Chapter 1

Introduction

Many real-world systems are represented as networks interpreting various entities of the system as nodes and interactions among the entities as connections. Examples include social networks such as acquaintance networks [48, 87, 120, 220], professional networks [59, 182], and collaboration networks [15, 139], biological networks such as protein-interaction networks [76, 211], gene networks [83, 106], and metabolic networks [72, 170], and ecological networks such as supply-chain networks [204]. Most real networks exhibit preferential connectivity [14], indicating disproportionate connections across various groups of nodes. Furthermore, the distribution of connections in the networks also relies on the phenomenon like cohesiveness [172] or homogeneity [49, 77] that naturally brings similar nodes close by establishing new connections. Presence of such densely connected latent groups in the network are termed as *communities*, *modules*, or *clusters*. Though formal definition of communities lacks generalized consensus, widely-accepted definition endorses connectivity within the community should be comparatively higher than the connectivity with rest of the network [88]. A schematic representation of a network with three communities is shown in Figure 1.1.

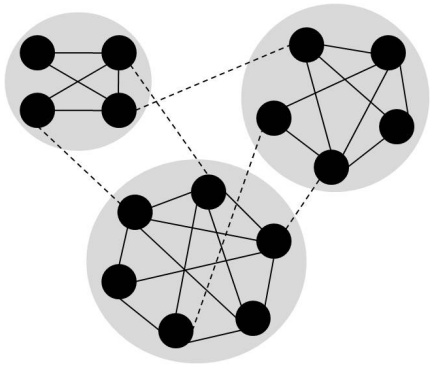

Figure 1.1: A schematic representation of a network with community structure.

1.1 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 (see [54] for the reviews). 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 are high [7, 148].

- Most community detection algorithms require prior information about communities (e.g. number of communities) [63, 186, 206]. The prerequisite inputs to an algorithm force limitations by default to the algorithm. Moreover, structure 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 [71]. 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.

- Real-world systems are complex as those cover wide range of aspects such as multiple relationships [9, 114, 164, 192], organizational hierarchy [131, 168, 226], or directional associations [125, 127, 146] 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 featured networks). 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 [7, 68, 148, 194]. Measuring quality incorporates edges, whereas measuring accuracy involves node labels. 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.

- Mostly community detection algorithms are random in nature (see [152] for the reviews). Different communities are identified in different executions of algorithms for the same network. Accumulation of outputs obtained in different executions of an algorithm during performance evaluation is another challenging issue.

1.2 Objectives

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, involvement 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, whereas 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. One of the complex form of networks is multiple featured networks that consists several networks. Thus, community detection in such networks is challenging as multiple networks have to be processed. Most community detection algorithms yield at least quadratic complexity in the networks where single connection exist among nodes. Often, meta-heuristic approaches such as nature-inspired evolutionary techniques are used for fast identification of communities. 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 to utilize this information further 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.

1.3 Contributions

Main contributions of the thesis are divided into five parts addressing the aforementioned seven objectives. Considering the first and fourth objectives two agglomerative algorithms are proposed. Both these algorithms do not require specifying number of communities in prior. An evolutionary algorithm is proposed covering the fifth objective. Considering the sixth objective, applicability of community structure into link prediction problem is investigated and proposed an algorithm involving community information. Covering the community evaluation related objectives i.e. second, third and seventh objectives, a set of three quality metrics and two community evaluation methodologies are proposed. Detail about the contributions are explained below.

1.3.1 Ego Network Based Community Detection

Considering the first objective, formation of social community is investigated from the perspective of ego network [11, 55, 129, 149, 207], which is a person centric network. Various levels of ego networks are discussed which include level 1.0 ego network, level 1.5 ego network and level 2.0 ego network. Level 1.0 ego network of a person or a node in the network is simply constituted with its neighbors, where node is the ego and its neighbors are alters. Incorporating the level 1.0 ego network, the notion of mutual interest in social relationship is introduced. Mutual interest involves two-way personal interests "from ego to alter" and "from alter to ego" respectively termed as forward interest and backward interest. It is believed that the formation of any social community has to be instantiated by a person or a group of persons anywhere within the network. Then the community grows as the new members joins. Naturally, the persons that are already member invite their colleagues and if any one agrees to join the community he becomes

a member. Here, sending an invitation is treated as forward interest, and agreeing to join the community is considered as backward interest.

Existing community detection algorithms mostly consider only forward interest in identifying communities. In existing algorithms, new members are identified based on the connectivity of community members. However, it has never been examined the likelihood of joining the community collectively with forward interest from the viewpoint of newly identified members. Primary focus of the work is on this direction i.e. how to incorporate the backward interest in community identification process. Involving backward interest not only explores the aspect of community formation that was missing earlier but also ensures greater role of individual nodes in community detection. A property called *Reachability* is defined to determine any node's the likelihood of joining a community. Another property called *Isolability* is defined to quantify the ability of any community to isolate its members from rest of the network. Based on these two properties an agglomerative algorithm called ENBC is designed. Complexity of the algorithm is shown to be $O(n^2)$. Wide range of empirical analysis on both real-world networks and synthetic networks shows ENBC identifies highly accurate communities compared to six state-of-the-art algorithms.

1.3.2 Agglomerative Fuzzy Community Detection

Following the forth objective, overlapping aspect of communities is considered in addition to disjoint communities. The concept of *Reachability* is extended further to uncover overlapping communities. Core nodes of communities are defined following the same principle as Reachability. Membership degrees of nodes to different communities i.e. the belongingness of nodes to different communities is defined involving the core nodes. Further exploring the first objective, equal opportunity is given to every node to form its

own community. The notion of self-membership is introduced for this purpose. If any node posses higher self-membership degree than the membership degree to other communities, the community associated with that node grows. The node having sufficient self-membership degree to its community is referred as an *anchor*. Communities grow in reference to the anchors they are associated with. The process of community identification follows agglomeration i.e. communities grow by accumulating new members in successive steps. An algorithm called FuzAg is designed to identify both disjoint and overlapping communities. The FuzAg algorithm first generates communities with membership degrees of nodes. Then a process called hardening of partitions [79] is applied to distinguish disjoint and overlapping communities. Empirical analysis on both real-world networks and synthetic networks show better performance of FuzAg algorithm compared to four state-of-the-art fuzzy community detection algorithms. Disjoint communities resulted by FuzAg algorithm are more accurate. Quality of overlapping communities are also better than other algorithms. Complexity of the algorithm is shown to be $O(n^2)$.

1.3.3 Community Detection in Multiple Featured Networks

To address the fifth objective, multiple featured network is considered for identifying communities. Connections in multiple featured networks are much more complicated than the networks where single connections exist among nodes. A multiple featured network contains several networks with same set of nodes. Each of those networks is an independent network involving specific feature. Thus, multiple connections present between two nodes in multiple featured networks, causing difficulty in identification of communities. Community detection in the networks having single connections among nodes is a NP-Hard [54, 156]. Often evolutionary techniques are considered for community detection in those networks by optimizing an objective function [26, 57]. This work studies community detection in multiple featured networks using PSO algorithm. An objective function

is defined based on Davies-Bouldin index [39]. A novel cognitive avoidance mechanism is introduced in the standard PSO, i.e. PSOCA (PSO with Cognitive Avoidance) . Empirical analysis on 25 benchmark functions shows the improved PSO results better solutions than other two variants of PSO. Optimizing the aforementioned objective function using PSOCA, communities are identified. Results on real-world networks indicate that community detection with PSOCA is much faster than other variants of PSO and identified better communities.

1.3.4 Applicability of Communities to Predict Missing Links

Prediction of missing links or future links in the network has great interest in several domains [4, 116, 121, 205]. Considering the sixth objective, this work studies the applicability of community structure. This work is primarily focused on two factors: the importance of the neighbors of a node in the network and the locality of the nodes in different communities. Three edge centrality measures: edge betweenness centrality, k-path edge centrality and spanning edge centrality are considered to define importance of neighbors of a node. Information about communities is utilized to influence the likelihood scores of missing links based on the localities of nodes associated with existing links or connections. Positive weight is assigned to an existing link if the associated nodes are members of same community for ensuring positive influence on the likelihood score, otherwise assign negative weight. A Community-based Link Prediction (CLP) algorithm is proposed. Performed a comparative analysis of three edge centrality measures with CLP algorithm on both real-world networks and synthetic networks. The analysis revealed all of the three centrality measures yield similar results in terms of quality and accuracy of missing links. However, CLP with k-path edge centrality predicts links in least time. Complexities of CLP with edge betweenness centrality, k-path edge centrality and spanning edge centrality are shown to be $O(n^2)$, $O(n^2)$ and $O(n^3 \log n)$ respectively.

1.3.5 Designing Validation Metrics and Evaluation Methodologies

Following the second objective related to community evaluation, a set of three quality metrics AVI, AVU and ANUI is proposed to assure accuracy of communities. The metric AVI is focused on intra-community connections, and it is designed utilizing isolation property that isolates the members of community from rest of the network. The metrics AVU is focused on inter-community connections, and it is designed by using unification property that unites members of smaller communities into one community. AVI should be high and AVU should be low for good communities since AVI and AVU involve intra-community and inter-community connections respectively. The metric ANUI is defined to balance both AVI and AVU. Empirical analysis on real-world networks as well as synthetic networks shows capability of proposed metrics in dealing with accuracy. Axiomatic analysis performed on theoretical ground shows proposed metrics satisfy all of the six quality related properties suggested by Van Laarhoven and Marchiori [196].

Considering the third objective, a framework is designed to analyze Relative Inclination Towards Accuracy (RITA) of a set community detection algorithms. The framework of RITA analysis utilizes Multiple Criterion Decision Making (MCDM) [12, 103, 219] technique to accumulate indications of various metrics into single score. RITA analysis involves both quality metrics and accuracy metrics in order to determine inclination of algorithms i.e. how likely an algorithm will identify accurate communities in comparison to other algorithms. RITA analysis is simple, it just requires to visually inspect the trend and height of the curves representing inclination of different algorithms. Empirical analysis on various real-world network shows the effectiveness of RITA analysis in indicating relative inclination of algorithms. Most importantly, with RITA analysis, the trade-off between quality metrics and accuracy metrics disappears during evaluation.

Lastly considering the seventh objective, another visual analysis methodology is proposed

to deal with *output variation* of community detection algorithms during evaluation. Solutions obtained with different algorithms (in terms of specific metric) are considered to compare performance of algorithms. Proposed methodology utilizes the concepts of quantile-quantile plot and simple regression analysis. Quantile-quantile plot ensures involvement of each individual solution in the evaluation process. With quantile-quantile plot, three kinds of dominance are defined in reference to X=Y line (referred as neutral line) to express how good or bad are the solutions of one algorithm compared to other algorithms. To accumulate overall dominance of all points in the quantile-quantile plot, linear regression analysis is performed. Then regression line dominance and shifting mechanism is developed, and incorporated into the analysis methodology.

With proposed methodology, algorithms are compared in terms of dominance of one algorithm over other algorithms. The dominance of an algorithm is determined simply by observing the angle between regression line and neutral line, and the position of intersection of regression line and neutral line. With proposed methodology, it can be easily expressed that how good or bad are the solutions in comparison to other algorithms, which otherwise was difficult to analyze. For instance, with existing analysis method such as non-parametric analysis, where mean or median are computed can only give only overall information regarding the distribution of solutions but cannot express how good or bad are the solutions in comparison to other algorithms. Quantile-quantile plot is used to define dominance of each point, which is nothing but comparison of good (or bad) solutions of one with good (or bad) solutions of other algorithms. Thus, with proposed methodology, it can be explained easily if an algorithm is performing better, actually, how good the solutions are compared to other algorithms.

1.4 Thesis Organization

The thesis is organized into eight chapters.

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

Chapter 3 investigates community structure from the perspective of a personalized network called ego network. Defined two properties *Reachability* and *Isolability*. Utilizing these two properties, ENBC algorithm is designed involving the role of nodes in the community detection process.

Chapter 4 deepens further the role nodes by introducing the notion of *self-membership*. A fuzzy agglomerative algorithm FuzAg is proposed to identify both disjoint communities and overlapping communities.

Chapter 5 deals with community detection in multiple featured networks. A novel *cognitive avoidance* mechanism is introduced in standard PSO algorithm, and then used to identify communities in multiple featured networks.

Chapter 6 studies prediction of missing links in the network as an application of community detection problem. A community-based link prediction scheme is proposed by incorporating the community structure and various edge centrality measures.

Chapter 7 details about proposed metrics and discusses important properties. A framework is designed to analyze relative inclination of community detection algorithms toward accuracy i.e. RITA analysis. Defined shifting mechanism of regression line and designed a methodology to compare performance of randomized community detection algorithms.

Chapter 8 concludes the contributions and details possible future directions in respect of each of the proposed works.