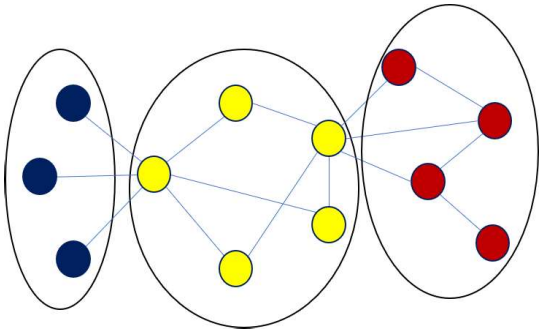
Chapter 1

Introduction

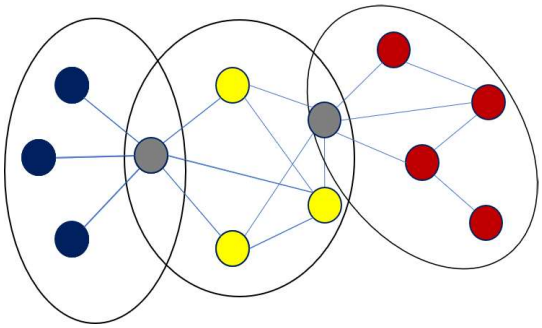
1.1 Background

Existence of network is prevalent in nature from living being to non-living objects. It's ubiquity provides greater opportunities to fetch meaningful properties and characteristics about it. Identifying groups, commonly known as communities, in network has gain a lot research attention. Weiss and Jacobson [1] are known to be the first who carry out a community analysis in 1955 within a government agency [2]. A community is a set of nodes sharing common behaviour. Hence providing meaningful structural/ behavioural information of any network. It is formed in such a way that the ratio of internal to external connections is high. A network graph consists of two elements: nodes and edges.

Communities can be identified on both levels, either by grouping nodes or edges. Edge-based communities are more common on graphs with attributed elements. Works presented here focuses on node-based grouping of non-attributed networks. A community can be classified based on the association of a node with a community as either disjoint or overlapping. When a node exclusively belongs to only one community, it is known as disjoint community structure. Whereas, if a node can be part of multiple community, it is known as overlapping community structure. The thesis focuses on disjoint community structure. Figure 1.1 presents a schematic representation of a network with three communities. Figure 1.1a shows a network with three disjoint communities where nodes with same color represent a community. An overlapping community structure is shown on



(A) Disjoint communities



(B) Overlapping communities

FIGURE 1.1: A schematic representation of a network with disjoint and overlapping community structure

same network in figure 1.1b where the nodes with multiple community association are in gray color.

Social networks are real-world networks that are pervasive in nature. Some typical examples of social networks that exist in our surrounding include research group collaboration networks [12] [114], acquaintance network [62] [100], company communication networks [159], online friendship networks [128], [59], [158] [39], behaviour-based networks of animals [108], biological network [23] [127] [153] [79] and others. It can be viewed as a graph, with nodes denoting network entities and edges denoting relations between them. Due to its extensive real-world applicability, it attracts significant research interest. Size of global datasphere had grown from 0.5 exabyte in 2009 to 18 exabyte in 2018. With the enormous expansion in internet-based services, the size of the global data sphere has grown exponentially, causing an increase in the size and nature of social networks. Now, not only do these networks have large data volume but also a tendency to change over time.

The continuous generation of data results in a dynamic network which evolves over time.

Nowadays most of the networks are dynamic in nature and possess no fixed topology. An evolving network continues to introduce more aspects requiring their consideration even while the process of deducing information from its former version is in progress. It is not practical for these networks to be stored at one place or to be processed as a whole in one run. An efficient methodology always needs data to be sampled into small sizes.

Traditionally, a network is represented as a graph with nodes and their in-between links. But this definition of network needs some modification to incorporate the other important dimension, time. The temporal dimension facilitates improvised understanding of network by embedding valuable information to it. The work presented here deals with this modified network definition. The available literature suggests two broad approaches for handling such evolving data. They are the temporal network approach and the snapshot network approach. The community structure of a network is updated on every change at next time unit in temporal network settings. On the contrary, in a snapshot setting the complete network graph present at current time is treated as a static network and the consecutive community structure is mapped to determine the progress. Temporal setting has been used in proposed algorithms.

Thesis aims to present new techniques to identify communities in dynamic social network. An abbreviation CD2N will be further used to refer to community detection in dynamic networks. Consider a dynamic network at time t represented as a graph $G_t(V_t, E_t)$ where V_t represent set of nodes or objects and E_t is a set of edges or events, community detection problem can be defined as:

Definition 1.1. Community detection aims to identify k(say) partitions $\mathbf{C_t} = \{C_1, C_2, ..., C_k\}$ of graph $G_t(V_t, E_t)$ representing the community structure of the network at time t where $C_i \in \mathbf{C_t}$ obeys following properties:

- $C_i \neq \phi$
- $C_i \cap C_j = \phi$
- $\cup_k C_i = V_t$

1.2 Open Issues

Identifying communities in dynamic social networks comes with several issues associated with data as well as computation. Some common issues are as follows:

- *Privacy and data consent*: Social networks are always linked with privacy concerns. Ethical obligations hinder the accessibility of these network data for further studies. Few anonymous datasets are made available by organizations, but it is still a significant issue in getting relevant and large data from social network websites.
- *Data collection*: Owing to the dynamic nature of data, it is not trivial to collect and monitor continuously generating data.
- *Nature of network*: A network consists of different entities having different relationships among them. Entities often have attributes, which also increases the data size. Extracting similar substructures from data that are diverse in nature is a tricky task.
- *Resolution limit*: The resolution limit problem occurs when small communities are not well detected due to limitations in the resolution of the community detection algorithms.
- *Noise and fluctuations*: Noise and fluctuations in dynamic networks can obscure community structures, leading to inaccuracies in community detection results.
- *Computational complexity*: Community detection in dynamic networks can be computationally intensive, especially for large-scale networks or when analyzing multiple time points.

1.3 Challenges in Community Detection in Dynamic Networks

CD2N is an emerging field of research in social network analysis. It comprises of tools and practices to handle dynamic networks. However, there are some open challenges which

required researchers' attention to lessen its limitations. This section attempts to enlist those issues and challenges. They are:

- *Data unavailability*: Dynamic network data are not readily available to researchers. Social network privacy policies restrict access to those network data and hence researchers have to rely on synthetic data. Although some anonymous data are available but with recent trends observed in network structure, real time data would help a lot in obtaining results from the present perspective.
- *Rate of change of network*: It is also hard to perform computations on network data, which is changing quickly. There are two possible options for researchers, either their algorithm's computation is fast enough or they must be able to map results of consecutive time stamps under observation.
- *Network type*: Sometimes network consists of multiple types of nodes and edges. Consider the example of a school network. A node can be a student, employer or employee and edges between them can also represent different relationships. In this scenario, identifying communities is challenging as one has to consider different types of nodes and edges. Thus, heterogeneity in the network imposes an additional challenge to the community detection problem.
- *Evolution of communities*: Keeping track of communities over time is a challenging task. A common approach found in the literature on community detection is to identify communities in a given snapshot and map them with previous snapshots to keep track of evolution of clustering. There are algorithms that tend to identify different communities on same network at different times. It imposes added complexity in tracking communities over time.
- *Evaluation of algorithms*: The evaluation of performance of an algorithm is also a challenge for researchers as there are no standard performance metrics available in literature. Most of the work evaluates their performance by using the quality and accuracy of communities obtained from algorithms. Some also consider running time as a metric. These metrics can be implementation and platform dependent.
- *Non-deterministic solutions*: Most of the algorithms often output different communities in different executions. This instability imposes greater concern in tracking the evolution of communities over time.

1.4 Applications of Community Detection

Community detection aims to extract meaningful information from network which can be used in further analyses. This section presents a brief discussion of various areas where community detection is used as a tool. It plays crucial role in network analysis such as in communication network, friendship network, biological network and social network. Some applications are listed below:

1. *Online social network:* Several social media (such as Facebook, Twitter, WhatsApp, YouTube, Skype and others) platforms facilitate web-based interactions among users, and it creates an online social network. Community detection is widely used in such network as a mining tool for analyses ranging from network reduction in large scale network to community identification in evolving and distributed networks.

   Ferrara in [42] conducted a study on Facebook network consisting of 500 million users in 2011. A web-based mining was performed for data collection which was further clustered using label propagation algorithm [52] and fast network community algorithm [69]. The network reflects some interesting properties such as, nodes tend to exhibit six degree of separation, power law degree distribution and communities with high similarities. A number of studies are also performed on Twitter [121] [36], Flickr [124] and others because they are widely used by people in diverse areas.

2. *E-commerce:* Now-a-days products and services are made available to customers via e-commerce websites. These online shopping platforms require effective optimization to attract online customers, provide them better recommendations and enhance the target audience advertisements [5]. Online networks of shoppers play important role in overall e-commerce market. Customers with similar background tend to purchase similar products. A product recommendation algorithm is proposed in [178] which applies community detection for classifying similar interest and recommend product based on it. Another approach is proposed by Reddy et al. [74] where they use geographical information of users and their friend circles for identifying groups and recommend products based on their friends' interests. Product based communities [136] are also found in literature which promotes an efficient shop layout and recommendations. Trust relationship among products is exploited by authors in [16] which presents an interesting result that customers with

similar rating belong to same community even though they share no other common background.

3. *Communication network:* The existence of community structure within communication networks has facilitated a broad range of applications, including design and optimization of search engines, efficient routing in Mobile Ad Hoc Networks (MANETs) and containment of worms in Online Social Networks [116]. The effectiveness of clustering algorithms applied to group objects within search engines directly impacts the visibility of searched items on the front page [96]. Soliman et al. [151] introduced an innovative approach to search engine results, focusing on the semantics of retrieved documents rather than the words contained within them. This approach, which resembles hierarchical clustering, aids in the clustering of documents based on semantic relevance. Routing algorithms based on community detection have demonstrated considerable enhancements over traditional routing methods [24][66].

4. *Healthcare:* Community detection is often used to identify certain groups susceptible to epidemic diseases [93]. A study in [143] presents the influence of communities on dynamics of diseases. Researchers have also applied community detection on molecular level for identifying abnormal patterns in human organs. It is also used to detect cancer cells and tumors. Bechtel et al. [14] introduced a community-based approach for the detection of lung cancer. Similarly, Haq and Wang [58] conducted a study utilizing genomic datasets to identify subgroups within twelve types of cancers. Their research delved into analyzing survival rates among these identified communities and the distribution of tumor types across them. Brain networks are also extensively analyzed using community detection [40][47]. Community structures are present in various biological networks like protein interaction network [37][27], metabolic network [182], gene structures [97], etc.

5. *Criminology:* Community detection also facilitates the identification of criminal activities associated with either real person or bots. Authors in Pinheiro (2012) employed a two-step algorithm to identify fraud events within a telecommunication network. Initially, communities were detected through data analysis of the telecommunication network. Subsequently, basic graph properties such as degree and betweenness were utilized as measures to identify abnormal nodes or outliers. Similarly, Gangopadhyay et al. (Gangopadhyay and Chen, 2016) devised two

methods to analyze various types of relationships, particularly small but exclusive, for the purpose of detecting fraud in healthcare. Some authors uses anomaly detection to identify criminal entities [10] [35]. They use community detection to group entities and consider outliers as criminal entities showing suspicious behaviors. A detailed discussion on bot identification is presented in [120].

6. *Others:* Application of community detection is not limited to above categories. It has a wide area of applicability. Many authors have used it as a tool for predicting future links [67][77] and influence maximization [149]. Authors in [170] presents a book recommendation algorithm which uses community detection for identifying readers of similar interest and provide recommendation based on it. Citation networks are also very popular among researchers, which helps in identifying patterns in inter-disciplinary publications [25] and co-authorship data. Stock market data are also analyzed for identifying communities in them where each stock represent a node and their market correlation represent edges in graph [55][166][6].

1.5 Objectives

The thesis aims to identify communities and track their evolution over time. We present six objective functions which collectively facilitate the primary goal of the thesis.

- ***Partitioning social networks into group of similar nodes***: Social networks are made of social actors (nodes) and their connections (edges) with each other. It can be partitioned into substructures, which consist of closely related nodes known as communities. Therefore, we consider the following objective function for this task:

  Objective 1: We aim to identify communities in a given social network.

- ***Handling dynamic behaviour of network***: Real world networks are complex networks. They are subject to topological variations with time. Continuous perturbations in network topology raise an essential concern for algorithms targeting these networks. The techniques presented in this thesis endeavour to explore dynamic social networks. We aim to model these networks in computationally

feasible settings with minimum compromise of temporal information. It can be achieved using the following objective function:

Objective 2: Represent networks into a mathematical model incorporating its dynamic nature.

- ***Utilizing properties of social network and behaviour***: Literature is evident with multiple studies claiming the effect of social networks' properties on their underlying community structure. We attempt to exploit some of these properties in proposed algorithms to identify the community structure of a given network. Social networks are made up of social actors and their respective interactions. Behavioural patterns of these actors also inspire the core idea of presented chapters. Subsequent objective functions are used to include these observations:

  Objective 3: Utilize the knowledge of neighbourhood of a node in community decision making.

  Objective 4: Use social network properties to leverage community decision making.

- ***Examine the belongingness of a node to a community***: The assignment of a node to a community is often considered a crisp phenomenon. However, it fails to address the extent to which a node can belong to a community. When dealing with disjoint partitions, a common assumption made by researchers is that they assign full membership to a community. However, it is much more practical to consider the fuzzy membership of a node to a community. This concept provides more flexibility in community assignment tasks. The following objective function covers this concept:

  Objective 5: Design an algorithm to cover the fuzzy membership of a node in a community.

- ***Regulate the performance of algorithms by using diverse evaluation standards***: Performance analyses of an algorithm are a challenging task. It depends on testing the algorithm on data by calculating a metric and comparing it with a state-of-the-art algorithm. Selection of data, metrics and state-of-the-art algorithms requires covering diverse aspects of empirical analysis. The thesis uses the following objective function for the empirical analysis of algorithms:

  Objective 6: Analyze the performance of algorithm against a diverse evaluation criterion.

1.6 Contribution

The thesis consists of three chapters in which we attempt to address previously mentioned objective functions. All three chapters address the first, second and sixth functions as they deal with broad ideas, their working setting and evaluation methods. A tree-based approach whose design focuses on objective four is proposed. Multi-objective optimization technique is exploited in another chapter considering objective 3. The last contribution focuses on objectives 3 and 5 to present a fuzzy-based approach for community detection. A brief discussion of each contribution is listed below.

1.6.1 Tree based Community Detection

The chapter presents a tree-based community detection in dynamic social networks (TCD2) algorithm, which exploits two important properties of social networks, connectedness and influence, for finding communities in the network. TCD2 uses a tree structure to maintain information on dynamically changing community structures in the network. The experimental results on real-world social networks, along with synthetic networks, validate the performance of TCD2. The tests also confirmed its superiority over the state-of-the-art algorithms. The results showed that the proposed algorithm achieves a significant trade-off between quality and accuracy.

1.6.2 Multi-objective based Community Detection

The algorithm proposed in this chapter uses three objective functions that are inspired by network properties. The community of a node corresponding to an input edge is updated by an algorithm based on its newness. The algorithm uses the Pareto front principle to identify the optimal community. The algorithm is evaluated over 12 datasets and compared to 10 state-of-the-art algorithms. It shows superior performance on real and connected datasets and also performs well for disconnected datasets. The algorithm is evaluated using both accuracy and quality metrics, with the quality metrics slightly outweighing the accuracy metrics.

1.6.3 Membership based Community Detection

The chapter presents a fuzzy-based approach to solving community detection problems in dynamic networks. A novel membership function inspired by literature observations is proposed in the work. Membership function is used for community assignment at decision step. The algorithm is capable of producing a set of disjoint communities of a dynamic network at a given time t. A detailed empirical analysis of the proposed algorithm MDCD is presented in the work. Evaluation incorporates four quality and three accuracy metrics' results which are further compared against ten state-of-the-art algorithms on real and artificial datasets. Results show that MDCD's overall performance is better for accuracy as well as quality metrics with bearable complexity.

1.7 Thesis Organization

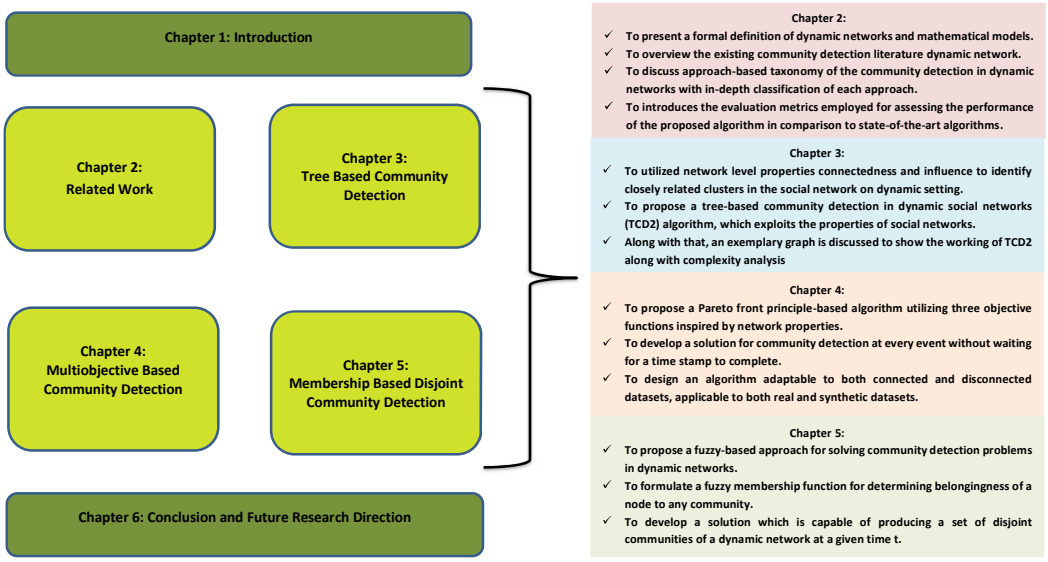


FIGURE 1.2: Explanation of thesis organization in relation to the four main contributory chapters

The structure of this thesis and the relation of four main chapters are illustrated in figure 1.2.

At first, community detection, its background, open issues and challenges are discussed in chapter 1. Six objectives addressed in the thesis are presented thereafter followed by contribution of three proposed works. The thesis organization is also outlined in this chapter.

Chapter 2 exhibits related work on community detection in dynamic networks. The chapter consist of three sections. Firstly, dynamic network modelling is explained, followed by a description of related literature, and the last section is dedicated to various explanations of evaluation metrics.

Chapter 3 address the community detection problem using connectedness and influence perspective. With these properties, we present a tree-based community detection algorithm on dynamic social network (TCD2).

Chapter 4 propose a Pareto front principle-based algorithm utilizing three objective functions inspired from real social behaviour of an entity.

A fuzzy-based approach for solving community detection problems in dynamic networks is proposed in Chapter 5. It also formulate a fuzzy membership function for determining belongingness of a node to any community.

Finally, Chapter 6 concludes the dissertation and presents some direction for future work.