Chapter 1

INTRODUCTION

1.1. Background

Real world complex systems can be represented in the form of networks. To understand the in-depth structure and detail function of those systems, it is important to study and analyze the networks (Mislove, et al., 2007). A trivial property of these networks is Community structure obtained by partitioning the network into several groups, within which connection between nodes are more dense than the rest of the network. The sets of This type of grouping is commonly referred as communities, but also known as clusters, cohesive groups, or modules as there is no globally accepted unique definition (Fortunato and Castellano, 2012). The concept of community detection is related to graph partitioning in some way; though it is very much dissimilar from graph partitioning. In case of graph partitioning, number of groups and the approximate size of those groups are known priory and the task is usually to divide the network into these many numbers of disjoint sub-graphs of almost same size, irrespective of whether a partition even exists. But in case of community detection, it is not known that how many communities are present in the network and it is not at all mandatory for them to be of same size. The community detection approach assumes that most of real world networks, divide naturally into groups of nodes (community) with dense connections internally and sparser connections between groups, and the experimenter's job is only to detect these already formed groups (Farrag and Nasr, 2017). The number of partitions and size of them are settled by the network itself and not set by the experimenter. So community detection is the technique which aims at discovering natural divisions of (social) networks into groups based on strength of connection between vertices' (Xie, et al., 2013).

Basically, community can be subdivided into two types; disjoint communities and overlapping communities. In disjoint communities nodes can be part of only a single community, but in overlapping communities partitions are not necessarily disjoint. There could be nodes that belong to more than one community (Lancichinetti, et al., 2009).

A social network is a collection of finite set of members (nodes) which can be a single person, a group, an organization; and relations (edges) among them may represent friendship, influence, affection or conversely, dislike, conflict or many other similar entities. In a social network, a community could be a group of people with common interest or location (Newman, 2012). Generally in any social network a person may be part of more than one different group or community, like a person can be part of his/her professional group and simultaneously can be part of his/her family group indicating overlap between the professional and family group. So for social networks, overlapping community detection technique should be considered over disjoint community detection technique. A schematic representation of a network with three communities is shown in Figure 1.1.

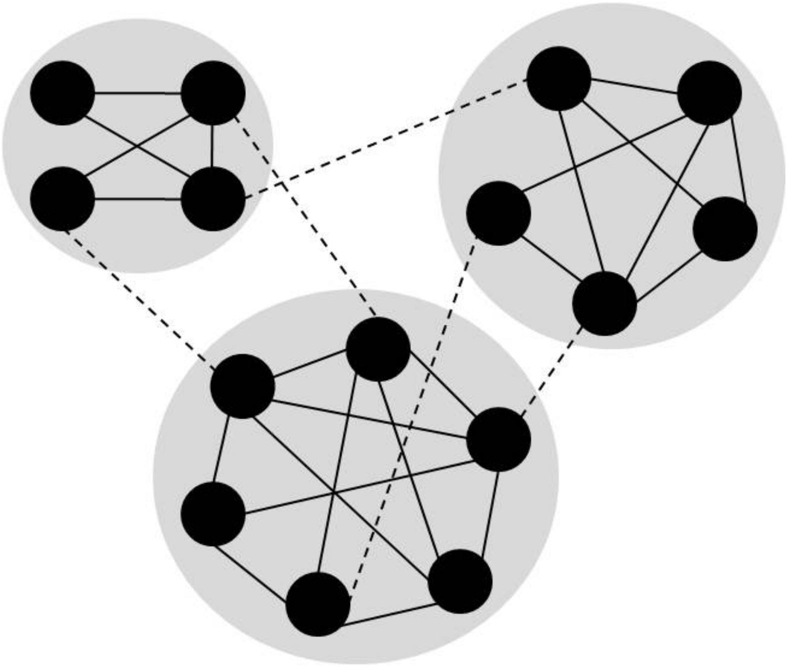

Figure 1.1 A schematic representation of a network with community structure.

1.2. Motivation for Work

The ability to find and analyze communities present in the network can provide help in understanding and visualizing the structure of networks. Social creatures interact in diverse ways by forming groups, sending messages, sharing items, joining in group discussion etc. Some of the interactions are accidental while others are a consequence of the underlying explicit or implicit social structures. In order to understand social interactions and the low-level structure of the network, it is therefore crucial to identify these social structures or communities. The well known structure of the network can have significant applications. For effective online marketing, such as placing online ads or deploying viral marketing strategies, known community structure in social network could often lead to more accurate targeting and better marketing results.

1.3. Major Issues

Community detection has significant importance in sociology, biology and computer science disciplines where systems are often represented as networks. Numerous techniques have been developed for community detection, yet the problem is not solved satisfactorily (Zhang and Yeung, 2012). There are several issues emerged along with the community detection problem, some of which are as follows:

- Identification of accurate communities is a major issue in community detection problem. Existing community detection algorithms mostly have to compromise on accuracy of communities even though the quality of identified communities is high (Chakraborty, 2015; Havemann, et al., 2011).
- Most community detection algorithms require prior information about communities (Gao, et al., 2010; Lancichinetti and Fortunato, 2009) (number of communities). The prerequisite input to an algorithm puts limitations by default to the algorithm. Moreover, structures of real networks are mostly unknown so any required prior information about communities has to assume. Those presumptions mislead to the identification of inaccurate communities. Hence, it is important to put a check on requirement of prior information about communities.
- In real-world scenarios, a person usually involves in different spheres of social relationship, splitting the time among a circle of friends, a club, and family. Thus, social networks contain disjoint communities as well as overlapping communities, where a node can appear in multiple communities with different belongingness (Ahn, et al., 2010). Therefore, community detection algorithms not only have to sense network structures but also quantitative affiliations to multiple communities. Hence, fuzzy membership of nodes to different communities has to be explored (Xie, et al., 2013).
- Real-world systems are complex as those cover wide range of aspects such as multiple relationships (Ahn, et al., 2010; Jacobson and Wilensky, 2006), organizational hierarchy or directional associations etc. To incorporate actual functionality and properties of real system, different kinds of system specific network representations are considered (eg. directed networks, weighted networks, multiple feature networks) (Barrat, et al., 2004; Boccaletti, et al., 2006;

Lancichinetti, et al., 2011). Identification of communities in those diverse networks with single algorithm is a challenging task.

- Assurance of both quality and accuracy is a major issue during the evaluation of communities (Sethi, 1979). Measuring quality incorporates edges, while measuring accuracy involves node labels (Bhagat, et al., 2009; Kohli and Torr, 2008). This fundamental difference between the two measures has led to the trade-off between accuracy and quality. Trade-off between quality and accuracy is a major issue during performance evaluation of community detection algorithms (Agarwal and Roth, 2002; Bagalà, et al., 2012).
- Mostly community detection algorithms are random in nature (Leskovec, et al., 2010). Different communities are identified in different executions of algorithms for the same network (Liu and Murata, 2010; Pons and Latapy, 2005). Accumulation of outputs obtained in different executions of an algorithm during performance evaluation is another challenging issue.

1.4. Objectives of the Thesis

The objective of this thesis is to effectively explore and identify the naturally formed communities in a network so that the in-depth structure of the network becomes understandable. Diverse overlapping and disjoint community detection algorithms need to be studied and implemented to investigate for the effective and better approach for this purpose. To identify the best and effective approach, they need to be evaluated by some means. Various overlapping modularity measures are needed to be considered to measure effectiveness of the approaches. The thesis is focused on seven objectives that are discussed below in five categories.

1) Assuring accuracy of communities: This goal is achieved either during identification of communities or during evaluation of communities. Community detection algorithms mostly explore dense connectivity to identify communities in the network. Dense connectivity holds no clear correlation with the nodes representing the original entities of real-world network, resulting in identification of inaccurate communities. There has a consensus process in forming communities of real-world network, those are not formed instantaneously, particularly the social networks. Forming a social community involves interaction of persons and their relationships. Thus, involvements of nodes also have to be considered to identify communities in real sense. Therefore, following objective is considered to assure accuracy during identification of communities.

Objective 1: Investigate the role of nodes in community formation in real networks and explore the possibilities of deepening the involvement of nodes in the community detection process.

On the ground of evaluation, both quality and accuracy are important for communities. Quality of communities is measured by considering the connectivity among community members. Accuracy of communities is measured by comparing members of communities with ground truth. Real-world networks mostly do not have ground truths so accuracy cannot be measured for those networks. However, quality can be measured easily since it does not require ground truth. Hence, it will be advantageous if accuracy can be ensured alternatively via quality measure. Though accuracy can be measured for the networks where ground truth is available, the evaluation process has to deal with the trade-off between quality and accuracy measures. Therefore, following two objectives are considered during evaluation of communities.

Objective 2: Define quality metrics for better assurance of accuracy.

Objective 3: Design effective evaluation methodology to mitigate the trade-off between quality and accuracy.

2) **Exploring the overlapping nature of community members:** In real-world scenarios, nodes exhibit degree of belongingness to different communities rather than having membership of single community. Identification of disjoint communities is not sufficient to meet the realities involving partial membership of nodes. Thus, the partial membership has to be investigated to uncover overlapping communities. Again, considering only overlapping communities will also be inappropriate as some nodes may engage only with single community. Therefore, involving both partial and full membership of nodes following objective is considered.

Objective 4: Develop an algorithm for uncovering both disjoint and overlapping aspects of communities.

3) **Dealing with diverse networks:** Various kinds of networks are developed to represent real systems, which include simplest form of networks to complex form of networks such as multiple featured networks. As discussed above, multiple featured networks consists several simple networks. Thus, community detection in such networks is challenging as multiple simple networks have to be processed. Most community detection algorithms yield at least quadratic complexity in simple networks. Often, meta-heuristic approaches such as nature-inspired evolutionary techniques are used to identify communities fast. Hence, the motive is to design suitable objective function and to incorporate evolutionary technique for

identifying communities in multiple featured networks. Therefore, following objective is considered.

Objective 5: Develop efficient community detection algorithm to handle complex relationships in multiple featured networks.

4) Application of communities: Once communities are identified in the networks, an immediate question may arise that how this information further utilize in different applications. Obviously, nodes within and outside of a community have different meaning from the viewpoint of applications. The nodes may influence the application based on their locality within the community they belong. Similarly, inter-community and intra community connections also may exhibit different influence on applications. Post-hoc analysis of communities has to incorporate those influences on different applications. Therefore, following objective is considered to study influence of nodes and connections on applications based on their locality in community structure.

Objective 6: Examine applicability of communities in the perspective of their influences.

5) Dealing with of output variations: Community detection algorithms that are of random nature identify different community structures in different execution of algorithms. Comparing performance of those community detection algorithms is challenging. Generally, various metric values obtained for identified communities are considered to compare performance of algorithms. Existing analysis methods such as non-parametric analysis, where mean, median or standard deviation are computed against different metrics can only give overall information regarding the distribution of metric values. However, if the communities are good (or bad) then computing mean, median or standard deviation cannot express how good (or bad) are the communities in comparison to other algorithms. Following objective is considered to overcome this drawback.

Objective 7: Develop an evaluation methodology for comparing different outputs of community detection algorithms.

1.5. Contributions

Main contributions of the thesis are divided into three parts addressing the aforementioned seven objectives. Considering the first and second part both are worked on Genetic algorithm. First part is based on the Genetic algorithm with OBL (opposition based learning) and modified the crossover. In GA we employed the matrix encoding technique and regenerative idea behalf of the mutation operation. In second part we used the GA for fuzzy community detection and evaluation. In third and the last part we used the Differential evolution algorithm for the community detection and improve the quality and accuracy of the found communities. Details about the contributions are explained below.

1.5.1. Community Detection Using Genetic Algorithm with OBL

In every optimization technique, initialization process plays an important role in convergence rate. In this work, Opposition based Learning (OBL) has been deployed for initialization. OBL not only check for the solutions on the initial guess, but also on the points opposite to the initial points. This approach ensures a faster convergence to solution and minimizing the probability of local minima/maxima.

To solve these shortcomings of GA while applying to community detection problem, Modified Crossover Opposition Based Genetic Algorithm (MCOBGA) has been proposed. To the best of our knowledge, it is the first time GA with dual initialization of population has been introduced for community detection. The other key contributions of this paper include: First, the design of an improved version of standard single point crossover in GA to transmit important information about the community structure during evolution of MCOBGA; Second, one of the important findings have been initialization by Opposition based learning. An intermediate algorithm based on OBL (OBGA) and the proposed algorithm MCOBGA showed significant improvement in convergence and quality and accuracy of algorithms. Third, the matrix encoding has been used for fast and effective process of real world networks data; Forth, a thorough evaluation of the performance of MCOBGA on real world social networks, which achieved better results than GA with vertex similarity (Li, et al., 2013) applied to community detection.

In this proposal, we proposed a regenerative genetic algorithm for detecting communities in social networks. Based on my literature survey of genetic algorithms for community detection reveal that quality and accuracy of the communities are not very well and the problem is initialization phase. So that once, crossover and mutation operations are done based on the random initialized solution members, i.e., if the crossover rate is (0.8) we cross the 80% of the fittest solution members and the mutation operation is done on rest 20% and the next generation of solution members is generated. This makes the approach little constrained and less efficient. The fitness of solution members is calculated through fitness function, a detailed description of fitness function is dealt in later sections of this paper. In our algorithm, we go for different approach regarding crossover and mutation operations on initial solution members. Solution members in our approach is initialized randomly first, and for suppose if the crossover rate is (0.8) then 80% of the fittest solution members are sent for cross over operation and for mutation we generate new population members again randomly and out of this 20 % of solution members are picked randomly and mutation is done on these 20% solution members. Total of newly generated 80% solution members after crossover operation and newly generated 20% solution members after mutation operation are sent for next iteration. This process is repeated until we get the optimal solution members for a given dataset. A detailed explanation of our algorithm is briefly portrayed in the later sections of this paper.

1.5.2. Fuzzy Based Community Detection

We use an evolutionary technique for detecting communities. After my long review (Newman, 2012; Porter, et al., 2009; Yang, et al., 2011) about the traditional algorithm and evolutionary algorithms, we will find much expectation for the genetic algorithm in further experiments. One of most precious and the widely used algorithm is the Genetic algorithm. This algorithm based on Darwin`s theory of evolution, Just as the chromosomes of different individuals are diverse and evolve with each generation with the theory of survival of the strongest. Here the individuals are operated upon using specific techniques, and then a solution set is created, in which the best of them is taken and again operated. An iteration of this takes place until we have a good enough solution set.

Till now much of the focus has been on the disjoint detection of communities. That is Individual communities having no relation with each other were being detected. The assumption made is that a network has dense connections around certain nodes and more connections there rather than any other nodes in the network. However, in real life, this is not the case. A person who has a family group can belong to other groups as well. These groups can be friends, workgroup, sports group, etc.

So basically real life network is not a simple collection of individual communities connected to each other, but a social network has nodes that belong to different communities. So an overlap is an important property that needs to be incorporate in our network detection. Hence in this paper, we talk about overlapping community detection algorithms that detect a set of groups that are not necessarily disjoint. There could be vertexes that belong to more than one community.

Community detection is an important research topic in the field of complex networks. Genetic algorithms have been used as an effective optimization technique to solve this problem. In this proposal, we employed the GAFCD (Su and Havens, 2014)with some modifications for the community detection. We choose that one because it's only algorithm for find the crisp and fuzzy communities in social network. We try to create a new algorithm just like a GAFCD and compare the GALS (Jin, et al., 2011) and MSFCM (Lin, 2014).

In the last section, we have employed the genetic algorithm for community detection in social networks. During this experiment, we are trying to do the best propose algorithm for both overlapped & disjoint community detection through a single algorithm. The first step of the planned methodology no needs to extra optimizing steps. It utilizes node similarity to initialize the population and performs simple crossover and mutation process, however reaches high accuracy. Therefore, it's a straight forward and efficient algorithm for disjoint community structure identification in social networks. The second step of the proposed method for overlapping community detection involves permanence based vertex replication algorithm that eliminates the need to develop a separate algorithm of community detection for overlapped communities. The new metric called permanence is using on every node, and overlapped vertices discovered from a disjoint community structure. This proposed algorithm name is NSGAP (Node similarity based Genetic Algorithm with Permanence concept).

1.5.3. Community Detection Using Differential Evolution Algorithm

The Differential Evolution (DE) algorithm is also a part of evolutionary techniques. DE differs from the other algorithms because it does not require prior information about the datasets and co-relation. It is a very useful property for the real-life datasets. In the recent years, DE is very popular as an optimization technique in the area of the social network. In this work, we want to try some experiment on DE for improvised the algorithm and find the most efficient and effective algorithm, So that, without modification in DE algorithm we replaced the fitness function and applying the different datasets. In this proposal, we totally focused on DE algorithm with multiple objective functions and found the best combination (variation) of the DE algorithm. After this experiment, hope we will find the best version of DE for community detection in the social network. We will analyze the new variation of DE for several parameters of the Accuracy and Quality wise. We have to used the 7 objective functions i.e. conductance, internal density, Average degree, Normalized cut, Expansion, cut ratio and modularity.

Detecting communities in networks helps us a lot in deriving essential information about interactions and relationship among the nodes. In past, many efficient algorithms have been proposed for community detection (Raghavan, et al., 2007; Shen, et al., 2009). It can be defined as a process which aims to identify clusters on the basis of different parameters and depends on the network's structure such as distance between nodes, relationships among them etc. We propose an improvised Differential Evolution approach based on the technique of vertex similarity (Li, et al., 2013), abbreviated as VSDE, for community detection in complex networks. In VSDE, the initialization step is based on the similarity between nodes in the network which exploits the initial structural similarity of the network to generate a better initial population and hence better community partition.

The concept of the opposition has an old history in sciences and other fields. This is the first time to contribute in social networks for community detection to enhance an optimizer. In this experiment, present a novel scheme to make the differential evolution algorithm faster. The proposed opposition based DE employs opposition based optimization (Wang, et al., 2011) for population initialization. Opposite numbers have been utilized to improve the convergence rate of traditional DE. The proposed algorithm also employs the tournament selection method for mutation and accelerates the differential evolution. We have to employed the combination of opposition based learning and Tournament selection method with DE for community detection in social networks. Combination of technique are optimized the results. New versions of proposed algorithm name are TOBDE, OBDE and TDE.

1.6. Thesis Organization

The thesis is organized into six chapters.

Chapter 2 provides literature survey on community detection approaches and their validation techniques.

Chapter 3 Identify community structures using Genetic algorithm with opposition based learning concept that is called the MCOBGA and another proposal is based on regenerative concept name is RGA (Regenerative Genetic Algorithm).Evaluation of the community structure with the help of accuracy and quality metrics.

Chapter 4 deepens further the role of Genetic Algorithm in the field of social network analysis with the help of community detection. We go further and find the disjoint and overlapped community detection using new generation of GA called FGA, MGAFCD. After this we include a new metric called permanence and algorithm name is NSGAP for the purpose of disjoint community and overlapped community detection in social network. This method is just like a GAFCD.

Chapter 5 deals with community detection using the Differential evolution algorithm. In this section, we employed the DE with multiple objective functions and find the improved version of DE algorithm for the different conditions and various type of datasets. After this we used the node similarity concept with DE algorithm and proposed algorithm name is VSDE. In the last section of this chapter we employed the OBL concept and Tournament method and proposed a TOBDE, OBDE and TDE algorithms and compare the results with some existing methods.

Chapter 6 concludes the contributions and details possible future directions in respect of each of the proposed works.