvåg, “Fast detection of size-constrained communities in large networks,” in *International Conference on Web Information Systems Engineering*, 2010, pp. 91–104.

[107] M. J. Barber and J. W. Clark, “Detecting network communities by propagating labels

under constraints," *Phys. Rev. E*, vol. 80, no. 2, p. 026129, 2009.

[108] F. Cheng, T. Cui, Y. Su, Y. Niu, and X. Zhang, "A local information based multi-objective evolutionary algorithm for community detection in complex networks," *Appl. Soft Comput. J.*, vol. 69, pp. 357–367, 2018, doi: 10.1016/j.asoc.2018.04.037.

[109] L. I. U. Han, Y. Fan, and L. I. U. Ding, "Genetic Algorithm Optimizing Modularity for Community Detection in Complex Networks," in *The 35th Chinese Control Conference*, 2016, no. 1, pp. 1252–1256.

[110] M. E. J. Newman, "Modularity and community structure in networks," in *Proceedings of the National Academy of Sciences PNAS*, 2006, vol. 103, no. 23, pp. 8577–8582.

[111] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *J. Stat. Mech. Theory Exp.*, no. 9, pp. 219–228, 2005, doi: 10.1088/1742-5468/2005/09/P09008.

[112] P. Wu and L. Pan, "Multi-objective community detection method by integrating users ' behavior attributes," *Neurocomputing*, vol. 210, pp. 13–25, 2016, doi: 10.1016/j.neucom.2015.11.128.

[113] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *J. Am. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.

[114] L. Hubert and P. Arabie, "Comparing Partitions," *J. Classif.*, vol. 2, no. 1, pp. 193–218, 1985.

[115] M. Meilă, "Comparing clusterings by the variation of information," in *Learning Theory and Kernel Machines*, 2003, pp. 173–187.

[116] S. Van Dongen, "Performance Criteria for Graph Clustering and Markov Cluster Experiments," *Natl. Res. Inst. Math. Comput. Sci.*, 2000.

[117] C. Shi, P. S. Yu, Y. Cai, Z. Yan, and B. Wu, "On Selection of Objective Functions in Multi-Objective Community Detection," in *the 20th ACM international conference on Information and knowledge management*, 2011, vol. 1, no. 2, pp. 2–5.

[118] V. Labatut, "Generalized Measures for the Evaluation of Community Detection Methods," *Int. J. Soc. Netw. Min.*, vol. 2, no. 1, pp. 44–63, 2016.

[119] G. Xu, J. Guo, and P. Yang, "TNS-LPA : An Improved Label Propagation Algorithm for

Community Detection Based on Two-Level Neighbourhood Similarity," *IEEE Access*, vol. 9, pp. 23526–23536, 2021, doi: 10.1109/ACCESS.2020.3045085.

[120] R. Mittal and M. P. S. Bhatia, "Classification and Comparative Evaluation of Community Detection Algorithms," *Arch. Comput. Methods Eng.*, vol. 28, no. 3, pp. 1417–1428, 2021, doi: 10.1007/s11831-020-09421-5.

[121] P. Nerurkar, M. Chandane, and S. Bhirud, "Community detection using node attributes: A non-negative matrix factorization approach," *Comput. Intell. Theor. Appl. Futur. Dir.*, vol. Volume I, pp. 275–285, 2019, doi: 10.1007/978-981-13-1132-1_22.

[122] J. C. Leão, M. A. Brandão, P. O. S. V. de Melo, and A. H. F. Laender, "Improving Community Detection by Mining Social Interactions," *arXiv Prepr. arXiv1810.02002*, 2018.

[123] S. Kumar and P. Kumar, "Upper approximation based privacy preserving in online social networks," *Expert Syst. Appl.*, vol. 88, pp. 276–289, 2017, doi: 10.1016/j.eswa.2017.07.010.

[124] X. Wen *et al.*, "A Maximal Clique Based Multiobjective Evolutionary Algorithm for Overlapping Community Detection," *IEEE Trans. Evol. Comput.*, vol. 21, no. 3, pp. 363–377, 2017.

[125] V. Labatut, "Generalised measures for the evaluation of community detection methods," *Int. J. Soc. Netw. Min.*, vol. 2, no. 1, pp. 44–63, 2015.

[126] J. Xie and B. K. Szymanski, "Towards Linear Time Overlapping Community Detection in Social Networks," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2012, pp. 25–36.

[127] A. Delmotte, E. W. Tate, S. N. Yaliraki, and M. Barahona, "Protein multi-scale organization through graph partitioning and robustness analysis: Application to the myosin-myosin light chain interaction," *Phys. Biol.*, vol. 8, no. 5, p. 055010, 2011.

[128] A. Karatas and S. Sahin, "Application Areas of Community Detection: A Review," in *International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism, IBIGDELFT 2018 - Proceedings*, 2018, pp. 65–70, doi: 10.1109/IBIGDELFT.2018.8625349.

[129] A. Karataş and S. Şahin, "A Review on Social Bot Detection Techniques and Research Directions," in *Int. Security and Cryptology Conference Turkey*, 2017, no. February, pp.

156–161.

[130] D. Sarma, W. Alam, I. Saha, M. N. Alam, M. J. Alam, and S. Hossain, “Bank Fraud Detection using Community Detection Algorithm,” in *The Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020)*, 2020, pp. 642–646.

[131] N. Haq and Z. J. Wang, “COMMUNITY DETECTION FROM GENOMIC DATASETS ACROSS HUMAN CANCERS,” in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2016, pp. 1147–1150.

[132] S. Javed *et al.*, “Cellular community detection for tissue phenotyping in colorectal cancer histology images,” *Med. Image Anal.*, vol. 63, p. 101696, 2020, doi: 10.1016/j.media.2020.101696.

[133] M. Ozer, N. Kim, and H. Davulcu, “Community detection in political Twitter networks using Nonnegative Matrix Factorization methods,” in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, 2016, pp. 81–88, doi: 10.1109/ASONAM.2016.7752217.

[134] A. Bakhthemmat and M. Izadi, “Communities Detection for Advertising by Futuristic Greedy Method with Clustering Approach,” *Big Data*, vol. 9, no. 1, pp. 22–40, 2021, doi: 10.1089/big.2020.0133.

[135] L. Zhang, J. Priestley, J. Demaio, S. Ni, and X. Tian, “Measuring Customer Similarity and Identifying Cross-Selling Products by Community Detection,” *Big Data*, vol. 9, no. 2, pp. 132–143, 2021, doi: 10.1089/big.2020.0044.

[136] C. Remy, B. Rym, and L. Matthieu, “Tracking bitcoin users activity using community detection on a network of weak signals,” in *Int. Workshop on Complex Networks and their Applications*, 2017, pp. 166–177.

[137] C. De Bacco, E. A. Power, D. B. Larremore, and C. Moore, “Community detection, link prediction, and layer interdependence in multilayer networks,” *Phys. Rev. E*, vol. 95, no. 4, p. 042317, 2018.

[138] Y. Qin, “ICN-driven group psychology visualization analysis mechanism using reinforcement learning,” *Internet Technol. Lett.*, vol. 4, no. 5, p. e284, 2021, doi: 10.1002/itl2.284.

[139] P. Ahlgren, Y. Chen, C. Colliander, and N. J. van Eck, “Enhancing direct citations: A comparison of relatedness measures for community detection in a large set of PubMed publications,” *Quant. Sci. Stud.*, vol. 1, no. 2, pp. 714–729, 2020, doi: 10.1162/qss_a_00027.

[140] T. Chakraborty and A. Chakraborty, “OverCite: Finding overlapping communities in citation network,” in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM*, 2013, pp. 1124–1131, doi: 10.1145/2492517.2500255.

[141] M. Gupta, C. C. Aggarwal, J. Han, and Y. Sun, “Evolutionary clustering and analysis of bibliographic networks,” in *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011*, 2011, no. July, pp. 63–70, doi: 10.1109/ASONAM.2011.12.

[142] A. E. Coca and L. Zhao, “Musical rhythmic pattern extraction using relevance of communities in networks,” *Inf. Sci. (Ny).*, vol. 329, pp. 819–848, 2016, doi: 10.1016/j.ins.2015.09.030.

[143] J. de Berardinis, M. Vamvakaris, A. Cangelosi, and E. Coutinho, “Unveiling the Hierarchical Structure of Music by Multi-Resolution Community Detection,” *Trans. Int. Soc. Music Inf. Retr.*, vol. 3, no. 1, pp. 82–97, 2020, doi: 10.5334/tismir.41.

[144] J. Zhang, P. S. Yu, and Y. Lv, “Enterprise community detection,” in *Proceedings - International Conference on Data Engineering*, 2017, vol. 2, pp. 219–222, doi: 10.1109/ICDE.2017.79.

[145] R. Hu, S. Pan, G. Long, X. Zhu, J. Jiang, and C. Zhang, “Co-clustering enterprise social networks,” in *International Joint Conference on Neural Networks*, 2016, pp. 107–114, doi: 10.1109/IJCNN.2016.7727187.

[146] H. Sarvari, E. Abozinadah, A. Mbaziira, and D. McCoy, “Constructing and analyzing criminal networks,” in *Proceedings - IEEE Symposium on Security and Privacy*, 2014, vol. 2014-Janua, pp. 84–91, doi: 10.1109/SPW.2014.22.

[147] C. André, R. Pinheiro, and R. De Janeiro, “Community Detection to Identify Fraud Events in Telecommunications Networks,” in *SAS Global Forum 2012 Customer Intelligence*, 2012, p. 106.

[148] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” in *Proceedings of the national academy of sciences*, 2002, vol. 99, no. 12, pp. 7821–7826.

[149] A. Lancichinetti, S. Fortunato, and F. Radicchi, “Benchmark graphs for testing community detection algorithms,” *Phys. Rev.*, vol. 78, no. 4, p. 046110, 2008, doi: 10.1103/PhysRevE.78.046110.

[150] W. W. Zachary, “An information flow model for conflict and fission in small groups,” *J. Anthropol. Res.*, vol. 33, no. 4, pp. 452–473, 1977.

[151] M. J. Herrgård *et al.*, “A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology,” *Nat. Biotechnol.*, vol. 26, no. 10, pp. 1155–1160, 2008, doi: 10.1038/nbt1492.

[152] Q. Wu, X. Qi, E. Fuller, and C. Q. Zhang, “‘Follow the leader’: A centrality guided clustering and its application to social network analysis,” *Sci. World J.*, vol. 2013, 2013, doi: 10.1155/2013/368568.

[153] J. Shuja, E. Alanazi, W. Alasmary, and A. Alashaikh, “COVID-19 open source data sets: a comprehensive survey,” *Appl. Intell.*, vol. 51, no. 3, pp. 1296–1325, 2020, doi: 10.1007/s10489-020-01862-6.

[154] B. Kleinberg, I. Van Der Vegt, and M. Mozes, “Measuring Emotions in the COVID-19 Real World Worry Dataset,” *Phys. Rev. Lett.*, vol. arXiv prep, 2020.

[155] J. M. Banda *et al.*, “A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research — An International Collaboration,” *Epidemiologia*, vol. 2, no. 3, pp. 315–324, 2021.

[156] S. Alqurashi, A. Alhindi, and E. Alanazi, “Large Arabic Twitter Dataset on COVID-19,” vol. arXiv prep, 2020.

[157] E. Chen, K. Lerman, and E. Ferrara, “Tracking Social Media Discourse About the COVID-19 Pandemic : Development of a Public Coronavirus Twitter Data Set,” *JMIR Public Heal. Surveill.*, vol. 6, no. 2, p. e19273, 2020, doi: 10.2196/19273.

[158] K. Zarei, R. Farahbakhsh, N. Crespi, and G. Tyson, “A First Instagram Dataset on COVID-19,” vol. arXiv prep, pp. 2–5, 2020.

[159] World Health Organization, “Contact tracing in the context of COVID-19: Interim guidance,” *Paediatr. Fam. Med.*, no. WHO/2019-nCoV/Contact_Tracing/2020.1, pp. 1–11, 2021, doi: 10.15557/PiMR.2020.0005.

[160] C. Shi, Z. Yan, Y. Cai, and B. Wu, “Multi-objective community detection in complex networks,” *Appl. Soft Comput. J.*, vol. 12, no. 2, pp. 850–859, 2011, doi: 10.1016/j.asoc.2011.10.005.

[161] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, “A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT , Drones , AI , Blockchain , and 5G in Managing Its Impact,” *Spec. Sect. Deep Learn. ALGORITHMS INTERNET Med. THINGS*, vol. 8, pp. 90225–90265, 2020.

[162] M. Shahroz *et al.*, “COVID-19 digital contact tracing applications and techniques : A review post initial deployments,” *Transp. Eng.*, vol. 5, p. 100072, 2021, doi: 10.1016/j.treng.2021.100072.

[163] M. Usman, W. Iqbal, Q. Mary, and J. Qadir, “Leveraging Data Science To Combat COVID-19 : A Comprehensive Review,” *IEEE Trans. Artif. Intell.*, vol. 1, no. 1, pp. 85–103, 2020, doi: 10.13140/RG.2.2.12685.28644/4.

[164] Authority, Information, and EGovernment, “BeAware Bahrain.” 2021.

[165] “Contact Tracing - Ministry of Health,” *Ministry of Health - Kingdom of Bahrain*, 2020. .

[166] L. WU, Q. ZHANG, C.-H. CHEN, K. GUO, and D. WANG, “Deep Learning Techniques for Community Detection in Social Networks,” *IEEE Access*, vol. 8, pp. 96016–96026, 2020, doi: 10.1109/ACCESS.2020.2996001.

[167] W. Viles and J. O’Malley, “Constrained Community Detection in Social Networks,” vol. arXiv prep, 2017.

[168] K. Nakata and T. Murata, “Fast Optimization of Hamiltonian for Constrained Community Detection,” in *Complex Networks VI*, 2015, pp. 79–89.

[169] M. Newman, “Fast algorithm for detecting community structure in networks,” *Phys. Rev. E*, vol. 69, no. 6, p. 066133, 2003.

[170] V. D. Blondel, J. Guillaume, and E. Lefebvre, “Fast unfolding of communities in large networks,” *J. Stat. Mech. theory Exp.*, vol. 10, p. 10008, 2008.

[171] C. K. Tsung, H. J. Ho, C. Y. Chen, T. W. Chang, and S. L. Lee, “Detecting overlapping communities in modularity optimization by reweighting vertices,” *Entropy*, vol. 22, no. 8, p. 819, 2020, doi: 10.3390/E22080819.

[172] Y. C. Chen, Z. Guan, Y. Peng, X. Shao, and M. Hasseb, “Technology and system of constraint programming for industry production scheduling — Part I_ A brief survey and potential directions,” *Front. Mech. Eng. China*, vol. 5, no. 1, pp. 455–464, 2010.

[173] Y. Peng, D. Lu, and Y. Chen, “A Constraint Programming Method for Advanced Planning and Scheduling System with Multilevel Structured Products,” *Discret. Dyn. Nat. Soc.*, 2014.

[174] G. Di Pillo, “Exact Penalty Methods,” *Algorithms Contin. Optim.*, vol. 434, no. Springer, Dordrecht, pp. 209–253, 1994.

[175] L. Danon, D. Albert, and J. Duch, “Comparing community structure identification,” *J. Stat. Mech. Theory Exp.*, vol. 9, p. P09008, 2005, doi: 10.1088/1742-5468/2005/09/P09008.

[176] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, “GemSec: Graph embedding with self clustering,” *Proc. 2019 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2019*, pp. 65–72, 2019, doi: 10.1145/3341161.3342890.

[177] P.-Z. Li, L. Huang, C.-D. Wang, and J.-H. Lai, “EdMot : An Edge Enhancement Approach for Motif-aware Community Detection,” in *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 479–487, doi: 10.1145/3292500.3330882.

[178] F. Ye, C. Chen, and Z. Zheng, “Deep autoencoder-like nonnegative matrix factorization for community detection,” *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 1393–1402, 2018, doi: 10.1145/3269206.3271697.

[179] R. George *et al.*, “ScienceDirect ScienceDirect A Comparative Evaluation of Community Detection Algorithms in Social Networks,” *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 1157–1165, 2020, doi: 10.1016/j.procs.2020.04.124.