



*Dissertation on*

**“Community detection in dynamic networks”**

*Submitted in partial fulfilment of the requirements for the award of degree of*

**Bachelor of Technology  
in  
Computer Science & Engineering**

**UE18CS390A – Capstone Project Phase - 1**

*Submitted by:*

<b>Mahammad Thufail</b>	<b>PES2201800646</b>
<b>Purushotham S</b>	<b>PES2201800480</b>
<b>Manne Vasanth</b>	<b>PES2201800425</b>
<b>Pulle Manikya Sri Manasa</b>	<b>PES2201800468</b>

*Under the guidance of*

**Prof. Sreenath MV**  
Assistant Professor  
PES University

**January - May 2021**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
FACULTY OF ENGINEERING  
PES UNIVERSITY**

(Established under Karnataka Act No. 16 of 2013)  
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India



## PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)

Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

### FACULTY OF ENGINEERING

# CERTIFICATE

*This is to certify that the dissertation entitled*

**‘Community detection in dynamic networks’**

*is a bonafide work carried out by*

<b>Mahammad Thufail</b>	<b>PES2201800646</b>
<b>Purushotham S</b>	<b>PES2201800480</b>
<b>Manne Vasanth</b>	<b>PES2201800425</b>
<b>Pulle Manikya Sri Manasa</b>	<b>PES2201800468</b>

In partial fulfilment for the completion of sixth semester Capstone Project Phase - 1 (UE18CS390A) in the Program of Study -Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Jan. 2021 – May. 2021. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 6<sup>th</sup> semester academic requirements in respect of project work.

Signature  
**Prof. Sreenath MV**  
Assistant Professor

Signature  
Dr. Sandesh B J  
Chairperson  
**External Viva**

Signature  
Dr. B K Keshavan  
Dean of Faculty

**Name of the Examiners**

**Signature with Date**

1. \_\_\_\_\_

\_\_\_\_\_

2. \_\_\_\_\_

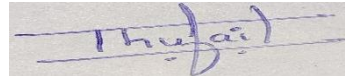
\_\_\_\_\_

## DECLARATION

We hereby declare that the Capstone Project Phase - 1 entitled “**Community detection in dynamic networks**” has been carried out by us under the guidance of Prof. Sreenath MV, Assistant Professor and submitted in partial fulfilment of the course requirements for the award of degree of **Bachelor of Technology in Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester January – May 2021. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

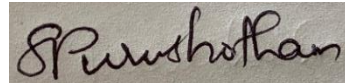
**PES2201800646**

**Mahammad Thufail**



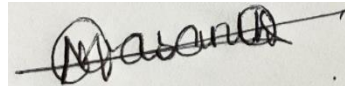
**PES2201800480**

**Purushotham S**



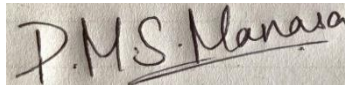
**PES2201800425**

**Manne Vasanth**



**PES2201800468**

**Pulle Manikya Sri  
Manasa**



## **ACKNOWLEDGEMENT**

I would like to express my gratitude to Prof. Sreenath MV, Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE18CS390A - Capstone Project Phase – 1.

I am grateful to the Capstone Project Coordinator, Dr.Sarasvathi V, Associate Professor, for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Sandesh B J, Chairperson, Professor, Department of Computer Science and Engineering, PES University, for all the knowledge and support I have received from the department. I would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University for providing to me various opportunities and enlightenment every step of the way. Finally, this Capstone Project could not have been completed without the continual support and encouragement I have received from my family and friends.

## **ABSTRACT**

The importance of social network research, both as a theoretical viewpoint and as a methodological toolkit, for understanding and evaluating terror groups, as well as developing counterterror policies and practises to identify and disrupt terror attacks, is a key issue in monitoring transnational terror patterns. Terrorist activities has led to the creation of a number of high-end methodologies for studying terrorist organisations and networks around the world. Social Network Analysis is one of the most powerful and predictive methods for combating extremism in social networks, according to existing studies. In terms of a global and regional context, the study examined various SNA steps for predicting the key players/main actors of terrorist networks. The applicability and viability of SNA tools for online and offline social networks were demonstrated in a comparative analysis of SNA tools. Data mining techniques can be used to integrate temporal analysis. It has the potential to improve SNA's ability to handle the complex behaviour of online social networks.

# TABLE OF CONTENTS

Chapter No.	Title	Page No.
<b>1.</b>	<b>INTRODUCTION</b>	<b>01</b>
<b>2.</b>	<b>PROBLEM DEFINITION</b>	<b>02</b>
<b>3.</b>	<b>LITERATURE SURVEY</b>	<b>03</b>
	3.1 Understanding the composition and evolution of terrorist group networks: A rough set approach	
	3.2 Finding influential nodes in social networks based on neighborhood correlation coefficient	
	3.3 Community detection in large-scale social networks: state-of-the-art and future directions	
	3.4 Hidden community detection in social networks	
	3.5 TI-SC: top-k influential nodes selection based on community detection and scoring criteria in social networks	
	3.6 Detection of Influential Nodes Using Social Networks Analysis Based On Network Metrics	
	3.7 Influential Nodes Detection in Dynamic Social Networks	
	3.8 Performance Evaluation of Modularity Based Community Detection Algorithms in Large Scale Networks	
	3.9 A comprehensive literature review on Community detection: Approaches and applications	
	3.10 Community detection in social networks	
	3.11 Analysis of the Dynamic Influence of Social Network Nodes	
<b>4.</b>	<b>DATA</b>	<b>23</b>
	4.1 Overview	
	4.2 Dataset	
<b>5.</b>	<b>SYSTEM REQUIREMENTS SPECIFICATION</b>	<b>24</b>

<b>6.</b>	<b>SYSTEM DESIGN</b>	<b>30</b>
<b>7.</b>	<b>IMPLEMENTATION AND PSEUDOCODE</b>	<b>33</b>
<b>8.</b>	<b>CONCLUSION OF CAPSTONE PROJECT PHASE - 1</b>	<b>50</b>
<b>9.</b>	<b>PLAN OF WORK FOR CAPSTONE PROJECT PHASE - 2</b>	<b>51</b>
	<b>REFERENCES/BIBLIOGRAPHY</b>	<b>52</b>
	<b>APPENDIX A DEFINITIONS, ACRONYMS AND ABBREVIATIONS</b>	<b>53</b>

## **LIST OF TABLES**

<b>Table No.</b>	<b>Title</b>	<b>Page No.</b>
<b>1</b>	<b>Sample value vector associated with features and interpretations</b>	<b>34</b>

## **LIST OF FIGURES**

<b>Figure No.</b>	<b>Title</b>	<b>Page No.</b>
<b>1</b>	<b>A sample graph with 3 shells</b>	<b>38</b>
<b>2</b>	<b>Adding edges/nodes</b>	<b>43</b>
<b>3</b>	<b>Removing edges/nodes</b>	<b>45</b>
<b>4</b>	<b>Adding Cross Community edges</b>	<b>48</b>
<b>5</b>	<b>Remove Edge to terminal nodes</b>	<b>49</b>

# CHAPTER-1

## INTRODUCTION

Terrorism is an organised type of violence that has a direct impact on stability, a country's or community's daily routine, and security, as well as a means of instilling fear in civilians. Terrorism is a fluid phenomenon, so equipping counter-terrorism operators with the resources they need to combat it is important.

The key aim of this study is to find a mechanism for eliciting information about perpetrators in terrorist incidents by looking at terror attacks over time. The aim is to build a sociogram, or criminal network, with nodes representing terrorist organisations and edges representing generic connections between two groups.

We used a tool that helps us to find terror organisations with common operational features in clusters. Specifically, we used open access data from terrorist activities worldwide since 1970 to create a network of terrorist organisations and related information on tactics, weapons, goals, and active areas.

Each partition is linked to the terrorist groups in our model. Later, we'll try to avoid attacks by identifying the most powerful party with the greatest number of connections to other networks. Community identification is a technique for identifying groups of nodes in which the connections between nodes within a group are greater than the connections between nodes in other networks.



## **CHAPTER-2**

### **PROBLEM DEFINATION**

Using a database of terrorism attacks, elicit terrorist organisation networks, finding most influential node and community detection.

We use rough set approach to understand the composition and evolution of terrorist networks. Later on we will find influential nodes based on neighborhood correlation coefficient. Finally for Community Detection we use Louvain algorithm using locality modularity optimization.

## **CHAPTER-3**

### **LITERATURE SURVEY**

#### **3.1 Paper 1: “Understanding the composition and evolution of terrorist group networks: A rough set approach”.**

##### **3.1.1 Introduction**

The key idea behind this research is to use GTD's historical data, which includes information on terrorist attacks that have occurred since 1970, to conceptualize terrorist groups' activities over time intervals. Following that, such conceptualizations are used to explain the similarities between terrorist groups and to evoke relationships to reflect terrorists networks. Through applying the method to various time points along the continuum and examining variations among the resulting networks, the above networks can be used to research the temporal evolutions of terrorist groups' behaviors. The method is focused on Rough Set Theory and Three-way Decisions Theory, and it generates an original similarity function based on boundary region description.

##### **3.1.2 Characteristics and Implementation**

The first step is to choose the correct attributes from the Global Terrorism Database's entire collection of attributes. The precise conceptualizations of terrorist groups and their approximations are given in the second step. The specification of the time interval for considering terrorist activities and the selection of terrorist organizations for which to analyze activity are the preliminary actions for the second level. The third step constructs the boundary regions for the above rough conceptualizations using the three-way decisions theory and the rough approximations given by the second step. The similarity matrix, which reports the similarities between all pairs of considered terrorist groups, is built in the fourth step, using the newly described similarity function. Finally, the fifth step entails converting the similarity matrix into a network graph that depicts the generic relationships between terrorist groups.

### 3.1.3 Features

The proposed method was demonstrated and evaluated by comparing the results to expert information obtained from the GTD database's resources and using a Python implementation of rough set operators (created by the authors) (START project). Using a similarity function based on rough set theory, the paper proposes an original method for eliciting terrorist group networks from a database of terrorist incidents. Several illustrative examples have been given to explain and demonstrate the method.

### 3.1.4 Limitations

Based on the positive results, the technique could be used in the testing frameworks used by counter-terrorism operators. Any of the shortcomings of the results presented in this report are as follows:

- A. The System lacks in dealing with time intervals automatically,
- B. There are no much tools to examine the temporal evolution of terrorist groups' networks.
- C. The lack of ability to establish the semantics of the elicited networks' connections.

All of the above concerns should be investigated, and the writers have suggested some changes to the proposed method:

- A systematic extraction protocol was used to remove 135 attributes from the original dataset .
- The idea of a semantic similarity attribute is essential to terrorist organizations.
- Many alternatives to traditional rough sets for conceptualizing terrorist organisation behavior, such as tolerance rough sets and neighbourhood rough sets, have been evaluated.
- The development of a tool for studying terrorist group networks' temporal evolution.

## **3.2 Paper 2: “Finding influential nodes in social networks based on neighborhood correlation coefficient”.**

### **3.2.1 Introduction**

Rapid dissemination of news and advertisements via social networks has created new opportunities for social media platforms to replace traditional forms of advertising such as radio, television, and banners. It is nearly impossible to reach out to all users directly, so a clever solution is to target a small number of influential users and communicate with them in the hope that they can maximize the desired influence on others. If the target users are carefully chosen, they can disseminate the messages more widely by spreading them through their friendship networks. Users are not all equal, and some have more clout because of their personal and social characteristics, as well as their placement in strategic locations. Influential users have a greater impact on the success of information dissemination than others. In recent years, there has been a lot of interest in defining users' influence ranges and ranking them based on their influence ranges. In many cases, the structure of the connection network is the only available information for identifying influential nodes. A number of techniques for locating influential users using network structure have been proposed in the literature.

### **3.2.2 Characteristics and Implementation**

The idea behind this method is that users who share more friends with their Neighbours tend to spread information over shorter distances. Only the number of friends shared by each user and their Neighbours is considered in the cluster rank method. In this paper, we look at the properties of the users' commonalities with their Neighbours. The proposed method estimates a node's influence score based on the similarity (or correlation) of the node's and its Neighbour connection structures. The proposed method is tested on a variety of synthetic and real-world networks, and it is compared to a variety of state-of-the-art influence ranking algorithms.

By taking into account the neighborhood's detailed correlation structure and actively using it to classify the most influential nodes, the proposed algorithm makes a substantial contribution. For this, the Pearson correlation is used. A variety of centrality tests have been proposed to quantify the influence of nodes in successful information dissemination. Degree centrality is the most basic criterion for measuring prominent nodes. Nodes with lower degrees, on the other hand, often have much greater power than nodes with higher degrees. Indeed, a low-degree node in a strategic network position can facilitate knowledge spread more than higher-degree nodes. In the entropy centrality process, a node's impact spectrum is calculated by the degree distribution. With respect to its first-order Neighbours and second-order Neighbours.

### **3.2.3 Features**

Social network analysis and mining have become more and more popular nowadays, there are a wide range of possible applications in a variety of industries. Influence maximization is one of the issues that has gotten a lot of coverage in this area. The proposed approach utilizes the similarity of relations between adjacent nodes which is purely dependent on the coefficient of local clustering. The method is derived on the basis of k-shell decomposition, in which the power of a node is measured by the number of mutual relations with respect to its neighbours. Any two nodes with Neighbors which are in the same areas of the network are termed as correlated nodes in the process that is proposed. Thus all the Correlated nodes have low spreading capacities than uncorrelated nodes since they occupy a wider region of the network. As a consequence, a node with a low connection with its Neighbours could be a stronger choice for being a strong node due to its ability to disseminate messages across the network with the help of its Neighbours.

### **3.2.4 Limitations**

The proposed method's main flaw is that it requires the processing of all data in order to locate the most influential node.

### **3.3 Paper 3: “Community detection in large-scale social networks:state-of-the-art and future directions”.**

#### **3.3.1 Introduction**

The discovery of the structure of a social network is an important research area in social network analysis, and community detection is one of the most important research areas in this field. In disciplines like sociology, biology, and computer science, where processes are often described as graphs, detecting societies is critical. This is an NP-hard problem that has yet to be solved satisfactorily. This is an NP-hard problem that has yet to be solved satisfactorily. Two major factors impede this computational complexity. The first aspect is the massive scale of today's social networks, such as Facebook and Twitter, which have billions of nodes. The second reason has to do with the complex existence of social networks, which change their structure over time. As a result, group detection in social network analysis is gaining a lot of scientific interest, and a lot of research has been done in this field. This paper's main aim is to provide a concise overview of group detection algorithms in social graphs. We include a taxonomy of current models based on their computational existence (either centralized or distributed) and thus in static and dynamic social networks for this purpose. We also have a detailed overview of current community identification applications in social networks. Finally, we discuss future research paths as well as some unresolved issues.

#### **3.3.2 Characteristics and Implementation**

This study will predict actor characteristics, the appearance of any connections, the diffusion arrangements in the network, and so on. Despite the fact that the graphs have no apparent structural properties, they all share one that defines them regardless of their specific material. Identifying communities helps one to get a macro view of complex structures, which is a useful method for understanding and analyzing them. Group detection is an NP-complete graph partitioning problem from a theoretical standpoint (Fortunato 2010; Xie and Szymanski 2013;

Ben Romdhane et al. 2013; Rhouma and Romdhane 2014). Communities are a key feature of the network that can be divided into two types: structure-based communities (a cluster of nodes with more connections to each other than other nodes in a graph) and semantic-based communities (a cluster of nodes with similar semantic meaning or a group of nodes with similar interests). Since both the meaning and the relationship of the nodes are used to detect semantic communities, the result may more effectively reflect community cohesion. It's impossible to classify both topological constructs and semantic meaning at the same time due to the differences in their characteristics.

### **3.3.3 Features**

The primary goal of this survey is to present, coordinate, and review existing models for assessing populations in large-scale networks. The below are the paper's key points:

- The aim of this project is to categories and compare theoretic group detection algorithms.
- Using two strategies: centralized and dispersed, combine techniques for characterizing, identifying, and extracting populations.
- The aim of this study is to look at the detection technique in relation to the dynamic of networks: whether they are static or changing over time.
- They used partition recognition metrics which are purely based on structural partition and semantic-and-structural partition.

### **3.3.4 Limitations**

Group identification is becoming increasingly difficult due to the heterogeneity of data and systems built in them, as well as their size. The dilemma gets even more complicated as we consider the complexities of such a large volume of data. Large-scale network distributed algorithms have caused an increase in interest as a result of these limitations. Both methods can be used to derive two broad meanings. A distributed detection method built on shared-memory parallelism, also known as "distributed computation," is the first and most well-studied. These parallelization approaches split the social network into non-overlapping subnetworks that are

done in parallel. In the other hand, the high cost of shared memory multiprocessor systems and the high degree of data dependency are seen as barriers to this approach. In this direction, efforts have been made to detect multiple communities in a hierarchical manner using an undistributed social network. We use the most up-to-date models for detecting parallel group structure to conduct a portion of our survey here. The quest for such groups has gotten a lot of attention in recent years, and many studies on community composition in graphs have been conducted. This scholars are undertaking to work with large volumes of structured data as well as the use of relational and semantic data in combination.

### **3.4 Paper 4: “Hidden community detection in social networks”.**

#### **3.4.1 Introduction**

Community identification has become an important role in network analysis in recent decades, providing insight into the underlying structure and possible roles of networks. Early research centered on finding disjoint communities that divide a network's collection of nodes. Researchers have recently noticed a multiplicity of interwoven community memberships and built algorithms to find overlapping groups. To deal with the overlapping case, certain partitioning techniques are also extended. Within these two groups, a dendrogram which is hierarchical in nature is centered on the granularity of the communities that are been detected can be constructed. Inspite the progress made, there are few new forms of group structure which are called as the secret community structure that have been receiving little coverage in the literature. The sparse community structure found in real-world networks, such as hidden societies or temporary communities, is highly weak than those with dense community structures found in families, close friends, and many more as confirmed by widely used group scoring metrics. When the major percentage of the members belonging to a less modular group are also the members belonging to denser groups, the communities are often ignored.



### 3.4.2 Characteristics and Implementation

The major goal of this paper is to determine the hidden structure of these communities. We determine the value of community's hiddenness as the percentage of nodes in stronger communities, and then they present Hidden Community Detection (HICODE), a meta-approach for identifying both the dominant and hidden structures of these networks. HICODE starts by using an existing algorithm to the network as a basic algorithm, it then weakens the structure of the network's belonging to the detected communities. In this same way, the community's weaker, secret structure is shown. All these steps are repeated until unless there are no further significant structures to be discovered. HICODE then weakens the secret group structure, which then results in a more reliable version of the structures with respect to dominant community.

### 3.4.3 Features

We assume the information we gleaned about secret group structure would be useful in future investigations. The following are the paper's key contributions:

- **Conception on Hidden Community:** We define a group's hiddenness and incorporate as the concept of Structures related to hidden communities, which is spread across the entire social networks.
- **Methods based on Hidden Community Detection:** HICODE is a very useful tool for identifying the two types of communities, the dominant and secret configuration of a hidden community. HICODE is implemented using many such group identification algorithms as the basic algorithm, as well as the structure weakening methods Remove-Edge, Reduce-Edge, and Reduce-Weight.
- **Validation based on the datasets of real world:** From the experiments performed on a number of networks from real-world, we have shown that as the hiddenness importance of a population becomes high, then it is more complicated for an algorithm to find that group, HICODE does not perform many state-of-the-art community's discovery methods to uncover hidden groups.

### 3.4.4 Limitations

- From the experiments performed on a number of networks from real-world, we have shown that as the hiddenness importance of a population becomes high, then it is more complicated for an algorithm to find that group, HICODE does not perform many state-of-the-art community's discovery methods to uncover hidden groups.

## 3.5 Paper 5: “TI-SC: top-k influential nodes selection based on community detection and scoring criteria in social networks”.

### 3.5.1 Introduction

The investigation of identifying powerful nodes is the most critical topics in community identification networks. An influential spreader detection problem occurs when a single influential node is detected without considering the positions of other influential particulars in the network. The influencing maximization issue, on opposite side, is described as identifying the set of influential nodes in terms of their topological effects on each other. The aim of influence maximization is to identify the most powerful people who can spread the most information. Diffusion is the mechanism by which network knowledge spreads from one node to another. Total influence time is critical in the diffusion process.

### 3.5.2 Characteristics and Implementation

This algorithm determines the strong nodes by comparing the connections between the core nodes and other nodes' scoring potential. To avoid seed node collection duplication, the scores are changed after each seed node is chosen. In networks with a large number of Rich-Club members, this algorithm performs well.

In this article, they proposed, for determining the top-K prominent nodes, an effective community-based algorithm with a ranking measure is used. This algorithm's scoring criteria excludes seed node duplication, resulting in the selection of the best effect distribution K-node.

### 3.5.3 Features

To summarize, the following are the main characteristics of this algorithm:

1. We investigate the impact maximization issue in the context of group structure, with the goal of reducing the search space for seed node selection.
2. We suggest using the scoring capacity of other nodes to minimize seed node overlap.
3. We use the relationship between core nodes to integrate communities with similar knowledge diffusion structures.
4. In terms of impact distribution, experiments on simulated and real-world networks show that the proposed algorithm outperforms the base algorithms. This algorithm is faster than standard algorithms.

### 3.5.4 Limitations

Influence maximisation is a optimization problem in which the aim is to find a subset of the network's seed nodes that have the most powerful on a propagation model. Seed node overlap and a lack of optimal seed node coverage aggravate the problem.

## **3.6 Paper 6: “Detection of Influential Nodes Using Social Networks Analysis Based On Network Metrics”.**

### **3.6.1 Introduction**

Social media is a form of communication and interaction between people in which they develop, share, exchange, and access data and ideas, forming groups (small nets) and networks in the process (net of individuals and groups). A social network is a reflection of social network research that is made up of different communicating actors. The study of social community issues. A study of community networking metrics, that involves a variety of criteria that define social media analysis, is becoming increasingly important in the modern age. Various approaches are used for optimizing and analyzing the structure of the design of complicated community networks. A model which connects various points to evaluate nodes which also evaluate relationships is known as social network analysis. When it comes to analyzing complex networks, finding the main character in an online community identification network is a most troubling aspect. The main player is the person who has the most clout in the social network. The identification of powerful nodes from a social network has gotten a lot of attention in the online social culture in upcoming centuries. The main hypotheses in the detailed analysis of social networks are network measures. The most critical metric for completely analyzing and behaving correctly in a team is centrality.

### **3.6.2 Characteristics and Implementation**

The identification of powerful nodes from a social network has gotten a lot of attention in the online social culture in recent decades. The main hypotheses in the detailed analysis of social networks are network measures. The importance of the centrality measure in analyzing organizational and team actions cannot be overstated. The suggested procedure is as follows:

- 1) The Degree Centrality is a network exposure index that counts the node's number of direct contacts has to determine how well it is connected to the network.

2) Centrality is a term used to describe the amount of connection and communication with people in a network

a) In a complex graph, a node's closeness, or normalised closeness centrality, is the average length of the network's shortest path between two nodes.

b) Betweenness Centrality is a quantifiable metric that allows a node in a social network to have complete control over its interpersonal interactions with two other nodes.

c) The calculation of a node's power in a social network is called Eigenvector centrality.

d) The Clustering Coefficient determines how likely nodes are to be associated with one another.

### 3.6.3 Features

- Explain the related work that reflects on the core features that we used to classify the key actors in social network research (influential nodes).
- The system we suggest is called Centrality.
- The outcomes of the experiments and statistics used to evaluate the output
- Validity of the proposed schema.

### 3.6.4 Limitations

With an increasing number of people entering social networks every day, identifying prominent nodes in such a network is a difficult challenge.

## 3.7 Paper 7: “Influential Nodes Detection in Dynamic Social Networks”.

### 3.7.1 Introduction

More scientists researching the impact maximisation issue have turned their focus to this field as social media sites such as Facebook and Twitter have grown in prominence in recent years. This topic has received a lot of interest, and many researchers have proposed various algorithms

for detecting powerful nodes, the majority of which are based on static social networks. In the other hand, true social networks change over time. Users make new connections, and some users lose communication. Complex social network research heavily relies on them so many influential nodes can be found by carefully analysing the relationships among the connections.

### **3.7.2 Characteristics and Implementation**

The key aim of SNDUpdate is to identify the most strong nodes in a dynamic social network. It takes advantage of the structural and semantic properties of the network. As a consequence, the main idea is to advocate for a two-phased solution. Indeed, the first step of SNDUpdate is concerned with the network's structural aspects, while the second is concerned with the semantic aspects. Each customer is described in the following section of this paper by a collection of interests expressed as an attributes vector. The weight of the relation between two nodes remains unchanged in the previous work because the powerful nodes are detected in a static social network. The network in this paper is dynamic, meaning that its configuration changes over time. As a consequence, the relation between two nodes belonging to a snapshot graph  $G_t$  can be omitted in the snapshot graph  $G_{t+1}$ , leading to a change in the sense of logical similarity between two nodes as well as a change in the set of leader nodes.

### **3.7.3 Features**

Similar to a graph structure, a social network is made up of nodes and edges, with nodes representing members and edges representing links between them. Users form relationships with one another over time, and their interactions change. An evolving social network consists of observations of social networks at various time stamps ( $G_1, G_2, \dots, G_n$ ) and offers not only a list of node relationships, but also information about how these relationships evolve over time.

This section examines a number of dynamic social network analysis findings. The challenge of identifying powerful nodes in social networks with moving edges is the subject of this article.

- 1) Non-linear Architecture-Based Methods
- 2) Metric-based methods
- 3) Diffusion Model-Based Methods

### **3.7.4 Limitations**

The aim of the influence maximization issue is to find powerful nodes that will help you achieve your viral marketing goals on social media. Previous research has primarily focused on static social network analysis and algorithm creation in this context. As the network evolves, however, certain algorithms must be modified.

## **3.8 Paper 8: “Performance Evaluation of Modularity Based Community Detection Algorithms in Large Scale Networks”.**

### **3.8.1 Introduction**

In the field of complex networks, community detection is a hot topic, and several studies have been done on it. A widely accepted definition of a group in a network is a subset of nodes with high internal density but low external density. Several studies evaluating and comparing various measures of partition efficiency can be found in the literature. In their article, Yang and Leskovec, for example, look at the suitability of certain steps for classifying ground-truth based populations. Moradi and colleagues Compare the ability of different consistency functions in an email network to differentiate between helpful and spam messages. According to Newman and Girvan, the most widely used criteria for determining the stability of groups in networks is modularity. Modularity is defined as the discrepancy between the fraction of edges within a group and the fraction expected by a random version of the network while keeping the degree distribution of the nodes in general.

### **3.8.2 Characteristics and Implementation**

The aim of this study is to investigate the computational challenges of large-network group detection methods. The suggested fine-tuning stage reduces the number of switched nodes, allowing for up to 50% quicker execution without losing the accuracy of the partitions obtained. The methods were chosen with the aim of identifying a set of computational resources capable of dealing with group detection modularity optimization in a variety of ways, including divisive agglomerative, heuristic solution, and relaxed solution. To make for a fair compare, both of the methods were applied using free software. To do so, a collection of suitable data structures and analytical methods were used to the extent possible.

### **3.8.3 Features**

The computational complexity of the studied methods is investigated in terms of their numerical application. The applied algorithms are used to compare the Newman spectral method and the CNM method on a qualitative and quantitative level, with a focus on their application to large-scale networks.

### **3.8.4 Limitations**

In almost all of the measured networks, when the number of moving nodes is increased to 20% (Newman-FT20%), the obtained modularity is almost equal to the value obtained when 100% of the nodes are transferred. In other words, in the fine-tuning stage, permuting more than 80% of the nodes almost always results in a huge loss of time and computing resources.



## **3.9 Paper 9: “A comprehensive literature review on Community detection: Approaches and applications”**

### **3.9.1 Introduction**

Since it helps to expose coherent and meaningful sub-graphs, recognise the features, functions, configuration, and dynamics of such complex networks, community recognition has been designed as an axial field of Complex Network Analysis (CNA). A number of methods and approaches have been developed in this regard over time to provide appropriate solutions to complex network paradigms, especially problems involving group detection. Meanwhile, scientists face a significant challenge in defining populations in a complex network, which necessitates a thorough review of the literature and a survey. Researchers must conduct reviews of the major papers relating to group recognition in complex networks in order to identify their major strengths and weaknesses, so in this report, we summarise the literatures on community detection for complex networks. We assume that this literature contribution, which does not include all submissions, would be a valuable resource for group detection practitioners. As a result, the importance of the selected approaches' contributions was more important to us than the publication's chronological order.

### **3.9.2 Characteristics and Implementation**

In this article, we examine and categorise a variety of group identification techniques and strategies in order to demonstrate their key strengths and weaknesses. Examine experiments that deal with application in a variety of domains as well.

Finally, we focus on English-language journal articles that offer valuable knowledge for clinicians involved in group identification, as well as the most current studies, which are available from online databases.

1. A method focused on non-overlapping static communities;

2. Approach focused on neighbouring populations that are static;
3. Static hierarchical communities as a starting point;
4. Dynamic groups focused on an approach.

### 3.9.3 Features

1. **Approach:** For group identification, a technique was used.
2. **Technical principal of approach:** The technical principle or algorithm of the implemented method is defined in this column.
3. **Network type:** This attribute includes "Weighted" if the edges connecting the studied network's nodes are weighted, and "Unweighted" if the edges aren't weighted.
4. **Directionality of the network:** Indicates "Guided" if hyperlinks between network nodes are directed, and "Undirected" if the hyperlink's directionality is ignored. If the solution can accommodate both structures, "all" is stated.
5. **Network nature:** If the network is static or dynamic is indicated by this attribute.
6. **Network size:** The network size is a critical parameter for the computing performance of the solution. This column shows the scale of the funded network.

### 3.9.4 Limitations

The field of group detection is still developing, and categorising a comprehensive set of methods and approaches continues to be a work in progress.

## 3.10 Paper 10: “Community detection in social networks”

### 3.10.1 Introduction:

Individuals social networks are shaped by their own encounters and interactions with other people. Social networks reflect and model people's social connections. The exponential growth of the

internet has resulted in a significant rise in web user activity. Many social networking sites have come up to help people communicate, such as Facebook and Twitter. The number of contacts has grown exponentially, making it incredibly impossible to keep track of these communications. People who share their interests and values are attracted to them. Social media enables user to extent their relations and connections between various people through a simple platform called social apps. It has grown its significance by making available of text feature which does not need to speak in front with the person rather it allows us to type the content to make connections between people. This also made us easier to make friends who also has a similar taste as us.

### **3.10.2 Characteristics and Implementation**

Communities are graph pieces with smaller connections to the rest of the graph and denser internal connections. Without some prior experience of the objects, unsupervised learning attempts to group them together. There is an issue called clustering impurity which refers to the similarity in features like topology and among the other graph properties. The two methods used for locating the social network graph in the literature called network partitioning which is feature in social network analysis and clustering which is used in many data structure solving problems. These are the things which are described briefly:

1. The process which divides the graph into already determined number of tiny components each having its unique set of characters is called partitioning the graph.
2. And another process which enables grouping of associated objects into groups knowingly clusters and the process is called clustering.

### **3.10.3 Features**

- i. Citation networks in academia are made up of citation connections among papers and researchers, so trend analysis in citation networks is essential.
- ii. Community Detection Improves Recommender Systems: Recommender systems (RS) provide recommendations based on data from similar users or goods.

- iii. As the number of social networking sites increases, the focus and distribution of sites are broadening and diversifying.

### **3.10.4 Limitations**

In social networks, communities are increasing at an exponential rate. Each algorithm for various types of group detection has its own set of drawbacks.

## **3.11 Paper 11: “Analysis of the Dynamic Influence of Social Network Nodes”.**

### **3.11.1 Introduction**

People's social interactions have changed dramatically in recent decades as a result of revolutionary advances in communication technologies. Milgram's small-world experiment in the 1960s demonstrated that the average distance between any two individuals on Earth is six, a phenomenon known as six degrees of separation. The average distance between Facebook network nodes was just 4.74 degrees in 2011, according to the results of a study of the friend networks of 750 million active users on Facebook. Finding out or mining which node has the greatest effect is crucial in social network research. As a result, a variety of metrics, such as degree centrality, betweenness centrality, closeness centrality, k-shell centrality, eigenvector centrality, and the PageRank algorithm, have been proposed to quantify the value of a node from various perspectives.

### **3.11.2 Characteristics and Implementation**

Most current measurements are based on statistical properties of network topology and do not account for the impact of changes in mutual confidence among nodes during information dissemination. A new scheme for measuring the complex power of nodes in a social network is

introduced in this paper. The modification of node trust value during information propagation is essential in this new scheme. The new scheme also takes into account the accumulated shift in node confidence value.

### **3.11.3 Features**

A new decision scheme for the complex power of social network nodes is introduced in this paper. A new calculation of node dynamic impact is proposed, taking into account the effect of changes in the information distribution process on confidence values. It is a step forward from previous algorithms. Finally, we examine the power of nodes based on network topology or statistical properties, and compare it to other classical algorithms to ensure the algorithm's validity and accuracy.

# CHAPTER-4

## DATA

### 4.1 Overview

The Global Terrorism Dataset (GTD) is the world's only unclassified terrorist activities database. The National Consortium for the Study of Terrorism and Responses to Terrorism (START) has made whole Global Terrorism Database accessible on this website to raise awareness of the problems created by these kind of terrorist gangs which has to be studied and defeated more quickly. The GTD is created by a devoted team of researchers and technical staff.

Since 1970, the GTD has gathered statistics on domestic and international terrorism attacks, and it currently includes over 200,000 cases. Every event has information about the date occurred, location of the bombing, the type of weapons used, the intent of the influenced target, the casualties number, and – where known – the party or responsible individual.

### 4.2 Dataset

Properties of the GTD:

- About 200,000 terrorist acts are recorded in this database.
- The world's most extensive unclassified database on terrorist attacks is currently available.
- Around 20,000 assassinations, 96,000 bombings, and 15,000 kidnappings and hostage cases have occurred since 1970.
- Each case has data on at least 48 variables, with presenting data on over 120 variables in more recent events.
- To gather event statistics, over 4,500,000 news stories and 35,000 news article sources were reviewed from 1998 to 2019.

## CHAPTER-5

### System Requirement Specification

#### 5.1 Product Perspective

By fusing international and domestic CT intelligence, providing terrorism research, exchanging information with stakeholders around the CT enterprise, and guiding whole-of-government action to protect our national CT priorities, we lead and integrate the national counterterrorism (CT) initiative.

##### 5.1.1 Product Features

Selecting relevant features, constructing rough conceptualizations of terrorist groups, constructing terrorist group boundary areas, and developing terrorist group networks are all steps in eliciting terrorist group networks.

Finding the most powerful nodes: For intelligence and security informatics, predicting terrorist networks and recognizing key players is critical. Using machine learning methods, we suggest a framework for analyzing social networks. To delete unnecessary and passive nodes from the entire network, the proposed technique employs the k-core principle. It then uses a hybrid classifier to classify the key actors by extracting multiple features. The proposed technique is put to the test on a publicly accessible dataset, and the results show that the method is efficient.

In various fields, communities are referred to as classes, clusters, coherent subgroups, or modules; community identification in a social network entails recognizing sets of nodes where the connections between nodes within a set are greater than the connections between nodes in other networks.

### 5.1.2 User Classes and Characteristics

The proposed methodology, which elicits terrorist groups' networks, finds the most powerful nodes, and tracks the temporal evolution of terrorist networks, is not an open-source model; rather, these data are shared with national intelligence management in order to protect the country from terrorist threats and carry out counter-terrorism operations. This has been put in place with the help of college professors. Data scientists, National Intelligent Management, and Government are among the users who can change the dataset and improve the algorithms.

### 5.1.3 General Constraints, Assumptions and Dependencies

- Regulatory policies

We use the START project's Global Terrorism Database , prediction models, for example, are one type of data that can be evaluated. The main goal of this study is to use GTD's historical data, which provides statistics on terrorist attacks since 1970, to conceptualise terrorist groups' actions over time.

- Hardware limitations.

Depending on the hardware we use, analyzing the dataset and implementing the algorithm we implement takes time. When we use any modern CPU, the task is completed faster. The dataset needs the least amount of storage possible.

- Safety and security consideration

GTD research and algorithm implementation should not be used for anything other than educational and development (counter-terrorism) purposes. This processed data is only available to national intelligent management in order to defend the country from terrorist attacks and carry out counter-terrorism operations.

- Assumption: Terrorism cannot be defeated



Terrorism is, without a doubt, one of the defining characteristics of our day. It hits the news regularly, threatening or targeting states, private businesses, and ordinary people. It has also become one of the most serious challenges to peace, security, and stability in many parts of the world.

#### **5.1.4 Risks**

Data leakage occurs when sensitive or otherwise confidential information leaves an organization's infrastructure, leaving it vulnerable to unauthorized disclosure or malicious use. Mitigating the risks of such data handling and leakage may be a costly endeavor.

### **5.2 Functional Requirements:**

Data is gathered from the Global Terrorism Database (GTD) as well as other sources. Steps for pre-processing have been completed. Rough Sets are used to approximate conceptualizations of terrorist group activities. Later, relevant features will be selected, rough conceptualizations of terrorist groups will be created, terrorist group boundary regions will be created, and terrorist group networks will be designed.

Identify more dominant and temporal nodes in most influencing networks using the neighbourhood correlation coefficient. The suggested method actually works based on a locality clustering coefficient and also assigned and takes help of the similarity of relations among the adjacent nodes. The process is completely dependent on a k-shell decomposition technique, which determines a node's power based on how it shares relationships with its neighbours.

This project's aim is to classify and compare theoretic group detection algorithms. Combine methods for characterising, distinguishing, and separating populations using two strategies: centralised and distributed. The aim of this study is to look at the detection technique in relation to the dynamic of networks: whether they are static or changing over time. To use partition recognition metrics such as structural-dependent partition and semantic-and-structural-dependent partition.

## **5.3 External Interface Requirements**

### **5.3.1 User Interfaces**

On the top of the project UI (web page), there are options such as checkout (gather the information given in the text field), view (display the network's visual output), and so on. In this project, we create a web page that displays the outcome of our work. Essentially, the user interface is a simple HTML page that allows the user to communicate with the project's outcome. The working area is in the middle of the tab, where the user can fill in the information and see the results. To communicate with the server, the project will use python and json. An error message will appear on the website if the user sends incorrect input or input format.

### **5.3.2 Hardware Requirements**

Any Intel(7th gen or higher) or AMD(2nd gen or higher) processor with a base clock speed of at least 3.5 GHz. On the computer, all of the project's results are shown. The TCP protocol is used to retrieve the result from the server. The TCP protocol is used for all XML requests and responses between the client and the server.

### **5.3.3 Software Requirements**

Python Version : 3.7 or higher

Operating Systems : Ubuntu 16.04 or higher, Windows 7 or higher, Mac OS 10 or higher

Tools and Libraries (Open-source):

igraph - The network research software. igraph is a group of network analysis tools that focuses on speed, portable, and flexibility to use.

NetworkX - NetworkX is a Python package that allows you to generate, extrapolate and formalize the structure and functionality of very complicated networks.

### 5.3.4 Communication Interfaces

To obtain the result from the server, as well as all XML requests and responses between the client and the server, we use the TCP protocol. To load a few image format outputs, the line speed should be at least 10 kbps. The application's predefined functions will handle the network's entire buffer size.

## 5.4 Non-Functional Requirements

### 5.4.1 Performance Requirement

**Efficiency:** A measure of network efficiency is the amount of nodes that can instantly reach a wide number of different nodes – sources of information, status, and so on – from a comparatively small number of connections. These nodes have nonredundant contacts added to them.

**Effectiveness:** The aim of effectiveness is to enter a cluster of nodes using non-redundant contacts. In the other hand, performance aids in reducing the amount of time and effort spent on redundant contacts. Each set of contacts is a self-contained information source. Since people who are connected want to hear about the same things at the same time, one cluster around this non-redundant node, no matter how big it is, is just one source of information.

### 5.4.2 Safety Requirements

In order to have a stable and healthy working atmosphere, we in the safety profession have had to reconsider our positions. Engineering protections, procedural methodologies, and technological obstacles, as well as ways to eliminate risks and even weapons of mass destruction, have all been considered.

If a catastrophic malfunction, such as a disc failure, results in substantial loss to a major portion of the database, the correcting process restores a previous version of the storage which was easily backed and

archived in the database and reconstructing the more current state by reapplying or reperforming the committed whole procedure from a archived log, until the failure of time.

### 5.4.3 Security Requirements

Since the model relies on the data for learning, it should not be tampered with or poisoned, since this might bring the system to a halt. Database storage is often needed by security systems.

## 5.5 Other Requirements

**Scalability:** We believe that the scope of this research is not limited to the Global Terrorism Database, but that it can be applied to other social media platforms as well.

**Maintainability:** The framework should be designed in such a way that it can be expanded in the future. It should be simple to add new feature specifications or accommodate changes to existing requirements.

## **CHAPTER-6**

### **System Design**

#### **6.1 Novelty**

For intelligence and security informatics, predicting terrorist networks and identifying key players is critical, and few studies address this topic. As a result, we suggest a framework for analysing social networks that makes use of machine learning techniques (ie.,K-Core Concepts). After the networks have been clustered appropriately, we use group identification methodology to evaluate the relationships between terrorist nodes within the same cluster as well as between terrorist nodes from different clusters.

#### **6.2 Innovativeness**

The methodology we suggest elicits terrorist group networks, identifies the most powerful nodes, and tracks the temporal evolution of terrorist networks, but it is not an open-source model; rather, these data are shared with national intelligent management in order to protect the country from terrorist threats and carry out counter-terrorism operations.

#### **6.3 Performance**

Since we quantify it by the number of nodes that can access a large number of different nodes with a very small number of connections, the way we suggested is more effective. To address the nodes, nonredundant contacts are used.

Some approaches are successful when applied to a group of nodes that can be accessed through non-redundant contacts. However, in our situation, continuity entails reducing the time and effort spent on repetitive connections. Each set of contacts is a self-contained information source. Since people who are

connected want to hear about the same things at the same time, one cluster around this non-redundant node, no matter how big it is, is just one source of information.

## **6.4 Reliability**

The methodology we use is capable of conducting operations using data from the terrorist database and generating results in a time frame that allows us to focus on other tasks. It's dependable for the data we use and the effective outcomes we get at the end.

## **6.5 Maintainability**

To ensure that the users see the correct results, we need to use good ranking methods and algorithms. We'd also give the nodes weights so that the value of defining the group changes over time. To ensure that the tool works properly, these ranking methods must be checked and revised on a regular basis. The underlying search engines results are retrieved using their respective APIs, which are all free of charge. If their respective policies change, maintenance would be needed.

## **6.6 Legacy to modernization**

To improve operational efficiency as part of the legacy modernization, we're upgrading and optimising business processes by giving government agencies graphical access so they can see each community's development and powerful nodes in a single graphical view. As a result, users are able to meet their needs in terms of their experience and are more readily adapted to newer technology platforms.

## **6.7 Application compatibility**

Since our project can run on a variety of operating systems and has a user-friendly environment, we can simplify the testing process, ensuring that all of the applications are checked for compatibility at the same time. To a certain degree, auto-removal of features that aren't enabled by the operating system is possible.

## 6.8 Resource Utilization

There is a lot of data to process, and much of it isn't necessary for the results we want. As a result, we only use the results that have a significant impact on terrorist growth and are also needed for future connection prediction among the groups. We can almost see the relationship between different groups and their development over time thanks to group detection. As a result, the data is used to meet our intermediate needs and forecasts.

## **CHAPTER-7**

### **IMPLEMENTATION AND PSEUDOCODE**

#### **7.1 A rough set guide to gain a better view of the structure and evolution of terrorist organisation networks.**

The overall approach's workflow

- Selecting relevant features
- Creating rough conceptualizations of terrorist groups
- Creating boundary regions for terrorist groups
- Building Similarity matrix
- Designing the terrorist groups network

##### **7.1.1 Selecting features from GTD**

The suggested strategy is based on the notion of conceptualising terrorist groups using information about the attacks they have carried out. As a result, it's essential to pick a relevant subset of the GTD's features. Such a subset must be appropriate for describing terrorist groups behaviour.

To assess a perpetrator's normal behaviour, we must summarise the behaviours shown by the same perpetrator over the course of a set of events. As a consequence, we now have a GTD view  $U[t_1, t_2]$  that indicates the Universe of Discourse, which encompasses all terrorist acts within a defined time frame.



Table 1: Features and meanings are associated with a sample value variable.

Feature	Value	Interpretation
attacktype	2	<i>Armed Assault</i>
weaptype1	6	<i>Explosives/Bombs/Dynamite</i>
suicide	0	<i>No suicide</i>
targetype1	3	<i>Police</i>
INT_LOG	0	<i>The nationality of the perpetrator group is the same of the location of the attack</i>
INT_IDEO	0	<i>Any and all nationalities of the perpetrator group are the same as the nationalities of the victims</i>
ishostkid	0	<i>No kidnapping</i>
crit1	1	<i>The incident meets criterion 1a</i>
crit2	1	<i>The incident meets criterion 2b</i>
crit3	1	<i>The incident meets criterion 3c</i>

### 7.1.2 Creating rough conceptualizations of terrorist groups

The rough sets operators are applied to the set  $V[t_1, t_2]$  yields the rough set  $(\text{lower}(V[t_1, t_2]), \text{upper}(V[t_1, t_2]))$ .

$V[t_1, t_2]$ , the approximation for the lower value, includes the attacks that the group  $g$  is almost likely to have carried out.  $V[t_1, t_2]$  is the approximation for upper value, which includes attacks that may or may not have been committed by the category  $g$ , as well as attacks that have been unquestionably committed by such a party. In other terms, the approximation for upper value represents both  $g$ 's determining acts and possible  $g$ -related behaviours.

### 7.1.3 Creating boundary regions for terrorist groups

When a rough compilation of a militant group's activities has been compiled, the well-known three way decisions theory can be used to separate the attacks into three categories: positive, negative, and boundary. The BND area contains enigmatic occurrences (due to uncertain data) that should be investigated further in order to learn more about the behavioural correlations between different populations.

$$POS = \underline{V}_{[t_1, t_2]}$$

$$BND = \overline{V}_{[t_1, t_2]} - \underline{V}_{[t_1, t_2]}$$

$$NEG = U_{[t_1, t_2]} - BND$$

The POS area encompasses all events that are certain to have been committed by the terrorist organisation in question and that define its activity;

The NEG area encompasses all incidents that were most likely not committed by the suspected terrorist group and may not characterise their conduct;

All acts that may or may not have been committed by the terrorist organisation in question and that may or may not constitute its actions are included in the BND zone.

### 7.1.4 Building Similarity matrix

Only the BND regions of two terrorist organisations can be used to compare their similarities. In fact, excluding the POS regions makes sense because they only contain actions that strongly characterises group behaviour. In other words, such distinguishing actions cannot be found in another terrorist group's POS zone. Regardless of the POS regions, BND regions include activities that are replicated across communities. As a result, BND regions are worth looking into in order to find the aforementioned parallels. A appropriate function must be used in order to calculate such a similarity factor. The Overlap Measure 1 appears to be the best fit among the similarity functions since it is unaffected by set cardinality.

Let  $G_1$  and  $G_2$  be two rough sets representing two terrorist groups conceptualised activities, and  $BNDG_1$  and  $BNDG_2$  be the first and second terrorist groups' respective boundary regions. The Overlap Measure 1 can now be used to define the similarity equation:

$$Sim_{TG}(G_1, G_2) = \frac{|BND_{G_1} \cap BND_{G_2}|}{\min(|BND_{G_1}|, |BND_{G_2}|)}$$

If there are more terrorist acts that match patterns that may be perpetrated by any of them, the function  $Sim_{TG}$ , which has a spectrum of  $[0, 1]$ , two separate terrorist organisations are more close.

$Sim_{TG}(G_1, G_2)$  must be computed for each  $(G_1, G_2)$  to build a network of possible connections out of a list of terrorist organisations  $g_1, g_2, \dots, g_n$ , where is the sequence of rough sets representing all terrorist groups conceptualizations  $g_1, g_2, \dots, g_n$ , in addition The rough collection corresponding to the definition  $g_i$  is called  $G_i$ . A similarity matrix can be generated using the measured similarity steps:

$$M = \begin{bmatrix} s_{1,1} & s_{1,2} & \dots & s_{1,n} \\ s_{2,1} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ s_{n,1} & \dots & \dots & s_{n,n} \end{bmatrix}$$

Where  $s_{i,j} = Sim_{TG}(G_i, G_j)$ , The similarity value between  $G_i$  and  $G_j$  is measured for all  $i, j = 1 \dots n$ , where  $G_i$  is the rough set calculated on the concept  $g_j$ , and  $G_j$  is the rough set obtained starting from the concept given.

### 7.1.5 Designing the terrorist groups network

The final move is to use the similarity matrix to construct the terrorist group's network while still considered a thresholding operation. The network-building algorithm is fairly straightforward. Assume  $W = (V, E)$ , with  $W$  representing the terrorist group's network,  $V$  representing the network's nodes, and  $E = V \times V$  representing the network's borders.

Algorithm for building terrorist groups network:

```

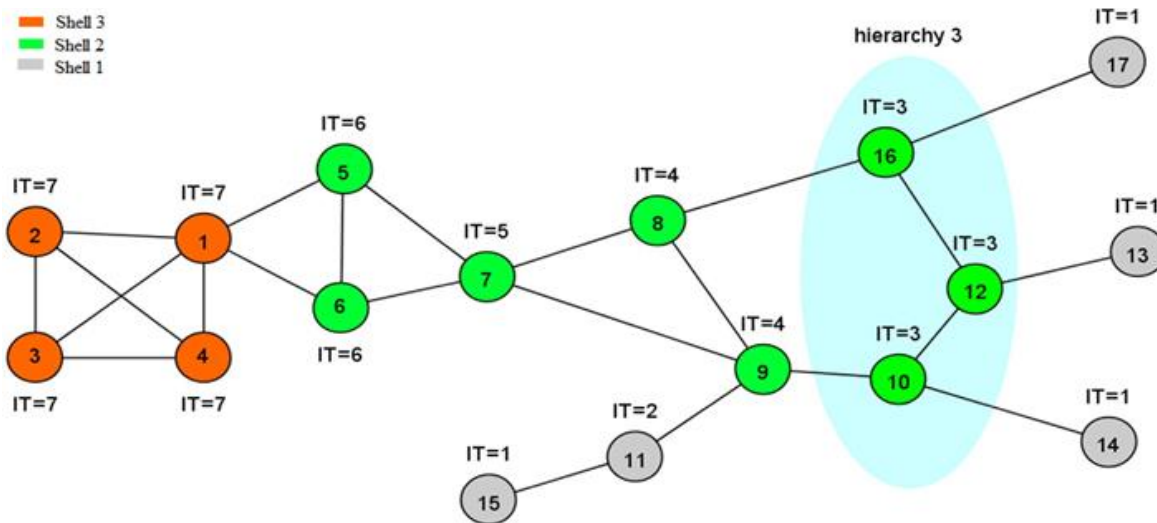
 $V \leftarrow \{g_1, g_2, \dots, g_n\}$ 
 $E \leftarrow \phi$ 
 $W \leftarrow (V, E)$ 
for  $i=1 \dots m$  do
  for  $j=1 \dots m$  do
    if  $s_{i,j} \geq \gamma$  and  $i \neq j$  and  $(g_j, g_i) \notin E$  then
       $E \leftarrow (g_i, g_j)$ 
    end if
  end for
end for

```

## 7.2 Using the neighbourhood correlation coefficient to find influential nodes in social networks.

We aim to find effective indices of relationships between nodes by examining what a node has in common with its surrounding nodes. In the implemented method, called Extended Cluster Coefficient Ranking Measure, the number of common neighbours is less important than the common hierarchy

between a node and its neighbours. To do so, the network must first be divided into parts. The hierarchy can be assumed if nodes in the same hierarchy are removed in the same iterations (IT) of the k-shell algorithm.



Take a look at the graph above. Three shells make up this graph, each with its own colour scheme. The iteration (IT) in which the nodes are removed is evaluated using the following form, which is implemented by the k-shell decomposition. The initial value of the IT counter is 1. The lowest-degree nodes are removed from the graph and given  $IT = 1$ , indicating that they belong to the first hierarchy. The IT counter is then increased by one, and the lowest-degree nodes are once again removed from the graph. Nodes that have been deleted are given  $IT = 2$  and placed in the second hierarchy at this stage. This process is repeated until the graph contains no more nodes, and at the end of each iteration, the IT counter is increased by one and allocated to the nodes removed in that step. Node 8 has a neighbour, 16, in hierarchy 3 ( $IT = 3$ ), as seen in the table.

A shell vector is computed for each node  $v_i$  based on the hierarchy of its neighbours after the network has been decomposed into different hierarchies.

Node  $v_i$ 's shell vector is defined as

$$SV_i = \left\{ \left| N_i^{(1)} \right|, \left| N_i^{(2)} \right|, \left| N_i^{(3)} \right|, \dots, \left| N_i^{(f)} \right| \right\}$$

Where  $|N_i^{(k)}|$  denotes the number of neighbours of node  $v_i$  who belong to hierarchy  $k$ , the network's maximal hierarchy is denoted by  $f$ .

Between the shell vectors, Pearson's correlation coefficient is used to assess node correlation, and ECRM is dependent on that. We use the Pearson correlation between their shell vectors to evaluate the degree of commonality between two nodes' neighbours, which is determined as:

$$C_{i,j} = \frac{\sum_{k=1}^f (SV_i^{(k)} - \overline{SV}_i)(SV_j^{(k)} - \overline{SV}_j)}{\sqrt{\sum_{k=1}^f (SV_i^{(k)} - \overline{SV}_i)^2} \sqrt{\sum_{k=1}^f (SV_j^{(k)} - \overline{SV}_j)^2}}$$

where  $SV_i^{(k)} = |N_i^{(k)}|$  denotes the value of the  $k$ th cell in vector  $SV_i$ , the mean value of the shell vector is denoted by  $\overline{SV}_i$ . Since the number of values in  $SV_i$  equals the degree of node  $v_i$ , we have  $\overline{SV}_i = d_i / f$ . For two vectors, the correlation coefficient is always between 0 and 1, with a value of 0 indicating no correlation, a value of -ve indicating indirect correlation, and a value of +ve indicating direct correlation.

Node  $v_i$ 's shell clustering coefficient is measured as

$$SCC_i = \sum_{v_j \in N_i} \left( (2 - C_{i,j}) + \left( 2 \frac{d_j}{\max(d)} + 1 \right) \right)$$

The equation above is divided into two sections. The correlation coefficient is included in the first section, while the degree is included in the second. A higher association between node  $v_i$  and each of its neighbours reduces  $v_i$ 's spreading capacity. As a consequence, node  $v_i$ 's shell clustering coefficient is determined ( $\sum_{j \in N_i} C_{i,j}$ ). Apart from the relationship between node  $v_i$  and its neighbour  $v_j$ , the degree of the latter influences the shell clustering coefficient. As a consequence, consider the expression ( $\sum_{j \in N_i} (d_j / \max(d)) + 1$ ), where  $\max(d)$  is the graph's maximum degree.

The neighbourhood rule would increase the accuracy with which the most influential nodes are identified. The neighbourhood rule is first used to quantify the Cluster Coefficient Ranking Measure (CRM) for each node in the proposed solution, given the shell clustering coefficients of the neighbours. By adding the CRMs of the neighbours, the ECRM is determined.

$$CRM_i = \sum_{v_j \in N_i} SCC_j$$

$$ECRM_i = \sum_{v_j \in N_i} CRM_j$$

Finally, the ranking is determined by sorting the nodes according to the ECRM index.

### Algorithm: Pseudo-code of ECRM method:

Input:  $G = (V, E)$  //  $G$  is undirected and unweighted graph with  $V$  as the set of nodes  
and  $E$  as the set of edges.

Output: A ranking list of nodes' influentially.

Begin

Assign  $IT(v_i)$  for each  $v_i \in V$  using the k-shell algorithm

For  $i=1$  to  $|V|$

Calculate  $SV_i$

End for

for  $i=1$  to  $|V|$

Set  $SCC_i = 0$

for each  $v_j \in N_i$

Calculate  $C_{ij}$

$SCC_i = SCC_i + (2 - C_{ij}) + ((2d_j / \max(d)) + 1)$

End for

End for

for  $i=1$  to  $|V|$

Set  $CRM_i = 0$ ;

for each  $v_j \in N_i$

$CRM_i = CRM_i + SCC_j$

End for

End for

for  $I=1$  to  $|V|$

Set  $ECRM_i = 0$ ;

for each  $v_j \in N_i$

$ECRM_i = ECRM_i + CRM_j$

End for

End for

Sort the nodes in descending order based on ECRM scores to obtain the ranking list.

End



## 7.3 Community Detection:

### 7.3.1 Louvain algorithm using locality modularity optimization:

To detect groups, we use the Louvain algorithm with local modularity optimization. This algorithm uses a greedy optimization approach to increase the modularity of a network partition iteratively.

Modularity, a metric of network structure, is used to determine the strength of division of modules. In networks with high modularity, connections between nodes are dense, but connections between nodes in various modules are sparse. It's mostly used in network community structure detection optimization methods.

The goal function is maximised in each iteration to calculate the populations. In step 1, small groups have been formed by maximising the modularity on a area level. Only local infrastructure upgrades are allowed at this time. In the following step, nodes that belonging to the same entity are combined into a unique node that is going to point out a community in a original aggregating network of gangs. Through the construction of a hierarchy of units, steps will get repeated iteratively until no further changes in modularity are legally possible.

The groups must be re-computation from the beginning if the original algorithm avoids the insertion or elimination of newly originated edges and nodes right after getting the correct structure of community.

### 7.3.2 Adding the edges/nodes:

Adding edges/nodes results in four types of effects at the community structure level in this method.

#### Cross-community edge:

When we try to join two nodes in a Cross-community edge that are already linked to other nodes, two things can happen. The community structure remains unchanged if the linking nodes belong to the same community. If the linking nodes belong to separate communities, the two communities are merged into one.

#### Inner community edge:

If the two nodes on the edge already exist and belong to the same party, inserting a new edge between them increases the community's internal links thus leaving the intercommunity connections unchanged. As a consequence, the organisational structure of the company remains intact.

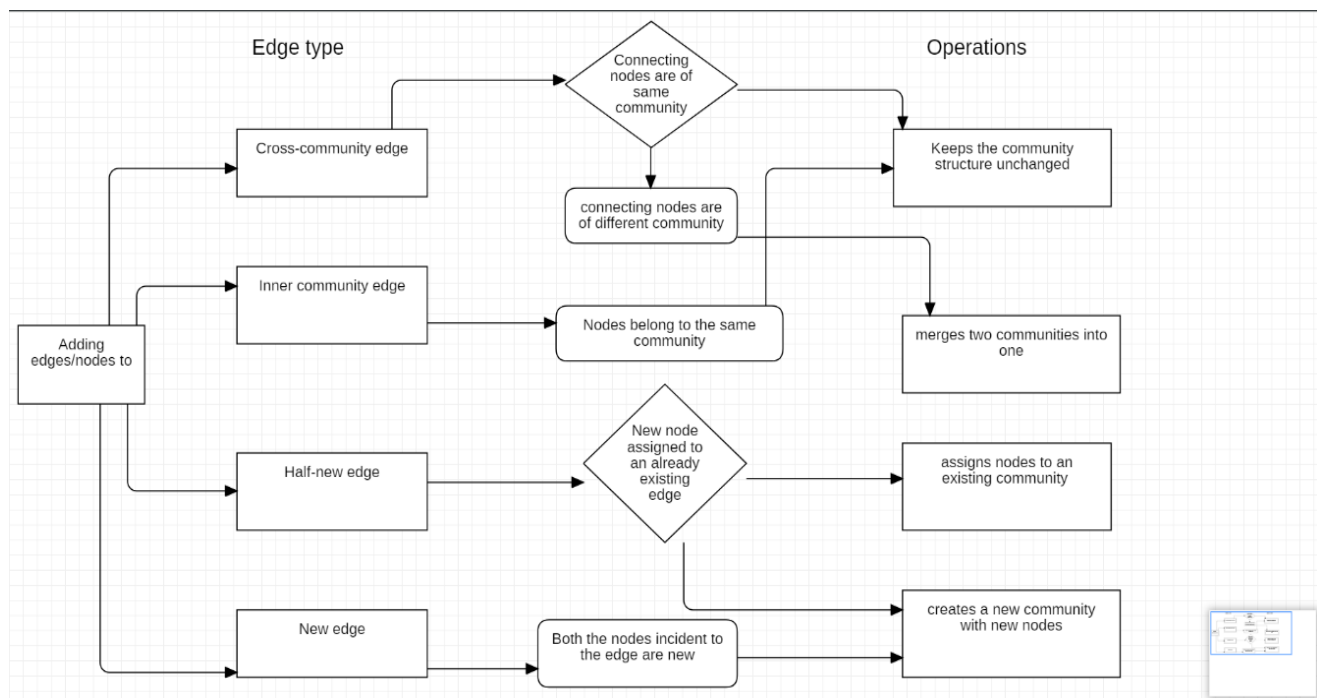


Fig.2 Adding edges/nodes

**Half-new edge:**

In this type of edge, the elaborating edge is a half new edge, which means one nodes in the structure is present in the network and another edge is newly originated. When a node is assigned to an existing community, the community configuration remains unchanged. Otherwise, a new group is generated with new nodes.

**New edge:**

In the new edge, all nodes adjacent to the edge are new. Here, there may happen two kind of things, that is nodes are assigned to the new community which is just generated or it may get assigned to two different communities where each one is for the each newly generated node.

### **7.3.3 Removing the edges/nodes:**

At the group structure stage, removing edges/nodes results in four types of effects:

**Cross-community edge:**

In a cross-community edge, two different nodes adjacent to a excluded edge assigned to varying communities. The inner links of the group are maintained while intercommunity relations are minimised by eliminating these types of edges. This operation would not result in any existing communities being merged, nor does it disband any communities in which the excluded edge is a part.

**Inner community edge:**

The two nodes in the inner-community edge that are nearby to a edge is assigned to the original same group. Through removing these types, the community's inner links are reduced while intercommunity bonds are maintained. The disbanding has very little issues to the whole networks if the nodes belonging to the excluded edge are connected to other edges; otherwise, the community is split into smaller categories and branches that join other established communities.

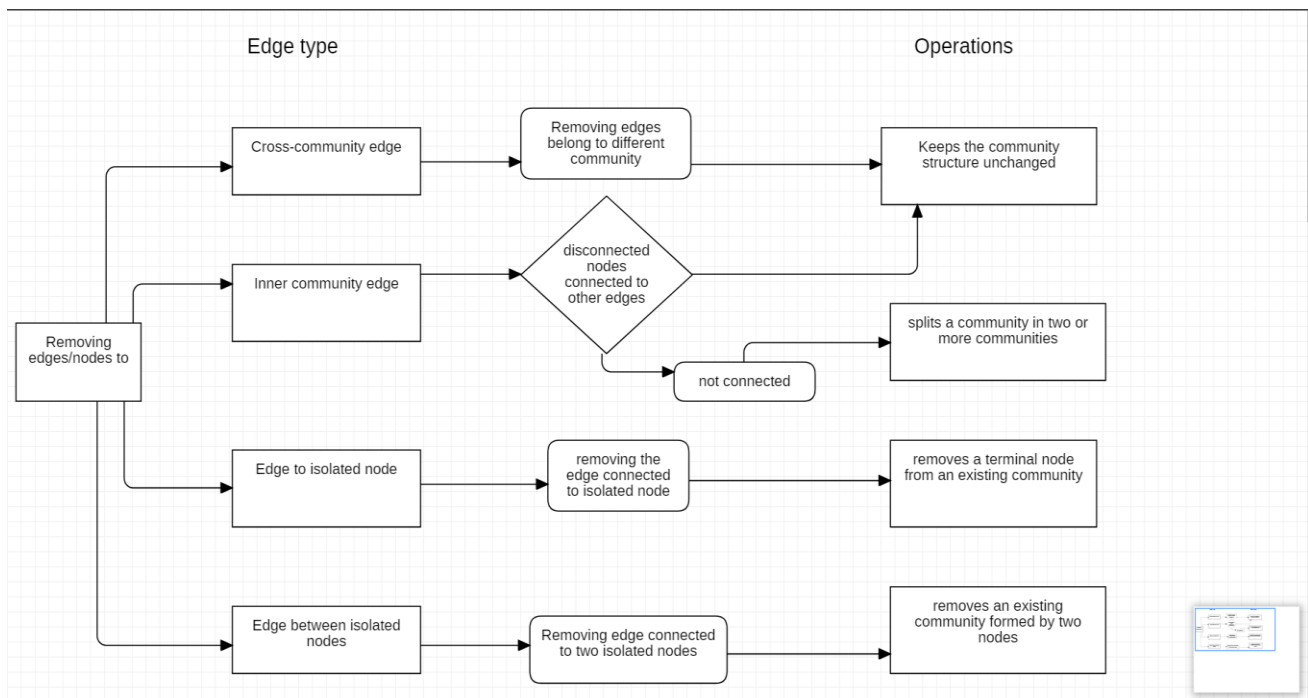


Fig.3 Removing edges/nodes

**Edge to isolated node:**

An isolated node is one of the nodes incident to the boundary, so disorienting these type of isolated edges leads to disorientation of the complete node itself. Since the deleted node is a node having single value, it has no impact on the group system and therefore on the community's inner relations.

**Edge between isolated nodes:**

The edge in Edge of isolated nodes that needs to be disoriented from the community is assigned to two different terminal nodes. By eliminating the edge, all nodes are removed as well, essentially ending the group or groups of which they belong. The remainder of the neighbourhood will be unaffected.

Two of the eight operations arising from the disorientation/assignment of nodes/edges reduce the number of communities, two increase the count of communities, and four keep the population structure unchanged.

The working feature of the algorithm are completely conceptual and collective sub procedures for completing particular tasks (i.e. adding edge to the CII). They're duplicated until the modularity needs to be increased, or the edges has to be disoriented or assigned.

Procedure P1a: The insertion of an edge to CII causes a list of impacted nodes and communities to be retrieved, as well as the edge itself being added to CII.

Procedure P1b: The edge has been replaced in the CII. When an edge is deleted, a list of damaged nodes and their communities is retrieved, as well as the edge itself being removed to CII.

Procedure P2: In CII., affected communities should be dissolved. Affected groups will be dissolved in CII depending on the AffectedByAddition() or AffectedByRemoval() lists of affected nodes and societies ().

Procedure P3: The improvements in CII should be reflected in Cul. To reassign the changes in group architecture to the Cul, use the AffectedByAddition() or AffectedByRemoval() lists of affected nodes and classes. It's worth remembering that the Cul, as well as the inserted or disabled edges, will be modified in this process.

Procedure P4: Cul will be used to complete Step 1 of the Louvain Algorithm and calculate improvements in group composition that could contribute to locally optimised modularity.

Procedure P5: Update CII with the populations that have changed by assigning the Louvain type of structure from Step 1 to Cul.

Procedure P6: Using the Cul to run the Louvain Algorithm Phase 2 and aggregate populations.

## Dynamic Community Detection Algorithm (Pseudo code):

```

V ← {u1, u2, ..., uv} , E ← {(i1, j1), (i2, j2), ..., (ie, je)}
A ← array{(i1, j1), ..., (im, jm)} , R ← array{(i1, j1), ..., (in, jn)}
procedure Main(G ← (V,E), A, R)
  Cll ← {C1, C2, ..., Cn}, Cul ← {}, Caux ← Cll
  InitPartition(Caux)
  mod ← Modularity(Caux), old mod ← 0
  m ← 1, n ← 1
  while (mod ≥ old mod ∨ m ≤ |A| ∨ n ≤ |R|) do
    Caux ← OneLevel(Caux)
    n, c CommunityChangedNodes(Cll, Caux)
    Cll ← UpdateCommunities(Cll, n, c)
    old mod ← mod, mod ← Modularity(Cll)
    Cul ← PartitionToGraph(Cll)
    if m ≤ |A| then 16: src, dest A[m]
    anodes ← AffectedByAddition(src, dest, Cll)
    Cll ← AddEdge(src, dest, Cll)
    Cll ← DisbandCommunities(Cll, anodes)
    Cul ← SyncCommunities(Cll, Cul, anodes)
  end if
  if n ≤ |R| then
    src, dest R[n]
    anodes ← AffectedByRemoval(src, dest, Cll)
    Cll ← RemoveEdge(src, dest, Cll)
    Cll ← DisbandCommunities(Cll, anodes)
    Cul ← SyncCommunities(Cll, Cul, anodes)
  end if
  Caux ← Cul, m ← m + 1, n ← n + 1
end while
end procedure

```

### Adding Cross Community edges:

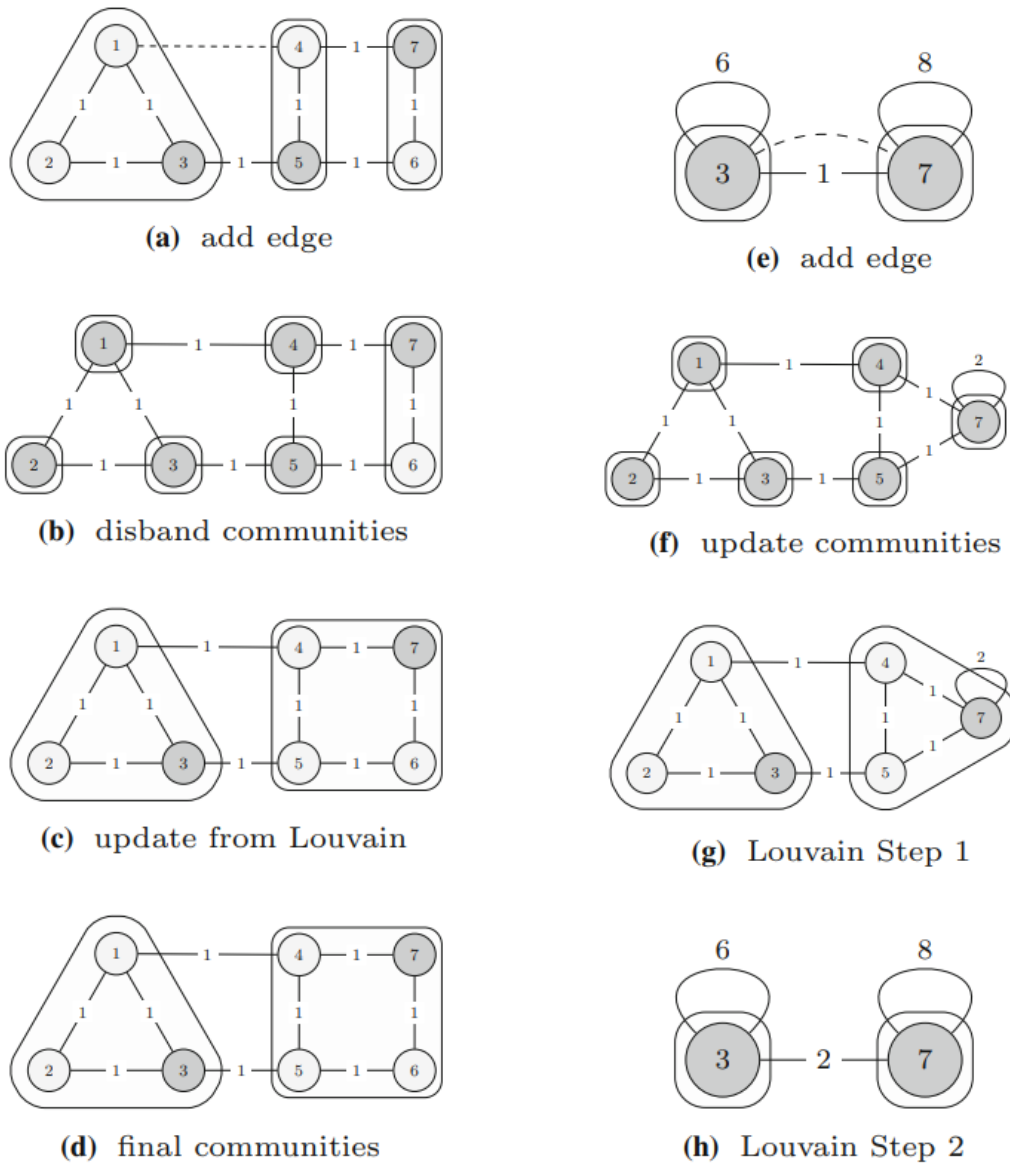


Fig.4 Adding Cross Community edges

### Remove Edge to terminal nodes:

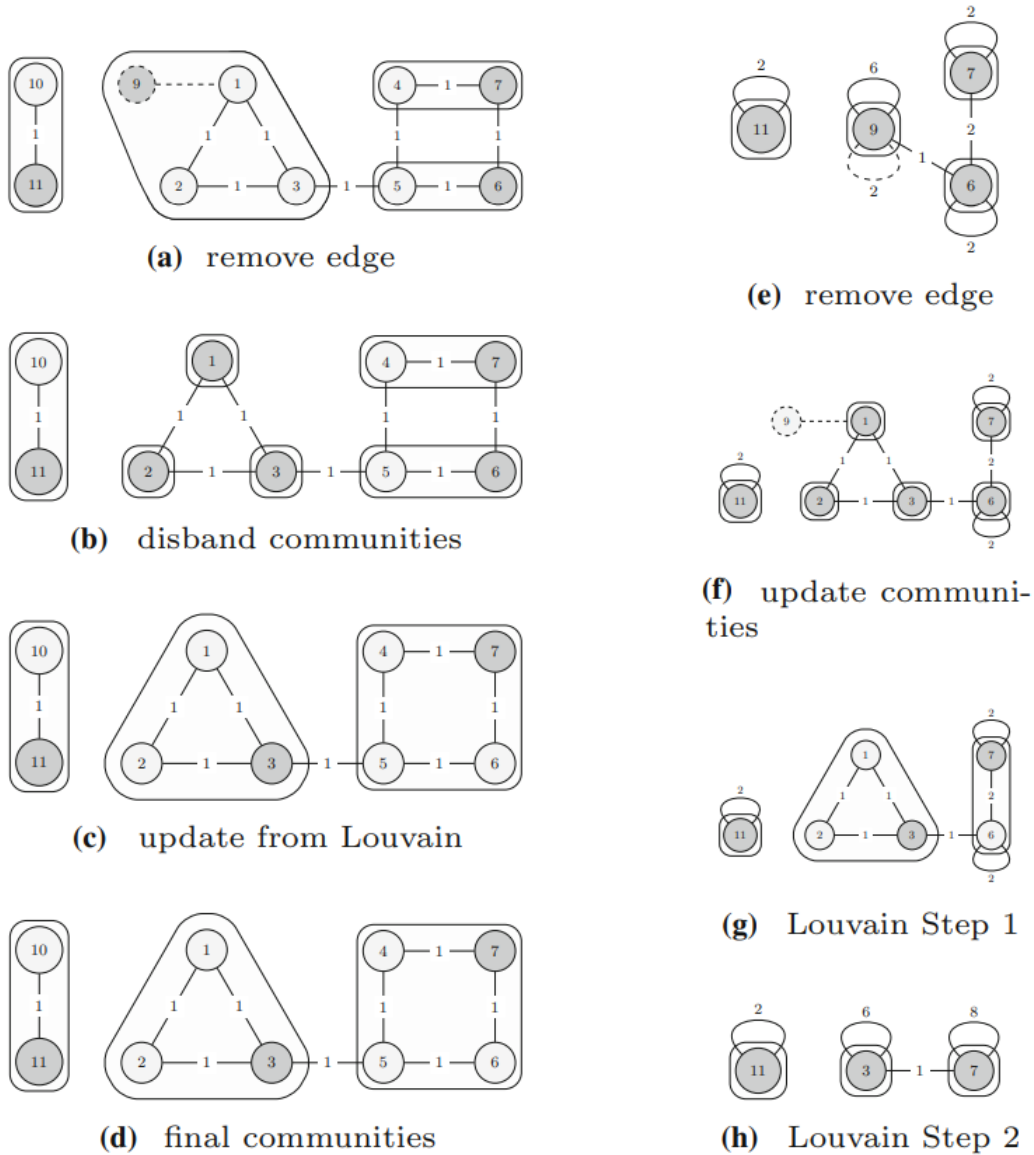


Fig.5 Remove Edge to terminal nodes



## **CHAPTER-8**

### **CONCLUSION OF CAPSTONE PROJECT PHASE – 1**

In order to solve unique problems in this field, several algorithms have been suggested. In this research, we suggested a system of low computational complexity that can be extended to large-scale networks. Where it came to the stability of the group configuration, the results showed that when just considering the locally modularity optimization for the regions where the network had nodes or edges inserted or deleted, there was no penalty on the calculated global modularity. The results show that the number of edges inserted or excluded can be changed according to the network's size, as expected. The number of edges moved in each iteration has a direct relationship with the probability of groups disbanding. The proposed methodology / approach is supposed to produce improved results because it uses algorithms with lower time and space complexity that perform well.

## **CHAPTER-9**

### **PLAN OF WORK FOR CAPSTONE PROJECT PHASE – 2**

Our preferred algorithm's final implementation. Get clarity on proposed system. Ensuring compatibility. Final integration step, integrating the implemented algorithms of understanding composition of networks, finding influential nodes and community detection. Present our final method and get started on our research report.

## REFERENCES/BIBLIOGRAPHY

- [1] Vincenzo Loia, Francesco, “Understanding the composition and evolution of terrorist group networks: A rough set approach.” Paper 2019,  
<https://www.sciencedirect.com/science/article/pii/S0167739X19307757>
- [2] Ahmad Zareie, Amir Sheikahmadi, Mahdi Jalili, Mohammad Sajjad Khaksar Fasaei, “Finding influential nodes in social networks based on neighborhood correlation coefficient.” Paper 2020,  
<https://www.sciencedirect.com/science/article/pii/S0950705120300630>
- [3] Mehdi Azaouzi, Delel Rhouma, Lotf Ben Romdhane, “Community detection in largescale social networks: stateoftheart and future directions.” Paper 2019,  
<https://www.sciencedirect.com/science/article/pii/S0020025517310101>
- [4] Kun Hea, Yingru Li a, Sucheta Soundarajanc, John E. Hopcroft , “Hidden community detection in social networks.” Paper 2019,  
<https://link.springer.com/article/10.1007/s12652-020-01760-2>
- [5] Hamid Ahmadi Beni, Asgarali Bouyer, “TI-SC: top-k influential nodes selection based on community detection and scoring criteria in social networks.” Paper 2020,  
<https://www.researchgate.net/publication/333197917>
- [6] Aftab Farooq,Muhammad Uzair,Gulraiz Javaid Joyia,Usman Akram , “Detection of Influential Nodes Using Social Networks Analysis Based On Network Metrics” Paper 2018,  
<https://ieeexplore.ieee.org/abstract/document/8346372>
- [7] Hong-Jian Yin, Hai Yu, Yu-Li Zhao, Zhi-Liang Zhu, Wei Zhang, “Analysis of the Dynamic Influence of Social Network Nodes.” Paper 2019  
<https://www.hindawi.com/journals/sp/2017/5046905/>

- [8] Nesrine Hafiene, Wafa Karoui<sup>1</sup>, and Lotfi Ben Romdhane, “Influential Nodes Detection in Dynamic Social Networks”. Paper 2019  
[https://link.springer.com/chapter/10.1007%2F978-3-030-20482-2\\_6](https://link.springer.com/chapter/10.1007%2F978-3-030-20482-2_6)
- [9] Mohamed EL-Moussaoui, Tarik Agouti, Abdessadek Tikniouine, Mohamed Eladnani ,  
“Community detection: Approaches and applications.” Paper 2019,  
<https://www.sciencedirect.com/science/article/pii/S1877050919305046>
- [10] Punam Bedi and Chhavi Sharma, “ Community detection in social networks ”. Paper 2016,  
[https://www.researchgate.net/publication/295395520\\_Community\\_detection\\_in\\_social\\_networks](https://www.researchgate.net/publication/295395520_Community_detection_in_social_networks)
- [11] Vinicius da Fonseca Vieira, Carolina Ribeiro Xavier, Nelson Francisco Favilla Ebecken and Alexandre Goncalves Evsukoff , “Performance Evaluation of Modularity Based Community Detection Algorithms in Large Scale Networks.” Paper 2014,  
<https://www.hindawi.com/journals/mpe/2014/502809/>

## APPENDIX A DEFINITIONS, ACRONYMS AND ABBREVIATIONS

GTD	-Global Terrorism Database
CRM	-Coefficient Ranking Measure
ECRM	-Extended Coefficient Ranking Measure
SV	-Shell Vector