CHAPTER 2

COMMUNITY STRUCTURE AND COMMUNITY DETECTION ALGORITHMS

2.1 INTRODUCTION

After the justification and brief statement of objectives of the study in the earlier chapter, this chapter is a presentation of the community structure and community detection algorithms. The latest swell of concentration in the properties of networks of many kinds pitches in as diverse fields as internet, World Wide Web, citation networks, transportation networks, software call groups, email networks, food webs and social and biochemical networks (Newman 2003). Many complex systems can be represented as networks, where the elementary parts of a system and their mutual interactions are nodes and links described by Lancichinetti & Fortunato (2010). Complex systems are usually organized in compartments which have their own roles and functions. In the network representation, such compartments appear as sets of nodes, with a high density of internal links and a comparatively lower density in links between compartments. These subgroups are called communities or modules and occur in a wide variety of networked systems.

2.1.1 Community – A Few Definitions

Community is defined in several ways as exemplified in Figure 2.1.

- A community is a number of representative authority nodes linked by an important node that shares a common node.
- A community is a highly linked bipartite sub-graph and has at least one core containing complete bipartite sub graph.
- A set of nodes that links more nodes in the community than those outside the community could be defined as a sub community.
- A research community could be based on a single most node and contains all nodes that link it.

![Figure 2.1 Definitions of Community](image)

While each of the above definitions characterizes some essential properties of a community, it makes the community mining task rather demanding because of the absence of a standardized definition.

### 2.2 COMMUNITY DETECTION

Identification and analysis of compartments or communities yield knowledge on the organization of complex systems and their functions. Therefore detecting communities in networks has become a fundamental problem in network science. Many methods have been developed, using tools and techniques from disciplines like physics, biology, applied mathematics,
computer and social sciences. However, it is not clear which algorithms are reliable and shall be used in applications. The question of the reliability itself is tricky as it requires the shared definitions of community and participation which are present still missing. This essentially denotes that, despite the voluminous research and findings in the field, there is still no consensus among the researchers on what a network with communities looks like.

Community or modular structure is considered to be a significant property of real-world social networks as it often accounts for the functionality of the system. Despite the ambiguity in the definition of community, numerous techniques have been developed for both efficient and effective community detection. Random walks, spectral clustering, modularity maximization, differential equations, and statistical mechanics have all been used earlier. Much of the focus within community detection has been on identifying disjoint communities described by Danon et al (2005).

However, it is well-understood that people in a social network are naturally characterized by multiple community memberships. For example, a person usually has connections to several social groups like family, friends, and colleagues; a researcher may be active in several areas. Further, in online social networks, the number of communities an individual could belong to is essentially unlimited because a person can simultaneously associate with as many groups as he wishes. This also happens in other complex networks such as biological networks, where a node might have multiple functions. In Kelley et al (2011) and Reid et al (2011), the authors showed that the overlap is indeed a significant feature of many real-world social networks.
2.3 BASIC CONCEPTS IN COMMUNITY DETECTION

Some basic concepts and details of community detection are essential before an endeavor is made in understanding its application in real life networks.

- Given a network or graph $G = (E, V)$, $V$ is a set of $n$ nodes and $E$ is a set of $m$ edges.
- For dense graphs, $m = O(n^2)$, but for sparse networks $m = O(n)$.
- The network structure is determined by the $n \times n$ adjacency matrix $A$ for an unweighted networks and weight matrix $W$ for weighted networks.
- Each element $A_{ij}$ of $A$ is equal to 1 if there is an edge connecting nodes $i$ and $j$; and it is 0 otherwise.
- Each element $w_{ij}$ of $W$ takes a non negative real value representing strength of connection between nodes $i$ and $j$.
- In the case of overlapping community detection, the set of clusters found is called a *cover* $C = \{c_1, c_2, \ldots, c_k\}$ Lancichinetti et al(2009), in which a node may belong to more than one cluster.
- Each node $i$ associates with a community according to a *belonging factor* (i.e., soft assignment or membership) $[a_{i1}, a_{i2}, \ldots, a_{ik}]$ Nepusz et al(2008), in which $a_{ic}$ is a measure of the strength of association between node $i$ and cluster $c$.

Without loss of generality, the following constraints are assumed to be satisfied

$$0 \leq a_{ic} \leq 1 \ \forall i \in V, \forall c \in C \quad (2.1)$$
and \[ \sum_{c=1}^{|C|} \alpha_i c = 1 \] (2.2)

Where \(|C|\) is the number of clusters. However, the belonging factor is often solely a set of artificial weights. It may not have a clear or unambiguous physical meaning in Ren et al(2009).

In general, algorithms produce results that are composed of either of the two types of assignments, crisp (nonfuzzy) assignment or fuzzy assignment (Gregory 2011). With crisp assignment, the relationship between a node and a cluster is binary. That is, a node \( i \) either belongs to cluster \( c \) or does not. With fuzzy assignment, each node is associated with communities in proportion to a belonging factor. With a threshold, a fuzzy assignment can be easily converted to a crisp assignment. Most detection algorithms give outputs as crisp community assignments.

### 2.4 ALGORITHMS FOR COMMUNITY DETECTION

Algorithms have been developed to identify the communities and their interrelationships. Some of these algorithms are listed below.

- Algorithm of Girvan and Newman
- Fast greedy modularity optimization by Clauset, Newman and Moore
- Exhaustive modularity optimization via simulated annealing
- Cfinder
- Markov cluster algorithm
- Structural algorithm by Rosvall and Bergstrom
- Dynamic algorithm by Rosvall and Bergstrom
- Spectral algorithm by Donnetti and Munoz
- Expectation-maximization algorithm by Newman and Leicht
- Potts model approach by Ronhovde and Nussinov

Some of these and the following algorithms are exclusive overlapping community detection algorithms.

- Clique percolation
- Line graph and link partitioning
- Local expansion and optimization
- Fuzzy detection
- Agent based and dynamical algorithms

2.5 FUZZY ALGORITHM FOR COMMUNITY DETECTION

As the present study pertains to the application of fuzzy logic algorithm in identifying the path in large road networks, the various facets and kinds of fuzzy algorithm are alone taken up for analysis.

Fuzzy community detection algorithms quantify the strength of association between all pairs of nodes and communities. In these algorithms, a soft membership vector or belonging factor (Gregory 2010) is calculated for each node. A drawback of such algorithms is the need to determine the dimensionality \( k \) of the membership vector. This value can be either provided as a parameter to the algorithm or calculated from the data.

Nepusz et al (2008) modeled the overlapping community detection as a nonlinear constrained optimization problem which can be solved by simulated annealing methods. The objective function to minimize is
\[ f = \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}(s_{ij}-s_{ij})^2 \]  

(2.3)

Where \( W_{ij} \) denotes the predefined weight, \( s_{ij} \) is the prior similarity between nodes \( i \) and \( j \), and the similarity \( s_{ij} \) is defined as

\[ s_{ij} = \sum_{c} a_{ic}a_{jc} \]  

(2.4)

Where the variable \( a_{ic} \) is the fuzzy membership of node \( i \) in community \( c \), subject to the total membership degree constraint in and a nonempty community constraint. To determine the number of communities \( k \), the value of \( k \) is increased until the community structure does not improve as measured by a modified fuzzy modularity, which, by weighting \( Q \) with the product of a node’s belonging factor, is defined as

\[ Q^{Ne} = \frac{1}{2m} \sum_{c} \sum_{i,j \in c} \left[ A_{ij} - \frac{ki_{ij}}{2m} \right] a_{ic}a_{jc} \]  

(2.5)

Where \( a_{ic} \) is the degree of membership of node \( i \) in the community \( c \).

White & Smyth (2005), Newman (2006) and Zhang et al (2007) proposed an algorithm based on the spectral clustering framework. Given an upper bound on the number of communities \( k \), the top \( k-1 \) eigenvectors are computed. The network is then mapped into a \( d \)-dimensional Euclidean space, where \( d \leq k - 1 \). Instead of using \( k \) means, Fuzzy C-Means (FCM) is used to obtain a soft assignment. Both detection accuracy and computation efficiency rely on the user-specified value \( k \). With running time \( O(mkh + nk2h + k3h) + O(nk2) \), where \( m \) is the number of edges, \( n \) is the number of nodes, the first term is for the implicitly restarted Lanczos method, and the second term is for
FCM, it is not scalable for large networks. An extended modularity that used the average of the belonging factor is also proposed as

\[
Q_{ov} = \sum_c \left[ \frac{A(V'_c, V'_c)}{A(V, V)} - \left( \frac{A(V'_c, V)}{A(V, V)} \right)^2 \right]
\]  

(2.6)

where \( V'_c \) is the set of nodes in a community \( c \), \( w_{ij} \) is the weight of the link connecting nodes \( i \) and \( j \).

\[
A(V'_c, V'_c) = \sum_{i,j \in V'_c} \frac{w_{ij}}{2} = A(V'_c, V'_c) + \sum_{i \in V \setminus V'_c} w_{ij} \frac{(a - (1-a)c)}{2}
\]

(2.7)

and \( A(V, V) = \sum_{i,j \in V} w_{ij} \)

Due to their probabilistic nature, mixture models provide an appropriate framework for overlapping community detection (Newman & Leicht 2007). In general, the number of mixture models is equal to the number of communities, which needs to be specified in advance.

In SPAEM6 Ren et al (2009), the mixture model is viewed as a generative model for the links in the network. Suppose that \( \pi_r \) is the probability of observing community \( r \) and community \( r \) selects node \( i \) with probability \( Br_i \). For each \( r \), \( Br_i \) is a multinomial across elements \( i = 1, 2, \ldots, n \), where \( n \) is the number of nodes.
Therefore,

\[ \sum_{i=1}^{n} B_{r,i} = 1 \]  

(2.8)

The edge probability \( e_{ij} \) generated by such finite mixture model is given by

\[ p(e_{ij} \mid \pi, B) = \sum_{r=1}^{k} \pi_r B_{r,i} B_{r,j} \]  

(2.9)

The total probability over all the edges present in the network is maximized by the Expectation-Maximization (EM) algorithm. As in Kim & Jeong (2011), the optimal number of communities \( k \) is identified based on the minimum description length. There is another algorithm called FOG Davis & Carley (2008) which tries to infer groups based on link evidence.

Similar mixture models can also be constructed as a generative model for nodes Fu & Banerjee (2008). In Sampled Spectral Distance Embedding (SSDE) Magdon-Ismail & Purnell (2011), the network is first mapped into a \( d \)-dimensional space using the spectral clustering method. A Gaussian Mixture Model (GMM) is then trained via the Expectation-Maximization algorithm.

The number of communities is determined when the increase in log-likelihood of adding a cluster is not significantly higher than that of adding a cluster to random data which is uniform over the same space.

Stochastic Block Model (SBM) by Nowicki & Snijders (2001) is another type of generative model for groups in the network. Fitting an empirical network to an SBM requires inferring model parameters similar to GMM. In Overlapping Stochastic Block Models (OSBM) by Latouche et al (2011), each node \( i \) is associated with a latent vector (i.e., community assignment) \( Z_i \) with \( K \) independent Boolean variables \( Z_{ik} \in \{0, 1\} \), where \( K \) is
the number of communities and $Zik$ is drawn from a multivariate Bernoulli distribution. $Z$ is inferred by maximizing the posterior probability conditioned on the presence of edges as in Ren et al (2009).

OSBM requires more efforts than mixture models because the factorization in the observed condition distribution for edges, given $Z$ is, in general, intractable. Model-based Overlapping Seed Expansion (MOSES) by McDaid& Hurley(2010) combines OSBM with the local optimization scheme in which the fitness function is defined based on the observed condition distribution.

MOSES greedily expands a community from edges. Unlike OSBM, no connection probability parameters are required as input. The worst-case time complexity is $O(en^2)$, where $e$ is the number of edges to be expanded.

Nonnegative Matrix Factorization (NMF) is a feature extraction and dimensionality reduction technique in machine learning that has been adapted to community detection. NMF approximately factorizes the feature matrix $V$ into two matrices with the no negativity constraint as $V \approx WH$, where $V$ is $n \times m$, $W$ is $n \times k$, $H$ is $k \times m$, and $k$ is the number of communities provided by users. $W$ represents the data in the reduced feature space. Each element $w_{ij}$ in the normalized $W$ quantifies the dependence of node $i$ with respect to community $j$. In Zhang et al (2007), $V$ is replaced with the diffusion kernel, which is a function of the Laplacian of the network. In Zarei et al (2009), $V$ is defined as the correlation matrix of the columns of the Laplacian. This results in better performance than Zhang et al (2007).

In Zhao et al (2010), redundant constraints in the approximation are removed, reducing NMF to a problem of symmetrical Nonnegative Matrix Factorization (s-NMF).
Psorakis et al (2011) proposed a hybrid algorithm called Bayesian NMF. The matrix $V$, where each element $v_{ij}$ denotes a count of the interactions that took place between two nodes $i$ and $j$, is decomposed via NMF as part of the parameter inference for a generative model similar to OSBM and GMM.

Traditionally, NMF is inefficient with respect to both time and memory constraints due to the matrix multiplication. In the version of Psorakis et al (2011), the worst-case time complexity is $O(kn^2)$, where $k$ denotes the number of communities.

Wang et al (2009) combined disjoint detection methods with local optimization algorithms. First, a partition is obtained from any algorithm for disjoint community detection. Communities attempt to add or remove nodes. The difference, called variance, of two fitness scores on a community, either including a node $i$ or removing node $i$, is computed. The normalized variances forms a fuzzy membership vector of node $i$.

Ding et al (2010) employed the affinity propagation clustering algorithm by Frey & Dueck (2007) for overlapping detection, in which clusters are identified by representative exemplars. First, nodes are mapped as data points in the Euclidean space via the commute time kernel (a function of the inverse Laplacian). The similarity between nodes is then measured by the cosine distance. Affinity propagation reinforces two types of messages associated with each node, the responsibility $r(i, k)$ and the availability $a(i, k)$. The probability for assigning node $i$ into the cluster represented by exemplar node $k$ is computed by equation $p(i, k) = e^{-r(i, k)}$, where $r$ is the normalized responsibility as in Geweniger et al (2009).
2.6 CONCLUSION

Thus it becomes clearly evident that community structure is an important property of complex networks. This characteristic is quite challenging in various disciplines such as computer networking, biology, social network, physics and transportation. Analyzing communities in networks emerges as the hottest topic since network representation can disclose some relevant features of the system in general, involving its structure, its function as well as the interaction between structure and function. These ideas can go a long way in assisting the optimization of large road networks in the identification of solution to various risk factors. The subsequent chapter is a detailed account of fuzzy logic algorithm, fundamental to the entire study.