Analyzing Networks with Multiple Links and Attribute Information using Stochastic Block Model

Anubhav Jain, B.Tech

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Data Science)

Supervisor: Dr. Arthur White

August 2018

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Anubhav Jain

August 29, 2018

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Anubhav Jain

August 29, 2018

Acknowledgments

I would like to thank my supervisor, Dr. Arthur White for his continuous support, guidance and patience. Without his support and motivation, this dissertation would not have been possible.

I would also like to thank my parents for continuously motivating me and helping me to proof read the dissertation.

Lastly, I would like express my gratitude towards my flatmates and Data Science batchmates for all the learnings and a delightful journey of my Masters.

Anubhav Jain

University of Dublin, Trinity College
August 2018

Analyzing Networks with Multiple Links and Attribute Information using Stochastic Block Model

Anubhav Jain, Master of Science in Computer Science

University of Dublin, Trinity College, 2018

Supervisor: Dr. Arthur White

Identifying clusters based on observed patterns or attributes and analyzing the interaction of actors within and outside these clusters allows to uncover hidden information about a network. Stochastic Block Model(SBM) is a generative model and one of the benchmark models for community detection. In this dissertation, Variational Expectation-Maximization algorithm approach of SBM is used to estimate the model parameters and Integrated Complete Data Likelihood(ICL) is used to compute optimal number of communities in the network. In real world, nodes are connected to other nodes with a string of multiple relationships, where each relationship defines its influence and meaning within a network. The SBM is used to fit all the links separately treating each link as a individual network. The resultant set of clusters and model parameters from the set of fitted models specific to each link are compared against each other and the node attributes to identify the connection between cluster formation, actor attributes and link type. This process is applied to Lawyer Lazega Dataset which is a network of lawyers working in a law firm connected to each other by multiple links- Friends, Advice and Work. The results show that there exists a relationship between the clusters formed across different link types; with many clusters behaving and containing same node attributes as of the clusters belonging to other link type.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Background Study

1.1 Introduction

Humans are meant to be naturally social, who look out for different mediums to connect and form a network. The network could be a social interaction, personal/professional relationship or people bonding just for a common interest. With the advent of social networks and other different ways to connect virtually, people from all round the world are connected directly or indirectly converging them into a network; these networks can be used to find communities, behaviour patterns and knowledge discovery.

In recent years, the area of Network analysis has gained immense popularity among Data Scientists, Researchers and Statisticians. Many active and interesting researches are being carried out in this area. Community Detection is one of the fundamental and critical problems in Network Analysis. Many algorithms and methodologies have been proposed for identifying communities within the network. This dissertation will focus on **Stochastic Block Model** which is one of the most popular and widely used clustering or community detection model. This model will be used to identify and study the clusters of a network containing multiple types of links among the nodes. Simultaneous and comparative analysis is performed on these networks containing same set of nodes with different relationships. Identified clusters from these networks with multiple links are compared against each other and with node attributes to analyze,

if some kind of connection exists between them. This study is useful to check the existence of clusters that contain nodes with similar characteristics irrespective of the link type connecting them and also to analyze the interaction between/within the clusters formed by multiple networks.

The paper is organized as follows: Chapter 1 introduces the concepts of Network Analysis, Community Detection and a brief overview graph theory along with node and edge properties. Followed by dataset analysis and description and limitation of Erdos Renyi Model. Chapter 2 states the literature review and the advancements of Stochastic Block Model. Chapter 3 explains the concepts behind the SBM model and Variational Expectation-Maximization(EM) Algorithm. Model inference is stated in Chapter 4 and application of the model on the datasets are given in Chapter 5. Followed by cluster comparison and analysis in Chapter 6. Lastly, theoretical concepts of Mixed Membership Model as a further extension to SBM are stated in Chapter 7.

1.2 Analysis of Networks

Network analysis means study and visualization of structures using methodologies and techniques of graph and network theory. It involves visualization and conceptualization of structures in terms of connected nodes or actors that are linked or tied together by some common entity or relationship. But questions like: Why do we need to represent the structure as a network or graph? How do we represent the structure as network? Characteristics and properties of networks? Steps to study and get meaningful insight from a network? All these question, their answers and their undergoing research make network theory, a very interesting and intriguing topic with wide area of applications and usages.

Facebook friends, Linkedin Professional, twitter followers-follows, protein enzymes, boss-employee and many other networks can give diverse amount of information and trends if analyzed, cleaned and visualized correctly. Network analysis has wide and varied applications in areas like Biology, Crime investigation, Intelligence and Military, Customer identification, targeting, friendship, kinship, work-partner, leader-follower, disease flow, root cause analysis etc. With the advent of self driven cars and connected cars, network analysis can also be used for faster route estimation, accident prevention, vehicle breakdown, vehicle tracking and theft prevention.

1.3 Community Detection

A community or a group represents a set of actors/nodes that may share common interests, skills, goals or a central attribute/characteristics that connects all of them. A network may or may not contain meaningful communities as some nodes relationship can be random and arbitrary. Link from one node to another may be directed or undirected. Example of directed network is twitter followers, i.e if one individual follows another individual, it doesn't mean the reverse is true. Facebook friends/ Linkedin Professional network is an example of undirected network which implies a two way behaviour. The general notion to establish and define communities within a network is that, nodes belonging to same community are densely connected and nodes belonging to different communities are sparsely connected. An node can belong to multiple groups or communities resulting in overlapping networks.

In few cases, the attributes and properties of the network are not revealed when viewed as a whole or average. Communities that make up a network can reveal missing data and other aspects of the network.

1.3.1 Assortative versus Disassortative Mixing

Assortative Mixing refers to the networks where similar nodes connect to each other i.e. higher degree actors connect to other higher degree nodes and lower degree actors connect with lower degree nodes. Higher degree actors form the core of the network while the lower degree actors lie on the boundary of the network.

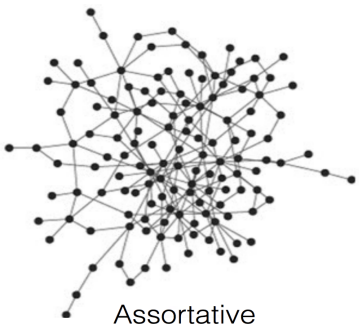


Figure 1.1: Assortative Mixing [Hao and Li 2011]

Disassortative Mixing refers to the networks where dissimilar nodes are connected to each other i.e higher degree nodes are connected to lower degree nodes. These type of network usually have a set of central actors or leaders that influence the whole network.

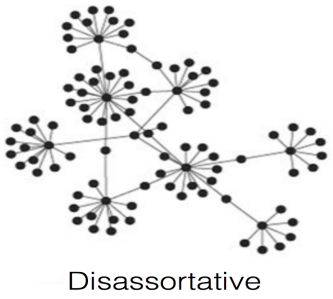


Figure 1.2: Disassortative Mixing [Hao and Li 2011]

1.4 Graph and Network Theory

This section lays down the basic foundation for the whole dissertation. Basic definitions, notations and terminologies of graph theory will be discussed here.

1.4.1 Basic Notations and Definitions

We are surrounded by networks, this can be defined as the relationship that combines different entities into groups or communities that can form a meaningful combined entity; giving trends, insights and the behavior of the group, individuals belonging to that group along with the intra and inter group interaction.

Network: Networks can be defined as collection of interacting entities.

Graph: Mathematical representation of a network can be termed as a graph.

A graph is represented as:

$G = \{V, E\}$, where

- **V:** $\{1, ..., n\}$ are the set of nodes/ entities or actors in a network
- **E:** $\{e_1, ..., e_p : e_k = (i_k, j_k) \in (V, V)\}$ are the set of edges or links that connect the nodes

(a) Undirected Edge (b) Directed Edge (c) Weighted Edge

Figure 1.3: Type of Edges in a Network

- **Order:** It is the number of vertices in a network; $N_v = |V|$
- **Size:** It is the number of edges in a network; $N_e = |E|$
- **Path:** A path in a graph can be defined as a flow across a sequence of vertices that allow us to traverse from vertex to other. In a directed graph; we are only permitted to traverse in particular direction across the edge.
- **Cycle:** A closed path with all different edges.

1.4.2 Properties of Nodes in a network

This section explains the important attributes and characteristics of nodes that determine and influence the various specifications of a network. These properties will be further used for **Network Analysis** and **Goodness of Fit Comparison**.

- **Degree Centrality:** This can be defined as the most basic node attribute which involves computation of degree for each node present in the network. Degree of a node as discussed in the previous section is the measure of number of edges/ links corresponding to that node in an undirected graph. In a directed graph it is a combination of in-degree and out-degree links to/from a node.

  Significance: It describes and helps to visualize all the one hops in a network. It depicts the most influential and popular nodes in the network, where most of the information flow takes place. Nodes with the highest degree are popular

and are centrally attracted in the network. Degree centrality can also highlight whether the network is uniformly, centrally or randomly distributed.

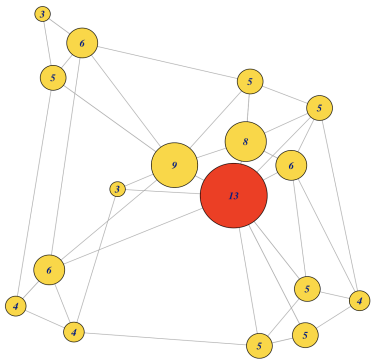


Figure 1.4: Degree Centrality in a network: size of vertex proportional to degree

- **Closeness Centrality:** It defines the closeness of a node with other nodes in the network. It is computed as reciprocal of sum of all the shortest paths from the node to all other nodes in the network. The node with maximum number of shortest path will have the maximum closeness to the other nodes.

  Significance: This property identifies the fastest nodes that influence the whole network i.e. in shortest amount of time they can connect to maximum number of nodes. The nodes with high closeness centrality are good broadcasters for out-degree links or good listeners for all in-degree links.

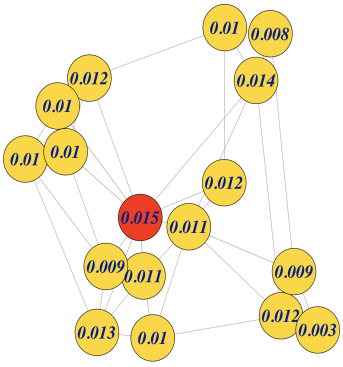


Figure 1.5: Closeness Centrality

- **Betweenness Centrality:** This centrality measure can be defined as the number of times a node occurs between the shortest path of any pair of nodes within the network.

  Significance: This property of node signifies flow of information within the network. The nodes with high betweenness centrality acts as a bridge for the information flow in the network.

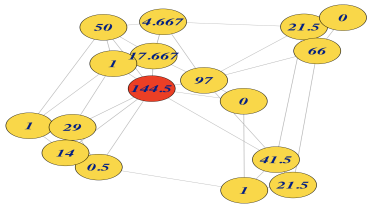


Figure 1.6: Betweenness Centrality

- **Eigen Vector Centrality:** It is an extension to degree centrality measure which considers only degree of a node as a sole centrality measure score without any considerations to type of nodes the actor is connected to. Eigen Vector Centrality assigns a high score to a node if it is connected to a another highly connected node. Thus, the score of a node not only depends upon the degree of a node but also depends on the degree of nodes that the node is connected to. Power iteration method is used to calculate Eigen vectors for the node.

  Significance: It gives a better overview of the network and the influence measure of an node on the whole network rather than only being restricted to neighbours.

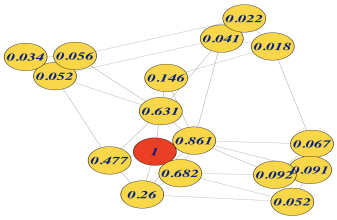


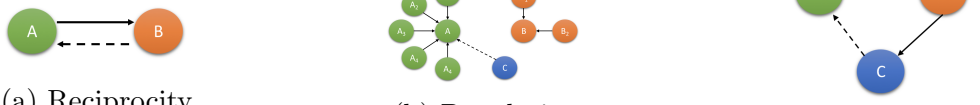
Figure 1.7: Eigen Vector Centrality

1.4.3 Properties of Edge or Relationships in a network

- **Reciprocity:** If $Actor_A$ has an link to $Actor_B$, then what is tendency that $Actor_B$ will have an link to $Actor_A$

- **Popularity:** Nodes with higher degree tend to be popular than other lower degree nodes and have a higher tendency of new connection. For example in LinkedIn, people who have large connections are more likely to make new connections.

- **Transitivity:** If $Actor_A$ has a link with $Actor_B$ and $Actor_B$ has a link with $Actor_C$; then what is tendency that $Actor_A$ will have a link with $Actor_C$. For example in case of Facebook there is a high probability making a connection with friends of your friend.



(a) Reciprocity (b) Popularity (c) Transitivity

Figure 1.8: Properties of edges -Reciprocity, Popularity, Transitivity

- **Clique:** It is a small knit of close friends who are relatively isolated from rest of the world. Any actor pair from the clique group will have an edge between them.

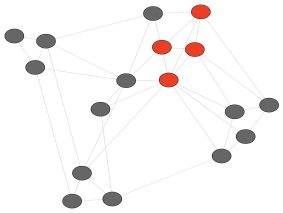


Figure 1.9: Clique: Actors marked in red form Clique

- **Star:** It represents a central node with many nodes connected to it forming a structure similar to a star.

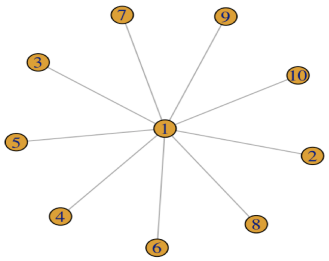


Figure 1.10: Star Formation

- **Geodesic Distance:** The shortest path between the two nodes is called the Geodesic Distance.
- **Diameter:** It is defined as the maximum value of all the shortest paths in the network. It depicts the speed of spread of information across the network.

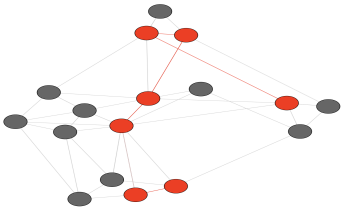
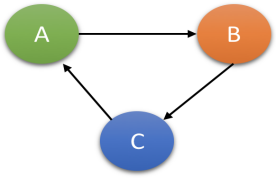
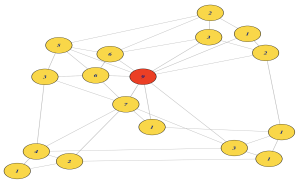


Figure 1.11: Diameter - marked in red

- **Triangle Count:** It is the number of triangle formations in the network. This property signifies the cycle of information. A triangle is formed if $Node_A$ is connected to $Node_B$, $Node_B$ is connected to $Node_C$ and lastly $Node_C$ is connected to Node A.



(a) Triangle from Node A to Node C



(b) Triangle Count

Figure 1.12: Triangle Formations in a network

1.5 Dataset Description

This section describes the datasets being used in this dissertation. Each dataset is subjected to exploratory Data Analysis and processing to gain information about properties of the network before fitting the model.

1.5.1 Zachary Karate Club dataset

Karate Club Dataset is one of the most popular dataset for network implementation, visualization, clustering model trial and validation. This dataset was formed by Wayne Zachary in 1977 by accumulating data from the University Karate Club. The dataset is majorly divided into two groups following a conflict and split between two leaders/teachers. One faction of the dataset is headed by the teacher **Mr. John A** and the other faction is headed by **Mr. Hi**. [Zachary 1977]. This dataset has an attribute **Faction** which divides the nodes of this network into two groups. The dataset has 34 node and 78 undirected edges. Karate Club Dataset has been packaged and ready too be used in Igraph package of R [Csardi 2015].

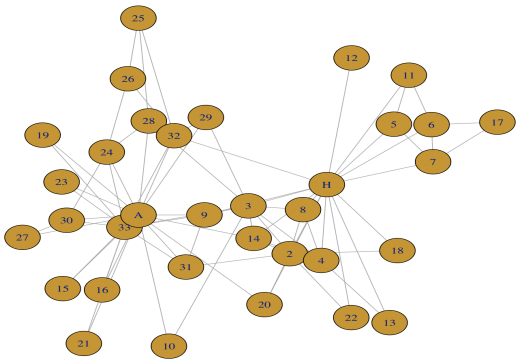
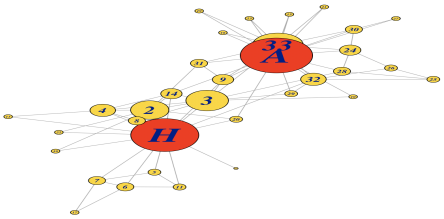


Figure 1.13: Network Visualization for Zachary Karate Club dataset

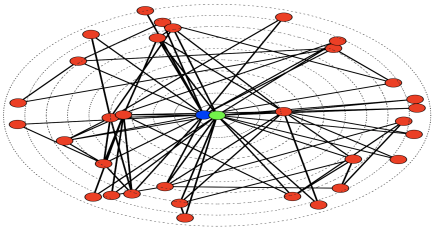
Karate Network Sumnmary Statistics

Karate Dataset is an example of Disassortative mixing dataset with **Assortative Degree** of -0.47(negative value indicates low assortativity), where nodes with different attributes connect with each other. Popular actors(Faction Leaders) are connected to unpopular actors and Faction leaders themselves don't interact with each other. This

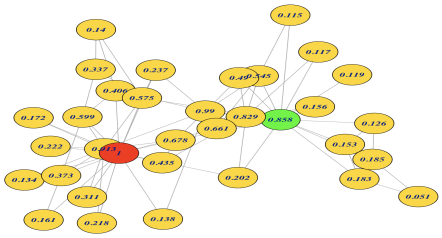
is a two-point influential dataset where whole influence of the network rests upon two actors(Mr. A and Mr. H). The dataset has an average path length of 2.4 and diameter of 13. Transitivity score in the network is 25% and the size of the largest clique is 5.



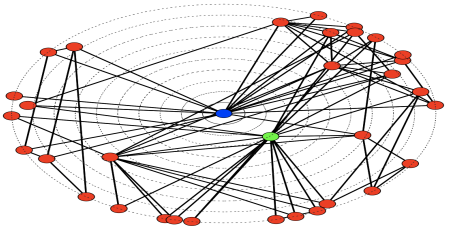
(a) Degree Centrality



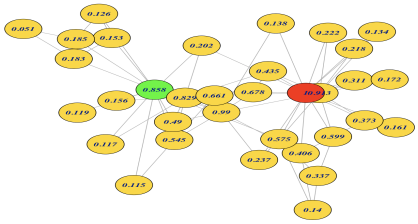
(b) Circle Target Plot for Degree Centrality



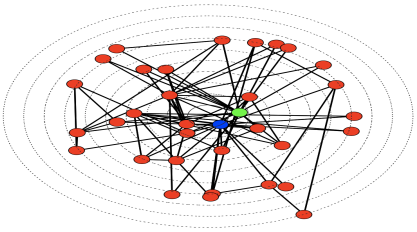
(c) Betweeness Centrality



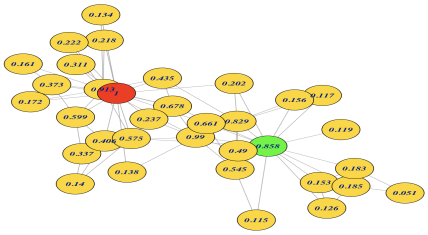
(d) Circle Target Plot for Betweeness Centrality



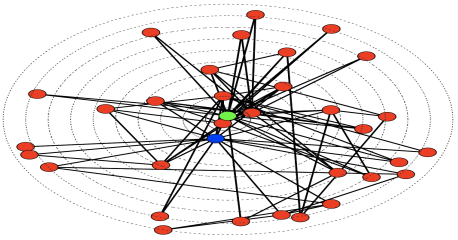
(e) Closeness Centrality



(f) Circle Target Plot for Closeness Centrality



(g) Eigen Vector Centrality



(h) Circle Target Plot for Eigen Centrality

Figure 1.14: Centrality Measures for Zachary Karate Club Network

All the plots in Figure 1.14 represents the centrality measures of the Karate Dataset. Colored circles represent the Faction Leader(Mr. H and Mr. A). Figure 1.14 (a) depicts the degree of actor by increasing the size of vertex with increase in degree of the vertex. All the target plots depict high centrality measures for the actors in the middle of the circle. Every plot shows that faction leaders exhibit a very high value of all the centrality measures with the highest score and centrally located in all the target plots.

1.5.2 Lawyers Lazega Dataset

This dataset has been accumulated from Northeastern US corporate Law Firm in New England from the year 1988-1991. The dataset is a combination of 71 nodes with different attributes and characteristics. The number of nodes of the dataset are connected to each other by three relationship types- Friendship, Advice, Work. With the same set of nodes, this dataset is a combination of three different networks that forms the core basis of analysis and study in the upcoming chapters of this dissertation. The dataset is appropriate for studying hierarchical working of a law firm and how different relationship affect the actors in the network. The dataset has been divided into following relationship networks:

- **Friendship Network:** The links in this network include all actors that socialize and are friends outside work. It does not include link between actors who are just-friends(on a friendly term only). This is a directed network.
- **Advice Network:** This network includes all the links between the actors who seek advice from other lawyers in the past year when the survey was conducted. This is also a directed network.
- **Work Network:** This network includes all the links where an actor has worked with another actor in atleast one case, has been assigned same case, used or read each others work. The work network is an undirected network, imply if lawyer 1 works with lawyer 2, it can be stated that lawyer 2 also works with lawyer 1.

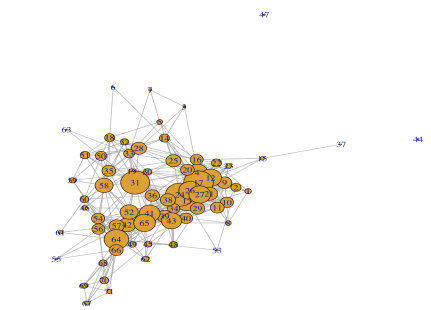
There are 8 actor attributes in the dataset:

- **Seniority:** This attribute defines the seniority level of the lawyer working in the firm where 1 being the highest seniority level and 71 being the lowest.
- **Status:** 1- partner(36/71 actors) and 2- associate(35/71 actors)
  Partners are the shareholder, operators and joint owners of the firm.
  Associates are employees in the firm that may become partner in the future.
- **Gender:** 1- Man(53/71 actors) and 2- Woman(18/71 actors)
- **Office:** 1- Boston(48/71 actors); 2- Hartford(19/71 actors); 3- Providence(4/71 actors)
- **Years in firm**: Mean working year in the firm is 10.5
- **Age**: Mean age of lawyers is 42
- **Practice:** 1- Litigation(41/71 actors) and 2- Corporate(31/71 actors)
  Litigation lawyers work with client to resolve any dispute or charges.
  Corporate practice lawyers are involved in corporate dealings. They are a part of acquisitions, mergers, corporate tax and other ventures. They represent corporate to other organizations and government institutes.
- **Law School:** 1- Havard/Yale(15/71 actors), 2- ucon(28/71 actors) and 3- others(28/71 actors)

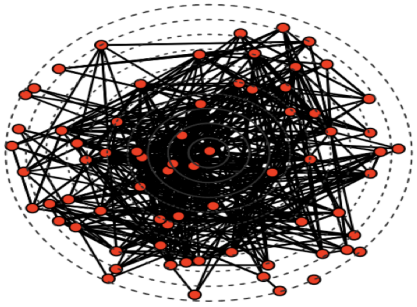
Friendship Lawyer Network Summary Statistics

Figure 1.15 shows all the centrality measure for the Friendship Network in the Lawyer Dataset. All actors in the centre of the plots have dense connection, thus all the plots have higher score for actors lying in the centre. There are two actors that have centrality measure as they are connected to any other actor in the network. The target plot for degree centrality Figure 1.15(b) shows that very few actors are in the middle of the target i.e. only few actors have high degree score and most of the actors have middle order degree score. The Betweenness centrality target plot Figure 1.15(d) shows that most of the actors are concentrated in the centre or middle which indicates most of the actors have high or average betweenness score. The Eigen Vector Centrality

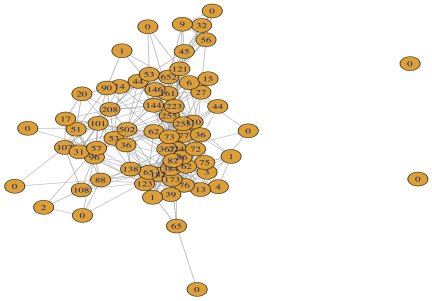
target plot Figure 1.15(f) is dispersed and distributed across all target indication range of different values assigned to the actors.



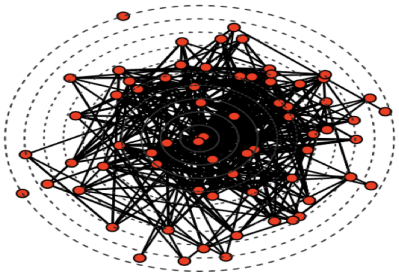
(a) Degree Centrality



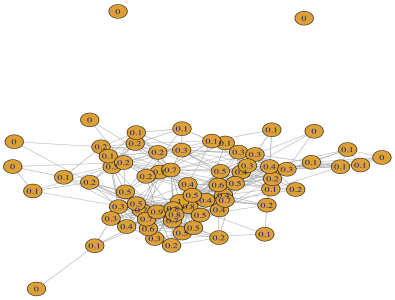
(b) Circle Target Plot for Degree Centrality



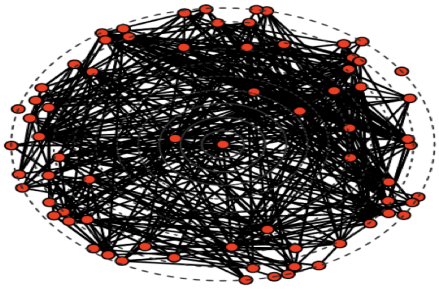
(c) Betweeness Centrality



(d) Circle Target Plot for Betweeness Centrality



(e) Eigen Vector Centrality



(f) Circle Target Plot for Eigen Vector Centrality

Figure 1.15: Centrality Measures for Lawyers Lazega Datasets

From Table 1.1: Advice network has the maximum number of edges. Advice and Friendship Network have high score Transitivity which makes sense if Actor A take advice from Actor B and Actor B takes advice from Actor C, there are high chances that Actor A will take advise from Actor C. Symmetric property of Work network is

	Friendship	Advice	Work
Number of Edges	575	892	756
Transitivity	0.44	0.47	0.30
Reciprocity	0.61	0.39	1
Average Path Length	2.5	2.4	2.1
Diameter	7	6	4
Assortativity	0.079	0.04	-0.17

Table 1.1: Network Summary Statistics for Friendship, Advice and Work Lawyers Network

validated by reciprocity of 1. Friends network has maximum average path length and diameter. Friendship network has positive value assortativity indicating lawyers make friends who are similar to them, whereas negative value of assortativity for the Work Network indicates it is a disassortative network with actors of different characteristics are working with each other and there are few actor that greatly influence the work network.

1.6 Erdös Renyi Model

This model is one of the simplest and basic probabilistic model where the presence and absence of edges are independent and identically distributed (i.i.d).

The edges in Erdös Renyi Model are formed with probability $p \in 0, 1$ and is independent of every edge. Let Y_{ij} follow Bernoulli Distribution, indicate the presence of edge between $actor_i$ and $actor_j$, such that

$$Y_{ij} = \begin{cases} 1 & \text{with probability p} \\ 0 & \text{with probability 1 - p} \end{cases}$$

Number of edges $\mathbb{E} = \mathrm{E}[\sum Y_{ij}] = \frac{N(N-1)p}{2}$

This model poorly fits the real world data as links or edges in the networks of real world are not independent of each other which violates the i.i.d constraint of Erdös

Renyi model. Some nodes in the network are more connected than other nodes of the network, the degree attribute along with other network characteristics are ignored by this model. The distribution of real world data is far from the Poisson Distribution which is the distribution followed by this model for the formation of edges.[Daudin et al. 2008]

Figure 1.16 shows the degree plot for Karate dataset containing 34 actors and 78 edges. The line with "+" denotes the predicted value from Erdos Renyi Model and the line with "o" denotes the observed karate dataset values. It is evident from the degree plot that the Erdös Renyi Model poorly fits the data. The observed and predicted lines are not overlapping and the model is overestimating in some cases while underestimating other points.

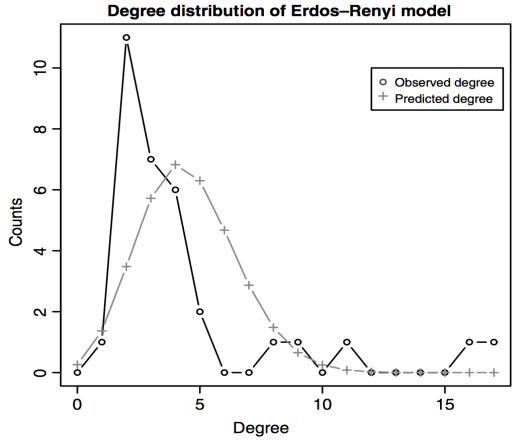


Figure 1.16: Karate Network- Observed vs Predicted(Erdos Renyi)[Salter-Townshend et al. 2012]

Chapter 2

Literature Review

Stochastic Block Model acts as a benchmark model for all the clustering and community identification models. The SBM model lays down the foundation for studying and analysis of communities within the data. There has been many advancements and ongoing research in the field of Social Network Analysis and Community Detection. The Stochastic Block Model is a vital part of the ongoing research, with many variants and adaptations applied to the basic model to improve the performance and optimally fit the data.

The General Stochastic Block model with Bayesian approach was proposed by [Daudin et al. 2008] to overcome the limitation of Erdos Renyi Model. Model parameters and selected number of classes are estimated given degree distribution and clustering coefficients. The paper states and studies the limitations of Erdos Renyi model's ability to map real world networks. It has been observed that the degree from Erdos Renyi models are far from Poisson Distribution and it assigns actors to unknown groups.

The paper of Abbe, Bandeira, Hall [Abbe et al. 2014] have focused on increasing the efficiency of the clustering algorithm and community recovery from existing bounds by identifying a threshold for the exact recovery. Threshold identified close to communities has been successfully recovered with semi-definite programming relaxation on the Maximum Likelihood.

General Stochastic Block Model assumes that the actor belongs to only one group but in real life situation, people may belong to multiple groups in the same network.

This variant of Stocchastic Block Model was proposed by [Airoldi et al. 2008] to accommodate Multi-Group membership for each actor called **The Mixed Membership Stochastic Block Model**. The paper proposed that each actor is assigned its own mixing parameter τ, denoting their tendency to belong to the assumed groups and two membership indicator vector Z^1 and Z^2 denoting the interaction between the sender and receiver. The paper also presented a variational inference algorithm for posterior inference approximation.

In most of the networks, the formation of communities and groups are not influenced by the density of links but also by the actor attributes in the network. Actor attributes are not taken into consideration in the generalized Stochastic Block Model. The variant **Mixed Membership of Experts Stochastic Block Model** proposed by [White and Murphy 2016] is an extension of Mixed Membership Stochastic Block Model to allow the influence of actor attributes into the model as a function of covariate data. In the proposed model the actor related attributes are incorporated in τ (Individual Mixing parameter) to prior distribution.

Degree Corrected Stochastic Block Model presented by [Gao et al. 2016] focuses on the limitation of SBM to map real world data, by considering nodes in the same community have same degree distribution and are interchangeable with each other; whereas in the real world dataset it is shown that actors belonging to same community can have different degree distribution. This heterogeneity of degrees among the actors is resolved by using the popularity of each actor as a set of degree corrected parameters $\theta_i = \theta_1, ..., \theta_n$ with the updated link distribution $Y_{ij} = Y_{ji} \overset{ind}{\sim} Bernoulli(\theta_i \theta_j B_{z(i)z(j)})$. The paper defines SBM as a special case of Degree Corrected SBM with zero degree corrections.

[Barbillon et al. 2015] proposed Multiplex Stochastic Block Model variant that can fit a Multi-Relationship Network data like the Lazage Lawyer Dataset with Friendship, Advice and Work Type between the actors. The adjacency matrix in this case is defined as an combination of Multiple Networks $Y_{ij}^{1..K} = Y^1, .., Y^K$ where Y^i is the individual adjacency matrix of the i^{th} network and K is the number for relationship types. Binomial edge Distribution of General Stochastic Block Model is replaced by Multinomial Edge Distribution in case of Multiplex SBM.

Chapter 3

Stochastic Block Model

Stochastic Block Model is a generative model for random graphs. The word **"Stochastic"** means statistical analysis and estimation of random distribution that may be difficult and practically impossible to estimate precisely. Stochastic Block Model is a community detection and cluster identification model which serves as a good benchmark to identify underlying groups within the network. In this chapter, we will study about this model, its concepts and notations and investigate implementation of the model.

3.1 Generative Model vs Discriminative Model

Generative Models tries to capture the underlying and hidden data generating mechanism of the give network dataset. These type of models lay their basis on estimations and parameters that map the distribution of the model. Once the data generation process of the dataset is captured, it can be used to simulate and synthesize demo datasets that would be similar in properties and attribute of the original dataset. Group distribution plays a very vital role in generative model. Naive Bayes Classifier is an example for this model.

Discriminative Models on the other hand lays its foundation on the data itself and not on its distribution or generation process. These type of models are heavily dependent on the amount of data, number of attributes and the quality of data like

number of missing variables, errors, noise. Logistic Regression is an example of this model.

Stochastic Block Model is a generative model, thus it tries to estimate the scheme behind the data generation process.

3.2 Specifications of Model

We consider a set of actors $a_1, a_2, a_3, ..., a_N$(where N is the total number of actors in the network) and edges or links $e_1, e_2, e_3, ..., e_E$(where E is the total number edges present in the network). For this dissertation we consider unweighted edges i.e. all edges are considered as same and binary where *1* signifies the presence of edge and *0* signifies the absence of edge and E signifies the total number of 1's or edges present in the network. Adjacency matrix for the dataset is represented as Y which is a matrix of dimensions NxN and it represents the presence and absence of edges as 0's and 1's between each actors in the network.

$$Y_{ij} = \begin{cases} 1 & \text{indicates the presence of link between actors } a_i \text{ and } a_j \\ 0 & \text{indicates the absence of link between actors } a_i \text{ and } a_j \end{cases}$$

Example of Adjacency Matrix Y of dimension 3x3, considering a network of 3 actors is given below:

$$Y_{ij} = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{array}$$

All the diagonals are 0 in the above adjacency matrix, this indicates absence of self loops on actors. Actor 1 in the first row has edge with actor 2 and actor 3, actor 2 has single edge with actor 3; and actors 3 has edges with both the actors.

For the purpose of the dissertation, we would be only considering networks with no self loops, thus the diagonal elements of the adjacency matrix would be zero.

3.2.1 Stochastic Block Model Notations and Symbols

Following are the notations and symbols along with their meaning that are used while defining the Stochastic Block Model:

- **N:** Total number of actors/nodes in the network
- **E:** Total number of edges in the network
- **G:** Number of Latent Groups underlying the data or number of clusters in the network data
- $\boldsymbol{Y_{ij}}$**:** Network represented as Adjacency Matrix of dimensions NxN

$$Y_{ij} = \begin{cases} 1 & \text{indicates the presence of link between actors } a_i \text{ and } a_j \\ 0 & \text{indicates the absence of link between actors } a_i \text{ and } a_j \end{cases}$$

- $\boldsymbol{\pi}$ **:** Interaction Matrix of dimension GxG. It depicts the interaction or presence of edges between the actors present in the same group and different groups. In other words, this matrix specifies the probability of interaction within the groups(intra-group) and outside the group(inter-group)

  π_{gh}- represents the probability of existence of link between actor belonging to Group g and actor belonging to group h. For undirected networks $\pi_{gh} = \pi_{hg}$

$$\pi_{gh} = \mathbb{P}(Y_{ij} |\, i\, in\, group\, g\, and\, j\, in\, group\, h)\, \forall i, j$$

$$\pi_{gh} = \begin{array}{c} \\ g_1 \\ g_2 \\ g_3 \end{array} \begin{array}{c} \begin{array}{ccc} g_1 & g_2 & g_3 \end{array} \\ \begin{pmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{pmatrix} \end{array}$$

  g_{11}, g_{22}, g_{22} represents the probability of connection within the groups 1, 2 and 3 respectively, higher probability signifies dense connection between the members of the groups; whereas other matrix elements indicates the probability of connection between members of two different groups like g_{12} denotes the probability of connection between members from group 1 and group 2.

- **$\boldsymbol{\alpha}$:** Class membership prior probability of dimension 1xG, it depicts the prior probability of actors belonging to group/class g

$$\sum_{g=1}^{G} \alpha_g = 1$$

- **Z:** Group Membership Indicator of dimension NxG, the symbol Z depicts the group to which actor belongs.

$$Z = \{Z_{ig}\}_{i=1,2,\ldots,N}^{g=1,2,\ldots,G} = \begin{cases} 1 & \text{if actor } a_i \text{ belongs to group g} \\ 0 & \text{otherwise 0} \end{cases}$$

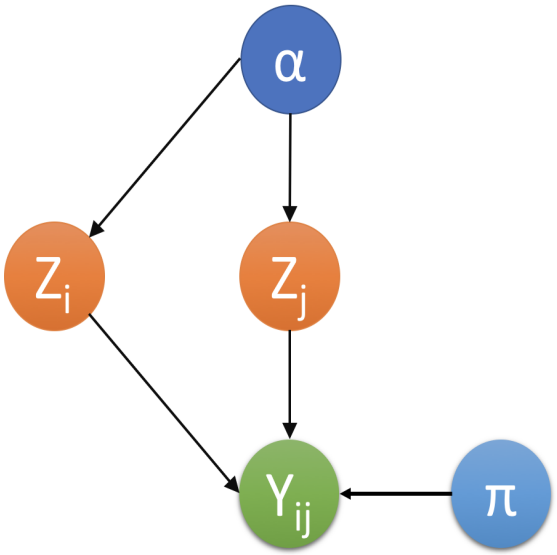


Figure 3.1: Parameters flowchart for SBM

In Stochastic Block Model, the group prior probability α and group membership indicator Z are the latent or hidden variables which would be estimated using the model.

3.2.2 Data Generative Process of Stochastic Block Model

The following figure depicts the data generative process of Stochastic Block Model as stated by [Snijders and Nowicki 1997:

- for i in 1 to N {
  $Z_i \sim Multinomial(\alpha)$
  }
- for i,j in 1 to N {
  $Y_{ij}|Z_{ig} = 1, Z_{jh} = 1 \sim Binomial(\pi_{gh})$
  }

Figure 3.2: SBM Data Generative Process

3.3 Expectation-Maximization(EM) Algorithm

Stochastic Block Model consists of latent variables and parameters α and π. Expectation-Maximization Algorithm tries to estimate the **Maximum Likelihood** using the model parameter following a series of iterations. EM algorithm uses the iterative and random assignment approach to estimate probabilistic divisions/clusters in the data. EM algorithm approach is very similar to that of K-means where the algorithm starts with random placement of cluster centres, then assigning each data point to their respective initial cluster, then again cluster centres are updated, followed by cluster reassignment for each node in the network till the time there is no change in cluster centre. This algorithm tries to capture the estimation of Maximum Likelihood as the precise value of Maximum Likelihood cannot be evaluated due to some missing data or incomplete information.

3.3.1 What is the significance of the Maximum Likelihood and Log Likelihood?

The aim of Maximum Likelihood is to estimate an optimal fit for the distribution of data. Parameter of the model are estimated in order to maximize the likelihood

of matching the data generation process of the network. Probability and estimation equations tend to become more complex and solving these complex equations is computational expensive and time consuming. Equations can be made simpler using the properties of Logarithmic function . The product to sum rule of log functions and easy differentiation of log terms makes the calculation easy and comprehensible. Logarithmic function is a monotonic increasing function which means that if the value inside the logarithmic function increases, the log value also increases, and this property is extremely useful while calculating the **Maximum Log Likelihood** as maximization of logarithm function also maximizes the original probability function.

3.3.2 Expectation Maximization Algorithm

The EM Algorithm involves two steps in a iterative manner: the first step is the **E-Step (Expectation Step)** where the current parameters of the model and adjacency matrix(observed network data) are used to find expected value of latent variables of the model; the second step is the **M-Step (Maximization Step)** where the likelihood function is maximized to estimate the model parameters of the model and latent variables/ missing data are assumed to be equal to the current iteration estimate.[Gupta and Chen 2011]

Let $\theta = \{\theta_1, \theta_2, ..., \theta_k\}$ be unknown parameters that needs to be estimated and calculated to optimize the random vector Y with distribution dependent on the values of θ. The values of θ needs to be estimated in such a way that Conditional Probability of Y given unknown parameter θ is maximized; which in turn is called maximum likelihood estimation. The goal is to calculate the **Maximum Log Likelihood** due to the benefits of using the logarithmic function over the normal Probability function as stated in the previous section. This can be represented by the following equation:

$$L_\theta = ln\, P(Y|\theta)$$

Strictly increasing property of Logarithmic function, will maximize the both L_θ and $P(y|\theta)$. Being an iterative algorithm, it will try to maximize the function L_θ at each iteration. Consider L_{θ_i} and $L_{\theta_{i+1}}$ log likelihood function at $i^{th}\, and\, i+1^{th}$ iteration respectively with θ_i and θ_{i+1} as the current of estimate of parameter θ. Thus to

maximize the likelihood function:

$$L_{\theta_{i+1}} > L_{\theta_i}$$

Putting $L_\theta = ln\, P(Y|\theta)$ in the above equation, it is equivalent to:

$$ln\, P(Y|\theta_{i+1}) > ln\, P(Y|\theta_i)$$

Therfore, we have:

$$L_{\theta_{i+1}} - L_{\theta_i} = ln\, P(Y|\theta_{i+1}) - ln\, P(Y|\theta_i)$$

Introducing the hidden/latent variable Z with the realization of z in the above equation to make θ likelihood tractable,

$$L_{\theta_{i+1}} - L_{\theta_i} = ln \sum_z P(Y|\theta_{i+1}, z) P(z|\theta_{i+1}) - ln\, P(Y|\theta_i)$$

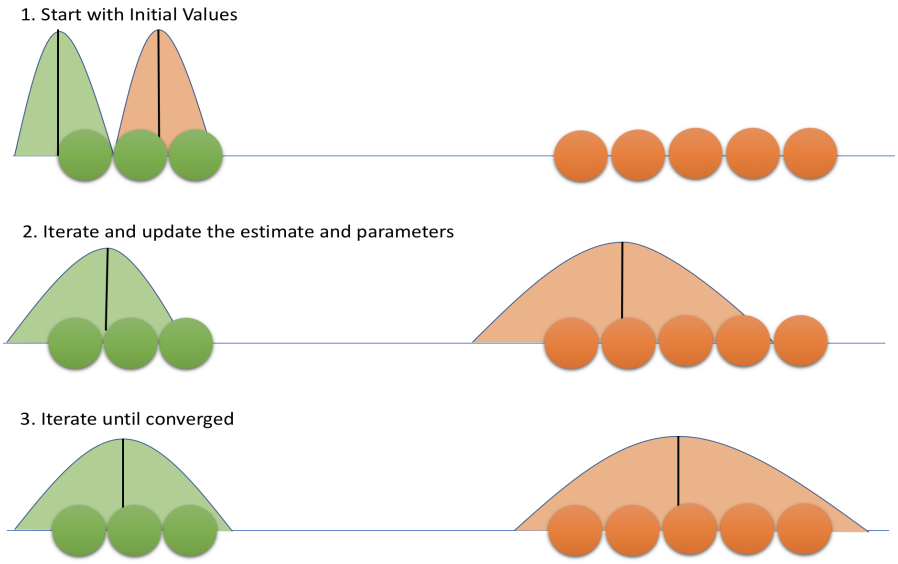


Figure 3.3: Illustrative Working of EM Algorithm

Algorithm:

1. **E(Expectation)-Step:** In this step we compute the function $F(\theta, \theta^i)$ where θ^i is the parameter estimate of θ for the iteration i

$$F(\theta, \theta^i) = Exp_{\theta^i}[L_\theta|x]$$

   Here $Exp_{\theta^i}[L_\theta|x]$ is expected value of log likelihood function over the given θ

2. **M(Maximization)-Step:** In this step value of θ^{i+1} is estimated by maximizing the function $F(\theta^i, \theta^{i-1})$ with respect to value of θ^i so that,

$$F(\theta^{i+1}, \theta^i) \geq F(\theta^i, \theta^{i-1})$$

3. **Repeat Step 1 and 2:** E-Step and M-Step are iterated to get the updated value of model parameters and log likelihood estimates until the model is converged or the difference between the values of log likelihood in subsequent iterations is sufficiently very small.

$$L_{\theta_{i+1}} - L_{\theta_i} \rightarrow is\, very\, small$$

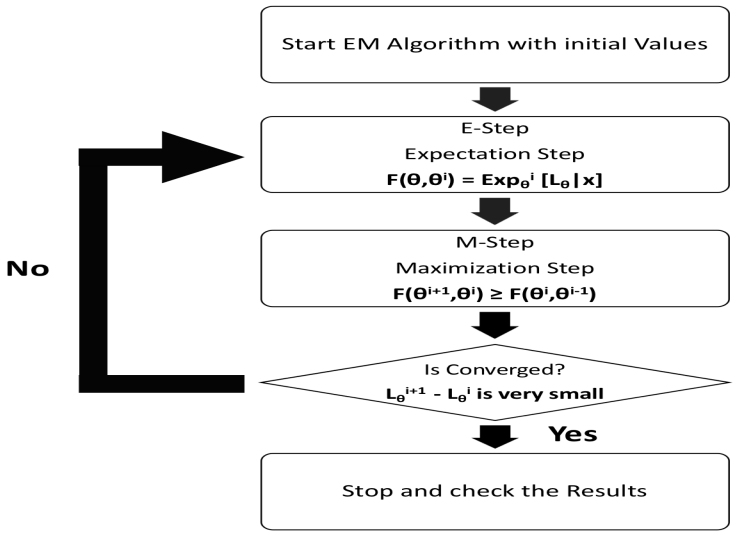


Figure 3.4: EM-Algorithm Flowchart

3.3.3 Advantages and Limitations of EM Algorithm

Advantages:

- Likelihood will increase at every iteration.
- The algorithm is guaranteed to converge at local optima.
- One of the popular and easy to use approach.

Some limitations of EM Algorithm:

- The algorithm may converge to local maxima rather than converging on global maxima.
- Due to large number of iterations involving high computation, the speed of the algorithm can be very slow.
- The algorithm works well on small dataset with less percentage of missing data and low dimensionality.

3.4 Log Likelihood for the Complete Data

The dataset by itself is incomplete, there are some latent or missing variables that define and influence the distribution of data and formation of groups within the network. Complete Data Log Likelihood as stated by [Daudin et al. 2008], considers Y as the adjacency matrix or the set of edges in the network $\left\{Y_{ij}\right\}_{i,j=1,2,...,N}$ and Z as the membership indicator with $\left\{Z_{ig}\right\}_{i=1,2,...,N}^{g=1,2,...,G}$,

$$logL(Y|Z) = \sum_{i=1}^{N}\sum_{g=1}^{G} Z_{ig}\, log\, \alpha_g + \frac{1}{2}\sum_{i\neq j}^{N}\sum_{g,h=1}^{G} Z_{ig}\, Z_{jh}\, log(\pi_{gh}^{Y_{ij}}(1-\pi_{gh})^{1-Y_{ij}})$$

Possible values of the latent variable Z can be used to find the summation of complete data likelihood which will give the values of likelihood of the given data defined by Y.

3.5 Variational EM Algorithm

The major limitation of the generalized EM is that it makes an assumption of conditional independence of network edges/links on the latent variables. The edges between $actor_i$ and $actor_j$ Y_{ij} is marginally dependent and conditionally independent on Z_{ig} i.e. the groups membership indicator of the $actor_i$. This assumption works well for smaller networks with less amount or vertices and edges but for larger networks, this makes the computation of complete data likelihood intractable. The dependency structure tends to become too complicated on integrating.

$$Pr(Z|Y) \textit{ is not computable as } Z_{ig}\textit{s are not independent}$$

Thus, Conditional Distribution $P(Z|Y)$ and Likelihood of the given data L(Y) is intractable; some other approach is needed to make the computation tractable.
In variational approach of EM Algorithm, **Mean Field Approximation** is used to ensure that the Group Membership Indicators Z are independent to each other given the observed data Y such that $Z_i \perp Z_j$ for $i \neq j$.

Lower Bound of Log Likelihood Log L(Y) $= J(R_Y) = log\, L(Y) - KL[R_Y(Z), Pr(Z|Y)]$

where KL is *Kullback-Leibler* divergence defined as,

$$KL[R_Y(Z), Pr(Z|Y)] = \sum_{z \in Z} R_Y(Z)\, log\, \frac{R_Y(Z)}{P(Z|Y)}$$

Kullback-Leibler Divergence is also known as Relative entropy which is a difference between two reference probability distribution.

$$\textbf{Kullback-Leibler} = \begin{cases} 1 & \text{distributions are different, first distribution approaches 0} \\ 0 & \text{distributions are similar, in above case } \mathrm{R}_Y(Z) = P(Z|Y) \end{cases}$$

The Lower Bound equation can be rewritten as,

$$J(R_Y) = \sum_{z \in Z} R_Y(Z)\ log\, P(Z|Y) - \sum_{z \in Z} R_Y(Z)\ log\, R_Y(Z)$$

Chapter 4

Model Inference

In this chapter, architecture, logic and code implementation of Stochastic Block Model will be discussed followed by a flowchart for fitting the dataset. Lastly model parameters will be evaluated and the methodology involved to study and analyze the goodness of fit will be explained.

4.1 Implementing SBM using Variational EM Algorithm

This section will depict the implementation details of Variational EM algorithm explained in the last chapter.

Symbol	**Description**
Z	Group Membership Indicator
α	Prior Group Probability
π	Group Interaction Matrix
Y	Adjacency Matrix
τ	E[Z] Expected value or approximation of Z

Table 4.1: Recap of SBM Symbols and Notations

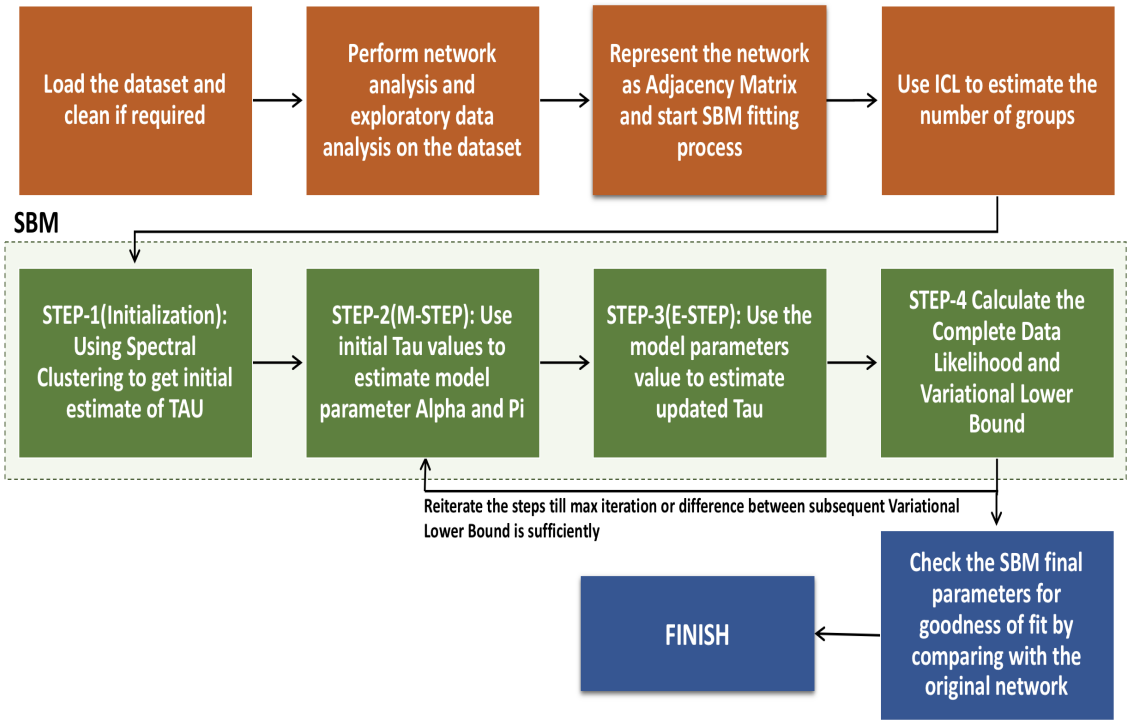


Figure 4.1: SBM Implementation Flowchart

4.1.1 Network Data Preparation, Pre-processing And Analyis

Pre-Processing

There may be some cases that the data in original form cannot be used for fitting the SBM model. The models expects the data in adjacency matrix format of Dimension NxN where N being the total number of actors:

- SBM expects the links to be in binary format i.e. 1 for presence and 0 for absence. If the network contains weighted links it should be converted to binary.
- SBM model cannot handle missing data, thus missing data should be handled or imputed before supply the values to the model.
- There are other variations of Stochastic Block Model that can handle actor and link attributes but the generalized SBM model does not support such attributes.

Exploratory Network Data Analysis

The network data should be subjected to analysis and visualization before fitting the SBM model. This will be beneficial and helpful to get an overview and gist of the data, important attributes, most influential nodes, most connected nodes, disconnected nodes and many other properties and characteristics that underly the data. Following properties are used and explored under this step, the basic notion and concepts behind these properties are discussed in Chapter 1:

- Degree Distribution
- Centrality Measures like degree centrality betweeness centrality, eigen vector centrality
- Network Graph Measure
- Geodesic distance measure

4.1.2 Groups Estimation Using Integrated Complete Data Likelihood

One of the most important assumptions and requirements for Stochastic Block Model is the number of Groups G that divide and cluster the data. The SBM model function requires number of latent groups as an input argument. But now the question arises, how do we estimate and choose the number of groups that will optimally fit the data. One way would be iterating the SBM function with different number of groups, keeping the input dataset same and then checking the optimal group size by evaluating goodness of fit for each group sizes. This process is time consuming, computational expensive and redundant as we are evaluating goodness of fit for all groups tested.

Form many mixture models, **BIC (Bayesian Information Criterion)** is used as a model selection metrics where higher value of BIC indicate better fit. Complexity of the model or number of parameters in the model is penalized by BIC.

$$BIC(G) = log\, L(Y) - \frac{V_G}{2} log\, N$$

where G is number of Groups, L(Y) is Likelihood of given data Y, N is number of actor and V_G number of parameters for the G Groups

There are limitations of using BIC in Gaussian mixture models as BIC assumes consistent and fixed number of components and works well for true distribution but this property of BIC may not scale well for grouping and clustering algorithms as it tends to overfit in majority of cases.[Nishii 1988]. And calculation of BIC involves computation of Data Likelihood for the given data which is intractable. Thus, [Baudry et al. 2010] proposed **Integrated Complete Data Likelihood(ICL)** as a appropriate model selection metric for clustering algorithms. ICL is equal to BIC but penalized by mean entropy.

$$ICL(G) = log\, L(Y, \hat{Z}|G, \hat{\theta}_G) - \frac{V_G}{2} log\, N$$

where $\hat{Z}$ is MAP estimate

computes Integrated Log Likelihood of the Complete Data which is tractable rather than computing likelihood of the given data. ICL favours better separated clusters which effectively means fewer groups are selected by ICL than BIC.

4.1.3 Principal Component Analysis using Spectral Clustering

Complex networks and dataset contain large number of variables and unnecessary information that have little or no influence on our target variable. **Principal Component Analysis or PCA** is a technique to retain the information and influence of vital variables in a dataset with large amount of variables but at the same time reducing the dimensionality i.e reducing from a large set of unnecessary variable/ dimensions to a set of important and vital variables. The goal of PCA technique is to transform the data from a large set of correlated variables to a set of **Principal components** (small set of vital uncorrelated variables). Principal components have decreasing order of variance with first principal component representing maximum variability in the data at that time, followed by second principal component representing the maximum

variability of the remaining data.

Spectral Clustering is used for Principal Component Analysis and initialization of SBM. This is a simple and widely used technique for exploratory data analysis and act as a perfect kickstarter to further algorithms. Spectral Clustering works with goal of divide and conquer. It places similar data points into same clusters and tries to maximize the dissimilarity between different clusters. Thus, at the end we have all the similar points in same cluster and dissimilar points in different cluster or group.

Spectral Clustering is used to estimate initial value of $\boldsymbol{\tau}$ (expected value of Group Membership Indicator Z), that is further used to estimate the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$. Thus, using this techniques, other model parameters and dependencies need not be assumed or randomly initialized, they can extracted from just a initial τ value estimate.

Algorithm:

1. Given an adjacency matrix Y as an input, we find the **Eigen Vectors** of the this matrix.

   Eigen Vectors play a important role in Machine Learning, Computer Vision and Principal Component Analysis. Linear transformation has no effect on the direction of eigen vectors. Linear Transformations can be expressed in terms of eigen vectors which gives scaling directions and eigen values which signifies the scaling factors.

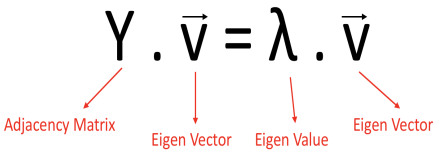


Figure 4.2: Eigen Vector and Eigen Values

Here, Y is NxN matrix, $\vec{v}$ *is* Nx1 vector and λ *is* scalar quantity

Figure 4.2 equation can be rewritten as,

$$(Y - \lambda.I)v = 0$$

2. Step 1 will give a matrix of Eigen vectors of dimension NxN: $Y_1, Y_2, ..., Y_N$. Step 2 involves selecting eigen vectors corresponding to the group size G such that the final selected Eigen Vectors are $Y_1, Y_2, ..., Y_G$.

3. Run **K-Means** Clustering Algorithm on the selected Eigen Vectors $Y_1, Y_2, ..., Y_G$ data points with clusters size as G. K-Means Algorithm will assign all the nodes in the dataset to their respective clusters.

4. Estimate $\boldsymbol{\tau}$ from the resultant cluster:

$$\tau_{ig} = \begin{cases} 0.8 & \text{if actor } a_i \text{ belongs to group g} \\ \frac{0.2}{G-1} & \text{otherwise} \end{cases}$$

Thus, the resultant τ matrix will have the elements containing group membership probability of 0.8 if the actor belongs to that group.

4.2 Estimation and Updates of Model Parameter using Variational EM Algorithm

This section will describe the evaluation approach and formulas used for various quantities and parameters that build up the SBM model. All the formulas and expressions given below need to be computed for the algorithm:

4.2.1 E-Step ($\boldsymbol{\tau}$ Update)

In E-Step the updated value of τ is computed on the basis of current estimate of model parameters α and π.

$$\hat{\tau}_{ig} = \alpha_g \prod_{i \neq j}^{N} \prod_{g,h}^{G} (\hat{\pi}_{gh}^{Y_{ij}} (1 - \hat{\pi}_{gh})^{1-Y_{ij}})^{\hat{\tau}_{jh}}$$

$\boldsymbol{\tau}$ values need to normalized before using further:

$$\hat{\tau}_{ig} = \frac{\hat{\tau}_{ig}}{\sum_h^G \hat{\tau}_{ih}}$$

4.2.2 M-Step (α and π Update)

In this step values of α and π are computed on the basis of current estimate of τ and adjacency matrix Y.

$$\hat{\alpha}_g = \frac{1}{N} \sum_i \hat{\tau}_{ig}$$

$$\hat{\pi}_{gh} = \frac{\sum_{i \neq j} \hat{\tau}_{ig} \hat{\tau}_{jh} Y_{ij}}{\sum_{i \neq j} \hat{\tau}_{ig} \hat{\tau}_{jh}}$$

4.2.3 Complete Data Log Likelihood Computation

$$logL(Y|Z) = \sum_{i=1}^{N} \sum_{g=1}^{G} Z_{ig} \, log \, \alpha_g + \frac{1}{2} \sum_{i \neq j}^{N} \sum_{g,h=1}^{G} Z_{ig} \, Z_{jh} \, log(\pi_{gh}^{Y_{ij}} (1 - \pi_{gh})^{1 - Y_{ij}})$$

4.2.4 Variation Lower Bound Computation

Variational Lower Bound is a check to test convergence of algorithm. With each iteration of the algorithm the value variational lower bound should increase. Maximization of lower bound maximizes the marginal probability. When the difference between the approximate distribution and true posterior distribution is minimum, it signifies that the lower bound has attained log probability.

$$J(\tau, \pi, \alpha) = \sum_{ig} \tau_{ig} \, log \alpha_g + \sum_{i<j,g,h} \tau_{ig} \tau_{jh} \pi_{gh}^{Y_{ij}} (1 - \pi_{gh})^{1 - Y_{ij}} - \sum_{i,g} \tau_{ig} log(\tau_{ig})$$

4.2.5 Integrated Complete Data Likelihood Computation

ICL is used to compare models fitted with different group numbers and find the optimal number of groups that best fit and describe the network communities. ICL is calculated by subtracting the penalty term from the likelihood. ICL penalty is a function of group number.

$$vICL(G) = \sum_{i,g} \hat{Z}_{ig} \, log \hat{\alpha}_g + \sum_{i<j,g,h} \hat{Z}_{ig} \hat{Z}_{jh} \hat{\pi}_{gh}^{Y_{ij}} (1 - \hat{\pi}_{gh})^{1 - Y_{ij}} - \frac{1}{2} \Bigg(\frac{G(G+1)}{2} \, log \frac{N(N-1)}{2} + (G-1) log(N) \Bigg)$$

4.3 Stopping Criteria for the model

Stochastic Block Model is an computational intensive algorithm, so iterating though the algorithm without any condition would be an overkill. Thus, a stopping criteria or condition is required to indicate that the model has finished publishing and fitting the data. Following are the two stopping criterion used in the SBM implementation:

- Limiting the maximum number of iteration depending on the size of the network dataset. The algorithm will stop when iteration count is equal to the allowed maximum iteration count.
- The Stochastic Block model EM algorithm is said to converge if there is no significant rise in Variational Lower Bound for two successive iterations i.e. difference between the two successive Variational Lower Bound is sufficiently small and tending to zero.

4.4 Goodness of Fit

Generative property of SBM is used to evaluate the model fit. The fitted model is used to simulate test networks that should map the data generative process of the original dataset. These set of simulated graphs are then compared against the network properties and statistics of the original dataset. The model is said to optimally fit the data if the network properties of the simulated graphs match the properties of original network. Network statistics used to evaluate goodness of fit in this paper are: Degree Centrality, Betweenness Centrality, Eigen Vector Centrality, Triangle count.

Chapter 5

Application of Stochastic Block Model to Datasets

This chapter presents the fitting process of Stochastic Block Model to Zachary karate club dataset and Lazega Lawyer Dataset. Lastly, it will discuss the goodness of the fit and present a detailed analysis of Stochastic Block Model output and study the identified clusters with respect to the network data.

5.1 Application to Zachary Karate Club Dataset

The Karate dataset has 34 actors and 78 edges.

5.1.1 Estimating number of Groups

Groups are estimated by running the ICL algorithm from 1 to 10 groups with 100 iterations of SBM model. Figure 5.1 shows a plot between group iterations from 2 to 10 and their corresponding ICL values. Higher the ICL value, better is the fitting capability of SBM model for that particular group. The figure shows a sharp peak at Group Size 4, which suggests that the Karate data is optimally divided into 4 groups. ICL increases from Cluster Sizes 2 to 3 with maximum at Group Size 4 and then falls continuously till the 10 group model. Decrease in ICL for larger groups is an indicator of higher penalties for higher order groups.

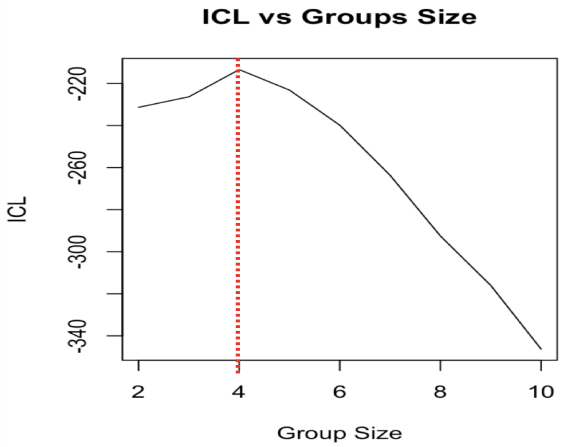


Figure 5.1: ICL Vs Groups size for Karate Network

Stochastic Block Model is fitted with **4 groups** in the Karate dataset.

5.1.2 Test of Convergence

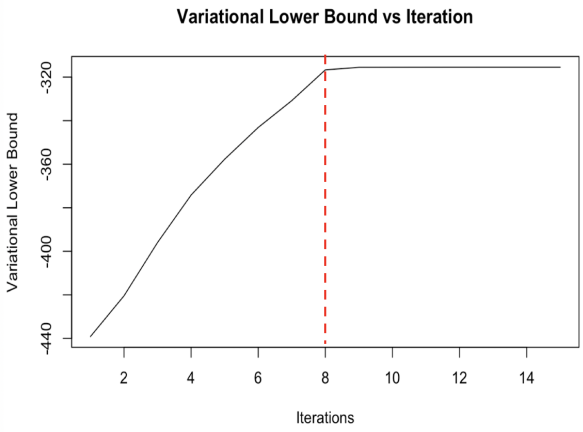


Figure 5.2: Variational Lower Bound Vs Iteration for Karate Dataset

Figure 5.2 shows that the SBM model with 4 groups converges at the 8^{th} iteration with Variation Lower Bound values becoming static after the iteration. The SBM algorithm reached its stopping criteria at the 14^{th} iteration with sufficiently small change in variational lower bound in successive iterations.

5.1.3 Network Visualization and Model Parameters Analysis

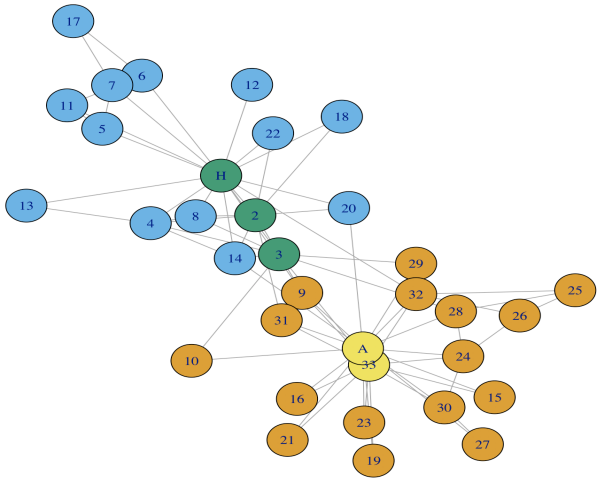


Figure 5.3: Network Visualization of Karate Dataset (4 groups)

Figure 5.3 show 4 clusters in Karate dataset color encoded according to their cluster membership.

Alpha	**Group 1**	**Group 2**	**Group 3**	**Group 4**
Group Membership	0.482	0.369	0.088	0.058

Table 5.1: Prior Group Membership parameter α for Karate Dataset

Comparing Figure 5.3 and Table 5.1 Group Membership α, it can be seen that Karate Network consist of two big groups on the boundary and two small groups in the centre of the network. The faction leaders Mr. A and Mr. H are part of two different smaller groups. As, it is already known that the faction leader have high degree centrality, it can be inferred that each small group will have high interaction with one of the big groups. Small Group color coded as yellow with Faction leader Mr. A have high interaction big group encoded in orange and similarly the other small group containing Mr. H interacts with blue coded big group.

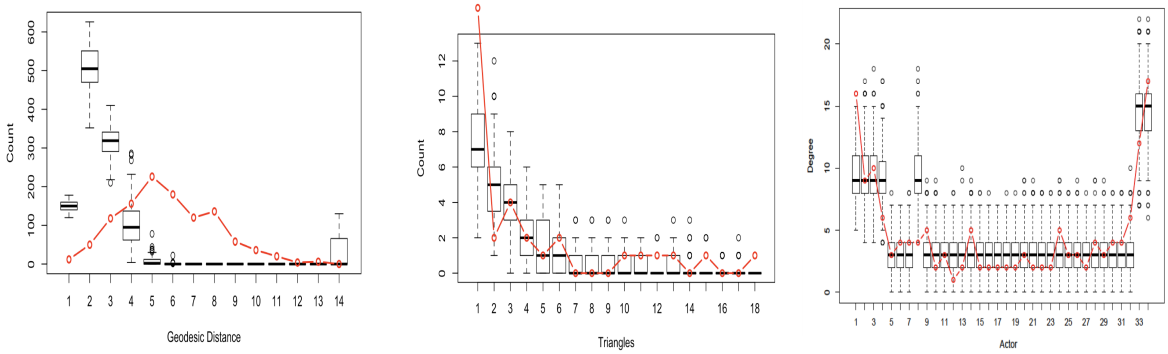
Table 5.2 and Figure 5.3 shows the value of interaction matrix, indicating high interaction of Group 1(big group - orange) with Group 4 which contains their Faction Leader and similarly Group 2(big group - blue) with Group 3 which also contains

Pi Group Interaction Matrix	Group 1	Group 2	Group 3	Group 4
Group 1	0.074	0.000	0.160	0.739
Group 2	0.000	0.113	0.530	0.069
Group 3	0.160	0.530	0.667	0.166
Group 4	0.739	0.069	0.166	0.500

Table 5.2: Group Interaction Matrix π for Karate Dataset

their Faction Leader Mr. H. Smaller groups with Faction Leader have high intra-group interaction.

5.1.4 Model Evaluation and Goodness of Fit



(a) Geodesic Distance Comparison (b) Number of Triangles Comparison (c) Degree Comparison

Figure 5.4: Goodness of Fit for Karate Network Dataset

Figure 5.4 plots convey that the simulated networks from SBM model are not able to map the **Geodesic Distance** of the original dataset depicted by red line on the plots. None of the Geodesic points for the original dataset lie inside the boxes in Figure 5.4 (a), whereas the model has mapped the generative process node degrees and triangle formation from the original dataset. Figure 5.4 (b) and (c) show that the majority red line points lie in the boxplot range for Degree and Triangle Plot. Thus, the model is able to simulate with approximately same degree and triangle but different shortest path.

5.2 Application to Lazega Lawyer Network Dataset

Lazega Lawyer Dataset consists of 71 actors spanning across 3 type of networks with different link types(Friendship, Network, Advice). The dissertation is implemented by keeping the Lawyer Friendship Network Dataset as a basis for training and fitting other network link types. The aim of this methodology is to bring the Friendship network, Work Network and Advice Network on the same level and lay down a basis for better comparison and contrast study. As the number and density of edges in three types of network differ from each other, we may estimate and compute different groups or cluster sizes from SBM model for each network type leading to a convoluted comparison process.

5.2.1 Estimation number of Groups for Friendship Lazega Dataset

To estimate the optimal number groups contained the dataset, Integrated Complete Data Likelihood function is executed for group size ranging from 1 to 10 along with 10 iteration of SBM algorithm for each group value.

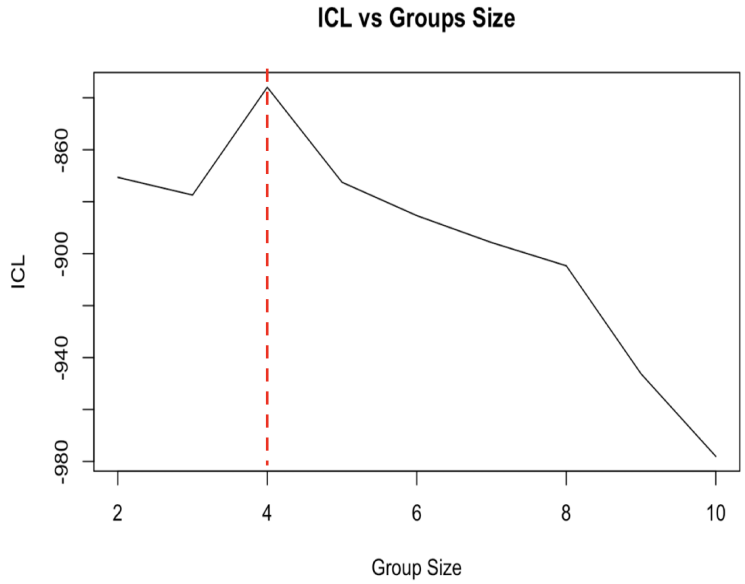


Figure 5.5: ICL Vs Groups size for Lazega Friendship Network

Figure 5.5 shows a plot between group iterations from 2 to 10 and their correspond-

ing ICL values. Higher the ICL value, better is the fitting capability of SBM model for that particular group. The figure shows a sharp peak at Group Size 4, which suggests that the friendship data is optimally divided into 4 groups. ICL increases from Cluster Sizes 2 to 3 with maximum at Group Size 4 and then falls continuously till the 10 group model.

As the maximum ICL value is achieved at 4 clusters, the Stochastic Block Model is fitted to the dataset assuming **4 groups** within the network.

5.2.2 Test of convergence

As discussed in previous chapters, the maximum iteration limit and successive variational lower bound differences govern the convergence and stopping criteria of the Variational EM Algorithm model. The Lazega Lawyer Network is tested for convergence with 10 maximum iteration limit and assuming number groups/clusters in the network as 4.

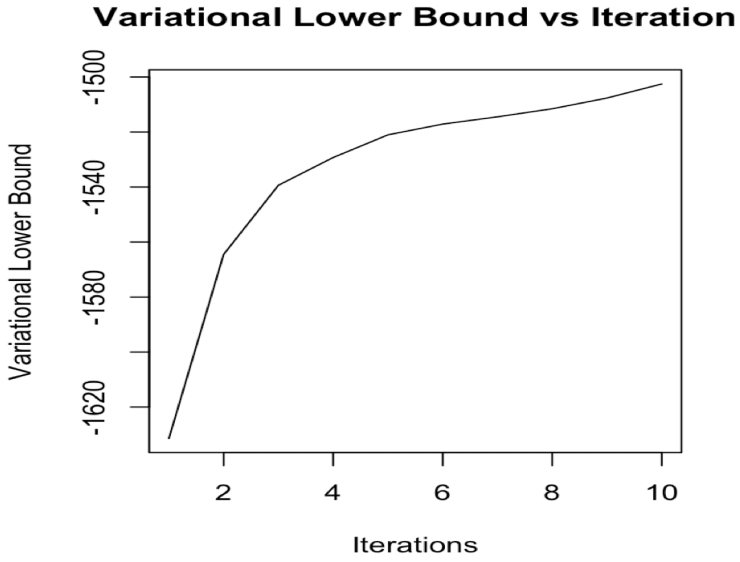


Figure 5.6: Variational Lower Bound Vs Iteration for Lazega Friendship Network

Figure 5.6 depicts the the values of model Variational Lower Bound with respect to the iterations. It is clearly visible that the model fails to converge under 10 maximum iterations and stopping criteria is reached before convergence. From the above graph it

can be inferred that the Variation Lower Bound increases exponentially till 4^{th} iteration and then it stabilizes showing an approximate horizontal level change till 10^{th} iteration. But the difference between any two successive iteration is not sufficiently small to trigger the stopping criteria and convergence.

Though the model has not converged under maximum iteration limit, but still we can consider partial convergence as the Variational Lower Bound change after 6^{th} iteration is almost the same and difference is minimal.

5.2.3 Network Visualization and Model Parameter Analysis

Figure 5.7 shows the visualization of Lazega Friendship Network with actors color encoded with respect to the clusters they belong.

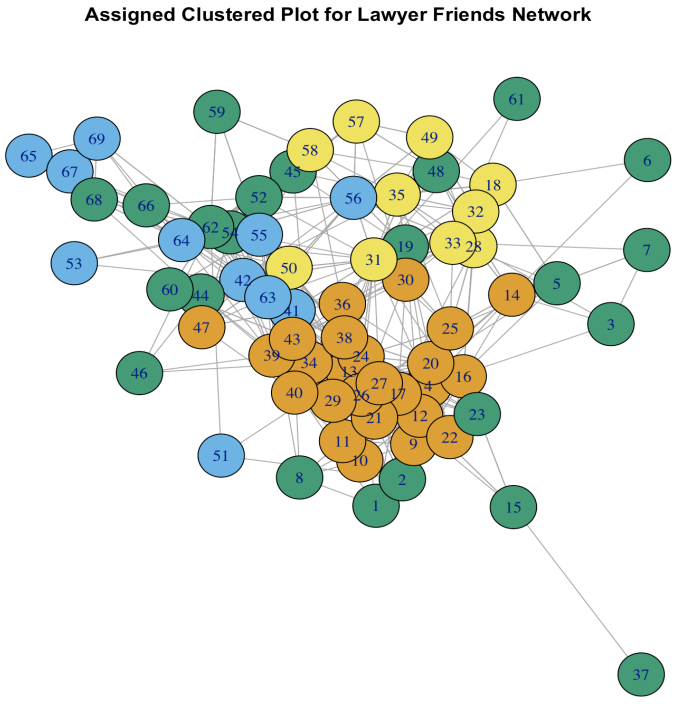


Figure 5.7: Network Visualization for Lazega Friendship Network

Actors in orange colored group are in the centre of the network with dense connections; whereas the actors belonging to green colored cluster are on boundary of the graph with low inter-connectivity. The network is distributed type of network with no central actor in any of the clusters.

Alpha Group Membership	Group 1	Group 2	Group 3	Group 4
	0.326	0.175	0.354	0.146

Table 5.3: Prior Group Membership parameter α for Friendship Network

Table 5.3 shows the Group Membership Probability of the network. It can be seen that there are two dominant groups with more than 30% probability of an actor belonging to that group and two smaller groups with similar probabilities. On comparing the group membership probability α table with the Network visualized in the figure 5.7, two dominant groups are color coded in green and orange colors and two similar groups have blue and yellow color.

Pi Group Interaction Matrix	Group 1	Group 2	Group 3	Group 4
Group 1	0.371	0.058	0.075	0.027
Group 2	0.094	0.462	0.093	0.086
Group 3	0.041	0.053	0.032	0.021
Group 4	0.095	0.094	0.048	0.548

Table 5.4: Membership Interaction parameter π for Friendship Network

Table 5.4 shows the value of parameter π(Group Interaction Matrix) for the network. The green colored diagonal elements show the interaction within the groups and rest other elements in the table show interaction between the marked groups. Interaction of each group would be examined by studying the Table 5.4 row-wise:

- **Group 1 Interactions:** Group 1 has dense interactions within the group(Group 1-Group 1: 0.371) which suggests that the lawyers belonging to this group are friends with each other and rest of the elements in the Group 1 row have very low probability which indicates low friendship outside the group. Thus, Group 1 lawyers are friends with each other and are not friendly outside the group.

- **Group 2 Interactions:** Group 2 members also have a high within group friendship (Group 2-Group 2: 0.462) and low friendship probability outside the group.

- **Group 3 Interactions:** All the elements in Group 3 row exhibit very low probability which indicates absence of links or friendship among the members of

Group 3 and even outside the group 3. This suggests that the group 3 members are aloof in the network and are not friendly in the network.

- **Group 4 Interactions:** Similar to Group 1 and Group 2; group 4 members are good friends with each other and exhibit low friendship outside the group.

5.2.4 Model Evaluation and Goodness of Fit

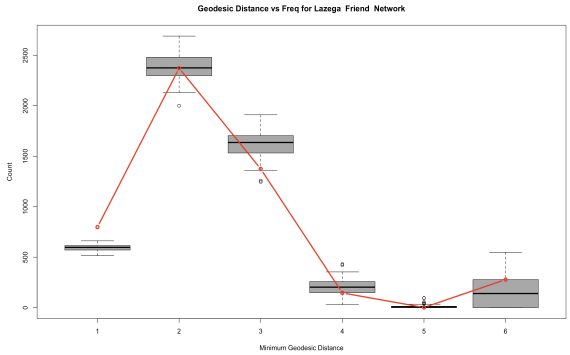
Variational EM Algorithm for the Stochastic Block Model has been fitted to all the three links of Lazega Lawyer Network. Friendship Network forms a basis for other two link networks i.e. estimated number of groups for the Advice and Work networks were same as computed for the Friendship Network by ICL algorithm evaluation. So the number of groups are assumed as 4 for all the three networks.

Goodness of fit for the model will be evaluated on the basis of network statistics and methodologies discussed in Chapter-3 Goodness of Fit Section. The network summary statistics would be compared against the actual network data and the simulated data. This section will describe the Goodness of fit for all the Lazega Lawyer Network types:

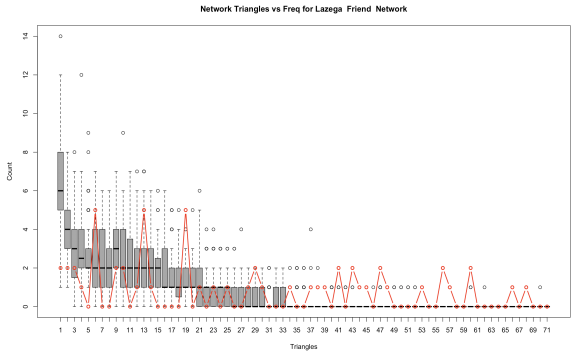
Lawyer Friendship Network

Figure 5.8 shows that the model optimally fits the Friendship Network Data. The gray colored box plots represent the spread of the data simulated from the model and the red line describes actual network data statistics.

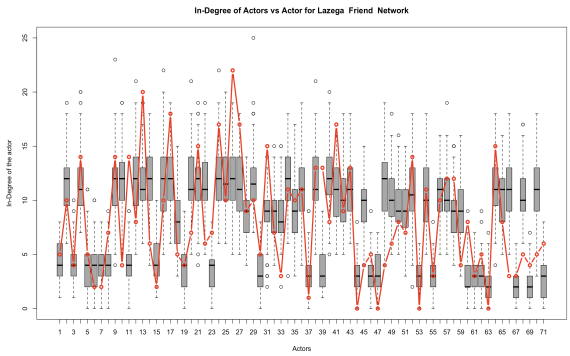
Figure 5.8(a) shows Geodesic comparison with simulated data boxeplots representing the actual data trends and following the red line pattern; though it can be seen at Geodesic Distance "1" the model underestimates the distance and overestimates at Geodesic Distance "3"; rest all the points lie inside the box representing a good fit.
Figure 5.8(b) represents the comparison of Triangle formations in the network between the simulated networks and actual network. The triangle formation statistics of the simulated network does not exactly map the actual data(red line) trends with only some points of the line falling into the box plots. The model is not able to predict the higher order triangle formation trend as the triangle count predicted by the model simulations after "28" degree are all zero.



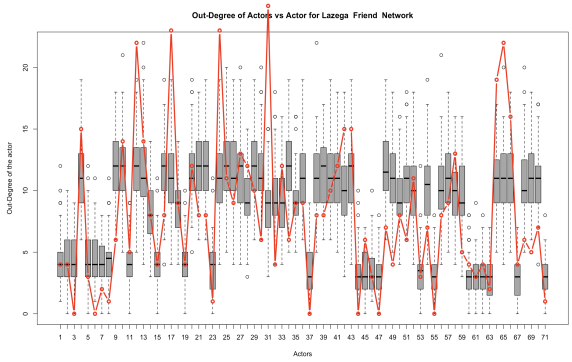
(a) Geodesic Distance Comparison



(b) Number of Triangles Comparison



(c) Degree(All) Comparison



(d) Betweeness Centrality Comparison

Figure 5.8: Goodness of Fit for Lazega Friendship Network Dataset

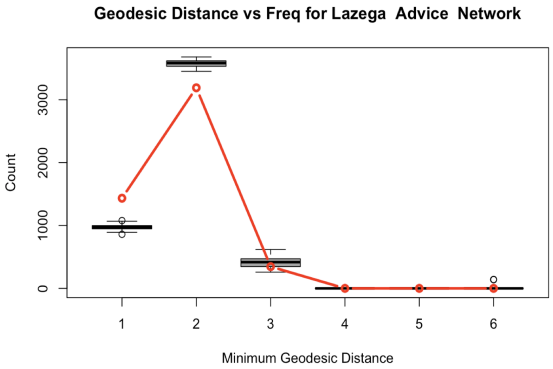
Figure 5.8(c) shows the in-degree comparisons, with the simulation approximately mapping the actual data trend. But the model underestimates and overestimates at various in-degree levels with only few points fitting between the upper and lower boxplot quartile.

Figure 5.8(d) describes the out-degree plot comparison with simulated model network poorly fitting and underestimating higher out-degree links of actual network and some points for the actors that have low out-degree links are between the upper-lower quartile of the boxplot. This means that the model is predicting low out-degree links correctly as compared to higher out-degree links.

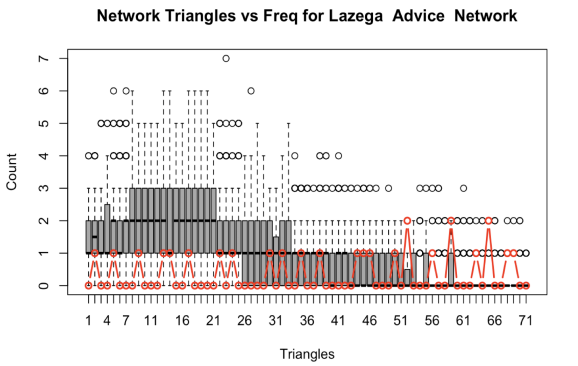
Thus, the simulated networks from the model is able to map and follow some of the network summary statistics of the actual data but it can be further optimized to include other network attributes and improve the performance.

Lawyer Advice Network

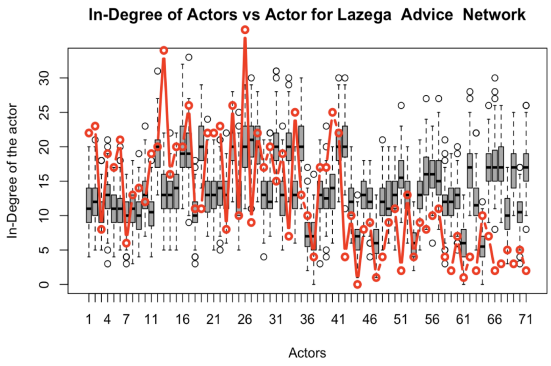
Figure 5.9 shows the comparison of actual network data and simulated network data from the model fitted on **Lazega Advice Network Dataset**. Figure 5.9(a) describes a will fitted Geodesic Distance plot with actual data point residing inside the boxplot at Geodesic distance "3". Simulated and actual network have zero count corresponding Geodesic of "4" and higher. Figure 5.9(b) depicts that our network is not optimally fitting the triangle count plot as the actual data triangle count is very low as compared to simulated boxplot which overestimates the count. The in-degree(Figure 5.9 c) and out-degree(Figure 5.9 d) plot shows an average fit of actual data with some points lying inside the boxes and most of the points lying outside the boxplot.



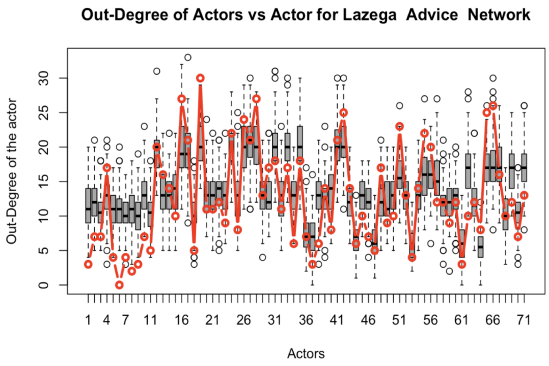
(a) Geodesic Distance Comparison



(b) Number of Triangles Comparison



(c) Degree(All) Comparison

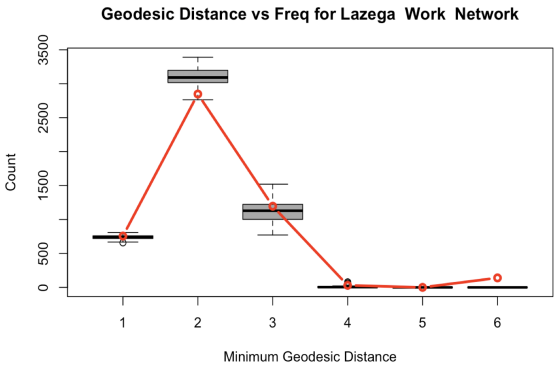


(d) Betweeness Centrality Comparison

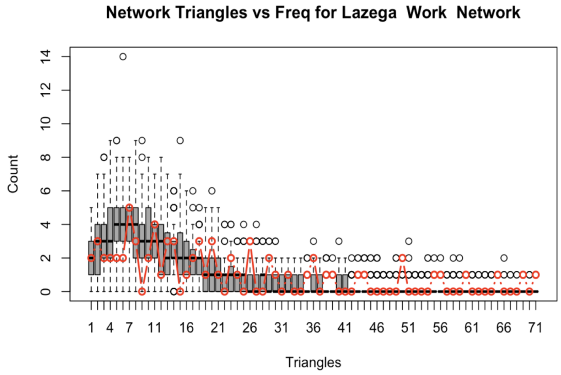
Figure 5.9: Goodness of Fit for Lazega Advice Network Dataset

Lawyer Work Network

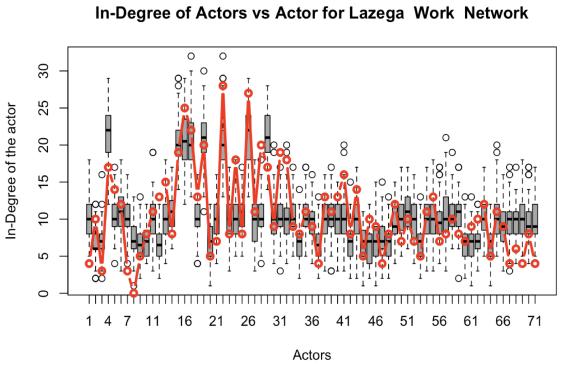
Figure 5.10 shows the Goodness of Fit for the Lazega Lawyer Work Network. Figure 5.10(a) shows that the model accurately simulates the Geodesic distance of the network and follows the trend of the actual data for the Geodesic Distance "1", "3", "4" and "5"; whereas it overestimates and underestimates for Geodesic Distance "2" and "6" respectively. Triangle Plot of simulated graphs fits the data very well with majority of points lying either near or inside the boxplot. Lazega Lawyer work network is a symmetrical network i.e. it has undirected edges(if $Actor_i$ works with $Actor_j$, it implies $Actor_j$ also works with $Actor_i$), so the out-degree and in-degree plots for the Work network will be the same. Both the plots optimally fit the network.



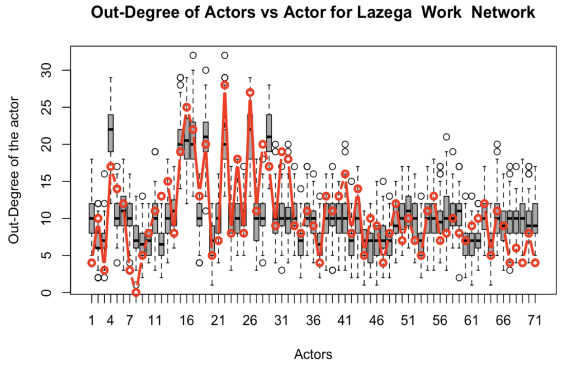
(a) Geodesic Distance Comparison



(b) Number of Triangles Comparison



(c) Degree(All) Comparison



(d) Betweeness Centrality Comparison

Figure 5.10: Goodness of Fit for Lazega Work Network Dataset

Chapter 6

Multiple Links Analysis and Comparison

This chapter presents the comparison approach for Multiple-Relationship Network(Friendship - Advice - Work) by comparing the clusters identified by fitting Stochastic Block Model separately on each network with different links. The comparison study is performed by assuming 4 number of clusters in each relationship graph. Even though all the networks have fixed number of clusters, still clusters need to be processed for comparison as clusters can be misplaced or dislocated from their position i.e. in one network first cluster can be the third cluster in other network. Thus, directly superimposing and comparing the clusters from the link networks would give wrong and misguided results. So, we need to be sure that the cluster indices in one network should approximately match the cluster indices of the other network.

The R package **e1071**[Meyer et al. 2017] is used to match clusters across the networks. The function **matchClasses** of this package tries different permutations and combinations to find similarity in a two way contingency table, mapping clusters from two different networks. This function iterates through all the mappings in the two groups to find the maximum possible matched pairs. Two way Contingency Table is a cross validation classification table that shows the count for each possible combination factor level. In the case of cluster analysis of two networks, it would be a 4x4 matrix assuming 4 number of groups, depicting maximum number of common actors matching in clusters of the two networks.

$$
ContingencyTable = \begin{array}{c} \\ W_1 \\ W_2 \\ W_3 \\ W_4 \end{array} \begin{array}{c} \begin{array}{cccc} F_1 & F_2 & F_3 & F_4 \end{array} \\ \left(\begin{array}{cccc} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{11} & c_{12} & c_{13} & c_{14} \\ c_{11} & c_{12} & c_{13} & c_{14} \\ c_{11} & c_{12} & c_{13} & c_{14} \end{array} \right) \end{array}
$$

Columns in the above matrix represent four cluster in Friends network and rows represent four clusters from the Work Network. Element c_{11} represents the maximum number of common actors in cluster 1 for both the networks, similarly c_{12} represents the common actors in cluster 1 of Work Network and cluster 2 of Friends Network. All the elements in above matrix represents total number of actors common in all possible cluster pair combinations from the two network. The sum of the diagonal elements of the matrix needs to be maximized to ensure that the clusters of two networks have maximum possible number of matching actors in corresponding clusters. **MatchClasses** function of *e1071* package in R tries maximize the diagonal elements of the matrix.

The cluster indices of Friendship Network are kept as the basis for matching cluster indices of Work and Advice Network.

		Z friend			
		Group 4	**Group 3**	**Group 2**	**Group 1**
Z Advice	**Group 1**	4	1	2	9
	Group 2	1	13	1	5
	Group 3	1	5	5	1
	Group 4	4	6	3	10

Table 6.1: Matched Clusters Cross Tabulation for Friend and Advice Lawyer Network

		Z friend			
		Group 2	**Group 4**	**Group 1**	**Group 3**
Z Work	**Group 1**	8	0	10	7
	Group 2	0	10	3	4
	Group 3	0	0	6	2
	Group 4	3	0	6	12

Table 6.2: Matched Clusters Cross Tabulation for Friend and Work Lawyer Network

Table 6.1 and 6.2 show the clusters matched and reordered to accommodate maximum matching of groups across the networks. The green colored marked cells are the diagnal elements with maximum sum that is achieved by computing **MatchClasses** function. The group indices are reordered to match the results obtained in above tables and subjected to further comparison.

6.1 Adjusted Rand Index Calculation

Similarity of Clusters between the networks can evaluated using a statistical measure called **Rand Index**. It is calculated as the number of node pairs that belong to same group or different groups in both the networks. The value of Rand Index is between 0(no similarity) and 1(perfect similarity).

$$Rand\,Index = \frac{Index - Expected\,Index}{Max\,Index - Expected\,Index}$$

Adjusted Rand Index is the corrected for chance version of Rand Index. Rand Index may give higher result in case of some random datasets and the value may not be constant all the time. This problem is resolved by using Adjusted Rand Index.

Adjusted Rand Index		
Friend-Work	**Friend-Advice**	**Work-Advice**
11.6%	8.7%	6.5%

Table 6.3: Adjusted Rand Index for Network Cluster Comparison

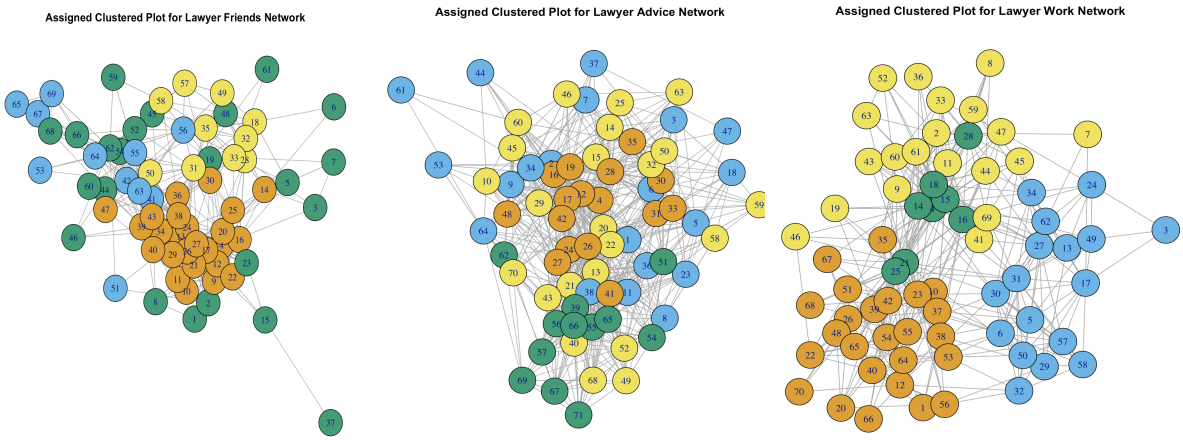
Table 6.3 shows that the clusters of Lawyer Friends Network and Lawyer Work Network have the highest percentage of Adjusted Rand Index than other network pairs. Thus the clusters of Work-Friend Network are more similar to each other. This can indicate that the lawyers in the firm tend to be friendly with the people they work with. Low value of Adjusted Rand Index in case of cluster matching for Work-Advice network may indicate that lawyers don't tend to take advice from the people they work with; this can contradict a popular notion "People usually take advice within their work-group". The cluster mismatch in Work-Advice network may also convey that the Work Network clusters contain lawyers belonging to only one particular seniority level.

For example: a group or cluster of all Junior level lawyer working with each other may take advice from other group containing all senior lawyers.

6.2 Comparison Plots for clusters of Friend-Advice-Work Network

This section will draw a comparison in terms of visualizations and plots between the clusters identified by fitting Stochastic Block Model separately for Friendship, Advice and Work networks. Network statistics, clusters and parameters evaluated from Stochastic Block Model would be plotted with the actors attributes(1.5.2) to draw a proper comparison. Characteristics of Lawyers in the network like seniority, status, age, practice, gender and others are grouped together with the respective cluster partitions to get an insight of their divisions and proportions within each cluster.

6.2.1 Network Cluster Visualization and SBM parameters Comparison



(a) Friendship Network (b) Advice Network (c) Work Network

Figure 6.1: Network Visualizations for Lawyers Lazaga Dataset(Friendship, Advice, Work) color encode by cluster

Figure 6.1 plot shows that the Work Network clusters are widely separated and distinguishable, whereas the clusters in other two networks are dispersed together.

$[\ Number\ of\ Edges: \ 575\ ]$	$[\ Number\ of\ Edges: \ 892\ ]$	$[\ Number\ of\ Edges: \ 756\ ]$
$friend$	$advice$	$work$

Table 6.4: Number of Edges for the Friends, Advice, Work networks

Table 6.4 shows that the Advice relationship type network is a more dense network as compared to other two networks as it has the highest number of edges with same number of actors.

$[\ 0.326 \quad 0.175 \quad 0.354 \quad 0.146\ ]$	$[\ 0.338 \quad 0.167 \quad 0.279 \quad 0.216\ ]$	$[\ 0.244 \quad 0.293 \quad 0.349 \quad 0.114\ ]$
α_{friend}	α_{advice}	α_{work}

Table 6.5: SBM parameter α values for the Friends, Advice, Work networks

The group membership probability parameter α is presented in Table 6.5. In the friends and advice network, $Group_1$ has the highest probability indicating that maximum number of actors lie in $Group_1$ cluster in both the networks. α_{friend} shows that two big groups($Group_1$ - 0.326, $Group_3$ - 0.354) and two small groups($Group_2$ - 0.175, $Group_4$ - 0.146) of approximately of same size underly the data; whereas in α_{advice} and α_{work} networks, there are one large group, two mid sized groups and one small group.

$\begin{bmatrix} 0.371 & 0.058 & 0.075 & 0.027 \\ 0.094 & 0.462 & 0.093 & 0.086 \\ 0.041 & 0.053 & 0.032 & 0.021 \\ 0.095 & 0.094 & 0.048 & 0.548 \end{bmatrix}$	$\begin{bmatrix} 0.155 & 0.037 & 0.136 & 0.263 \\ 0.270 & 0.424 & 0.121 & 0.221 \\ 0.066 & 0.001 & 0.097 & 0.095 \\ 0.328 & 0.041 & 0.305 & 0.453 \end{bmatrix}$	$\begin{bmatrix} 0.475 & 0.034 & 0.043 & 0.131 \\ 0.034 & 0.104 & 0.031 & 0.415 \\ 0.043 & 0.031 & 0.264 & 0.235 \\ 0.131 & 0.415 & 0.235 & 0.624 \end{bmatrix}$
π_{friend}	π_{advice}	π_{work}

Table 6.6: SBM parameter π values for the Friends, Advice, Work networks

Table 6.6 depicts the value of parameter π (Group Interaction Matrix) computed from the Stochastic Block Model for all the network types. Cluster 3 in π_{friend} and π_{advice} (3^{rd} row in both the matrices) have very low probabilities of interaction within and outside the groups indicating the actors beloging to cluster 3 neither exhibit a friendly behaviour nor they take advice from other lawyers within or outside their clusters. High probability values across the diagonal in π_{friend} and π_{advice} indicates that all the clusters except cluster 3 are friends with other actors within the same group and also seek advice from the members belonging to their own group. The group interaction

matrix of Work Link Network(π_{work}) is symmetrical as seen in the Table 6.6 where $\pi_{ij}^{work} = \pi_{ji}^{work}$. The links are undirected for the work network.

π	Friend	Advice	Work
Group 1	has friends within the group and less friends outside the group	Take advice form Group 3 and within the group	Work within the group
Group 2	has friends within the group and less friends outside the group	Take advice within the group and from Group 1	Work majorly with group 4
Group 3	no friendship within and outside the group	No advice taken within and outside the group	Work within the group and with the memebers of Group 4
Group 4	has friends within the group and less friends outside the group	Take major advice within the group and from Group 1 and 3	Work Majorly with the people in the same group and also with Group 2 members

Table 6.7: Group Interactions within and outside the group for the Lawyer Dataset

Thus, members of **Group 1** work, take advice, be friends with the member of their own group. **Group 2** members make friends and take advice within their own group but majorly work with Group 4 members. **Group 3** is an isolated group with fewer links between the members in terms of friendship and advice network. They work within the same group or with the members of Group 4. Similar to Group 1, members of **Group 4** have dense connections within the group for advice, work and friendship.

6.2.2 Clusters comparison with Actor Attributes

Comparing Cluster with respect to Seniority

Figure 6.2(given below) shows the box plots for all the relationship types in the Lawyer Network Dataset against the Seniority of a lawyer(1 being the highest and 71 being the lowest) on Y-axis and the assigned cluster on the X-axis. The plot for Advice and Friend network shows that members of Cluster 2 are younger and have have low seniority level in the firm because their box plot is concentrated towards low seniority range of the graph. Cluster 3 in all the relationship networks has wide spread of members across all seniority levels. Cluster 1 and Cluster 4 group members usually

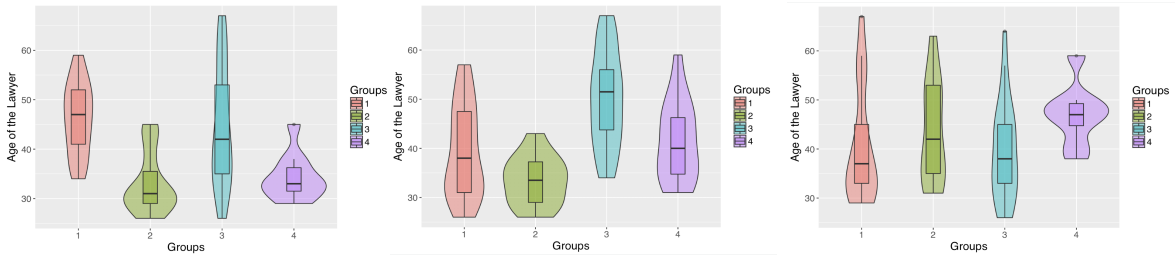
have high seniority level across all the link types Referring to Group Interaction Table 6.6, it can be inferred that the low seniority members of Cluster 2 seeks advice and works with high Seniority Clusters(Group 1 and Group 4).

(a) Friendship Network (b) Advice Network (c) Work Network

Figure 6.2: Box Plot Seniority vs Clusters for Friends, Advice, Work Network

Comparing Cluster with respect to Age

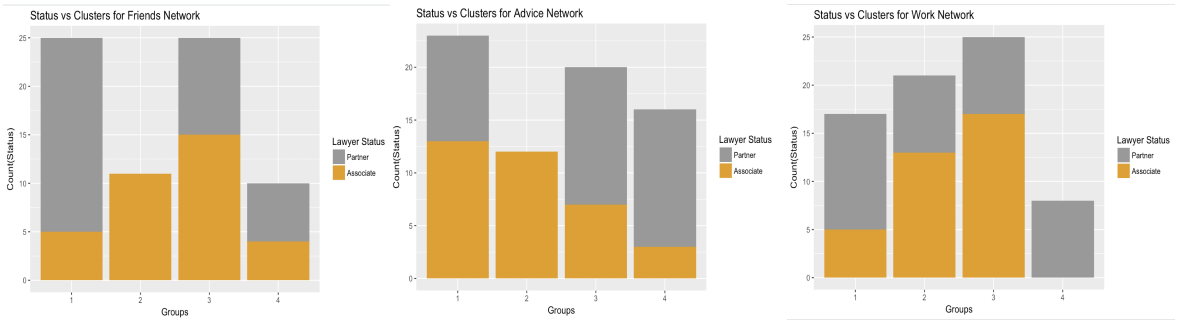
Figure 6.3 shows the Violin Plot overlapping the box plots of Age versus the assigned Cluster Index. Comparing Figure 6.2 and Figure 6.3, low seniority of Cluster 2 can be accounted to their young age group(Green Box Plot in Friends and Advice network). Similar to the Seniority Box Plot, Cluster 3 in Age box plot also has a wide spread of actors across several age groups i.e. older and younger actors are part of CLuster 3. For the Advice network, it can be seen from the box plot(Figure 6.3 b) and π_{advice} matix in Table 6.6 that in general, members of Cluster 3 are elder, and they usually don't take advice from other groups but other group members consult Cluster 3 for advice. Figure 6.3 (c) shows that most of the actors in Cluster 4 are above 45 years of age and have high Seniority Levels.



(a) A subfigure (b) A subfigure (c) A subfigure

Figure 6.3: Violin Plot Comparison Age vs Clusters for Friends, Advice, Work Network

Comparing Cluster with respect to Status



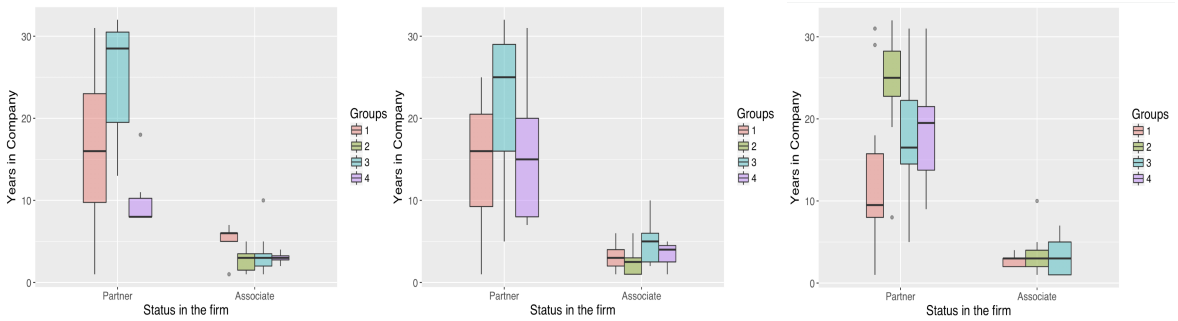
(a) A subfigure (b) A subfigure (c) A subfigure

Figure 6.4: Bar Plot Comparison Status vs Assigned Clusters for Friends, Advice, Work Network

Figure 6.4 compares the division of Partners and Associate across the clusters of different link networks. Members of Group 2 cluster are young and junior members who are all associates as implied by the full yellow bar for Group 2 in Friendship and Advice network; and 70:30 ratio of Associate to Partner in Work Network. Bar plot for Group 1 and Group 4 in Friendship, Work and Advice network indicate that most of the actors in these groups are Partners, which aligns and validates their high Seniority Status. Referring back π matrtix in Table 6.6, Associates Group 2 look for advice from senior Partner members in Group 1 and Group 4 and they usually work with Partners in Group 4.

Comparing Cluster with respect to Status and Years

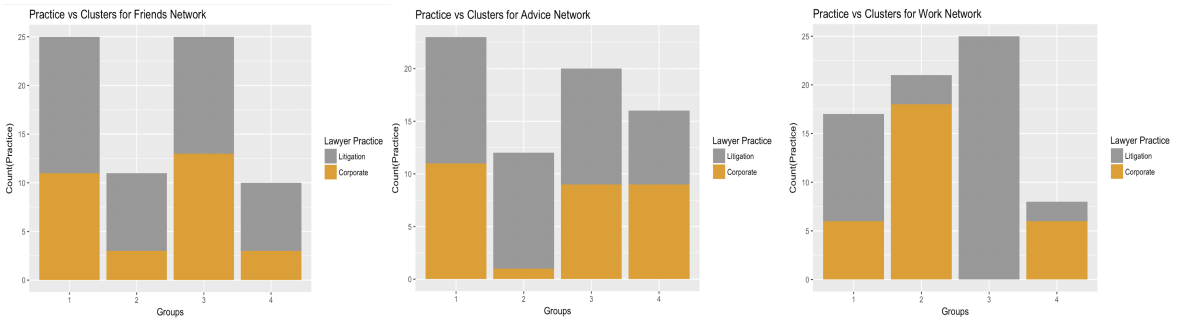
Figure 6.5 represents the analysis of cluster division with respect to Status and Year attribute of the dataset. In all the three sub-figures it is a clear distinction between the number of years, a partner has worked in the firm and the number of years, an associate has worked in the company. Huge margin in their plots indicate that Associates are the new joinees and Partner lawyers have been the part of firm from a very long time. All the box plots for Associate Status lawyers are small and concise indicating good agreement with respect to number of years in the company; whereas for Partner lawyers, the box plots are spread across a huge span of Years indicating that the lawyer with Partner Status have spent different amount of time in the firm.



(a) Friendship (b) Advice (c) Work

Figure 6.5: Box Plot Comparison for Status vs Years vs Assigned Clusters for Friends, Advice, Work Network

Comparing Cluster with respect to Practice



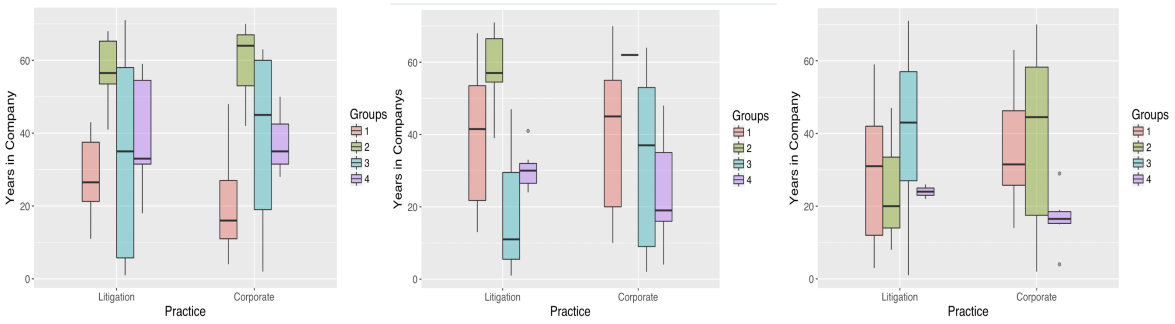
(a) Friendship (b) Advice (c) Work

Figure 6.6: Bar Plot Comparison for Practice vs Assigned Clusters for Friends, Advice, Work Network

Figure 6.6 suggests that clusters from Multiple Links Network does not seem to be similar and does not satisfy each other when plotted against Law Practice(Litigation or Corporate). Cluster 1 and Cluster 3 in Friends and Advice network have approximately equal number of Litigation and Corporate Lawyers. Cluster 3 in work network contains all Litigation lawyers. Figure 6.4 (a) and (b) show that Cluster 2 has only Associate Lawyers, thus comparing it with Figure 6.6 (a) and (b), it can be said most of the Associates in Group 2 practice litigation.

Comparing Cluster with respect to Years and Practice

There was a clear distinction between the Seniority/Years of Partner and Associate lawyers in the plot of Seniority versus Lawyer Status but in case of Practice versus Years plot, there is no such distinction i.e. anyone in the firm whether old or new can practice Litigation and Corporate. In Friends and Advice Network Cluster Litigation Lawyers are high in experience. Cluster 3 box plot is spanning across a range of years for both Litigation and Corporate in all the network types.



(a) Friendship (b) Advice (c) Work

Figure 6.7: Box Plot Comparison for Practice vs Years vs Assigned Clusters for Friends, Advice, Work Network

Chapter 7

Further Extension - Mixed Membership Stochastic Block Model

The main limitation of Stochastic Block Model is that it limits the actor to belong to only one group whereas in real life networks, actors may belong to multiple groups at once. The Mixed Membership Stochastic Block Model(MMSBM) is an extension to Stochastic Block Model, providing further flexibility and allowing actors to become part of multiple groups. Each actor in the model is associated with Membership Probability Vector.

The MMSBM is described and implemented as defined by [Airoldi et al., 2008]:

The model considers a graph G = N, Y where N is the total number of actors and Y is a binary valued adjacency matrix. The graph can be directed or undirected. The network is assumed to be divided in K factions or latent groups. Each node in the group has a vector $\vec{\pi}_i$ where, $\vec{\pi}_{ig}$ denotes the probability with which $actor_i$ belongs to group g. Thus, each actor in the network can belong to multiple groups with different propensity. Group Membership Indicator for each actor $\vec{z}_{i\to j}$ denotes the membership link from $actor_i$ to $actor_j$ and $\vec{z}_{j\to i}$ denote the membership link from $actor_j$ to $actor_i$.

Data Generative Process for MMSBM:

- for i in 1 to N {
  $\vec{\pi}_i \sim Dirichlet(\vec{\alpha})$
  }
- for i,j in 1 to N {
  $Membership\,Indicator\,(Sender)\,\vec{Z}_{i\to j} \sim Multinomial(\vec{\pi}_i)$
  $Membership\,Indicator\,(Receiver)\,\vec{Z}_{j\to i} \sim Multinomial(\vec{\pi_j})$
  $Y_{ij} \sim Bernoulli(\vec{Z}^T_{i\to j}\,B\,\vec{Z}_{j\to i})$
  }
  where symbol B parameterizes the distribution

The group membership for each node is context dependent i.e. the group membership will change if neighbors of the actor are changed. Thus, the group membership is highly influenced by the actors surrounding the node.

Paramter Estimation for Variational EM MMSBM:

- **E-Step:** Expected value Group Membersip Indicator $\vec{Z}_{i\to j}$ and $\vec{Z}_{j\to i}$

$$\hat{\phi}_{i\to j,g} \propto e^{\mathbb{E}_{[log\,\pi_{ig}]}} \prod (B_{gh}^{Y_{ij}}(1-b_{gh}^{1-Y_{ij}}))^{\phi_{j\to i,h}}$$

$$\hat{\phi}_{j\to i,h} \propto e^{\mathbb{E}_{[log\,\pi_{jh}]}} \prod (B_{gh}^{Y_{ij}}(1-b_{gh}^{1-Y_{ij}}))^{\phi_{i\to j,g}}$$

- **M-Step:** Estimated value of B(Distribution Parameter) and ρ(Sparsity Parameter)

$$\hat{B}_{gh} = \frac{\sum_{ij} Y_{ij}\phi_{i\to j,g}\phi_{j\to i,h}}{(1-\rho)\sum_{ij}\phi_{i\to j,g}\phi_{j\to i,h}}$$

$$\hat{\rho} = \frac{\sum_{ij}(1-Y_{ij})\sum_{gh}\phi_{i\to j,g}\phi_{j\to i,h}}{\sum_{ij}\sum_{gh}\phi_{i\to j,g}\phi_{j\to i,h}}$$

Chapter 8

Conclusion

This dissertation presents and investigates the network having Multiple Relationships among the same set of actors using Stochastic Block Model. It describes the motivation for this study and the importance of network analysis and community detection followed by a brief comparison between Assortative and Disassortative Mixing. Basic concepts about Graph Theory and properties of networks are discussed. Exploratory Data Analysis on the Karate Dataset revealed that the faction leader and nodes directly linked to them have higher score degree, closeness and betweenness. Karate Dataset is a disassortative network with popular faction leaders interacting with less popular members. Work network has the lowest assortativity score than Friendship and Advice, whereas the average path length of the Friendship networks is the maximum.

Theoretical concepts of Stochastic Block Model are explained with model specifications and data generative process. For this project Variational EM Algorithm SBM is implemented and two datasets (Karate, Lawyer Dataset) have been fitted into the model. The Lazega Lawyer Dataset, being a multiple link dataset, has been fitted by considering each relationship as a separate network. Integrated Complete Data Likelihood is used to estimate the optimal number of groups and the number of groups with highest value of ICL is chosen.

The model fitted for Karate dataset reached its maximum ICL value at 4 number of clusters which divides the network into 2 big and 2 small groups. The faction leader Mr. A and Mr. H belong to the small sized group. Simulated graphs from the model correctly map the degree and triangle count but fail to map Geodesic property of the

original network.

The resultant optimal number of groups computed by evaluating ICL for the Friendship Network, is used as the number of groups for the other two networks. The clusters from fitting all the three networks separately are reordered, aligned and matched for comparison. Goodness of fit evaluation for the networks reveal that the model is not able to capture all the key aspects of the generative process as some of the plots wrongly estimate the networks summary statistics.

Comparison study between the clusters reveal that a connection exists between these networks of varying links. First cluster in any of the networks, usually interact with the members within the group, whether the link is for friendship, advice or work. Group 1 and Group 4 contain high number of senior members of the firm. There are also some cases where the cluster mappings across the links does not make any sense and no information can be extracted out.

8.1 Future Work

For the future scope for this dissertation, it may be interesting to compare and contrast, approach of fitting the multiple links separately in a model versus using a Multiplex Stochastic Block Model[Barbillon et al. 2015] to fit all the links in one model. It can be used to analyze the differences in clusters and their interactions. To include actor attributes and to allow members to be part of multiple groups, Multiplex Mixed Membership of Experts Model[White and Murphy 2016] can be implemented to see the influence of multiple links with actor attributes. For the purpose of this dissertation, two other datasets(Colorado Springs Project 90[Morris and Rothenberg 2011], Yelp Dataset[17]) were prepossessed to be used as input to the model, but due to time constraints we were limited to use only Karate and Lawyer Dataset. It may be an interesting research to identify cluster and communities in HIV Transmission Network Metastudy Project(Colorado Springs Project 90) combined with actor attributes. Moreover, the computational processing and performance of SBM can be improved and enhanced for faster processing by using scaling algorithms and parallel processing[Bianfang et al. 2014].

Bibliography

[1] Dapeng Hao and Chuanxing Li. The dichotomy in degree correlation of biological networks. *PLOS ONE*, 6(12):1–13, 12 2011. doi: 10.1371/journal.pone.0028322. URL https://doi.org/10.1371/journal.pone.0028322.

[2] M. Salter-Townshend, A. White, I. Gollini, and T. B. Murphy. Review of statistical network analysis: models, algorithms, and software. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(4):243–264, 2012. doi: 10.1002/sam.11146. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11146.

[3] W.W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

[4] Gabor Csardi. *igraphdata: A Collection of Network Data Sets for the 'igraph' Package*, 2015. URL https://CRAN.R-project.org/package=igraphdata. R package version 1.0.1.

[5] J. J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, June 2008. ISSN 0960-3174. doi: 10.1007/s11222-007-9046-7. URL http://dx.doi.org/10.1007/s11222-007-9046-7.

[6] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *CoRR*, abs/1405.3267, 2014. URL http://arxiv.org/abs/1405.3267.

[7] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, June

2008. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1390681.1442798.

[8] Arthur White and Thomas Brendan Murphy. Mixed-membership of experts stochastic blockmodel. *Network Science*, 4(1):4880, 2016. doi: 10.1017/nws.2015.29.

[9] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Community Detection in Degree-Corrected Block Models. *ArXiv e-prints*, July 2016.

[10] P. Barbillon, S. Donnet, E. Lazega, and A. Bar-Hen. Stochastic Block Models for Multiplex networks: an application to networks of researchers. *ArXiv e-prints*, January 2015.

[11] Tom A.b. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14 (1):75100, Jan 1997. doi: 10.1007/s003579900004.

[12] Maya R. Gupta and Yihua Chen. Theory and use of the em algorithm. *Found. Trends Signal Process.*, 4(3):223–296, March 2011. ISSN 1932-8346. doi: 10.1561/2000000034. URL http://dx.doi.org/10.1561/2000000034.

[13] R Nishii. Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis*, 27(2):392 – 403, 1988. ISSN 0047-259X. doi: https://doi.org/10.1016/0047-259X(88)90137-6. URL http://www.sciencedirect.com/science/article/pii/0047259X88901376.

[14] Jean-Patrick Baudry, Adrian E. Raftery, Gilles Celeux, Kenneth Lo, and Raphael Gottardo. Combining Mixture Components for Clustering. *Journal of Computational and Graphical Statistics*, 19:332–353, 2010. URL https://hal.inria.fr/inria-00321090.

[15] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017. URL https://CRAN.R-project.org/package=e1071. R package version 1.6-8.

[16] Martina Morris and Richard Rothenberg. Hiv transmission network metastudy project: An archive of data from eight network studies, 1988–2001, 2011.

[17] Yelp Dataset, 2013. URL `https://www.yelp.com/dataset_challenge/`.

[18] Chai Bianfang, Jian Yu, Cai-Yan Jia, and Jing-Hong Wang. Fast algorithm on stochastic block model for exploring general communities. 24:2699–2709, 11 2014.