EXPLORING NETWORK MODELS UNDER SAMPLING

by

SHU ZHOU

B.S., Xiamen University, 2007

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2015

Approved by:

Major Professor
Perla Reyes

Copyright

Abstract

Networks are defined as sets of items and their connections. Interconnected items are represented by mathematical abstractions called vertices (or nodes), and the links connecting pairs of vertices are known as edges. Networks are easily seen in everyday life: a network of friends, the Internet, metabolic or citation networks. The increase of available data and the need to analyze network have resulted in the proliferation of models for networks. However, for networks with billions of nodes and edges, computation and inference might not be achieved within a reasonable amount of time or budget. A sampling approach seems a natural choice, but traditional models assume that we can have access to the entire network. Moreover, when data is only available for a sampled sub-network conclusions tend to be extrapolated to the whole network/population without regard to sampling error.

The statistical problem this report addresses is the issue of how to sample a sub-network and then draw conclusions about the whole network. Are some sampling techniques better than others? Are there more efficient ways to estimate parameters of interest? In which way can we measure how effectively my method is reproducing the original network? We explore these questions with a simulation study on Mesa High School students' friendship network. First, to assess the characteristics of the whole network, we applied the traditional exponential random graph model (ERGM) and a stochastic blockmodel to the complete population of 205 students. Then, we drew simple random and stratified samples of 41 students, applied the traditional ERGM and the stochastic blockmodel again, and defined a way to generalized the sample findings to the population friendship network of 205 students. Finally, we used the degree distribution and other network statistics to compare the true friendship network with the projected one.

We achieved the following important results: 1) as expected stratified sampling outperforms simple random sampling when selecting nodes; 2) ERGM without restrictions offers a poor estimate for most of the tested parameters; and 3) the Bayesian stochastic blockmodel estimation using a stratified sample of nodes achieves the best results.

Table of Contents

List of Figures

List of Tables

Dedication

I would like to thank my major professor Dr. Perla Reyes for her help and advice when doing this report and all through my master study in this program. I would not have been able to complete this work without her guidance and help. I really want to thank her for meeting me every week and explaining all those papers, theories and presentation skills to me with great patience. I learnt a lot from her. I would also like to give many thanks to my report committee members, Dr. Christopher Vahl and Dr. Suzanne Dubnicka for their always quick replies to my endless emails.

My mother and father have given me great love and support as usual when writing this report. I would not even have attended K-State without them. I want to thank my husband Bo Tong for his support and encouragement during my research and study. I always remember those days that we spent discussing interesting ideas, learning codes, figuring out equations and graphs and so much more. Finally, I would like to thank my all my dear friends especially Nan An, Ran Zhao, Chuyuan Wang and Zhouzhou He, Huan Wang, Qianli Pan for all the happy and memorable moments we hang out together.

Chapter 1

Introduction

1.1 Motivation

Networks (also known as "graphs") are easily seen in everyday life: a network of friends, the Internet, companies or organizations networks, human's neural networks, metabolic networks, books or papers citation networks. With the growing popularity of social networks many social media websites, such as Facebook, Linkedin or Twitter, have gained thousands of millions of users. The topology of these social networks helps users promote connections and exchange of information. It has also become an attractive way to reach target populations, for uses as diverse as the people engaged in social media, from marketing companies trying to sell their products to government and health officials looking to estimate the size of populations at risk (Handcock [1]) or to understand how information sharing in online communities may affect health behaviors (Balatsoukas [2]).

We will define Network as a set of items and their connections. The interconnected items are represented by mathematical abstractions called vertices (or nodes), and the links that connect pairs of vertices are called edges. On one hand, if edges point in a certain direction such as the relationships in a prey-predator network, it is called directed network. On the other hand, when all the edges are bidirectional or when the direction is of no interest, it is called undirected network. Moreover, one network may contain both directed

and undirected edges. Figure 1.1 shows the two types of edges: on the right edges run in only one direction, while in the left edges do not have a directional property.

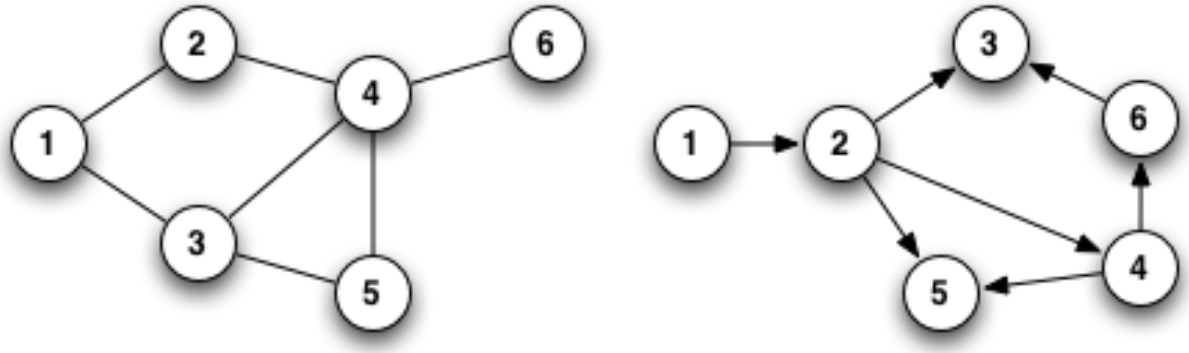

Figure 1.1: *Undirected Network (left) and Directed Network (right).*

Social networks share some common node and edge properties such as degree distribution, mean shortest distance (between vertex pairs, clustering coefficients, and community structure, which will be defined and explained in detail in Chapter 2.2. By studying those properties, we can investigate the correlation between nodes and structures, estimate node activity, predict future emerging edges or find hidden edges and so on. However, for networks with billions of nodes and edges, computation and inference might not be achieved within a reasonable amount of time and money when dealing with the complete network database. When data is too massive to be processed thoroughly, a sampling approach seems a natural choice. In addition, for many cases as pointed out in Shalizi and Rinaldo [3], data is only available for a sampled sub-network. The increase of available data and the need to analyze it have resulted in the proliferation of models for network data [4, 5, 6]. In terms of sampling methodologies, we can go as far back as Goodman's [7] snowball sampling, which evolved into response driven sampling (RDS). RDS is still the most used way to investigate and draw conclusions about hard to find populations. More recent examples of efforts into sampling networks include Blagus et al. [8] and Rezvanian et al. [9]. Typically, however,

models are developed assuming we have access to the entire network and conclusions based on a sampled sub-network are generalized to the whole network/population. Shalizi and Rinaldo [3] showed that the assumption of consistency under sampling that is required to make such generalizations is violated by many popular models. They discussed how the popular class of exponential random graph models (ERGM) and other similar models require strong assumptions to be able to project sampled data into the population network. One key issue that Shalizi and Rinaldo [3] left unanswered is how to obtain information from a part of the network and then draw conclusions about the whole population. Which sampling mechanism will make a network sample more "representative" of the whole network? After inference has been made using a sample from a network, how can we project it? In other words, how or what conclusions can be made about to the population network? How can we compare the estimated population network parameters with their true values? The main focus of this work is to explore these questions with a simulation study on a high school student friendship network.

In the following section, we will introduce the friendship network that will be used. We will define networks, network statistics of interest and traditional network models, such as ERGM in detail in Chapter 2. Chapter 3 will cover an alternative Bayesian method to model networks that we expect has the potential of satisfy the assumption of consistency under sampling. Finally, Chapter 4 describes sampling techniques and inference approaches that we explored to assess whether network information can be estimated using only a sub-network.

1.2 Faux Mesa High Dataset

For our empirical study, we used an example of a typical social network using the "Faux Mesa High" data set of Resnick et al [10], which is built in the R package called "ERGM" as a network object. It represents a simulation of an in-school friendship network. The school community is in the rural western U.S., and has a student body largely Hispanic and

Native American. The network has 205 vertices (students, in this case) and 203 undirected edges (mutual friendship). Furthermore, for each student three attributes are known: grade, gender and race. Grade has values 7 through 12, indicating each student's grade. Gender has two values: male and female. Race is based on the answers to two questions, one on Hispanic identity and one on race, and takes six possible values: White (non-Hisp), Black (non-Hisp), Hispanic, Asian (non-Hisp), Native American, and Other (non-Hisp). The basic information of the dataset is listed in Table 1.1.

Table 1.1: *Student Distribution by Attributes*

Distribution by Grade

Grade	7	8	9	10	11	12
Student	62	40	42	25	24	12
Percent	30.24	19.51	20.49	12.20	11.71	5.85

Distribution by Gender

Gender	female	male
Student	99	106
Percent	48.29	51.70

Distribution by Race

Race	Black	Hisp	NatAm	Other	White
Studuent	6	109	68	4	18
Percent	2.93	53.17	33.17	1.95	8.78

Graphs of the friendships network showing grade, gender and race are displayed in Figure 1.2.

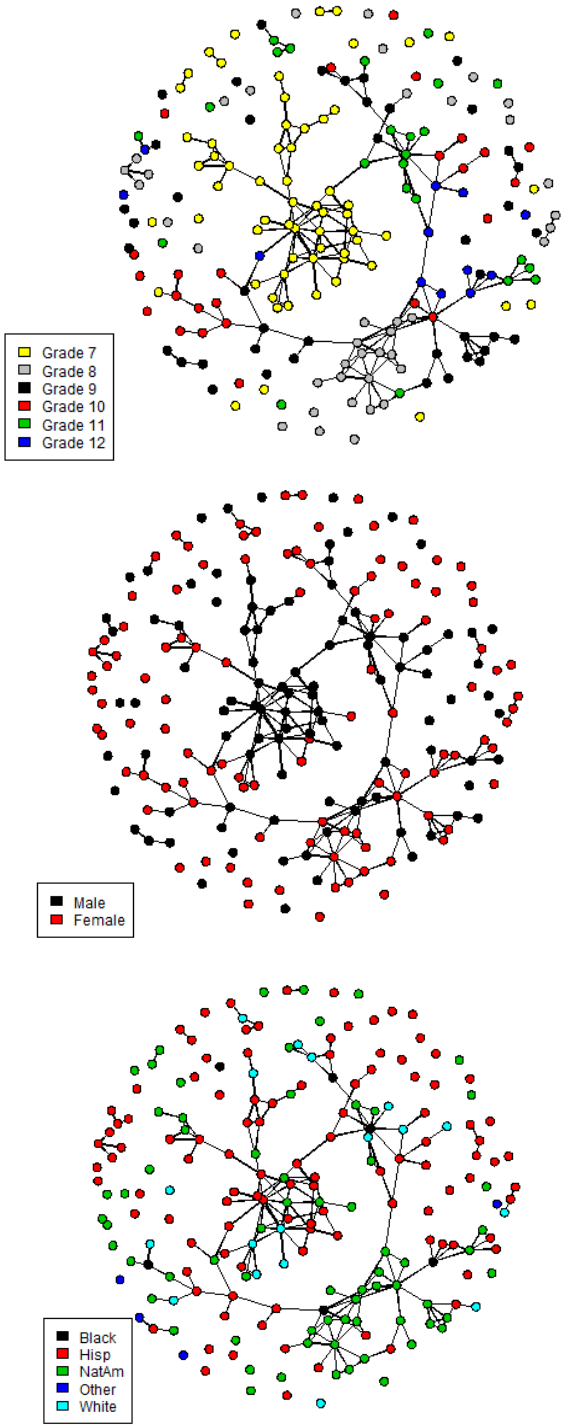

Figure 1.2: *Friendship Network Showing Grade (top), Gender (middle) and Race (bottom).*

Chapter 2

Traditional Network Modeling

2.1 Networks in the Real World

For networks introduced in Section 1.1 such as network of friends, the Internet, company or organization networks, human's neural networks and many others, we can calculate their properties and model their vertices and edges statistically. This chapter reviews some common properties in many of these networks and describe works on mathematical modeling of networks.

Watts and Strogatz [11] studied networks from different fields of application. Their findings on the common properties and mathematical models to simulate those properties are described in their ground-breaking paper. Inspired by their work, we followed the same way of dividing real world networks into four loose types as Newman [12]: social networks, information networks, technological networks, and biological networks.

A social network is a set of people or groups of people who have some patterns of interactions between them, such interactions could be friendship, business relationships, intermarriage between families and so on. An information network is sometimes called a "knowledge networks". For example, the network of citations between academic papers. Citations form a network where articles are vertices, and "article A cites article B" means that we have a directed edge from A to B. One thing worth mentioning is that a citation network is always

acyclic: newer papers can cite older papers, but older papers can never cite those that have yet to be written. A technological network is a man-made network designed for the distribution of commodity or resources, for example, the electric power grid, networks of roads, telephone networks and the Internet. One common feature of technological networks is that their structure is governed by space and geography to some degree. Nodes are connected by edges when they are technologically desirable and geographically feasible. The fourth category, biological network, is any biological systems with sub-units that are linked into a whole big unit. Classic examples of this kind are metabolic pathways, genetic regulatory network, neural network and the food web.

2.2 Networks Properties

Rapoport [13] was one of the first theorists that found the common properties of these networks and modeled them mathematically. He studied the degree distribution in all kinds of networks using random graphs, the simplest model of a network. A random graph is a graph in which properties such as the number of graph vertices, graph edges and connections between them are determined in some random way, for instance, edge probabilities between two vertices can distribute uniformly in the $(0, 1)$ interval. We will give a more detailed mathematical definition of random graph in Section 2.3. In this Section, we will discuss some important network properties defined in Newman [12] that are observed in many of those mentioned in Section 2.1.

2.2.1 Mean Shortest Distance between Vertex Pairs

Consider an undirected network with a fixed n number of vertices, there are $\frac{1}{2}n(n+1)$ possible edges in this network. If we treat the distance from each vertex to itself as zero, then the mean shortest distance between pairs is defined as:

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij} \tag{2.1}$$

where d_{ij} is the shortest distance from vertex i to vertex j, and n is the total number of vertices in the network. In this report, for distance we mean that the geodesic distance between vertex i and vertex j. It is the shortest path (in the number of edges) through the network from one vertex to another.

One problem with the quantity ℓ in equation (2.1) arises when networks have more than one component. In graph theory, a component of a network is a subgraph that can be reached from a vertex by paths running along edges of the graph. Therefore when there is more than one component, we could have vertex pairs with no connecting path. If one assigns infinite shortest distance d_{ij} to such pairs, then the value of ℓ becomes infinite. A way to avoid this kind of problem is to define ℓ to be the "harmonic mean" shortest distance between all pairs, i.e., the reciprocal of the average of the reciprocals:

$$\ell^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1} \tag{2.2}$$

In equation (2.2) infinite values of d_{ij} contributes nothing to the sum, thus nothing to the quantity ℓ.

2.2.2 Clustering Coefficient

Different from the random behavior of a random graph, network clustering is a commonly seen property. In many such networks, it is found that if vertex A is connected to vertex B and vertex B is connected to vertex C, then there is a higher probability that vertex A is also connected to vertex C. For instance, in a social network, it simply means that the friend of your friend is also likely to be your friend. To interpret it mathematically, clustering means to measure the number of triangles - set of three connected vertices forming a triangle. The clustering coefficient C is defined as:

$$C = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}} \tag{2.3}$$

where a "connected triple" means we have one single vertex with edges going to the other two vertices. For example, in the network figure 2.1, this network has one triangle and eight connected 3-vertex-loops, therefore it has a clustering coefficient of $3 \times \frac{1}{8} = \frac{3}{8}$. Similarly the clustering coefficient of a node is the number of triangles that pass through this vertex, relative to the maximum number of 3-vertex-loops that could pass through the node. It is always a number between 0 and 1.

Alternatively, we can define a clustering coefficient for specific vertex i as:

$$C_i = \frac{\text{number of triangles to vertex i}}{\text{number of triples centered on vertex i}} \tag{2.4}$$

If both numerator and denominator are zero, we put $C_i = 0$. Then the average of the clustering coefficients for all vertices in the network is:

$$C = \frac{1}{n} \sum_i C_i \tag{2.5}$$

Normally equation (2.4) is easier to calculate by hand. However, equation (2.5) is easier to obtain for a computer and is widely used in data analysis. Generally speaking, no matter which formula is used, the clustering coefficient measures the density of triangles in a network. Also the coefficient of a real world network tends to be higher than that of a random graph with similar number vertices and edges.

2.2.3 Degree Distribution

A vertex in an undirected network has degree k if the number of edges connected to that vertex is k. The vertex degree distribution of an undirected network gives the number (or fraction in some formulas) of vertices with degree k for $k = 0, 1, ...$ In a directed network, the in-degree of a vertex k is the number of incoming edges and the out-degree is the number

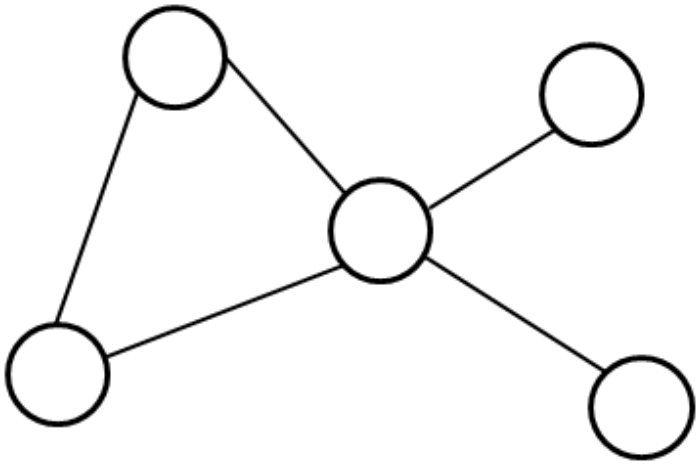

Figure 2.1: *Illustration of the Clustering Coefficient: In this case* $C = 3 \times \frac{1}{8} = \frac{3}{8}$

of outgoing edges. In this report, we will focus on the undirected network case.

Usually p_k is defined to be the fraction of vertices with degree k. It is interpreted as the probability that a randomly chosen vertex has degree k. A plot of p_k for a network can be obtained by drawing a histogram of the vertex degrees. This histogram is the degree distribution of the network.

A popular formula to describe the degree distribution is to use the plot of the cumulative distribution function:

$$P_k = \sum_{k'=k}^{\infty} p_{k'}, \tag{2.6}$$

which is the probability that the degree is greater than or equal to k. For example, in a random graph defined by Erdős and Rényi [14], each edge is either present or not with constant probability 0.5. Therefore the degree distribution of that random graph is binomial, or Poisson in the limit of large graph size. Real world networks are hardly random, hence it is unlikely for us to find its degree distribution strictly following binomial or poisson distributions. They are highly right-skewed most of time. Many degree distributions follow power law in their tails, i.e., $p_k \sim k^{-\alpha}$ for some constant exponent α. Networks with power-law degree distributions has been studied extensively in literature. They are sometimes

referred to as *scale-free networks*. A cumulative degree distribution following power law can be written as:

$$P_k \sim \sum_{k'=k}^{\infty} (k')^{-\alpha} \sim k^{-(\alpha-1)}. \tag{2.7}$$

2.2.4 Community Structure

A network is said to have community structure if the vertices of the network can be easily grouped into sets of vertices such that each set of vertices is densely connected internally. Most social networks show this kind of community structure since we can tend to find groups of vertices (i.e., groups of people) having a higher density of edges within them than between them. Groups of people can be divided based on some common characteristics such as age, gender, company, sorority and so on. Identifying those community structures in a network would provide useful insights into the process driving the network. The traditional method is called cluster analysis or sometimes called hierarchical clustering. This method requires defining a similarity measurement between any two vertices and then grouping similar vertices into communities according to this measurement. In social network literature, the so-called block models are basically divisions of networks into communities or blocks based on some criterion. Two vertices are said to be structurally equivalent if two vertices have the same neighbors. However, exactly the same structural equivalence is hard to find, but approximate equivalence is often used for doing hierarchical clustering.

2.3 Random Graph Models

In this section, we briefly discuss two random graph models: the classic Poisson random graph of Solomonoff and Rapoport [15], and Erdős and Rényi [16]; and the generalized random graph whose degree distribution follows a power law.

2.3.1 Poisson Random Graph

Paul Erdős and Alfréd Rényi's [16, 14, 17] model consists of n vertices joined by edges which are chosen and placed between vertices uniformly at random. They defined the random graph as $G_{n,p}$, in which each possible edge is present with a probability p. The average number of edges on the graph as a whole is $\frac{1}{2}n(n-1)p$, and the average number of ends of edges is twice of this because each edge has two ends. Thus the average degree of a vertex is:

$$z = \frac{n(n-1)p}{n} = (n-1)p \simeq np,$$

where the last approximate equality holds when n is large. If we assume n is fixed, p is proportional to z. The degree distribution of a random graph is also pointed out in the paper of Barabaási and Albert [18]. The probability that a vertex has degree k, p_k is given by the binomial distribution:

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-k-1}. \tag{2.8}$$

In the limit of large n and holding the mean degree $z = p(n-1)$ constant, equation (2.8) becomes:

$$p_k = \frac{z^k e^{-z}}{k!}, \tag{2.9}$$

which is the well-known Poisson approximation to the binomial distribution, thus, the model is also called "Poisson random graph".

The structure of the random graph varies with the value of p. The model shows a *phase transition* as z increases, which means from a low-density, low-p state where all components are small to a high-density, high-p state where a large fraction of all vertices are joint together to form a giant component. There is a critical value of z above which the giant component in the graph contains a finite fraction S of all vertices. The phase transition where the giant component forms occurs at $z = 1$.

2.3.2 Generalized Random Graph

As we have mentioned in Section 2.1, most real world networks or graphs have heavy-tailed degree distributions. We can improve our random graph model by incorporating non-Poisson degree distribution, which leads us to the "configuration model". Degree distribution is still denoted as p_k, and the number of edges k originating from a given vertex i (i.e., k_i) can follow a degree sequence, which does not have to be Poisson distribution. If k exhibits a power law distribution, then:

$$p_k = \begin{cases} 0 & \text{for } k = 0; \\ k^{-\alpha}/\zeta(\alpha) & \text{for } k \geq 0. \end{cases}$$

A power-law network can be constructed by progressively adding vertices to an existing network. The probability of vertex j being connected to a new vertex depends on its own degree k_j. So the higher its degree, the more likely it will get a new connection.

$$p_j \sim \frac{k_j}{\sum_i k_i}$$

This type of networks occur in many areas of sciences, including the internet and the World Wide Web. Only a few web pages have quite a large number of links, whereas most other pages (more than 80%) are only connected with four or less links.

2.3.3 Exponential Random Graph Models

Strauss [19] considers exponential random graph, also called p^* models. Exponential random graph models ($ERGM$) represent a general class of models based on exponential theory for specifying the probability distribution underlying a set of random graphs or networks. Instead of modeling the edges, $ERGM$ treats the whole graph as a random variable $\boldsymbol{Y}$ and defines a probability model for $P(\boldsymbol{Y} = \boldsymbol{y})$, which is defined in equation (2.10). The support of $\boldsymbol{Y}$ is the space with all possible graphs among n vertices. Within this framework, one can obtain maximum-likelihood estimates for the parameters of a specified model for a given

network data; simulate additional networks with the underlying probability distribution implied by that model and perform various types of model comparison.

The basic expression for the $ERGM$ class can be written as:

$$P(\mathbf{Y} = \mathbf{y}) = \frac{\exp(\boldsymbol{\theta}' g(\boldsymbol{y}))}{k(\boldsymbol{\theta})}, \tag{2.10}$$

where $\mathbf{Y}$ is the random variable for the state of the network (with realization $\mathbf{y}$), $g(\boldsymbol{y})$ is the vector of model statistics for network $\mathbf{y}$. $\boldsymbol{\theta}$ is the vector of coefficients for those statistics, and $k(\boldsymbol{\theta})$ represents the quantity in the numerator summed over all possible networks (typically constrained to be all networks with the same node set as $\mathbf{y}$).

This can be re-writen in terms of the log-odds of a single actor pair given the rest:

$$\text{logit}\,(Y_{ij} = 1 | y_{ij}^c) = \theta' \delta(y_{ij}),$$

where Y_{ij} is the random variable for the state of the actor pair i, j (with realization y_{ij}), which means the presence or absence of an edge between vertex i and vertex j. y_{ij}^c denotes the complement of y_{ij}, i.e., all dyads in the network except y_{ij}. That means all the random variables associated with potential pairs in the network except y_{ij}. The variable $\delta(y_{ij})$ equals $g(y_{ij}^+) - g(y_{ij}^-)$, where y_{ij}^+ is defined as y_{ij}^c along with y_{ij} set to 1, and y_{ij}^- is defined as y_{ij}^c along with y_{ij} set to 0.

That is, $\delta(y_{ij})$ equals the value of $g(\boldsymbol{y})$ when $y_{ij} = 1$ minus the value of $g(\boldsymbol{y})$ when $y_{ij} = 0$, but all other dyads are as in $g(\boldsymbol{y})$. This emphasizes the log-odds of an individual tie conditional on all others. We call $g(\boldsymbol{y})$ the statistics of the model, and $\delta(y_{ij})$ the "change statistics" for actor pair y_{ij}.

In this report, we consider the simplest possible model, the Bernoulli or Erdős-Rényi model, which contains only an edge term and therefore is estimated by a essential log-linear model. When covariate information about nodes, also known as attributes, is available a linear function $\boldsymbol{X\beta}$ can be included in $g(\boldsymbol{y})$.

2.3.4 Faux Mesa High ERGM

The ERGM package in R the user to fit exponential-family random graph (ERG) models to network datasets. These models, also known as p*, are described in Section 2.3.3. Let us fit the ERGM of the Mesa High School friendship network, and see how the probability of connection between student i and student j are determined in this model. The fitted model is:

$$logit(Y_{i,j} = 1) = -10.01277 + 3.23105grade + 1.19646race + 0.88438gender \tag{2.11}$$

where grade, race and gender are all categorical variables. All four terms in equation (2.11) are significant. How should we interpret these coefficients? One can interpret the coefficients of this model in terms of the log-odds of different types of links: the log-odds of a link that is completely heterogeneous (the two members differ from each other in race, sex, and grade) is -10.01; the log-odds of a link that is homogeneous by race only is -8.82 ($= -10.01 + 1.20$, with rounding error); the log-odds of a link that is homogeneous in all three attributes is -4.70 ($= -10.01 + 3.23 + 1.20 + 0.88$) and etc. The probability between two students that corresponds to the log-odds is $exp(-10.01)/(1 + exp(10.01) = 0.00044946$ if student i and student j are completely heterogeneous. From the coefficients we can see that grade has a larger influence on the probability of friendship than gender or race.

Chapter 3

Stochastic Blockmodels

As discussed in the previous chapter, a network describes a relational structure on a set of vertices. Each edge in the network describes a relationship between two vertices it connects. The network can be undirected, indicating symmetric relationships between vertices, or it could be directed, which means relationship from vertex i to vertex j does not necessarily imply the same relationship from vertex j to vertex i. Person A states that person B is his or her friend and hence there is a direction to the ties between individuals. It may also be that person B states that person A is his or her friend, but it does not have to be the case.

This chapter will discuss a model when community structure, such as the ones defined in Section 2.2.4, are relevant. In social science, a social network consists of a group of people, variously referred to as vertices or actors, connected by social interactions or ties of some kind. Those relationships between actors are defined by social interactions such as friendship, acquaintance, cooperations, and so on. In this chapter, we consider networks whose ties represent friendship. Friendship networks have been the subject of scientific study since at least the 1930s. A classic example can be found in the studies by Rapoport and collaborators [20] of friendship among schoolchildren in the town of Ann Arbor, MI, in the 1950s and 1960s, in which the authors distribute questionnaires among the students in a school asking them to name their friends. Many similar studies have been done since then, with different levels of complexity, but most employ a similar questionnaire-based

methodology.

In the social network analysis friendship clusters are observed for some reason. This analysis depicting social interactions between people have been studied by Scott [21] and Wasserman and Faust [22]. The patterns of friendship relationships between the actors are often affected by the attributes of those actors, e.g., age, gender, race, income among others. Actors having the same attributes tend to gather together and form friendship, and thus a group. These attributes enhance our interpretation of network structure, and they enable use to study subsections of the network. For example, in the paper of Reddy et al. [23] identifying clusters of customers with similar interests in the network of purchase relationships between customers and products of online retailers (e.g., www.amazon.com) enables the cyber-market to set up efficient recommendation systems that better guide customers through the list of items of the retailer and enhance the business opportunities. Clusters of large graphs can be used to create data structures in order to efficiently store the graph data and to handle navigational queries, like path searches, as studies in the paper by Agrawal and Jagadish [24], and another by Wu and Huberman [25].

Block modeling is a common approach in statistics and social network analysis to decompose a graph in classes of actors with common properties. Actors are usually grouped in classes of equivalence. In terms of equivalence, we mean that pairwise relationships within the group exhibit similar structure, while relationships between different groups exhibit different structures. Revealing those hidden structures of a network is the heart of many data analysis problems. In this report, we will build a hierarchical Bayesian model for an exchangeable network to identify grouping patterns, in specific, a type of stochastic blockmodel. Given a network, the goal in stochastic blockmodels is to divide the vertices such that pairs of vertices are grouped together if their connecting pattern to the other groups in the network is similar.

3.1 Exchangeability

Relationships between two actors are assumed binary: an edge joining a pair of actors either exists or not. $y_{i,j}$ are the relation between actor i and actor j, therefore:

$$y_{i,j} = \begin{cases} 1 & \text{i } and \text{ j } are\ friends \\ 0 & \text{i } and \text{ j } are\ not\ friends \end{cases}$$

The adjacency matrix is a (0,1)-matrix with zeros on its diagonal. If the graph is undirected, the adjacency matrix is symmetric. Based on this definition, we can draw the adjacency matrix for the Mesa High School friendship network:

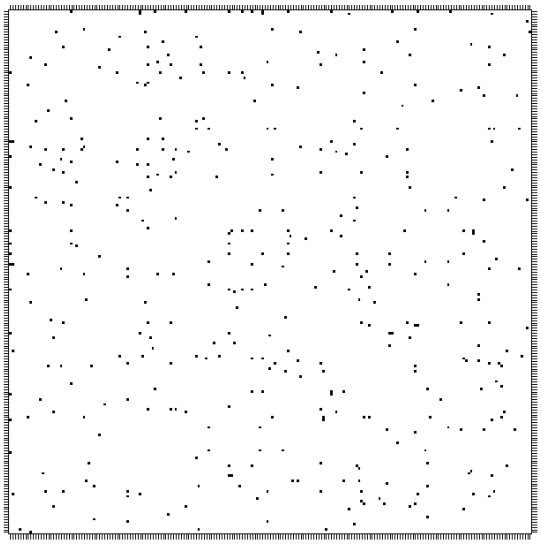

Figure 3.1: *Adjacency Matrix of Mesa High School Friendship Network*

In Figure 3.1 those black dots are those "1"(friendship connections) and those black spaces are those "0"(no friendship connections) in 205 high school students' friendship network.

A stochastic blockmodel is a generative model for blocks, groups, or communities in networks. The group structure and the pattern of the edges between groups are supposed to capture the main features of an empirical network.

Fienberg and Wasserman [26] defined a stochastic blockmodel as a probability distribution (or family of distributions) for networks of which the vertex set is partitioned into subsets called blocks, which have the property that the probability distribution for the network is invariant under permutations of vertices within blocks. The property is called exchangeablity, which means that the vertices in the same block are stochastically equivalent in the following way. Consider there is a block B_i and any vertex j in the network. The likelihood of the pattern of edges with vertex j is the same for all vertices in this block B_i. That is, if i and i' are two actors in B_i. Exchanging $Y_{i,j}$ and $Y_{i',j}$ will not change probability model about $\boldsymbol{Y}$. Before setting up the stochastic blockmodel, we will first discuss exchangeability on two dimensions.

For illustration we will use the following toy example: assume a network of 4 people illustrated by the 4×4 adjacency matrix of 1's and 0's below:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \\ . & & & \end{bmatrix}$$

If one is to switch the position of row 1 and row 3 of it. A exchangeable permutation in this case requires that one has also to switch the position of column 1 and column 3 in

order to preserve the joint probability distribution and the matrix becomes:

$$\begin{bmatrix} a_{33} & a_{32} & a_{31} & a_{34} \\ a_{23} & a_{22} & a_{21} & a_{24} \\ a_{13} & a_{12} & a_{11} & a_{14} \\ a_{43} & a_{42} & a_{41} & a_{44} \\ . & & & \end{bmatrix}$$

3.2 Building the Hierarchical Stochastic Blockmodel

After defining exchangeability in the above section, which is the root of the non-parametric Baysian model discussed in this report, then we can continue to build the hierarchical stochastic blockmodel. Let I be the number of actors in the network and K the number of groups or blocks. Each I actor is assigned to one of the K blocks, groups, or communities. I is assumed to be known when we are given the dataset. Furthermore, relationships between two actors are assumed to be binary, i.e., an edge joining a pair of actors either exists or not. Since a network can be represented as an $I \times I$ adjacency matrix Y, such that $y_{i,j}$ are the relation between actor i and actor j, therefore:

$$y_{i,j} = \begin{cases} 1 & \text{i } and \text{ j } are\ friends \\ 0 & \text{i } and \text{ j } are\ not\ friends \end{cases}, \tag{3.1}$$

K is a latent random variable with a given prior distribution. For given values of I and K, the stochastic blockmodel describes a random process for assigning the actors to groups and then generate the whole network. The latent clustering structure which indicate the groups among network actors is represented by a random vector ξ of length I such that $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_I)'$ and $\xi_i \sim_{iid} multinomial(w_1, w_2 \ldots, w_k)$. Here w_i is the probability of an actor being assigned to cluster i ($\sum_{k=1}^{K} w_k = 1$).

The matrix $\Theta = [\theta_{k,l}]$ is a $K \times K$ matrix, where $\theta_{k,l} \sim_{iid} H^{\lambda}$, a parametric distribution indexed by the hyperparameter λ. $w = \{w_1, w_2, \ldots, w_K\}$ is such that $\sum_{k=1}^{N} w_k = 1$. $w = \{w_1, w_2, \ldots, w_K\}$ and it tells how those I vertices are assigned to the K clusters.

The unique group indicators $\xi_i \in 1, ..., k$ are independently sampled from a Chinese Restaurant Process (CPR) process:

$$\xi_i|\mathbf{w} \sim_{iid} \sum_{k=1}^{N} w_k \delta_k, \qquad w_k = u_k \prod_{s<k}(1-u_s), \qquad u_k \sim_{iid} \mathsf{Beta}(\beta_2 + \eta\alpha). \tag{3.2}$$

Equation (3.2) implies a joint distribution for $\xi = (\xi_{k,1}, ..., \xi_{k,J})$ represented by a predictive distribution with $\xi_{k,1} = 1$ and:

$$\xi_{k,i}|\xi_{k,i-1}, \ldots, \xi_1 \sim \sum_{l=1}^{L^{i-1}} \frac{n_{k,l}^{i-1}}{\eta + i - 1}\delta_l + \frac{\eta}{\eta + i - 1}\delta_{L^{i-1}+1}, \quad 2 \le i \le I, \tag{3.3}$$

where δ_a is the degenerate probability distribution on a, $K^{i-1} = \max_{j<i}\{\xi_j\}$ is the number of unique values among $\xi_1, \ldots, \xi_{i-1}$, $m_k^{i-1} = \sum_{j=1}^{i-1} \mathbf{1}_{(\xi_j = k)}$ is the number of indicators equal to k among $\xi_1, \ldots, \xi_{i-1}$, and $\eta > 0$ is a constant. This sequence of predictive distributions is known as the Chinese restaurant process.

The CRP places a probability distribution on all possible partitions of I actors, whose shape is controlled by the parameter η and implies that $\mathsf{Pr}(\xi_i = \xi_j) = \sum_{k=1}^{\infty} \mathsf{E}\{w_k^2\} = 1/(1+\eta)$ for all i and j.

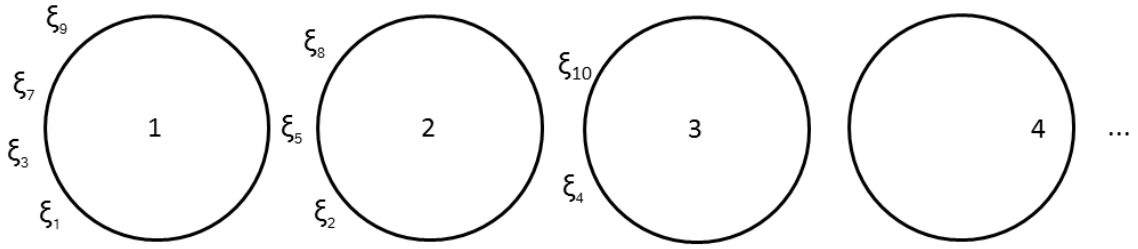

Figure 3.2: *The Chinese restaurant process. Circles denote infinite number of tables and the letters around them are the customers sitting at that table.*

Because the seating arrangement showed in Figure 3.2 can be described using the analogy of sitting customer at a Chinese Restaurant. ξ_i means the table is occupied by customer i. Customer 1 sits at table 1; customer i sits at any of the occupied tales with probability proportional to the number of customers sitting at that table, and sits at a new table with probability proportional to η. In Figure 3.2, customer 6 who is missing would sit at table 1 with probability $4/(\eta+9)$, at table 2 with probability $3/(\eta+9)$, at table 3 with probability $2/(\eta+9)$, and at table 4 with probability $\eta/(\eta+9)$. Any seating arrangement creates a partition.

Based on Bayes' theorem, the hierarchical priors described above, times the observed relationships between any two actors i and j, will determine the posterior probability of connection between pairs of actors. The observed relationships between actor i and j, the $y_{i,j}$ are assumed to be conditionally independent such that:

$$y_{i,j} \sim_{iid} \psi(y_{i,j}|\theta_{\xi_i,\xi_j}), \tag{3.4}$$

where ψ is a parametric distribution associated with the network, $\theta_{k,l}$ is the parameter that controls the rate of interaction among factions k and l in network, and ξ_i is the faction membership indicator for actor i in network.

On a Bayesian framework, the Stochastic Blockmodel produces an estimation of the parameter distributions instead of just a point estimate. We chose a Bernoulli-Beta model for $y_{i,j}$. The joint posterior distribution of all the parameters in the model can be describe by the equation below:

$$p(\boldsymbol{\Theta}, \boldsymbol{\xi}, \boldsymbol{\lambda}, \eta|\boldsymbol{Y}) \propto \prod_{i=1}^{I} \prod_{i'=1, i\neq i'}^{I} \psi(y_{i,j}|\theta_{\xi_i,\xi_j}) p(\boldsymbol{\Theta}|\boldsymbol{\lambda}) p(\boldsymbol{\xi}|\eta) p(\boldsymbol{\lambda}) p(\eta), \tag{3.5}$$

where $\boldsymbol{\lambda} = (a_D, b_D, a_{OD}, b_{OD})$, $\psi(y_{i,j}|\xi_i, \xi_j, \boldsymbol{\Theta})$ is assumed $Bernoulli(\theta_{\xi_i,\xi_j})$; and for the prior $p(\boldsymbol{\Theta}|\boldsymbol{\lambda})$, $\theta_{l,l} \sim_i idbeta(a_D, b_D)$ for diagonal elements, and $\theta_{l,k} \sim beta(a_{OD}, b_{OD})$ for off-diagonal elements. $p(\boldsymbol{\xi}|\boldsymbol{\eta}) \sim CRP$ introduced in equation (3.3). Hyperparameters a_{OD}

and a_D follow a $gamma(\alpha_a, \beta_a)$. b_{OD} and b_D follow a $gamma(\alpha_b, \beta_b)$. Finally, $p(\boldsymbol{\eta}) \sim gamma(a, b)$.

Since the posterior distribution does not have a closed form we used an MCMC sampler to explore the joint posterior distribution from equation (3.5). The MCMC uses a Gibbs sampler to iteratively draw from the following full conditionals:

1. $p(\xi \mid \lambda, \eta, Y)$.
2. $p(\Theta \mid \xi, \lambda, \eta, Y)$
3. $p(\lambda \mid \xi, Y)$.
4. $p(\eta \mid \xi)$.

3.3 Faux Mesa High Posterior Grouping

We applied the algorithm described in Section 3.2 under three different initial settings for ξ_0: 1) 205 students are in 205 different groups, i.e., $\xi_0 = 1, 2, ..., 205$; 2) 205 students are in the same group, i.e., $\xi_0 = 1, 1, ..., 1$; and 3) initial groups are sampled from the Chinese Restaurant Process prior. We ran $10,000$ iterations with 1000 burn-in. Initial values of α_a, β_a, α_b, β_b are all set to be equal to 2 in our code. The marginal likelihood of Y seems to be stable after first 2000 iterations, as shown in Figure 3.3 below.

Figure 3.4 shows the resulting group or block structures for the high school friendship network. There are 10 main communities found by running the MCMC algorithm after we tune parameters. To better see the relational structure of each community, a pivot table with different summary statistics is given in Table 3.1.

Based on Table 3.1, we can see differences by race, Hispanic students are in every group except group 10 and account for 60% of all students group 1. Second largest proportion of student in group 1 are Native American with 28%. All "Other" races only stay in group 1. Four Black students are evenly split in group 1 and 3, the rest two stay in group 4 and group 5 separately. White students are only in group 1, 2, 3, 6 and 8.

While if we look at the grade attribute of the 10 groups, students from grade 8, and 9 are the two largest parts. Group 2 is only consist of students from grade 7. Around 70% of students in group 5 is from grade 8. Six students of grade 9 construct the whole group 8 and 9 students of grade 7 construct the whole group 6. Group 9 has only one student of grade 7 while group 10 has only one student of grade 10.

As for the gender attribute, except for group 9 and 10 which has one student in each group, group 1 has almost twice more males than females and group 2 has more than twice females than males.

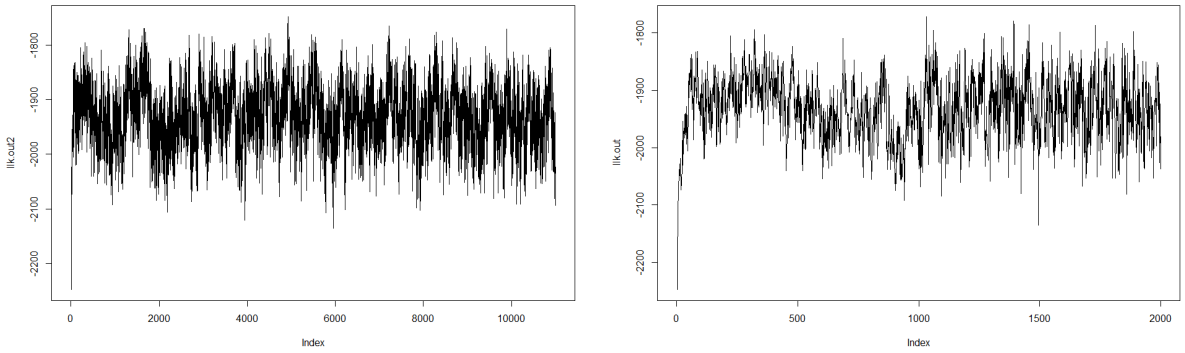

Figure 3.3: *Marginal likelihood of Y (left), Marginal likelihood of Y after thinning (right)*

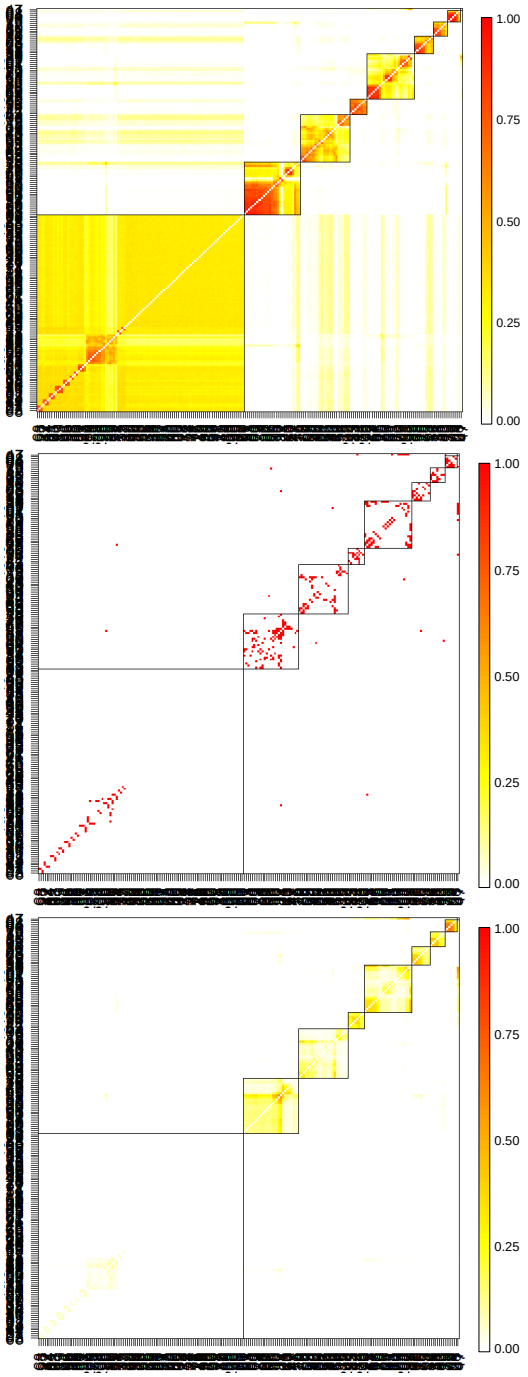

Figure 3.4: *Estimated Posterior Probability.*
Estimated Posterior Probability that Two Actors Belong to the Same Group (*top*); Estimated Posterior Probability of Adjacency Matrix (*middle*); Estimated Posterior Probability of a Connection between Student i and Student j (*bottom*).

Table 3.1: *Posterior Group Distribution: frequencies, row percentages between parenthesis*

	Race					Gender		
Group	Black	Hisp	NatAm	White	Other	Female	Male	Total
1	2 (2.0)	60 (60.0)	28 (28.0)	6 (6.0)	4 (4.0)	38 (38.0)	62 (62.0)	100
2		16 (61.5)	6 (23.1)	4 (15.4)		19 (73.1)	7 (26.9)	26
3	2 (8.3)	12 (50.0)	4 (16.7)	6 (25.0)		14 (58.3)	10 (41.7)	24
4	1 (12.5)	5 (62.5)	2 (25.0)			3 (37.5)	5 (62.5)	8
5	1 (4.4)	4 (17.4)	18 (78.3)			11 (47.8)	12 (52.2)	23
6		6 (66.7)	2 (22.2)	1 (11.1)		7 (77.8)	2 (22.2)	9
7		4 (57.1)	3 (42.9)			3 (42.9)	4 (57.1)	7
8		1 (16.7)	4 (66.7)	1 (16.7)		3 (50.0)	3 (50.0)	6
9		1 (100.0)				1 (100.0)		1
10			1 (100.0)				1 (100.0)	1
Total	6 (2.9)	109 (53.2)	68 (33.2)	18 (8.8)	4	99 (48.3)	106 (51.7)	205

	Grade					
Group	7	8	9	10	11	Total
1	19 (19.0)	23 (23.0)	26 (26.0)	18 (18.0)	10 (10.0)	100
2	26 (100.0)					26
3		1 (4.2)	6 (25.0)	5 (20.8)	9 (37.5)	24
4			1 (12.5)		4 (50.0)	8
5		16 (69.6)	3 (13.0)	1 (4.3)	1 (4.4)	23
6	9 (100.0)					9
7	7 (100.0)					7
8			6 (100.0)			6
9	1 (100.0)					1
10				1 (100.0)		1
Total	62 (30.2)	40 (19.5)	42 (20.5)	25 (12.2)	24 (11.7)	205

Chapter 4

Network Models for Sampled Data

In the previous chapter, we have shown that stochastic blockmodels can find useful grouping information from a network. However, in reality, for large or hard to find population of actors, it might be difficult to get information on all actors or all links between them. For example, in Ribeiro and Towsley [27], networks from Flickr and Youtube were studied having millions of vertices and edges. The large size of these social networks makes it costly querying the entire network, particularly if the goal is to monitor these networks regularly over time. In addition, only few people or organizations have complete access to the data. Without knowing the true underlying structure of a population network, sampling becomes a natural way to solve this issue. A further statistical question in such case emerges: how well the properties of the true network can be modeled from those of the sampled network. In what follows, we will explore some traditional sampling techniques. Furthermore, we provide some evidence on whether and how a sampled network can be used to estimate the true population network and to what extent the degree distribution of the estimated network reflects that of the true network.

Different sampling methods can be applied based on how we can access the network data and what is the goal of sampling. In some cases, the entire network data could be accessed fully then a random edge or vertex can be selected. It could also be accessed restrictively when the network is hidden but allows analyzing (Handcock [1]).

4.1 Simple Random Sampling

Simple random sampling consists of directly selecting a random sample from the total population. A random sample of 41 students was drawn from the population of 205 student in Mesa High School. Assuming that, we asked all 41 students whether they were friends with each other. We applied the algorithm described in Chapter 3 to the 41×41 adjacency matrix and two different initial settings of ξ_0 are introduced: 41 students are in 41 different groups, i.e., $\xi_0 = 1, 2, ..., 41$; 41 students are in the same group, i.e., $\xi_0 = 1, 1, ..., 1$. We ran $10,000$ iterations with 1000 burn-in. The 41×41 posterior mean probability matrix θ is:

$$\begin{bmatrix} 0 & 4.68\text{E}{-3} & \dots & 4.45\text{E}{-3} \\ \vdots & \vdots & \ddots & \vdots \\ 4.45\text{E}{-3} & 4.26\text{E}{-3} & \dots & 0 \end{bmatrix}_{41\times 41}$$

where each element θ_{ij} represent the posterior mean probability that actor i and actor j are friends in this sampled friendship network of 41 students. Diagonal elements are assumed to be 0 because students are not friends of themselves (no self-loop).

Next, this 41×41 matrix is enlarged into a 205×205 matrix Θ, whose element θ_{ij} represents posterior mean probability of actor i and actor j are friends in the 205×205 estimated friendship network of 205 students. The enlargement method of the 41×41 matrix θ is as follows: for each off-diagonal element θ_{ij}, we extend it into a 5×5 matrix with the same values as element θ_{ij} since the new "cloned" students are assumed to have the same probabilities of making friends as the original elements. For each diagonal element θ_{ii}, we extend it into a 5×5, whose diagonal elements are still 0 based on the no self-loop assumption and off-diagonal element values are the same as the highest values in the same row. For example, extending the 41×41 matrix into a 205×205 matrix, applying this rule

gives:

$$\begin{bmatrix}
0 & 4.43\text{E}{-3} & 4.43\text{E}{-3} & 4.43\text{E}{-3} & 4.43\text{E}{-3} & \dots & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} \\
4.43\text{E}{-3} & 0 & 4.43\text{E}{-3} & 4.43\text{E}{-3} & 4.43\text{E}{-3} & \dots & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} \\
4.43\text{E}{-3} & 4.43\text{E}{-3} & 0 & 4.43\text{E}{-3} & 4.43\text{E}{-3} & \dots & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} \\
4.43\text{E}{-3} & 4.43\text{E}{-3} & 4.43\text{E}{-3} & 0 & 4.43\text{E}{-3} & \dots & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} \\
4.43\text{E}{-3} & 4.43\text{E}{-3} & 4.43\text{E}{-3} & 4.43\text{E}{-3} & 0 & \dots & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} & 4.45\text{E}{-3} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\
4.45\text{E}{-3} & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0
\end{bmatrix}_{205\times205}$$

By the binary network assumption in Chapter 3, $y_{i,j} \in 0, 1$. Therefore $y_{i,j} = 1$ if student i is friend with student j, and $y_{i,j} = 0$ otherwise. Thus, $y_{i,j}|\Theta_{205\times205} \sim\ Ber(\theta_{i,j})$, by which we can generate an estimated population network of 205 students. Finally, we compute some of the network properties introduced in Section 2.2 for both estimated network and true network to compare them.

Each 41×41 simple random sample of the true 205×205 true friendship matrix can give an estimated 205×205 friendship matrix for us to compare with the true one. In this report, we draw 100 different simple random samples of 41 students and fit the stochastic blockmodel in Equation (3.5). Each is used to generate 100 realizations of the posterior probability matrix θ, we extend them to Θ and generated an adjacency matrix $\boldsymbol{Y}$. We randomly pick up the $47th$, the $74th$ and the $89th$ 41×41 simple random samples to show the results. The 100 realizations of the posterior degree distribution are shown in the left three columns of Figure 4.1. To better see the posterior degree distribution of each of the three simple random samples, we also draw their posterior mean of degree distributions, the true degree distribution of friendship network and the 95% credible intervals respectively. Plots are shown in the right three columns of Figure 4.1. In order to check the stability of our method, the 95% posterior credible interval of degree distribution of the 100 estimated networks is shown in Figure 4.2.

From both Figure 4.2 and Figure 4.1, we can tell that the 95% posterior credible interval of the estimated population network includes the true degree distribution of the students' friendship network. However, if we look at some randomly chosen samples, some of them

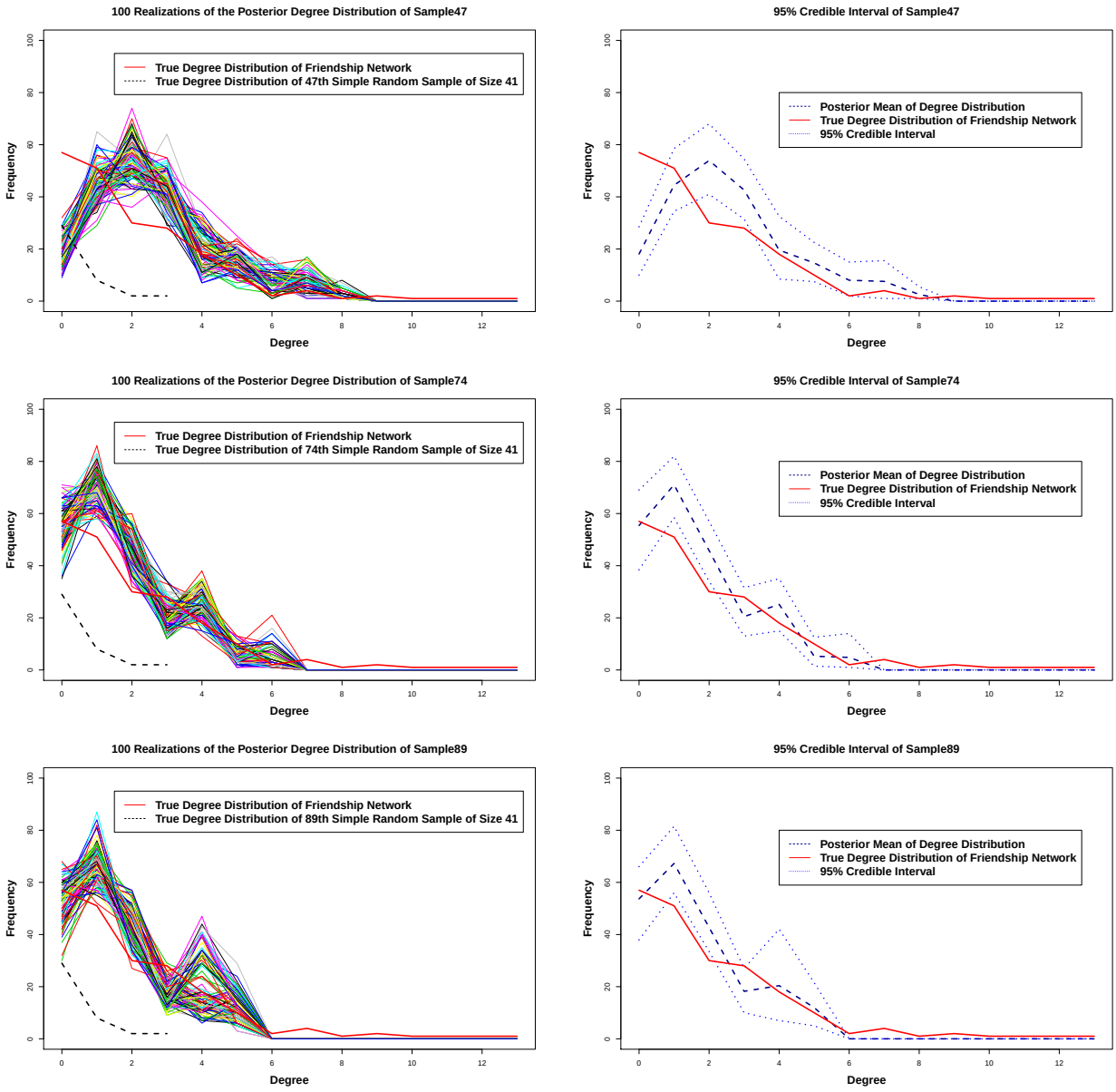

Figure 4.1: *Degree Distribution Inferences for Three Simple Random Samples.*
Left: 100 Realizations of the Posterior Degree Distribution; True Degree Distribution (*red solid line*); True Degree Distribution of 47*th*, 74*th* and 89*th* Simple Random Sample of Size 41(*black dashed line*);
Right: 95% Credible Interval (*blue dotted line*) of Degree Distribution of Sample 47, 74, and 89; Posterior Mean Degree Distribution (*blue dashed line*); True Degree Distribution (*red solid line*).

can estimate the true degree distribution well, while some of them does not.

4.2 Stratified Sampling

Stratified sampling is a probability sampling technique wherein the researcher divides the entire population into different subgroups or strata. The strata are homogeneous and represent possible subpopulation that may be of interest. The sample is drawn within each strata to guarantee an adequate representation of each subpopulation. Moreover, as with blocking of experimental units stratification tends to reduce standard errors. Suppose the population consists of N elements and they are divided into S strata. Each element of the population can be assigned into one and only one stratum. Therefore, the number of observations within each stratum N_s is known and $N = N_1 + N_2 + \ldots + N_S$. We draw students proportional to N_s of each stratum. Stratified sampling offers us several advantages over simple random sampling. Stratified sample can provide greater precision than a simple random sample of the same size. It can help avoid those "unrepresentative" simple random samples, for example, an all-male sample from a mixed-gender population. For Mesa High School data, the following strata are defined: students from grade 7 and 8 are "younger", grade 9 and 10 are "medium" and grade 11 and 12 are "older". Students whose races are black, white or others are combined into an "others" group to have three levels of race: Hispanic, Native American and Others. This leads to a total of $2 \times 3 \times 3 = 18$ strata to draw sample from. In Table 4.1 is the number of observations N_s and sample size n_s from each stratum.

Based on stratification, a proportional stratified sample 41 students are drawn and similar steps as in Section 4.1 are followed to get the estimated population networks. We draw 100 stratified samples and apply the MCMC algorithm with two initial ξ_0 settings as in Section 4.1. An MCMC chain of $10,000$ replicates was used to estimate the posterior mean probability that actor i and actor j are friends, which is a 41×41 matrix. This matrix then is expanded into a 205×205 matrix using Section 4.1 method. For each expanded matrix,

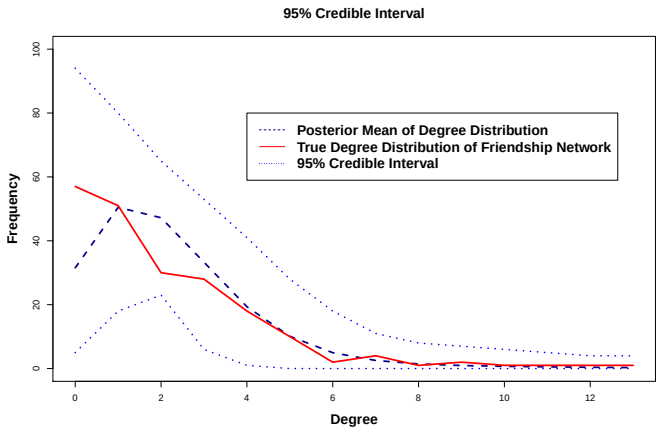

Figure 4.2: *Degree Distribution Inferences for Simple Random Sampling Method.* 95% Credible Interval (*blue dotted line*) of Degree Distribution; Mean Posterior Means of Degree Distribution (*blue dashed line*); True Degree Distribution (*red solid line*).

Table 4.1: *Strata Summary*

Gender	Race	Grade	N_s	n_s
Female	Hisp	older	11	2
		medium	16	3
		younger	28	6
	NatAm	older	4	1
		medium	7	1
		younger	17	3
	Other	older	2	0
		medium	9	2
		younger	5	1
Male	Hisp	older	8	1
		medium	17	3
		younger	29	6
	NatAm	older	8	2
		medium	14	3
		younger	18	4
	Other	older	3	1
		medium	4	1
		younger	5	1
Total			205	41

we generated 100 realizations of the posterior connection probability matrix, which is a 205×205 matrix Y of 0's and 1's. To better see the estimated results of stratified sampling, we also present the 95% credible intervals and the spaghetti plots of 100 realizations from the three random picked samples of size 41. In particular, the $13th$, $28th$ and $86th$ stratified samples, which are listed in Figure 4.3.

Again to check the stability of our method, we also draw 95% credible intervals for the degree distribution we obtained from the 100 estimated networks in Figure 4.4. Figure 4.4 shows that the 95% posterior credible interval is narrower than that of the simple random sample case and that the estimated mean degree distribution is closer to the true one at degree 0, 1 and 2.

By comparing Figure 4.1 with Figure 4.3, we can find that stratified sampling produces better estimation for the true friendship network of the 205 students in Mesa High School than simple random sample. Possibly because there is a greater chance to get a more representative sample in stratified sampling, if the strata are associated with the variable of interest. In this case such variable is the network and the results validates the assumption that grade, gender and race may influence the formation of friendship ties.

4.3 Egocentric Sampling

In this section, we will use the traditional network estimation method introduced in section 2.3.3 to estimate the friendship network of 205 students.

In many empirical contexts, it is not feasible to collect a network census or even an adaptive (link-traced) sample. Egocentrically sampled data, the data comprising information about respondents (egos) and their immediate partners (alters), are much easier to collect and may contain temporal information about the network ties, in the form of each respondents past history and duration of ongoing ties. Examples include the National Health and Social Life Survey (NHSLS) by Laumann, Gagnon, Michael, and Michaels [28] and Wave III of the National Longitudinal Study of Adolescent Health by Harris, Florey,

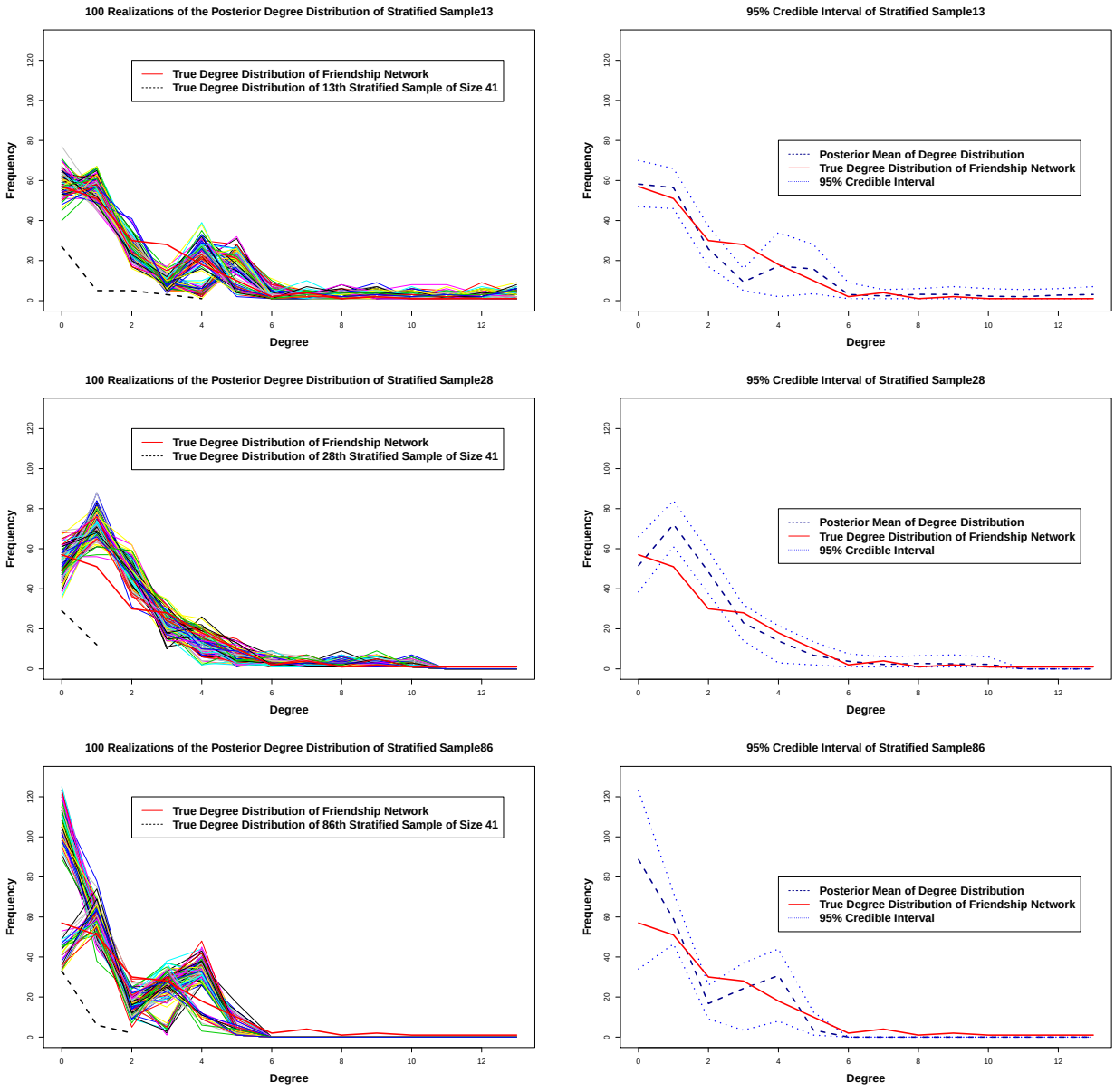

Figure 4.3: *Degree Distribution Inferences for Three Stratified Samples*
Left: 100 Realizations of the Posterior Degree Distribution; True Degree Distribution (*red solid line*); True Degree Distribution of 13*th*, 28*th* and 86*th* Simple Random Sample of Size 41(*black dashed line*); *Right*: 95% Credible Interval (*blue dotted line*) of Degree Distribution of Sample 13, 28, and 86; Posterior Mean Degree Distribution (*blue dashed line*); True Degree Distribution (*red solid line*).

Tabor, Bearman, Jones, and Udry [29].

Assuming that information of the population network is unobservable, a simple random sample of 41 student has summary statistics listed in Table 4.2 and Table 4.3.

Table 4.2: *Degree Distribution of 41 Students' Friendship Network*

Degree	Frenquency	Fraction	Links
0	28	0.68	0
1	8	0.20	8
2	5	0.12	10
Total	41	1.00	18

Table 4.3: *Mixing Matrices of 41 Students: Friendship Links between Gender, Grade and Race Seperately*

Gender			Grade				Race			
	Female	Male		Younger	Medium	Older		NatAm	Hisp	Other
Female	3	5	Younger	5	10	10	NatAm	10	15	10
Male	5	1	Medium	10	5	15	Hisp	15	5	5
			Older	10	15	0	Other	10	5	0

We can use an ERGM model introduced in Section 2.3.3 to fit the parameters associated with these observed statistics, then use the fitted model to simulate the population network of size 205 whose degree distribution is centered around these statistics. The output is shown in figure 4.5.

Regardless of the type of comparison, 95% credible interval or the spaghetti plot fitting a ERGM model does a worse job than the stochastic blockmodel from Sections 4.1 and 4.2 to replicate the degree distribution. The simulated credible band does not cover the true degree distribution of the population on the majority.

We have seen that using totals from the sample are not enough, but sometimes we may obtain summary statistics of the true population network in Table 4.4 and Table 4.5. *ERGM* introduced in Section 2.3.3 can be used to fit the parameters associated with these

observed statistics, then we use the fitted model to simulate the population network of size 205 directly. The output is shown in figure 4.6.

Table 4.4: *Degree Distribution of 205 Students' Friendship Network*

Degree	Frenquency	Fraction	Ties
0	57	0.28	0
1	51	0.25	51
2	30	0.15	60
3	28	0.14	84
4	18	0.09	72
5	10	0.05	50
6	2	0.01	12
7	4	0.02	28
8	1	0.00	8
9	2	0.01	18
10	1	0.00	10
13	1	0.00	13
Total	205	1.00	406

Table 4.5: *Mixing Matrices of 205 Students: Friendship Links between Gender, Grade and Race Seperately*

Gender Mixing Matrix

	Female	Male
Female	82	71
Male	71	50

Grade Mixing Matrix

	Younger	Medium	Older
Younger	108	7	5
Medium	7	39	16
Older	5	16	28

Race Mixing Matrix

	NatAm	Hisp	Other
NatAm	46	41	23
Hisp	41	53	31
Other	23	31	9

Figure 4.6 fits the true population's degree distribution better than Figure 4.5, which is an ERGM model based on a 41 sample of the population. It fits degree 0 better than stochastic blockmodel but fits the rest of the degree distributions worse. Its 95% credible interval is not as good and inclusive as the one by stochastic blockmodel.

Furthermore, another test of whether a model fits the data is how well it reproduces some other network properties such as mean shortest distance, diameter, transitivity that

are introduced in Section 2.2. We do this by comparing the values of these statistics observed in the true network with the values we get in simulated networks from our different models. We compute and report those properties introduced in Chapter 2: diameter, density, mean shortest distance, and transitivity. For example, the diameter of a network is the length of the longest geodesic path between any two vertices. Density is calculated by $m/n(n-1)$, where m is the total number of edges and n is the total number of vertex in a network. Besides the true 205×205 students' friendship network, we investigate four simulated networks: the 205×205 friendship network simulated from a simple random sample of 41×41 friendship network using stochastic blockmodel (SBM_{srs}); the 205×205 friendship network simulated from a stratified sample of 41×41 friendship network using stochastic blockmodel ($SBM_{stratified}$); the 205×205 friendship network fitted from the egocentric data of a 41×41 sample network using ERGM ($ERGM_{41}$); and the 205×205 friendship network fitted from egocentric data of a 205×205 friendship network using ERGM ($ERGM_{205}$). For network SBM_{srs} and network $SBM_{stratified}$, we construct 100 realization networks based on the posterior probability of connection between actor i and actor j. Therefore the diameters, densities, mean shortest distances and transitivities for them in Table 4.6 are the averages of those corresponding properties of 100 networks. Properties for those two $ERGM$ columns are computed directly according to equations in Chapter 2. The results are listed in columns of table 4.6.

From Table 4.6, we can see that stochastic blockmodel estimations offer better measures of the true population network than $ERGM$ estimation. In particular, stochastic blockmodel based on a more representative stratified sample gives better measure statistics than that of a simple random sample, in terms of all properties listed in the table.

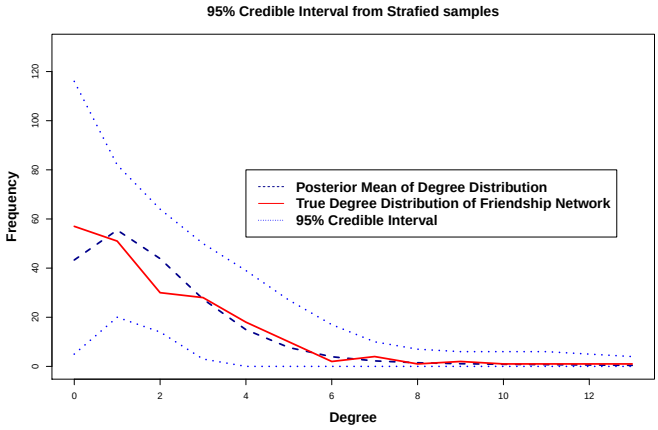

Figure 4.4: *Degree Distribution Inferences for Stratified Sampling Method.* 95% Credible Interval (*blue dotted line*) of Degree Distribution; Mean Posterior Means of Degree Distributions (*blue dashed line*); True Degree Distribution (*red solid line*).

Table 4.6: *Comparison Inference of Network Metrics*

	TRUE	SBM_{srs}	$SBM_{stratified}$	$ERGM_{41}$	$ERGM_{205}$
Mean Shortest Distance	6.8098	4.3890	4.4839	0.0498	0.0648
Diameter	16.0000	11.0370	11.1300	1.1100	1.1500
Transitivity	0.2822	0.2662	0.1365	0.0064	0.0020
Density	0.0097	0.0103	0.0154	0.0001	0.0001

1. Column one uses the true 205×205 network information to compute network properties.
2. Column two uses a 205×205 friendship network simulated from a simple random sample of 41×41 friendship network using stochastic blockmodel (SBM_{srs}).
3. Column three uses a 205×205 friendship network simulated from a stratified sample of 41×41 friendship network using stochastic blockmodel ($SBM_{stratified}$).
4. Column four uses a 205×205 friendship network fitted from the egocentric data of a 41×41 sample network using ERGM ($ERGM_{41}$).
5. Column five uses a 205×205 friendship network fitted from egocentric data of a 205×205 friendship network using ERGM ($ERGM_{205}$).
6. For network SBM_{srs} and network $SBM_{stratified}$, we construct 100 realizations based on the posterior probability of connection between actor i and actor j. Therefore the diameters, densities, mean shortest distances and transitivities for them in Table 4.6 are the averages of those properties of these 100 networks. Properties for $ERGM_{41}$ and $ERGM_{205}$ are computed directly according to equations in Chapter 2.

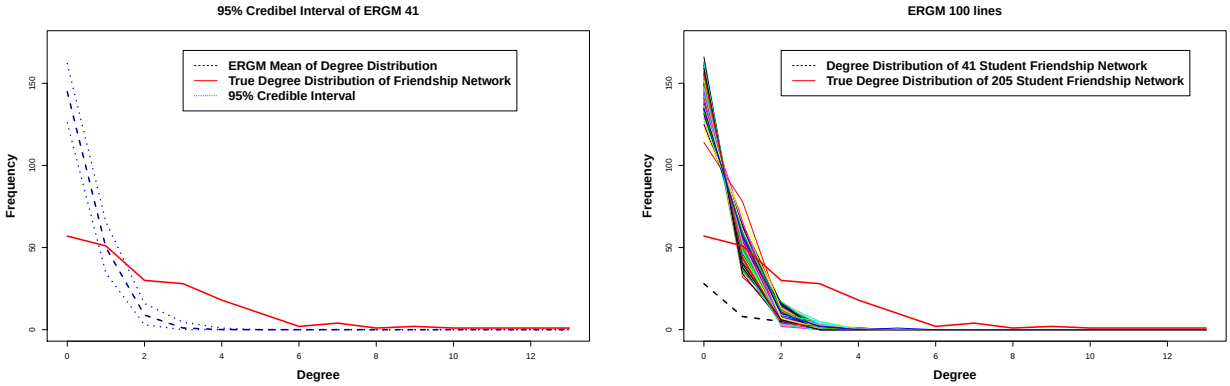

Figure 4.5: *Degree Distribution Inferences for ERGM Method based on a Sample of 41 Students*

Left: 95% Credible Interval (*blue dotted line*) of Degree Distribution of the Simulated Networks; Mean Degree Distribution (*blue dashed line*); True Degree Distribution (*red solid line*); *Right*: Spaghetti Plot of Degree Distribution of the Simulated Networks; True Degree Distribution (*red solid line*); True Degree Distribution of the Sample Network (*black dashed line*).

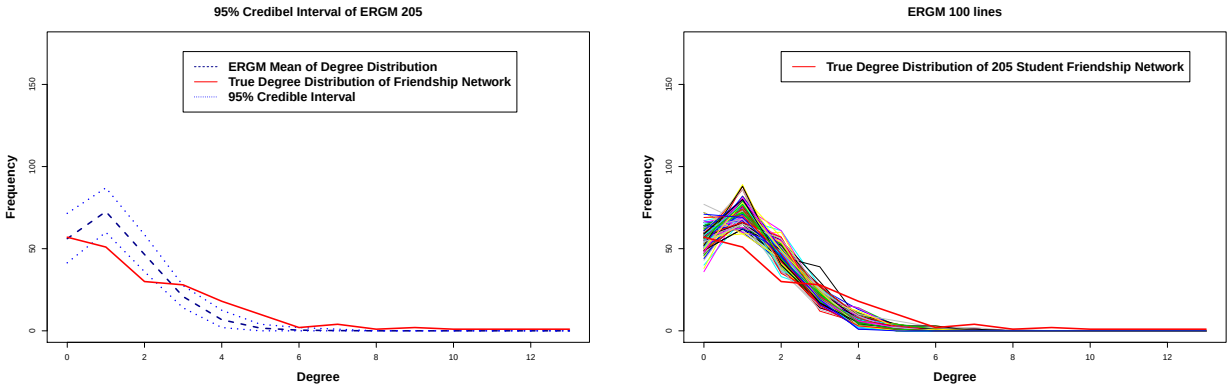

Figure 4.6: *Degree Distribution Inferences for ERGM Method based on 205 Students.*

Left: 95% Credible Interval (*blue dotted line*) of Degree Distribution of the Simulated Networks; Posterior Mean Degree Distribution (*blue dashed line*); True Degree Distribution (*red solid line*); *Right*: Spaghetti Plot of Degree Distribution of the Simulated Networks; True Degree Distribution (red solid line).

Chapter 5

Conclusion and Future Work

In this report we explore the statistical challenge of dealing with sampled network data when collecting true population network is impossible. We presented the concept of network, talked about different properties of complex networks and compared the population network estimations based on stochastic blockmodel and $ERGM$.

We began by talking about our motivation and introducing the Faux Mesa High School's friendship network. Then in the first part of Chapter 2, we discussed some important network properties that are calculated in this report, such as mean shortest distance, clustering coefficient, degree distribution, community structure and etc. What followed in the second part is a discussion on different random graph models. Particularly we describe the Exponential Random Graph Models and showed its application on analyzing the Faux Mesa High School friendship network.

Chapter 3 includes basic concepts and application of stochastic blockmodels, which are used to decompose a network in classes of actors with common properties so certain grouping patterns can be found. By applying the stochastic blockmodel on our high school friendship network, we found 10 main communities after tuning parameters. Each of these 10 groups shows a different grade, gender and race proportions from the whole network.

Then in the next Chapter 4, we explore some traditional sampling techniques when assuming the true network information is not available. We provide some evidence on

whether and how a sampled network can be used to estimate the true population network and to what extent the degree distribution of the estimated network reflects that of the true network. Stochastic blockmodels and ERGM are both used to model the sampled network. We expand both models to simulate the whole 205×205 population network and stochastic blockmodel seems to give better results in terms of degree distribution, mean shortest distance, transitivity, and density, which are the properties discussed in Chapter 2.

We found that different sampling methods can be applied based on how we can access the network data and what is the goal of sampling. In some cases, the entire network data could be accessed fully then a random edge or vertex can be selected. It could also be accessed restrictively when the network is hidden but allows analyzing. Applied social network analysis often use graphs constructed from data collected from a sample of nodes. We have seen that when nodes are selected purely randomly, the less representative sampling induces biased estimates of population network. $ERGM$ based only on degree distribution and egocentrically sampled network data offers a poor estimate. In our application, we have shown that the stochastic blockmodel method from a more representative stratified sample gives a better estimation. We could have some valid inference for the properties of the network based on its stratified sample. Network-based applied work must proceed cautiously, paying close attention to network data quality. From an application perspective, researchers should be careful to work with network analysis results when facing sampled data.

However even using the stochastic blockmodel with stratified sample, our simulated population network stills show some bias when estimating the 0 degrees of the true network's degree distribution. Our future work includes finding a network dataset with more links and adjusting parameter settings to better fit the stochastic blockmodel.

Bibliography

[1] Mark. S. Handcock, K. J. Gile, and C. M. Mar. Estimating the size of populations at high risk for hiv using repondent-driven sampling data. *Biometrics*, 71:258–266, 2015.

[2] Panos. Balatsoukas, C. M. Kennedy, Iain. Buchan, John. Powell, and John. Ainworth. The role of social network technologies in online health promotion: a narrative review of theoretical and emipirical factors influencing intervention effectiveness. *Journal of Medical Internet Research*, 17:141–155, 2015.

[3] R. C. Shalizi and A. Rinaldo. Consistency under sampling of exponential random graph models. *Annals of Statistics*, 41:508–535, 2013.

[4] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about A Highly Connected World*. Cambridge University Press, 2010.

[5] M. E. J. Newman. *Networks: An Introduction*. Oxford Unversity Press, 2010.

[6] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2:1–117, 2009.

[7] L. A. Goodman. Snowball sampling. *Annals of Mathematical Statistics*, 32:148–170, 1961.

[8] Neli. Blagus, L. Subelj, G. Weiss, and M. Bajec. Sampling promotes community structure in social and information networks. *Physica A Statistical Mechanics and Its Applications*, 432:206–215, 2015.

[9] A. Rezvanian and M. R. Meybodi. Sampling promotes community structure in social

and information networks. *Physica A Statistical Mechanics and Its Applications*, 424: 254–268, 2015.

[10] M. D. Resnick, P. S. Bearman, and R. W. Blum. Protecting adolescents from harm: findings from the national longitudinal study on adolescent health. *Journal of the American Medical Association*, 278:823–832, 1997.

[11] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[12] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45: 167–256, 2003.

[13] A. Rapoport. Contribution to the thoery of random and biased nets. *Bulletin of Mathematical Biophysicas*, 19:257–277, 1957.

[14] P. Erdős and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–61, 1960.

[15] R. Solomonoff and A. Rapport. Connectivity of random nets. *Bulletin of Mathematics Biophysics*, 13:107–117, 1951.

[16] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[17] P. Erdős and A. Rényi. On the strength of connecedness of a random graph. *Acta Mathematica Scientia Hungary*, 12:261–267, 1961.

[18] A. L. Barabási and R. Albert. Emergence of scaling in random network. *Science*, 286: 509–512, 1999.

[19] D. Strauss. On a general class of models for interaction. *SIAM Review*, 28:513–527, 1986.

[20] A. Rapoport and W. J. Horvath. A study of a large sociogram. *Behavioral Science*, 6: 279–291, 1961.

[21] J. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, 2000.

[22] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge Unversity Press, 1994.

[23] K. P. Reddy, M. Kitsuregawa, P. Sreekanth, and S. S. Rao. *A graph based approach to extract a neighborhood customer community for collaborative filtering*. Springer-Verlag, 2002.

[24] A. Rakesh and H. V. Jagadish. Algorithms for searching massive graph. *IEEE Trans. Knowl. Data Eng.*, 6:225–238, 1994.

[25] F. Wu and B. A. Huberman. Finding communities in linear time: a physics approach. *The European Physical Journal*, 38:331–338, 2004.

[26] S. E. Fienberg and S. Wasserman. Categorical data analysis of single sociometric relations. *Sociological Methodology*, 81:156–192, 1981.

[27] B. F. Ribeiro and D. F. Towsley. Estimating and sampling graphs with multidimensional random walks. *CoRR*, 28:1002–1751, 2010.

[28] E. O. Laumann, J. H. Gagnon, R. T. Michael, and S. J. Michaels. The social organization of sexuality. *The Journal of the American Medical Association*, 274:535–538, 1995.

[29] K. M. Harris, F. Florey, J. Tabor, P. S. Bearman, and J. Jones. The national longitudinal study of adolescent health: Research design. National Institute of Child Health and Human Development, 2003. (Study Design).