Chapter 2

LITERATURE SURVEY

This chapter surveys the related works covering various objectives of the thesis. Generally, two kinds of communities are identified: 1) disjoint communities where nodes can have the membership of only one community and 2) overlapping communities where some nodes can have membership of more than one community. Techniques of identifying both disjoint and overlapping communities are discussed. Traditional methods and especially evolutionary algorithms are also discussed in this chapter. Various metrics to evaluate both kinds of communities as well as evaluation methodologies are described. Generic datasets that are widely used to validate communities are discussed. Various applications involving post-hoc analysis of communities are also studied.

2.1 Elements of Community Detection

The elements of community detection have been well documented by Fortunato (Lancichinetti, et al., 2011) .The three primary elements of community detection are: Computational complexity, Communities and Partitions.

2.1.1 Computational Complexity

Efficiency is a critical issue for clustering algorithms because of the enormous amount of data on current real-world networks (Liu and Yu, 2005). The computational complexity of an algorithm is an estimate of the amount of resources (time and space) required by the algorithm to perform its specific tasks. Time taken is estimated by the number of computation steps performed by the algorithm and space consumed is estimated by the number of memory units that are needed by the algorithm (Lee, et al., 2010). Expressing the scalability of these demands with the size of the problem being studied is a standard technique for analyzing algorithms. In dealing with a graph, the size is expressed by the number of nodes n and/or the number of edges m.

Many clustering algorithms or problems related to clustering are NP-hard (Guha and Mishra, 2016). This means that it is futile to use exact algorithms for obtaining the solution, which could be used only for very small systems. Also, even if an algorithm has a polynomial complexity, it may still be too slow to actually practically work for large systems of interest. In all such cases it is common to use approximation algorithms. These are algorithms that produce an approximate solution instead of an exact one, with the benefit of a lower complexity. The goal is to deliver a solution which differs by a constant factor from the optimal solution. Approximation algorithms are very often used for optimization problems, in which one wants to find the maximum or minimum value of a given cost function over a large set of possible system configurations (Fortunato, 2010).

2.1.2 Communities

The first issue in community detection is to look for a quantitative definition of a community. No definition is universally accepted. In fact, the definition frequently depends on the system at hand and/or the application one has in mind. An intuitive idea is that there must be more edges inside the community than edges linking nodes of the community with the rest of the graph (Leskovec, et al., 2009). This notion is at the basis of most community definitions. But other alternative formulations are also possible. Also, in most cases, communities are algorithmically defined, i.e. they just happen to be the final product of the algorithm, without a concrete a priori definition. In Figure 2.1, there are different categories of the communities in given below:

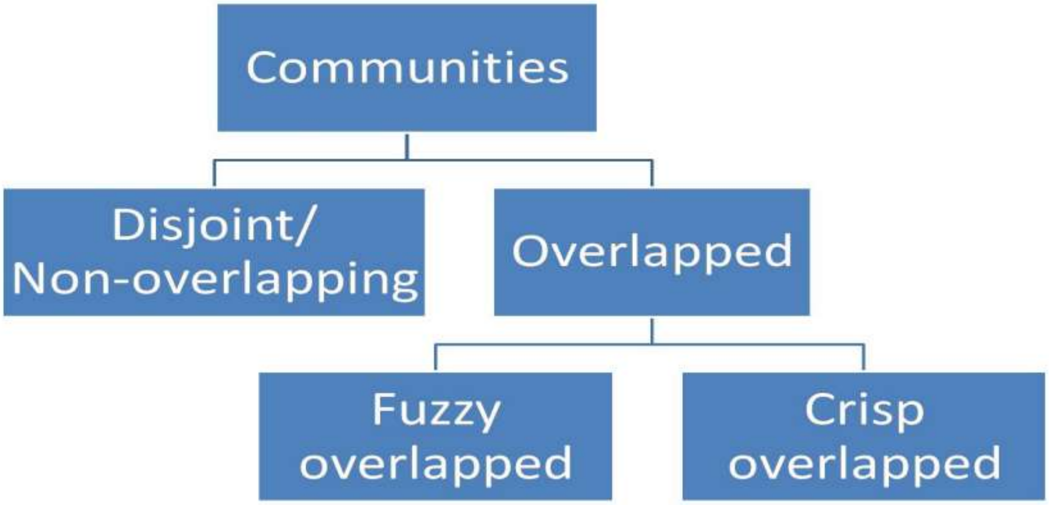


Figure 2.1 Categories of communities

Many local definitions (e.g. clique, strong community, weak community, etc.), global definitions (e.g. ones based on modularity, etc.) and definitions based on vertex similarity (e.g. ones based on euclidean distance, cosine similarity, etc.) have been proposed.

2.1.3 Partitions

A partition is a division of a graph in communities, such that each vertex belongs to one community. In real systems, however, a vertex may belong to multiple communities (Nepusz, et al., 2008) (e.g. a person can belong to multiple social circles in a social network). A division of a graph into overlapping communities is called a cover. Partitions can be hierarchically ordered, when the graph has different levels of organization at different scales (Lancichinetti, et al., 2009) . In such a case, clusters in turn display community structure, with smaller clusters inside, which may again contain smaller clusters, and so on.

Many clustering algorithms are able to identify multiple meaningful partitions (Jain, 2010). As not all the partitions are equally good, it is helpful to have a quantitative criterion to judge the goodness of a partition. A quality function is a function that assigns a value (a number) to each partition of a graph. The number is a measure of the goodness of the partition. We can then rank partitions based on their quality function value (Yang and Leskovec, 2015). Nevertheless, it should be kept in mind that the question of whether a partition is better than another one is ill- posed, and the answer depends on the specific concept of community considered and/or quality function used.

Many quality functions for determining the quality of a partition have been proposed, the most famous one being Network Modularity proposed by Girvan and Newman (Pons and Latapy, 2006).

2.2 Community Detection Techniques

Community detection algorithms are primarily focused on finding disjoint communities. However, often pointed out that it is common to see overlapping communities rather than disjoint communities in social networks (Ball, et al., 2011; Lee, et al., 2010; Xie, et al., 2013). Gregory has discussed two kinds of overlap: crisp overlapping and fuzzy overlapping (Gregory, 2011). In crisp overlapping, a node can belong to more than one community, but the node can have membership degree either 0 or 1. In fuzzy overlapping, membership degree of a node is in-between the range [0, 1]. The sum of membership degrees of a node to different communities is normalized to 1. Recent surveys have classified both disjoint and overlapping community detection algorithms based on the followed methodological principles, properties incorporated, input and output, or community definition (Akoglu, et al., 2015; Harenberg, et al., 2014; Sobin, et al., 2017). Fortunato divides disjoint community detection algorithms based on methodological principles (Plantié and Crampes, 2013), while categorize overlapping community detection algorithms based on community definition (Coscia, et al., 2011; Loe and Jensen, 2015). Xie et al. categorized solely the overlapping community detection algorithms based on the methodological principles covering both crisp and fuzzy overlapping communities (Rossetti and Cazabet, 2017; Xie, et al., 2013). Tang and Liu [188] have divided community detection algorithms based on the properties incorporated in the algorithm covering both kinds of communities (Tang, et al., 2017). Another recent survey also classified both disjoint and overlapping community detection algorithms from the perspective of input and output (Plantié and Crampes, 2013). In Figure 2.2, there are number of community detection techniques in given below with graphical format.

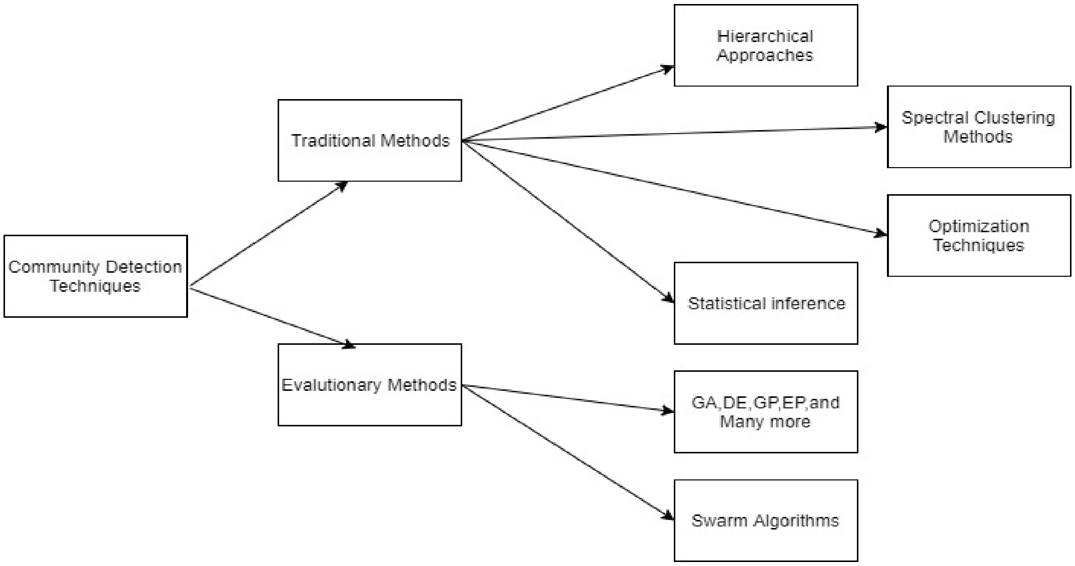


Figure 2.2 Type of community Detection Techniques

2.2.1 Hierarchical Approaches

Hierarchical approaches grow in two directions, divisive and agglomerative approaches. Both kinds of approaches also incorporate the properties described above for community exploration. Divisive methods follow top-down approach and agglomerative methods follow bottom-up approach. Agglomerative methods are mostly engaged with node level properties and assimilate nodes into communities starting from ground level entity such as nodes expanding through communities to entire network. Since, agglomerative approaches follow natural process of grouping specially the real world social network, communities detected are found to be more logical and accurate.

2.2.1.1 Agglomerative Techniques

One of the popular agglomerative algorithm is Fast Unfolding (FastU) proposed by Blondel et al., which is widely known as Louvain method (Blondel, et al., 2008). The FastU algorithm aggregates communities at different level to optimize modularity. Initially, individual nodes are considered as communities, aggregate these communities to get next level communities and so on until attain optimal modularity. Another popular agglomerative algorithm is (Xu, et al., 2007) , which also detects hubs and outliers with respect to communities. Unlike FastU algorithm, SCAN does not have clear levels of agglomeration. It merges nodes based on structural reachability and structural similarity. At the initial stage, the core nodes are identified and all those nodes that are structurally reachable from the core nodes are assigned to respective core nodes and form clusters. The process continues until all the nodes are either classified or marked as non-member. The nodes those identified as non-members are further classified as hubs and outliers.

Leader based agglomerative approaches such as Top Leader [100], Leader-Follower (LeadF) (Tolstedt and Anderson, 2013) and LICOD (Yakoubi and Kanawati, 2014) are developed in recent years. Leader based algorithms follow the concept that a community is constituted by a leader node followed by many other follower nodes. Prior to detect communities, these approaches first identify potential leaders of respective communities based on node and community level properties. Remaining nodes are assigned to one or more leaders depending on requirement of non-overlapping or overlapping communities respectively. Similar to leader based approaches, community expansion from pre-specified seeds is also introduced in recent years. Moradi et al. developed a local seed selection algorithm for overlapping community detection (Moradi, et al., 2014). The algorithm first selects some initial seeds and nodes are assigned based on their locality in reference to seeds. Recently, Zhang et al. have proposed another seed based agglomerative approach motivated by label propagation called Membership Degree Propagation (MDP) for fuzzy community detection (Alathel, 2015). Some seed nodes are considered based on local centrality. The algorithm iteratively propagates membership degrees of all nodes and each seed grows community around it. New seeds are considered iteratively and some seeds are deleted to improve the partitioning.

Besides exploring nodes as in leader and seed based approaches, connections are also prioritized in some approaches such as random walk (RandW) (Pons and Latapy, 2006) , where nodes are assigned to the same community if those fall within the randomly performed short walk. A similarity matrix is prepared to keep track of frequency every pair of nodes that appear in these random paths and using that matrix communities are formed. Wang et al. have proposed HC-PIN, where connections are grouped based on node level properties and associated nodes are assigned accordingly to the respective communities. Initially all the nodes are considered as to form different communities (Paxton, et al., 2015). Then Edge Clustering Value (ECV) is calculated for each connection and arranged in a decreasing order in a list (Cataldi and Sapino, 2010). A higher value signifies a greater tendency to be included in the same community.

2.2.1.2 Divisive Techniques

The philosophy of divisive techniques begins with the proposal of Girvan and Newman (Newman and Girvan, 2004). They primarily focused on edge between centrality since inter-community connections are supposed to have a large value of the edge betweenness. The idea was to remove all connections from the network starting with highest edge betweenness centrality. Tyler et al. proposed modification of Girvan-Newman algorithm to improve efficiency (Tyler, et al., 2003). The improved algorithm computes the contribution to edge betweenness following a sort of Monte Carlo estimation only from a limited number of centers that are randomly chosen. Rattigan et al. proposed another fast version of Girvan-Newman algorithm by approximating the edge between centrality (Mahajan and Kaur, 2015). Chen and Yuan pointed out a drawback of Girvan-Newman that counting all shortest paths to compute edge betweenness may led to unbalanced community sizes and proposed to consider only non redundant paths (Duan, 2012). Holme et al. another divisive algorithm, where nodes rather than edges are removed. They introduced a vertex centrality measure similar to edge betweenness and used that in the proposed algorithm (Azizifard, 2014). Girvan-Newman approach has been also modified by Gregory for overlapping community detection (Zhang, et al., 2013).

2.2.2 Optimization Techniques

The notion of optimization comes up with two pre-occupied questions. First question is about how to design a suitable objective function for obtaining good communities. Follow-up second question is about how to optimize the defined function. One approach for resolving the first question is through rigorous study of network and required communities. Another alternative, easy, and widely utilized approach is simply use the quality metrics (discussed in section 2.2.2) defined for evaluating communities. Numerous techniques have been developed to optimize the community related objective functions, which are discussed below.

2.2.2.1 Modularity Maximization

Newman-Girvan defined modularity as a stopping criterion for their community detection algorithm (Yang, et al., 2013). Afterward modularity rapidly becomes an essential element of many community detection methods. Modularity is by far the most used and best known quality function for community detection. Newman devised a greedy approach to maximize modularity (Newman, 2012). Later on, Clauset et al. proposed more efficient data structure like max-heaps to make Newman's algorithm faster (Lu, et al.). However, Wakita and Tsurumi noticed that the fast algorithm by Clauset et al. is inefficient due to the bias towards large communities (Labatut and Balasque, 2012). Schuetz and Caflisch proposed an efficient approach to avoid the formation of large communities. In order to favor small communities, Danon et al. suggested normalizing the modularity variation produced by the merger of two communities by the fraction of edges incident to one of the two communities (Liang, et al., 2014).

Lancichinetti and Fortunato raised limitation to modularity maximization regarding biases: the tendency to merge small communities and to split large ones (Cerina, 2015). They also pointed out that it is usually very difficult, and often impossible, to tune the resolution such to avoid both biases simultaneously. Arenas et al. adopted multi-level resolution version of modularity to overcome resolution problem of modularity (Carvalho, et al., 2014). Cafieri et al. proposed a divisive approach to locally optimize modularity (Cafieri, et al., 2014). Recently, Costa et al. proposed another divisive approach to optimize modularity (Bansal, et al., 2011). Besides these, modularity maximization within the framework of mathematical programming is proposed by Agarwal and Kempe (Chan and Yeung, 2011). Chen et al. used integer linear programming to transform the initial graph into an optimal target graph consisting of disjoint cliques, which effectively yields a partition (Linderoth, et al., 2001). Yazdanparast and Havens proposed positive programming approach to maximize modularity (Yazdanparast and Havens, 2017).

2.2.2.2 Heuristic Approaches

Community detection problem is a NP-Hard problem (Fortunato, 2010). Furthermore, Brandes et al. shown modularity maximization is NP-Complete problem (Brandes, et al., 2008). Several heuristic approaches have been developed to deal with the complexity of community detection problem, particularly nature-inspired evolutionary computation techniques are popular. Tasgin et al. proposed Genetic Algorithm based approach to optimize network modularity (Jia, et al., 2012). Pizzuti proposed GA based algorithm GA-net using newly defined function community score (Pizzuti, 2012). Pizzuti proposed another multi-objective GA based approach that optimizes two objective functions (Pizzuti, 2012). Gong et al. proposed a multi-objective Evolutionary Algorithm (EA) based on decomposition (Gong, et al., 2008). Shang et al. proposed hybridized GA with simulated annealing to improve the stability and accuracy of community detection (Manjarres, et al., 2013). Wen et al. proposed a maximal clique based multi-objective EA for overlapping community detection (Wen, et al., 2017). Li and Liu proposed multi agent based GA for community detection (Li and Liu, 2016).

Besides GA, another evolutionary optimization technique called Differential Evolution algorithm. Differential evolution algorithm is a very simple yet efficient evolutionary algorithm proposed by storn and price in 1995 (Price, et al., 2006; Qin, et al., 2009). DE algorithm is evolving from the very basic evolutionary algorithm like genetic algorithm with the help of some very important modifications e.g. Crossover operation, fitness function, mutation operation, biased process and clean-up operation similar to improve the quality of the individual in the population (Roberts, et al., 2011). In order to achieve better scalability to handle large-scale networks Therefore, Qiang Huang et al. make some change in DECD e.g. bias grouping, global network mutation and divide & conquer strategy is used (Shakya, et al., 2014). In this strategy divide a large scale problem into sub-components and evolve those subcomponents independently and co-adaptively.

Besides DE, another nature inspired optimization technique called Particle Swarm Optimization (PSO) is widely used for community detection (Sousa, et al., 2004). Xiaodong et al. proposed a PSO based web community detection approach (Xiaodong, et al., 2008). Gong et al. identifies communities by multi-objective discrete PSO (Gong, et al., 2014). Cai et al. used discrete to identify communities in signed networks (Gong, et al., 2014). Rees and Gallagher explored overlapping communities using PSO with decentralized multi- threading processing and label propagation. Another swarm intelligent technique called Ant Colony Optimization (ACO) is gaining attention in community detection (Ji, et al., 2013; Mandala, et al., 2013; Moradi and Rostami, 2015). Extremal optimization (Arenas, et al., 2007) is also used to maximize modularity for community detection. Furthermore, memetic algorithm is also used for community detection. Gong et al. used memetic algorithm for community detection (Gong, et al., 2011). Later on, they have proposed an improved version by incorporating level propagation and elitism strategy. In the same line of work, Gach and Hao proposed another memetic algorithm based community detection approach (Gach and Hao, 2012). Recently, Ma et al. developed a multi-level learning based memetic approach for community detection (Ma, et al., 2014).

2.2.3 Spectral Clustering

Spectral clustering is one of the traditional clustering techniques. The algorithms that are designed to cluster set nodes by using spectrum of matrix (better known as Eigen values) or matrices derived from eigen value are referred as spectral clustering techniques. Often, graph Laplacian matrices are considered as key tools for spectral clustering (Belkin and Niyogi, 2003). Normalized graph Laplacian matrices are used extensively in spectral clustering. Shi and Malik developed two-way normalized spectral clustering algorithm (Shi and Malik, 2000). Minimization of normalized cut is the basis of their algorithm. Following same principle, Ding et al. formulated min-max cut algorithm (Ding, et al., 2001). Ng et al. proposed kway normalized spectral clustering algorithm (Ng, et al., 2002). However, Nadler and Galun discussed the limitations of these approaches such as they cannot successfully cluster datasets that contain structures at different scales of size and density (Nadler, et al., 2008). Newman proposed recursive two-way spectral clustering algorithm considering modularity maximization instead of various graph cuts (Shiga and Mamitsuka, 2011). Verma and Meila argued multiway spectral clustering algorithms are expected to perform better if easily traceable structures are present in the data. Meila et al. [130] explored spectral clustering in the view of random walks (Pentney and Meila, 2005). Filippone et al. studied the unification of kernel and spectral clustering. Recently, Joseph et al. carried out wide range of study addressing the impact of regularization in spectral clustering (Walsh, et al.).

2.2.4 Statistical Inference

Statistical inference in graph data aims fit statistical models to the actual graph based on hypotheses on how nodes are connected to each other. Community detection problem is expressed as an inference or maximum likelihood problem then preferably Bayesian inference (Ronquist and Huelsenbeck, 2003) or other models such as Potts model (Reichardt and Bornholdt, 2004), block modeling (Zhao, et al., 2012), information theory (Rosvall and Bergstrom, 2007) etc. are adopted.

2.2.4.1 Generative models

Bayesian inference has been used extensively in community detection, where the best fit is obtained through the maximization of a likelihood (generative models). Hastings

(Peixoto, 2013) chose a planted partition model of network with communities. Reichardt and Bornholdt proposed fuzzy community detection algorithm based on Potts model (Duch and Arenas, 2005). (Gao, et al., 2017) proposed a generic Bayesian approach to identify communities inferring modules of the network. Newman and Leicht proposed a method based on mixture model and expectation-maximization technique (Newman and Leicht, 2007). (Zhou, et al., 2007) developed another similar method based on the group fractions. (Copic, et al., 2009) defined an axiomatization approach for community detection problem using maximum likelihood estimation. (Psorakis, et al., 2011) proposed a method based on Bayesian non-negative matrix factorization for detecting overlapping communities.

2.2.4.2 Information theoretic approach

The modular structure of a graph can be treated as a compressed description of the graph. Rosvall and Bergstrom utilized this concept to approximate the whole information contained in the adjacency matrix that represents graph (Gong, et al., 2012). They envisioned a a communication process in which community in the graph represents a synthesis of the full structure that tries to infer the original graph topology. Sun et al. also followed the same idea for bipartite graphs evolving in time (Getoor and Diehl, 2005). Rosvall and Bergstrom utilized further the notion of describing a graph by using less information to optimally compress the information needed to describe the process of information diffusion across the graph (De Domenico, et al., 2015). (Chakrabarti and Faloutsos, 2006) has applied the minimum description length principle to put the adjacency matrix of a graph into the (approximately) block diagonal form for a good compression of the graph topology, and having very homogeneous blocks, for a compact description of their structure. Ziv et al. proposed a principled information theoretic community detection approach on basis of module discovery in the network (Cannistraci, et al., 2013). (Buhmann, 2010) proposed a information theoretic model validation approach for clustering.

2.2.5 Other Approaches

Some algorithms that do not fit into the above mentioned categories are discussed below. Zhang et al. proposed an iterative process that reinforces the network topology and propinquity that is interpreted as the probability of a pair of nodes belonging to the same community (Chakraborty, 2015). The propinquity between two vertices is defined as the sum of the number of direct links, number of common neighbors and the number of links within the common neighborhood. Gregory proposed an overlapping community detection algorithm, where each node updates its belonging coefficient by averaging the coefficients from all its neighbors at each time step in a synchronous fashion (Xie, et al., 2013). (Xie and Szymanski, 2012) developed another overlapping community detection algorithm based on general speaker-listener information propagation process. A game-theoretic framework was proposed by (Chen, et al., 2010), in which a community is associated with a Nash local equilibrium.. (Bruggeman, et al., 2012) studied communities in the networks, where connection weights are both negative and positive.

In recent years, many community detection methods and survey have been introduced with each such methods being classified according to its algorithms. Most survey classifies research papers and methods according to the type of community detection algorithm. some useful survey are given below:

In survey by (Fortunato and Barthélemy, 2007) is exhaustive with respect to many community detection methods and has been based on a graphic representation (Plantié and Crampes, 2013). In survey conducted by (Porter, et al., 2009)only includes graph partitioning approaches and offers insight into graphical techniques through citing the first survey.

In survey by (Allahyari, et al., 2017) is quite exhaustive relative to all techniques relying on graphical representation and produces a good overview of the field through classifying all techniques in a tree structure.

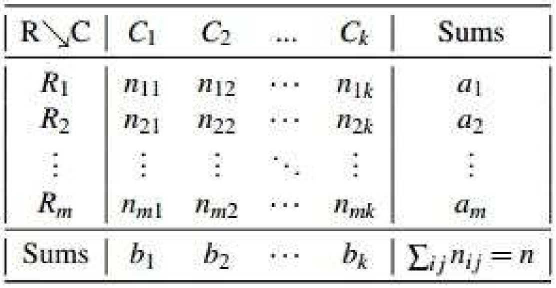
(Estrada, 2011) conducted a partial survey analyzing hierarchical type community detection methods and provides a number of leads for future community detection approaches conducted a partial survey analyzing hierarchical type community detection methods and provides a number of leads for future community detection approaches. In another survey by (Plantié and Crampes, 2013) incorporates several community detection methods and classifies them into five different families' i.e. classical approaches, separative approaches, agglomerative approaches, random walk type algorithms and miscellaneous approaches. (Papadopoulos, et al., 2009) classifies community detection techniques in five methodological categories' i.e. Cohesive group discovery (Wang, et al., 2005), Vertex clustering (Low and Tan, 1997), Community quality optimization (Huang, et al., 2013), Divisive (Savaresi, et al., 2002), Model based method (Chen, et al., 2012). (Danon, et al., 2005) primarily focuses on the performance of each type of algorithm.

2.3 Validation Metrics

Evaluation of predicted community structure is another aspect of community detection. If ground truth community structure of a network is available, then simply predicted communities are compared with that of ground truth to measure accuracy. However, it is difficult to collect ground truth communities for most of the real-world networks. Therefore, community evaluation has to rely on the quality related measures that are designed based on connectivity pattern of communities. In this section, first various accuracy metrics are detailed, and then briefed some popular quality metrics.

2.3.1 Accuracy Metrics

Communities predicted by the algorithm are evaluated in terms of various accuracy metrics as follows. Let C = (C_1, C_2... C_k) be the set communities obtained with a community detection algorithm applied to any network of n nodes. Let R = (R_1,R_2...R_m) be the real community structure. Here, C_k and R_m are interpreted as set of nodes in the respective communities. Overlapping of nodes in C and R can be summarized with a contingency table as presented in 2.1. Contingency table can also be presented with 2 × 2 matrix as shown in 2.2. Entries in the tables are decision pairs of two different nodes. If any pair of nodes are predicted to belong in the same community C_i and these two nodes are actually lie in same community R j then entry for the pair of nodes will be True Positive (TP). If nodes pair is predicted to lie in same community, but actually the pair belong to different community the entry will be False Positive (FP). If nodes pair is predicted to lie in different community and the pair actually belong to different community, then the entry will be True Negative (TN). If nodes pair is predicted to lie in different community, but the pair actually belongs to the same community the entry will be False Negative (FN).



R↘C	C_1	C_2	...	C_k	Sums
R_1	n_{11}	n_{12}	$\cdots$	n_{1k}	a_1
R_2	n_{21}	n_{22}	$\cdots$	n_{2k}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
R_m	n_{m1}	n_{m2}	$\cdots$	n_{mk}	a_m
Sums	b_1	b_2	$\cdots$	b_k	$\sum_{ij} n_{ij} = n$

Table 2.1: Contingency table Here, each n_{ij} denotes the number of nodes in common between communities C_i and R $_j$: $n_{i\,j}$ = |C_i ∩ R $_j$|.

2.3.1.1 ARI

Adjusted Rand Index (ARI) is corrected version of Rand Index (Vinh, et al., 2009) that measures the degree of overlapping between two partitions. As Rand Index suffers scaling problem and its expected value between two random partitions does not take a constant value (Meila, 2003). In proposed a corrected version as in the form of 2.1.

$$Ad\,justedIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex}$$

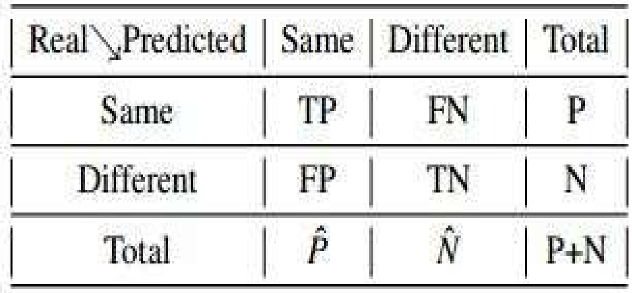
……. (2.1)

More specifically, with overlapping entries of different communities of C and R in contingency table as shown in Table 2.1, ARI can be computed as follows:

$$ARI(R,C) = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}$$

……. (2.2)

where, $n_{i\,j}$ is the number of nodes that are present both in communities C_i and $R_{\ j}$, a_i is the summation of all $n_{\ i\ j}$ corresponding to any R_j of R and all C_i of C, and b_j is the summation of all n_{ij} corresponding to any C_i of C and all R_j of R. Rand Index can take a value between the range [0,1], while the ARI may take negative values if the index is less than the expected index. Even though ARI takes negative values, generally considered range for ARI is [0,1]. ARI value 0 indicates a real and predicted community does not agree on pairing, ARI value 1 indicates real and predicted communities both represent the same communities.



Real↘Predicted	Same	Different	Total
Same	TP	FN	P
Different	FP	TN	N
Total	$\hat{P}$	$\hat{N}$	P+N

Table 2.2: Contingency table

2.3.1.2 NMI

Normalized Mutual Information (NMI) (Vinh, et al., 2010) is an information theoretic approach to measure shared information between two data distribution. In information theory, the information contained in a distribution is called entropy. In perspective of communities of a network with n nodes, the entropy H(R) of real partitioning R and the entropy H(C) of predicted partitioning C can be expressed as follows:

$$H(R) = -\sum_{i=1}^{m} \frac{n_i^R}{n} \log\left(\frac{n_i^R}{n}\right)$$

.... (2.3)

$$H(C) = -\sum_{j=1}^{k} \frac{n_j^C}{n} \log\left(\frac{n_j^C}{n}\right)$$

.... (2.4)

Where, m and k are the number of communities presents in R and C respectively, n_i^R represents the number of nodes in community $R_i \in R$ and n^C_j represents the number of nodes in community $C_j \in C$. Again mutual information share between R and C is expressed as:

$$I(R,C) = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{n_{ij}^{RC}}{n} \log\left(\frac{\frac{n_{ij}^{RC}}{n}}{\frac{n_i^R}{n} \times \frac{n_j^C}{n}}\right)$$

.... (2.5)

The definition of mutual information demonstrates that $I(R, C) \leq (H(R) + H(C))/2$. Thus, NMI between R and C can be defined as:

$$NMI(R,C) = \frac{2 \times I(R,C)}{H(R) + H(C)}$$

.... (2.6)

After simplifying Equation 2.6 by placing values of H(R), H(C) and I(R, C) NMI can be derived as follows:

$$NMI(R,C) = \frac{-2 \times \sum_{i=1}^{m} \sum_{j=1}^{k} n_{ij}^{RC} \log\left(\frac{n_{ij}^{RC} \times n}{n_i^R \times n_j^C}\right)}{\sum_{i=1}^{m} n_i^R \log\left(\frac{n_i^R}{n}\right) + \sum_{j=1}^{k} n_j^C \log\left(\frac{n_j^C}{n}\right)}$$

.... (2.7)

NMI takes values in the range [0,1]. Value 1 indicates maximum mutual information share between R and C or in other words both represents same set of communities. Value 0 indicates no information sharing between R and C or in other words both represents two different set of communities.

2.3.1.3 Purity

Purity, which is very simple and clear measure, considers only correctly assigned nodes to any community. Each community is assigned a community label which is most frequent, and then count correctly assigned nodes with respect to ground truth communities. This number is then divided by the total number of nodes in the network. Purity of R and C is calculated as follows (Yang, et al., 2002):

$$Purity(R,C) = \frac{1}{n}\sum_{m} max_k(R_m \cap C_k)$$

……… (2.8)

Actually purity is the average of total correctly assigned nodes to different communities. Purity takes values in the range [0,1]. Higher purity value indicates high accuracy and lower value indicates bad communities are identified.

2.3.1.4 F-Measure

Harmonic mean of Precision and Recall is referred as F-measure, also known as F-score or F_1-score.

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

……… (2.9)

This is the balanced version F-measure (Amigó, et al., 2009), where Precision and Recall given equal weight. Actually this is balanced version of general F_β-measure, where Precision given specific non negative weight as given below:

$$F_\beta - measure = (1+\beta^2) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$$

…….. (2.10)

As F_1-score takes its best value at 1 and worst score at 0, also more popular in evaluating communities than F_β-score, so we consider F_1-score. When we use F-measure it will refer to F_1-score only.

Precision: Precision is a measure of exactness of the predicted communities with respect to real communities. It evaluates the proportion of true positives against all the positive results. Precision can be expressed in terms of elements of contingency table (Table 2.2) as follows (Leskovec, et al., 2009):

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{\hat{P}}$$

……. (2.11)

Recall: Recall is a measure of completeness of the predicted communities with respect to real communities (Wang, et al., 2011). Recall also known as the true positive rate, or the recall rate, or sensitivity. It evaluates the proportion of true positives against actual true positives results. Recall can be expressed in terms of elements of contingency table (2.2) as follows:

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{P}$$

.... (2.12)

2.3.1.5 Entropy

Similar to NMI, entropy is another information theoretic concept which measures the information content of messages (Li and Huang, 2002). In community evaluation, entropy measures how different predicted communities share the distribution of nodes of real communities. Entropy of predicted communities C with respect to real communities R is defined as:

$$H(R,C) = \sum_{i=1}^{k} \frac{n_i^C}{n} \left(-\frac{1}{\log m} \sum_{j=1}^{m} \frac{n_i^j}{n_i^C} \log \frac{n_i^j}{n_i^C} \right)$$

.... (2.13)

where, n is total number of nodes, k and m are number of communities present in predicted and real partitioning, $n^C{}_i$ is the number of nodes in the community i of predicted communities, n_i^j is the number of nodes of real community j is assigned to the predicted community i. Lower value of entropy indicates better partition of predicted communities.

2.3.2 Quality Metrics

2.3.2.1 Modularity

Modularity (Q) is the most widely used metric designed especially for the purpose of measuring quality of predicted communities. This metric has got wide acceptance over the years for evaluating communities when ground truth is not known. Modularity can be computed as follows. Let us consider an algorithm predicted a clustering C with k communities for a network. Define a k × k symmetric matrix e whose element e_{ij} is the fraction of all edges in the network that connect nodes in the community i to nodes in community j. The trace of this matrix tr (e) = $\sum_i e_{ii}$ gives the fraction of edges in the network that connect nodes in the same community. Whereas, the row (or column) sums ai = $\sum_j e_{ij}$ represents the fraction of edges that connect to nodes in community i. With these tr (e) and a_i modularity can be defined as:

$$Q(C) = \sum_i \left(e_{ii} - a_i^2\right) = tr(e) - \|e^2\|$$

.... (2.14)

Where, kxk represents the sum of the elements of matrix x. Thus, Q effectively measures the fraction of edges in the network that connect nodes in the same community by subtracting the expected value of this quantity from it if the edges were placed at random. If the number of connections between nodes of same community is less than random, we will get Q = 0 and value approaches Q = 1 if connections between nodes of same community is higher.

2.3.2.2 Coverage

Coverage (Chockler, et al., 2003) of any clustering C is the fraction of edges that connect two nodes of same community within the total no of edges present in the network. Coverage can be computed as follows:

$$Coverage(C) = \frac{\sum_{i=1}^{k} \sum_{j=1,i=j}^{k} e_{ij}}{m}$$

...... (2.15)

Where, k is the number of communities present in C and m is total number of edges present in the network. Intuitively, higher value of coverage indicates better quality communities. If all communities predicted has only one node (i.e. n = k, n is number of nodes in the network) coverage will become 0. If number of communities is one, coverage takes value 1, since all edges of the network will lie within the community.

2.3.2.3 External Density

External density (Coscia, et al., 2011) of a network partitioning is defined as the ratio of edges that connect two different communities to the maximum number of edges possible that connects two different communities. It is the ratio of inter community edges to the maximum number of inter community edges possible. External density can be computed as follows:

$$ExtD(C) = \frac{\{(u,v) \mid u \in C_i, v \in C_j, i \neq j\}}{n(n-1) - \sum_{i=1}^{k} (|C_i| (|C_i| - 1))}$$

...... (2.16)

Where, u and v are any pair of nodes, C_i and C_j are communities, n is the number of nodes in the network, C is the predicted community list which comprises k communities. Lesser value of external density indicates better quality communities.

2.3.2.4 Fuzzy Modularity

Membership of nodes with different belongingness to multiple communities are handled using fuzzy membership of nodes. To measure the quality of communities having fuzzy membership of nodes, fuzzy version of modularity has been developed. (Havens, et al., 2013) proposed a generalized fuzzy modularity as follows:

$$Q_g = \frac{tr\left(UBU^T\right)}{||W||} \qquad \text{....... (2.17)}$$

Where, W is weighted adjacency matrix of the graph, U is partition matrix for membership of nodes to communities, $B = \left[W - \frac{(m^T m)}{||W||}\right]$, and $m = (m_1, m_2, \ldots, m_n)^T$.

2.4 Datasets

Evaluation of communities requires benchmark networks. There are widely used real world networks compiled by network scientists specifically for community detection. Besides these networks, synthetic networks are also used by imitating the properties of real-world networks. Some popular real-world networks and synthetic networks used for community evaluation are discussed below.

2.4.1 Real-world Networks

Karate: This data set is about study of a karate club network by (Zachary, 1977). The network consists of 34 members of a karate club as nodes and 78 connections among members representing friendships in the club which was observed over a period of two years. Due to a disagreement between members of the club's administrator and the club's instructor, the club was split up into two groups. Originally, Zachary constructed friendship network among members of the club with various measure of ties. Here, we consider simply unweighted version of the network used by Girvan and Newman for evaluating communities.

Strike: (Michael, 1997) studied labor strike patterns in a wood-processing facility after new management proposed compensation packages. The study was based on age and ethnic group. Set of 24 labors were grouped into three such groups depending on 34 associations among labors during different schedules of strikes.

Football: Another popular data set used for community evaluation is the network of United States college football, which was also used firstly by (Girvan and Newman, 2001). Network was represented with the schedule of Division I games for the 2000 season. Nodes in the network represent teams and connections represent regular-season games between two connected teams. The network contains 115 teams and 616 games were played among different teams. The teams are divided into conferences containing around 8-12 teams. However, considered 12 conferences for community evaluations.

Dolphin: (Lusseau, et al., 2003) network study of dolphins living in Doubtful Sound, New Zealand is also used for evaluating communities. They had studied behavior of about 62 bottlenose dolphins for over seven years. The network was divided into two groups depending on the association patterns of dolphins.

Sawmill: Data were collected from a mid-sized softwood sawmill to examine the network of communications among employees (Yang, et al., 2011). This data was collected in order to analyze the communication structure among the employees after a strike. An edge in the network means that the two connected employees have discussed the strike with each other very often. The network of communications among employees consists 36 employees and 62 communicating links.

Poll Books: A network of books about recent US politics sold by the online bookseller amazon.com. Edges between books represent frequent co-purchasing of books by the same buyers. The network was compiled by (Harvey, et al., 1996). The network contains 105 nodes and 441 connections.

Jazz: List of edges of the network of Jazz musicians. (Gleiser and Danon, 2003) studies the collaboration network of jazz musicians. Two different levels of networks are prepared. First the collaboration network between individuals, where two musicians are connected if they have played in the same band. Second the collaboration between bands, where two bands are connected if they have a musician in common. The later network contains 198 bands and 2742 connections, and is used widely to evaluate community detection algorithms.

LesMis: Les Miserable is a French historical novel, written by (Hugo, 1862) Victor Hugo and published in 1862. The co-appearance of weighted network of characters in the novel has been extracted by Knuth. The nodes are the characters of the novel and an edge indicates that the two characters appear together in the same chapter of the novel, at least once. There are 77 nodes representing different characters and 254 connections in the network.

Email: (Gleiser, 1973) studied a social network constructed from email communications within a medium sized university with employees. This is the email communication network at the University Rovira i Virgili in Tarragona in the south of Catalonia in Spain. Nodes are users and each connection represents that at least one email was sent. The direction of emails or the numbers of emails are not stored. The network was constituted with 1133 users and 5451 interchanges of emails as connections.

Words: Newman compiled a network representing just a positions of words in a corpus of English text, in this case the novel (Dickens, 1838). To construct this network, the 60 most commonly occurring nouns in the novel and the 60 most commonly occurring adjectives were considered. The nodes in the network represent words and an edge connects any two words that appear adjacent to one another at any point in the book. Eight of the words never appear adjacent to any of the others and are excluded from the network, leaving a total of 112 vertices and 425 connections.

GR-QC: Arxiv GR-QC (General Relativity and Quantum Cosmology) collaboration network (Lyth, et al., 2005) was prepared from the e-print arXiv and covers scientific collaborations between authors and papers submitted to General Relativity and Quantum Cosmology category. The data covers papers published within the period from January 1993 to April 2003 (124 months). Network was constructed with authors as nodes and co-authorship among author were represented with connections.

HEP-TH: Arxiv HEP-TH (High Energy Physics - Theory) collaboration network was prepared from the e-print arXiv and covers scientific collaborations between authors and papers submitted to High Energy Physics-Theory category. Network was constructed with authors as nodes and co-authorship among author were represented with connections (Strominger and Vafa, 1996).

Wiki-Voter: Administrator selection in Wikipedia was done through voting among users and volunteers around the world. Using the Wikipedia page edit history in (Bordier, 2012) extracted all administrator elections and vote history data. Nodes in the network represent Wikipedia users and voting of any user for other users are represented with a directed connection. In our case we have considered all connections as undirected.

2.4.2 Synthetic Networks

LFR Graphs Lancichinetti, Fortunato and Radicchi defined benchmark graphs for evaluating community detection algorithms, which are referred as LFR graph (Fortunato and Lancichinetti, 2009) . Benchmark graph are generated based on power law distribution of both node degree and community size. The graphs are generated according to various parameters such as number of nodes (n), average degree, maximum degree, exponent degree distribution and community size distribution, range of community sizes and mixing parameter μ for controlling the neighbors in other communities. Mixing parameter μ determines intra-community and inter-community connections. To evaluate performance of community detection algorithms variety of LFR graphs are generated. LFR graph generated with variation of mixing parameter, number of nodes and average node degree are popularly utilized in community evaluation. Source Code of LFR graphs available online.

2.5 Evaluation Methodologies

Popularly used community evaluation method is value based analysis, where predicted communities are evaluated in terms of various metrics discussed above (see section 2.2). Value based comparison of community detection algorithm is rather common strategy in order to determine performance of algorithms in the ground of quality and accuracy. Accumulation of indications of different metrics is a major difficulty in value based analysis. Another disadvantage is value analysis unable to determine whether communities are good or bad when there is a trade-off between quality metrics and accuracy metrics. (Kou, et al., 2011) developed MCDM technique to evaluate community detection algorithms, where a single comparative score is generated accumulating the values obtained for different metrics. Although with the methodology proposed by Kou et al. resolves the trade-off between quality and accuracy metrics, but it cannot express explicitly how likely the algorithm will identify accurate communities or good quality communities.

There are some other aspects of community evaluation, which are discussed in different literature. Visual inspection of predicted community structure in reference to ground truth is often considered. However, the visual inspection of community structure is suitable only for small networks. Sensitivity of community detection algorithm with variation of certain network parameters such as size of network (i.e. number of nodes), inter-community connections, and intra-community connections are analyzed. Use of mixing parameter μ of LFR graph for variation of intercommunity and intra-community connections is popular in community evaluation. Statistical measures such as mean and standard deviation of different metrics are used when community detection algorithms are randomized i.e. the algorithm detect different community structures in different execution of algorithm. With this methodology, information about the distribution of metric values computed for various community structures can be obtained. However, it cannot express good (or bad) communities are how much good (or bad) in comparison to the communities identified by other algorithm.

2.6 Applications of Community Structure

In this section, various applications that incorporate predicted communities are studied and discussed how the community information are utilized in those systems.

2.6.1 Community Structure in Link Prediction

The information of community of nodes is leveraged in the task of link prediction. (Ding, et al., 2015) have proposed a multi-resolution community division based link prediction approach. They considered the property of hierarchical organization of communities to extract different levels of communities. Then, a simple frequency statistical model is used to compute frequency of node pairs in different resolution of community, and likelihood of missing links are generated. Recently, (Ding, 2011)proposed another link prediction algorithm by using the latent information between different communities that introduces a community similarity feature called community relevance. (Akoglu, et al., 2015) also proposed a similarity based algorithm exploring herd phenomenon in different communities to predict missing links. (Dai, et al., 2017) proposed a generative model for text documents using the notion of community to identify missing links. (Sachan and Ichise, 2010) proposed to build a link predictor in a co-authorship network, and showed that the knowledge of a pair of researchers lying in the same dense community can be used to improve the accuracy of predictor further. (Shahriary, et al., 2015) have explored communities to predict links and sign of the links in signed networks.

2.6.2 Community Structure in Information Diffusion

Communities are central for efficiently disseminating news, rumors, and opinions in human social networks. Dissemination of information in the network in the context of community structure has been studied extensively. (Shahriary, et al., 2015) studied the role of communities in target oriented information diffusion, and they showed exploring community information highly efficient strategy can be achieved. Epidemic spreading in weighted scale-free networks with community structure is investigated by (Chu, et al., 2009). The impact of overlapping community structure on susceptible-infected-susceptible (SIS) epidemic information diffusion model is investigated by (Shang, et al., 2015). (Wang, et al., 2010) used community analysis within the independent cascade model to find influential nodes for information diffusion on a social network. (Nematzadeh, 2017) proposed a linear threshold model for systematic analysis of community structure influence on global information diffusion. (Weng, et al., 2013) developed predictive models of information diffusion based on interactions among communities. (Shang, et al., 2015) considered both overlapping and non-overlapping communities to examine the affect in epidemic spreading.

2.6.3 Community Structure in Recommendation Systems

Community detection algorithms are also useful in the development of network based recommendation system. (Kawahara, et al., 1999) explored user preference projected by the community in which the user belongs and introduced the notion of partially similar interest of users for recommendation system. (Lops, et al., 2013) proposed a recommender system based on user community behavior to recommend a set of relevant keywords for the resources to be annotated. (Zhuhadar, et al., 2012) proposed visual recommender system using community information for recommending resources to cyber learners that belong to same community. (Kim and Ahn, 2008) proposed another approach for community specific recommendation system using Bayes modeling consideration changes in the behavior of users over time. A community based social recommender system is proposed by (Fatemi and Tokarchuk, 2013), where social data is utilized for person specific recommendation based on the communities constructed from user interactions over the time.

2.7 Summary

Community detection problem is studied in three perspectives: identification communities, evaluation of detected communities, and post-hoc analysis leading to applications. First the study covers various community detection algorithms categorizing those based on their methodological principles. Second discussed aspects of community evaluation, which include validation metrics, datasets and evaluation methods. Lastly, various applications of community structure are discussed. Community detection problem has been studied several inter-disciplinary domains, yet the problem is not solved satisfactorily and leaves us with number of challenging issues.