Title

Understanding Networks with Exponential-family Random Network Models

Author

wang, zeyi

Publication Date

2023

UNIVERSITY OF CALIFORNIA

Los Angeles

Understanding Networks with Exponential-family Random Network Models

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Zeyi Wang

2023

ABSTRACT OF THE THESIS

Understanding Networks with Exponential-family Random Network Models

by

Zeyi Wang
Master of Science in Statistics
University of California, Los Angeles, 2023
Professor Mark Handcock, Chair

The structure of many complex social networks is determined by nodal and dyadic covariates that are endogenous to the tie variables. While exponential-family random graph models (ERGMs) have been very successful in modeling social networks with exogenous covariates, they are often misspecified for networks where some covariates are stochastic. Exponential-family random network models (ERNMs) are an extension of ERGM that retain the desirable properties of ERGM, but allow the joint modeling of tie variables and covariates. We compare ERGM to ERNM to show how conclusions of ERGM modeling are improved by consideration of the ERNM framework. In particular, ERNM simultaneously represents the effects of social influence and social selection processes while commonly used models do not. We also look into the latent class models for group clustering problems in social network. The random nodal attributes can be observed or latent in ERNM. When the attributes is treated as latent, we can investigate the probability of cluster group. Stochastic Block Models are a well-known statistical models for group classification, however, they rely on the relational data and omit the nodal characteristics. We compare these two models and provide a case study to illustrate the main points.

The thesis of Zeyi Wang is approved.

Oscar Hernan Madrid Padilla

Nicolas Christou

Xiaowu Dai

Mark Handcock, Committee Chair

University of California, Los Angeles

2023

To my parents and Qing . . .

For those days and nights...

accompany and support...

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Social network analysis has been highly valued in the social sciences in recent years. Statistical models are widely used in various fields to represent network structure. The well-known exponential-family random graph model (ERGM) is widely applied, where random graphs consist of a selection of nodes with some fixed nodal or dyadic covariates and random connections (edges) between nodes (Frank and Strauss, 1986; Hunter and Handcock, 2006; Lusher et al., 2013). A problem is that the assumption of fixed nodal covariates is sometimes inappropriate. The processes of edge and covariate formation commonly occur simultaneously (Leenders, 1997).

A generalization of ERGM called the exponential-family random network model (ERNM) was developed in Fellows (2012) and Fellows and Handcock (2012). It represents both "social selection" and "social influence" processes, where the former states that the social connections are determined by the nodal attributes (Robins et al., 2001a; Friemel, 2015) and the latter holds that the nodal attributes are determined by the social connections (Robins et al., 2001b). ERNM represents a joint exponential-family model, where some or all the nodal/dyadic attributes and social connections (edges) are treated as endogenous.

This paper is structured as follows. In the next two chapters (chapter 2 and 3), we introduce the ERGM and ERNM classes with their model specifications, and we discuss their model interpretations. Chapter 4 focuses on some interesting network statistics and model estimation. In chapter 5, we compare ERGM and ERNM conceptually. A case-study is conducted in chapter 5, which includes a detailed modeling and analyzing study using both ERGM and ERNM for an adolescent health dataset (Harris et al., 2007). Chapter 6 discusses the Latent ERNM and two case studies. Chapter 7 discusses the results of the comparisons and concludes the paper.

CHAPTER 2

ERGM and ERNM Classes

2.1 Construction of Social Network

A social network describes the network structures with n social actors (nodes), together with a set of q actor variates, and the social connections between each pair of actor. The social connections can be directed or undirected, which leads to either a directed social network or an undirected social network.

We consider the situation where the network is the result of a social process, modeled stochastically. For a social network (X, Y), with n nodes, $Y \subset \mathbb{R}^{n\times n} \in \mathcal{Y}$ is the graph with $X \subset \mathbb{R}^{n\times n\times q} \in \mathcal{X}$ as dyadic attributes. The space of tie variables, $\mathcal{Y}$, can be arbitrary although here we focus on binary tie variables:

$$Y_{ij} = \begin{cases} 1 & \text{if actor } i \text{ is connected to actor } j \\ 0 & \text{otherwise} \end{cases},$$

$i, j = 1, \cdots, n$. For undirected networks, $Y_{ij} = Y_{ji}$. Y is often called an adjacency matrix. The dyadic covariates, X_{ijk}, are measures on the $(i,j)^{th}$ pair. An important special case is covariates that depend only on i or j, that is nodal covariates denoted by X_{ik} or X_{jk}. Examples of dyadic covariates include a homophily term:

$$X_{ijk} = \begin{cases} 1 & \text{if actor } i \text{ and actor } j \text{ have identical values} \\ & \text{on the } k^{th} \text{ nodal characteristic } (X_{ik} = X_{jk}) \\ 0 & \text{otherwise} \end{cases},$$

$k = 1, \cdots, q$, for each of q separate nodal characteristics.

2.2 ERGMs Specification

Random graph describes the probability distribution on the graphs. Consider a network with n nodes: a random graph add edges between nodes successively at random. The basic construction of an ERGM includes a graph $Y \in \mathcal{Y}$ that can be explained by some sufficient statistics defined by a d-vector valued function $g(\mathcal{Y})$. A general form of ERGMs that describes a probability distribution of undirected graphs with n nodes:

$$P_\eta(Y = y) = \frac{1}{c(\eta, \mathcal{Y})} \exp\{\eta \cdot g(y)\} \quad y \in \mathcal{Y}, \tag{2.1}$$

where $\eta \in \mathbb{R}^d$ is a parameter vector associated with a d-vector valued function $g(\mathcal{Y})$ and $c(\eta, \mathcal{Y})$ is the normalizing constant which ensures that this is a proper probability distribution. The maximum entropy of the exponential-family is guaranteed by the mean constraint on $g(y)$, where $\mathbb{E}_\eta[g(y)] = \mu$. Different choices of $g(\mathcal{Y})$ determine different models under ERGMs class.

Different from traditional statistical models that measure observations with some predefined response variables and explanatory variables separately, exponential-family random graph models (ERGMs) consist of explanatory variables that are functions of response variables themselves. More specifically, in a network, the response variables are typically defined as the state of a tie y – either formation or dissolution. In general ERGMs (2.1), the graph statistics $g(y)$ are configurations of ties, where $g(\mathcal{Y})$ are jointly sufficient for the model. The observations in network data also consist of nodal attributes x, for example, the age of nodes. The nodal attributes can be included in ERGMs as exogenous predictors (Fienberg and Wasserman, 1981; Wasserman and Pattison, 1996). By defining the nodal covariates as explanatory variables, a type of ERGM class that allows Markov dependence with nodal attributes is

$$P_\eta(Y = y|X = x) = \frac{1}{c(\eta, x, \mathcal{Y})} \exp\{\eta \cdot g(y|x)\} \quad y \in \mathcal{Y}, x \in \mathcal{X}, \tag{2.2}$$

where $\eta \in \mathbb{R}^d$ is a d-vector of parameters. $g(y|x)$ is a d-vector of graph statistics, where $g(\mathcal{Y}|x)$ are jointly sufficient statistics. The normalization constant is $c(\eta, x, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp\{\eta \cdot g(y|x)\}$.

Choices of network statistics of interest depend on common knowledge and the social context. Morris et al. (2008a) provides examples of the common features.

2.3 ERNMs Specification

Exponential-family random network models generalize ERGMs by treating the nodal attributes as endogenous variables (Fellows and Handcock, 2012). Inspired by Leenders (1997), which argued that the process of social selection and social influence are simultaneous. ERNMs model the joint relationship between edges and nodal variates.

Definition 2.3.1 (ERNM, Fellows and Handcock (2012))**.** *For a social network (X, Y), with n nodes, where $Y \subset \mathbb{R}^{n \times n}$ is the graph with $X \subset \mathbb{R}^{n \times q}$ as nodal attributes, the multivariate distribution of Y and X can be written as:*

$$P_\eta(Y = y, X = x) = \frac{1}{c(\eta, \mathcal{N})} \exp\{\eta \cdot g(y, x)\}, \qquad (y, x) \in \mathcal{N}, \tag{2.3}$$

where $\mathcal{N}$ is the sample space of Y and X, $\eta \in \Lambda$ is a q-vector of parameters, $g(y, x)$ is a q-vector of network statistics, with $g(Y, X)$ jointly sufficient for the model, and $c(\eta, \mathcal{N})$ is the normalisation constant. The formal definition of $c(\eta, \mathcal{N})$ is given in Fellows and Handcock (2012): Let $(N, \mathcal{N}, P_0)$ be a σ-finite measure space with reference measure P_0, then, a probability measure with respect to this space is an ERNM if it is dominated by P_0. The normalisation constant is defined as

$$c(\eta, \mathcal{N}) = \int_{y, x \in \mathcal{N}} \exp\{\eta \cdot g(y, x)\} dP_0(y, x), \tag{2.4}$$

where $\Lambda \subset \{\eta \in \mathbb{R}^q : c(\eta, \mathcal{N}) < \infty\}$.

2.4 Model Interpretation

To interpret the coefficient of ERGM, consider the logit form of exponential family models (2.2):

$$\text{logit}\left(P_\eta(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)\right) = \eta \cdot \left(g(y_{ij}^+) - g(y_{ij}^-)\right), \tag{2.5}$$

where y_{ij}^c is the set of tie values $y \backslash y_{ij}$, y_{ij}^+ and y_{ij}^- correspond to the graphs $(y_{ij}^c, y_{ij} = 1)$ and $(y_{ij}^c, y_{ij} = 0)$, respectively, Y_{ij}^c is the random variable $Y \backslash Y_{ij}$. This is often referred to as the conditional log-odds of a tie Y_{ij}. We see that η has the interpretation of the change in conditional log-odds of a tie Y_{ij} per unit change in the graph statistics were y_{ij} toggled from zero to one.

The interpretation of graph statistics Y_{ij} of ERNM is very similar to ERGM. The only difference is that ERNM models need to consider the covariates X. From (2.3) we have:

$$\begin{aligned} &\text{logit}\left(P_\eta(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c, X = x)\right) \\ &= \eta \cdot \left(g(y_{ij}^+, x) - g(y_{ij}^-, x)\right), \end{aligned}$$

so that the ERNM parameter has a closely allied interpretation to that of the ERGM parameter as the distribution of Y explicitly conditional on X is the same as the ERGM with implicit conditioning on X. This implies that we are able to interpret the parameter of ERNMs in a similar fashion as ERGMs. See Appendix A for a derivation.

There are other interpretations as well. Fellows and Handcock (2012) discusses the way to interpret coefficients of dyadic variables of ERNM with logistic regression. The dyadic attributes X in equation (2.3) can be partitioned into two parts, Z and T by defining $Z \in (0, 1)$ as a binary dyadic variate of interest (that is an outcome variable) and T as a matrix of regressors, where $Z, T \subset X$. We can rewrite the equation (2.3) with the new definition:

$$\begin{aligned}&P_\eta(Y=y,Z=z,T=t)\\=&\frac{1}{c(\eta,\mathcal{N})}\exp\left\{\alpha\cdot g(y,t)+\lambda\cdot h(y,z)+z\cdot t\beta\right\}\\&(y,x)\in\mathcal{N},\end{aligned}\tag{2.6}$$

where $\eta = (\alpha, \beta, \lambda)$ are parameters, $g(y,t)$ and $h(y,z)$ are network statistics, $z \cdot t\beta$ is the relationship of T to Z. We can then derive the logit form of the distribution of z_{ij} from equation (2.6) condition upon the rest of the network (proof details in Appendix B):

$$\begin{aligned}&\operatorname{logit}\left(P_\eta(z_{ij}=1|z_{ij}^c,t_{ij},Y=y)\right)\\&=(t_{ij}\beta)-\left(\lambda\cdot(h(y,z_{ij}^-)+h(y,z_{ij}^+))\right),\end{aligned}\tag{2.7}$$

where z_{ij} and t_{ij} are the measures on the $(i,j)^{\text{th}}$ pair of Z and T, z_{ij}^c is the set of variants $z \setminus z_{ij}$, z_{ij}^+ and z_{ij}^- correspond to the variant of z_{ij} where $z_{ij} = 1$ and $z_{ij} = 0$, respectively. Suppose the matrix of regressors t_{ij} changes to $t_{ij}^{'}$, with all other variables and networks remaining fixed, then the logarithm of odds ratio (R) is

$$\ln R = \beta \cdot (t_{ij} - t_{ij}^{'}).\tag{2.8}$$

Therefore, the coefficients of the outcome variable z may be interpreted as a conditional logistic regression model. For one unit change in t_{ij}, the log-odds changes by β, keeping all other variables constant.

CHAPTER 3

Specification and Estimation for ERGM and ERNM

Choices of statistics for modeling are very flexible and case-based. Statistics like edges, mutuality, homophily, and transitivity are primary choices to be included in the model to grasp the major characteristics of the network. For example, the R package ergm contains over one-hundred "terms", each being a coherent set of graph statistics (R Development Core Team, 2022; Handcock et al., 2021; Morris et al., 2008a).

The set of network statistics for ERNM includes those for the ERGM, with the difference that they have different roles in the model due to the the endogeneity of nodal attributes. Moreover, some statistics, for example, those that involve the nodal characteristics but not the tie values, are specific to ERNM but not to ERGM. An example of such a statistic, one with an important role below in this paper, is the number of students who are smokers in Section 5.2.

3.1 Primary Network Statistics for ERGM and ERNM

Different to traditional statistical models that measure observations with some predefined response variables and explanatory variables separately, exponential-family random graph models (ERGMs) consist of explanatory variables that are functions of response variables themselves. More specifically, in a network, the response variables are typically defined as the state of a tie y – either formation or dissolution. In general ERGMs (2.1), the graph statistics $g(y)$ are configurations of ties, where $g(\mathcal{Y})$ are joint sufficient for the model.

The maximum entropy of exponential-family guaranteed by the mean constraint on $g(y)$, where $\mathbb{E}_\eta[g(y)] = \mu$. The observations in network data also consist of nodal attributes x, for example age of nodes. And the network statistics $g(y|x)$ in ERGMs conditioning on nodal attributes are functions of ties given exogenous covariates. Choices of network statistics of interest depend on common knowledge and the social context. Morris et al. (2008b) provides examples of common features including both graph statistics and statistics conditioning on nodal covariates. ERGM terms are also classified as dyadic dependent, dyadic independent and curved exponential-family (CEF) terms in Hunter et al. (2008). Transitivity and homophily are two interesting features that will be focused on.

3.1.1 Dyadic Independence vs Dyadic Dependence

ERGMs are defined by different choices of sufficient statistics $g(y|x)$, which are explanatory variables to the model. The way to calculate graph statistics relates with a term called *change statistics*:

Definition 3.1.1 (**Change Statistics**, Hunter et al. (2008))**.** *Given a network y (either directed or undirected) and a dyad with a pair of nodes (i, j), the change of statistics under the context of ERGMs (2.1) and a vector of statistics $g(y)$ is defined as*

$$\delta_g(y)_{ij} = g(y_{ij}^{+}) - g(y_{ij}^{-}), \tag{3.1}$$

where $g(y_{ij}^{+})$ and $g(y_{ij}^{-})$ correspond to the graph statistics when $y_{ij} = 1$ and $y_{ij} = 0$. More specifically, change of statistics measures the change in graph statistics $g(y_{ij})$ when y_{ij} is toggled from 0 to 1.

A dyadic independent term does not have direct dependence on dyads, and its definition relies on the definition of change statistics in Definition 3.1.1.

Definition 3.1.2 (**Dyadic Independence**, Hunter et al. (2008))**.** *A term in an ERGM is dyadic independence if the updates of the corresponding network change statistic $\delta_g(y)_{ij}$ does not require information of other dyads.*

Concrete examples of dyadic dependence and dyadic independence are transitivity and homophily. A consequence of transitivity is a closure of a triangle, however, the triad-closure of homophily is not necessary (Figure 3.1).

Transitivity The transitivity (clustering coefficient) measures the tendency of the people to cluster together. A famous analogy in social sciences illustrates this as “the friend of my friend is my friend”. Hence, this term exhibits a triad-closure feature. The state of one dyad depends stochastically on the state of other dyads, thus it is a dyad dependent term.

Homophily Homophily of a certain attribute measures the tendency of people to make connections within the specified attributed group. It may or may not leads to a triad-closure, and the state of a dyad does not depend on state of other dyads. Therefore, it is a dyad independet term. There are two types of homophily: uniform and differential. Uniform homophily counts the total number of edges with same attributes, whereas differential homophily counts each unique value of the number of edges with same attributes. Those are measured with *nodematch* terms in ERGM and ERNM.

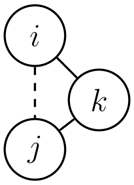

Figure 3.1: Transitivity vs Homophily: three vertices are all from the same value of the attribute

Informally, transitivity is the tendency of people to cluster together. A famous analogy illustrates this as ”the friend of my friend is my friend”. Hence, this term exhibits a triad-closure feature. This tendency can be quantified: consider a three-number summary of the triads in an undirected graph or network being the number of ties, $\frac{1}{3}\sum_{ijk} y_{ij} + y_{jk} + y_{ik}$, the number of two-stars, $\sum_{ijk} y_{ij}y_{ik}$, and the number of triangles, $\sum_{ijk} y_{ij}y_{ik}y_{jk}$. A highly transitive graph would have a lot of triangles relative to the number of two-stars. It would seem natural to include these three statistics in a model and use them to measure transitivity. However, models with these terms in them have been shown to have bad statistical properties, referred to as model degeneracy. So we will instead use a measure of transitivity that does have better properties: the geometrically weighted edgewise shared partner distribution.

3.1.2 Curved-Exponential Family (CEF) Terms

In the last section, we discussed the problem of degeneracy in ERGMs brought by the dyadic dependence terms, especially the single triangles. A Shared partner distribution manages to handle this problem introduced under the context of curved exponential family (Snijders et al. (2006); Hunter and Handcock (2006); Goodreau et al. (2008)).

Two nodes are edgewise shared partner (ESP) if they are connected, and each is also connected with another node. A triangle is formed as a result of a tie between them. It is also a measure of transitivity. The edgewise shared partner statistic is defined as ESP(k), meaning that the number of unordered pairs $\{i, j\}$ such that an edge exists between i and j, i and j have exactly k common neighbors, under the context of undirected network. Figure 3.2 illustrated the edgewise shared partner distribution under an undirected network. The measure ESP for edge $(i-j)$ counts the number of nodes k such that i and j share edges with. In this network, the edge $(i - j)$ has three edgewise shared partners: k_1, k_2, k_3. All other six edges have exactly one edgewise shared partner. Hence, the edgewise shared partner distribution of this network is: ESP(1)= 6, ESP(2)= 0, ESP(3)= 1.

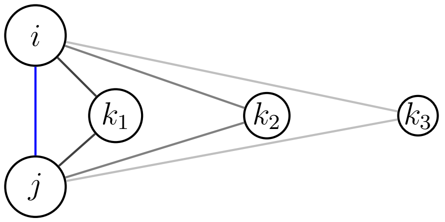

Figure 3.2: Undirected ESP Distribution

Though it is straight-forward to understand one pair of edgewise shared partnerships, the distribution of edgewise shared partner is not uniform. A tie that closes triangles is more likely to form than a tie that do not close triangles. Geometrically-Weighted Edgewise Shared Partnerships (GWESP) is developed for this purpose.

Definition 3.1.3 (**GWESP**, Hunter (2007))**.** *For a network with fixed number of nodes, n, and adjacency matrix $\boldsymbol{Y}$, the geometrically-Weighted Edgewise Shared Partnerships statistics is described as*

$$\mathrm{GWESP} = e^{\theta} \sum_{k=1}^{n-2} [1 - (1 - e^{-\theta})^{i}] \mathrm{ESP(k)}, \tag{3.2}$$

where θ is the decay parameter. $n-2$ is the maximum number of edgewise shared partners for any pair of nodes in the network. The geometrically (G) here implies the statistic is based on the geometric sequence $(1 - e^{-\theta})^{i}$.

1. As more shared partners of an edge exist, the effect of GWESP decays as the result of geometric sequence in equation (3.2). This adjustment prevents the explosion of triangles in the network. 2. GWESP in directed network, similar as ESP, provides different results given the different choices of two-path statistics. Both outgoing two paths and incoming two paths may be considered in real applications.

In summary, homophily measures the tendency of individuals to connect with similar individuals. GWESP measures the tendency of transitivity, which is an important feature of human social networks. We will manipulate with the homophily and GWESP terms under exponential-family random graph models and exponential-family random network models in some real cases analysis in Chapter 5. The performance in capturing transitivity and homophily of the two models will be analysed.

3.2 Degeneracy and MCMC Diagnostics

The inference on ERGM parameters typically employs MCMC procedure to compute the maximum likelihood estimation (MLE) (Hunter and Handcock, 2006). The inference on the likelihood function of ERGMs with only dyadic independence terms can be simplified to a standard logistic regression form. For models with dyad dependence, as for the models considered here, certain combinations of terms result in the model not placing sufficient probability mass on realistic graphs and networks. This is called the model degeneracy problem (Handcock, 2003a,b; Schweinberger, 2011; Schweinberger et al., 2020).

CHAPTER 4

Comparing ERGM to ERNM Conceptually

A main goal of social network analysis is to model the relationship between social ties in the context of nodal attributes. Two types of processes are commonly considered: social selection and social influence. In social selection processes, individuals form social ties on the basis of attributes, theirs and others (Robins et al., 2001a; Friemel, 2015). In social influence processes, the direction is reversed, where the network structure influences the attributes of the individuals in the network; that is, an individual's attributes may be changed by other individuals whom they share social ties with (Robins et al., 2001b). We follow the definition of the social selection and social influence processes in Leenders (1997):

1. Social selection process: Conventional network statistical models represent the network structure stochastically, measure the dynamic change of the network, and treat the nodal attributes as independent variables (or explanatory variables), usually stable and fixed. The nodal attributes in this context refer to some fixed actor characteristics, such as age, gender, and race. The involved process is called the social selection process, where the network structure is determined by some fixed actor attributes.

2. Social influence process: Conversely, actors may alter their attributes because of the influence of the network structure they are embedded in. In other words, the network structure is conditioned on and regarded as exogenous and is invariant over time, whereas the actor attributes are modeled stochastically. This process is called the social influence process (some literature uses the term ”contagion”).

The two processes can be illustrated using network examples shown in Figure 4.1 and Figure 4.2. Both figures contain networks with four nodes (actors), A, B, C, D, and social ties. In Figure 4.1, two new ties are formed between nodes with the same color (A and B, C and D). It is the tendency of which an individual makes connections with other individuals that share the same attributes, such as age or habits, which is referred to as social selection. In Figure 4.2, node A and node D adjust their color to the nodes which they share ties with (B and D, respectively). Hence, social influence processes describe the tendency that individual may change their attributes influenced by other related individuals.

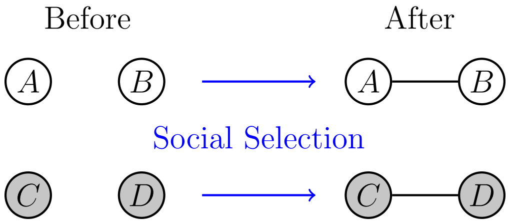

Figure 4.1: Illustration of Social Selection: Color of nodes: nodal attributes

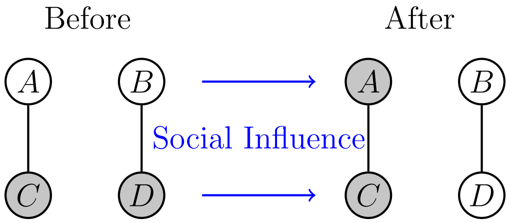

Figure 4.2: Illustration of Social Influence: Color of nodes: nodal attributes

It is essential to note that these are rarely disjoint processes. Specifically, we would expect that a mixture of social selection and social influence processes occur simultaneously. It is extensively argued that the two processes are not mutually exclusive: the social ties affect the nodal attributes and vice versa (Erickson, 1988; Leenders, 1997). As we shall see, this is precisely the situation that ERNM represents, while ERGM and autologistic actor attribute models (ALAAM) (Robins et al., 2001b; Lusher et al., 2013, Chapter 9) do not. ALAAMs are developed as alternatives to ERGMs to capture social influence processes.

ALAAMs model how network relationships influence the nodal attributes, for instance, how friendship of adolescents may influence their smoke behaviors (nodal attributes). Hence, ERNM can be thought of as jointly ERGM and ALAAM, plus more. Explicitly, ERGM and ALAAM are each given by conditional views of ERNM. ERGM and ALAAM are each special cases of ERNM, a point we make here and one developed more fully in Fellows and Handcock (2012). ERNM represents the joint connection between the social selection and social influence processes. Consequently, both nodal attributes and social connections are treated as endogenous and stochastic variables, reflecting reality. Moreover, despite what many network models assume, it is implausible to have invariant nodal attributes in the network, as ERGM assumes. Although some reference nodal attributes in social selection processes, such as sex, age, and race, are invariant in social influence processes, many other attributes, such as smoking and drinking behaviors, may be altered by the network structure. Despite this, the two social network processes are widely studied in the literature separately, the mutual interdependence between the two processes being ignored.

ERNMs are exponential-family graph models that model the joint behavior of edges and nodal attributes. The model (2.3) can be rewritten as

$$P_\eta(Y=y, X=x) = P_\eta(Y=y|X=x)P_\eta(X=x), \tag{4.1}$$

where

$$P_\eta(X=x) = \frac{c(\eta;x)}{c(\eta;\mathcal{N})} \quad x \in \mathcal{X}. \tag{4.2}$$

The first component of (4.1) can be viewed as an ERGM that is conditional on nodal attributes X (Frank and Strauss, 1986; Hunter and Handcock, 2006):

$$\begin{aligned} &P_\eta(Y=y|X=x) \\ =&\frac{1}{c(\eta;\mathcal{N}(x),x)} \exp\{\eta \cdot g(y,x)\} \quad y \in \mathcal{N}(x), \end{aligned} \tag{4.3}$$

where $\mathcal{N}(x) = \{y : (y,x) \in \mathcal{N}\}$. The second component $P_\eta(X=x)$ is the marginal distribution of the nodal variate X, which does not necessarily belong to a non-trivial exponential family. The rewritten form (4.1) illustrates the difference between ERNM and ERGM: the former models the joint behavior of Y and X, whereas the latter models the conditional distribution of Y given X.

In the ERGM (2.2), the graph statistics $g(y|x)$, equivalent to $g(y,x)$ in (4.3), model the network conditioning on the nodal attributes. In other words, the formation or dissolution of an individual's social ties is influenced by other individuals' fixed nodal attributes. Hence, the nodal attributes are treated as exogenous to the model, and in many real situations, this assumption is inappropriate. On the contrary, ERNMs use the network statistics $g(y,x)$, which brings more flexibility to the modeling. Moreover, ERNMs take care of both nodal attributes and dyadic variables, different from ERGMs, which stochastically model the tie variables only. As a consequence, ERNMs are conceptually able to model both the social selection and the social influence processes with endogenous nodal attributes. To further illustrate the two models under the context of the two social processes, consider homophilous selection, which is measured by homophily terms in ERGM and ERNM:

ERGM homophily selection:

$$\begin{aligned}&P_\eta(Y=y|X=x)\\ =&\frac{1}{c(\eta;\mathcal{N}(x),x)}\exp\{\eta\cdot\text{homophily}(y|x)\}\quad y\in\mathcal{N}(x).\end{aligned}\tag{4.4}$$

ERNM Homophily selection:

$$\begin{aligned}&P_\eta(Y=y,X=x)\\ =&P_\eta(Y=y|X=x)P_\eta(X=x)\\ =&\frac{1}{c(\eta;\mathcal{N}(x),x)}\exp\{\eta\cdot\text{homophily}(y|x)\}E_Y[P_\eta(X=x|Y)]\qquad (y,x)\in\mathcal{N},\end{aligned}\tag{4.5}$$

where the last equality comes from

$$\begin{aligned}
P_\eta(X = x) &= \int_{\mathcal{N}(x)} P(Y = y, X = x)dy \\
&= \int_{\mathcal{N}(x)} P_\eta(X = x|Y = y)P_\eta(Y = y)dy \\
&= E_Y[P_\eta(X = x|Y)].
\end{aligned}$$

Both ERGM and ERNM capture social selection. By controlling for other alternative mechanisms, we can achieve a more accurate homophilous selection result (Steglich et al., 2010). Because ERGM treats the nodal attributes as exogenous, it does not represent any social influences. ERNM is able to reflect the social influence at the same time since the nodal attributes are free to vary on the basis of the fixed network structure:

$$\begin{aligned}
P_\eta(Y = y, X = x) &= P_\eta(X = x|Y = y)P_\eta(Y = y) \\
P_\eta(X = x|Y = y) &= \frac{1}{c(\eta; \mathcal{N}(y), y)} \exp\{\eta \cdot g(y, x)\}, \qquad (4.6) \\
x \in \mathcal{N}(y) & \qquad\qquad (y, x) \in \mathcal{N}
\end{aligned}$$

where $\mathcal{N}(y) = \{x : (y, x) \in \mathcal{N}\}$. The term (4.6) represents the ALAAM class (Robins et al., 2001b). This decomposition makes the relationship between ALAAM and ERNM transparent. ALAAM is the ERNM conditional over the network tie structure. So it is a special case of ERNM.

The differences between ERNM and ERGM go beyond homophily terms involving the endogenous covariates. It applies to all terms in the ERNM/ERGM (e.g., k-stars, degrees, GWESP, GWDSP). This is direct for terms involving endogenous covariates, but also indirectly via interactions between model statistics. For example, the presence of direct terms changes the interpretation and coefficients of the other terms in the model.

It is essential to note that ERNM models the association between ties and nodal attributes, and is not a causal model. With cross-sectional data and no causality specified, we cannot preclude social selection from other mechanisms (Steglich et al., 2010). For example, if we are trying to model the adolescent connections on smoking behavior, the homophilous selection modeled on the smoking attribute may preclude the transitivity or reciprocity processes. Other mechanisms, like similarity in drinking behaviors, may also be masked. Although we present a way to jointly model the social selection and social influence processes in a cross-sectional context, in order to disentangle the two processes, longitudinal data is needed (Leenders, 1997).

The ERNM model is a complete class model and can model any specific joint model of Y and X by appropriate choice of network statistics. However, this is not anywhere as strong as it seems as the statistics are unknown and could be of arbitrary complexity and number. Because of this, there is a great need and opportunity for highly structured models that represent much of the complex structure of the network in a relatively simple fashion. For example, Almquist and Butts (2014) introduce a vertex process temporal ERGM for modeling joint edge and behavior dynamics but make limiting assumptions so that the model is tractable. Another is Weng (2020) who develops a separable model for the tie variables and endogenous covariates and introduces individual-specific random effects to represent individual unobserved heterogeneity influencing both network formation and the covariates. This model makes quite different assumptions than ERNM, and a comparison would need to be in-depth.

CHAPTER 5

Case-study Indicating the Need for ERNM over ERGM

5.1 Introduction to the Adolescent Health Data

Much network data on school friendships were collected by the National Longitudinal Study of Adolescent Health (Harris et al., 2007) (Add Health). This nationally representative study includes a longitudinal sample of more than 20,000 adolescents in grades 7 to 12 who were surveyed with in-school questionnaires in the US in 1994 and 1995. Their smoking behaviors are recorded by asking whether they have ever smoked at least once. In our study, we used four networks of grades 9 to 12 (Clark and Handcock, 2022). The smoking behavior is coded as a binary variable, where 1 was used for students who reported that they have ever smoked at least once and 0 otherwise.

Students were asked to nominate up to 5 other students who were their best female friends and also up to 5 other students who were their best male friends. We build the network of weak friendship ties, that is, an undirected tie if either student nominated the other. A visualization of the networks is shown in Figure 5.1. The edges are undirected, as a connection between nodes A and B may represent A nominated B, B nominated A, or both. Table 5.1 shows the summary of each network. It is clear that the higher the grade, the greater the proportion of smoking adolescents, which is to be expected.

We are specifically interested in smoking behavior, especially its interconnection to the network structure. Unlike other measurements like sex, age, and race that are exogenous to

	Network Summary				
	Nodes	Edges	Non-smoker	Smoker	Smoking ratio
Grade 9	256	617	168	88	0.3438
Grade 10	228	498	138	90	0.3947
Grade 11	192	416	118	74	0.3854
Grade 12	193	413	101	92	0.4767

Table 5.1: Network Summary

the network, smoking behavior may be influenced by social connections and is expected to be endogenous.

5.2 Models

We fitexttted the same model terms for both ERGM and ERNM. The computational aspects are discussed in Section 5.3. In the nomenclature of that section, these are:

- ERGM: `Network` $\sim$ `edges` + `ESP(0, 1, 2)`
  + `GWESP(0.5)` + `GWDegree(0.5)`
  + `nodefactor(smoke)` + `nodematch(smoke)`

- ERNM: `Network` $\sim$ `edges` + `ESP(0, 1, 2)`
  + `GWESP(0.5)` + `GWDegree(0.5)`
  + `nodefactor(smoke)` + `nodematch(smoke)` | `smoke`

For grade 12, only `ESP(0)` is used. In the first analysis, all the statistics included in ERGM and ERNM are the same. What causes the difference between the two models is that ERNM treats the smoke indicator variable as stochastic, whereas ERGM treats it as fixed. Because of this feature of ERNM, we can, and do, fit another model to ERNM by adding the node

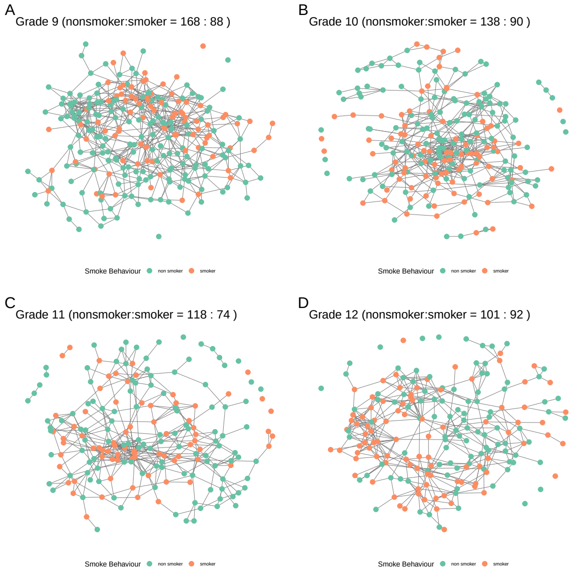

Figure 5.1: Addhealth Network Visualisation

count of smokers statistic to the first model:

- ERNM-Count: $\texttt{Network} \sim \texttt{edges} + \texttt{ESP(0, 1, 2)}$
  $+ \texttt{GWESP(0.5)} + \texttt{GWDegree(0.5)}$
  $+ \texttt{nodefactor(smoke)} + \texttt{nodematch(smoke)}$
  $+ \texttt{nodecount(smoke)} \mid \texttt{smoke}$

The `edges` term represents the overall density of the network. The `ESP` term models the lower end of the shared partner distribution. The `GWESP` term with decay parameter 0.5 is a much more robust measure of transitivity than triangles (as discussed in Section 3.1.2). The `nodefactor` term counts the number of times a node appears in an edge for each value of the atttribute. `GWDegree` stands for geometrically weighted degree distribution, specifically representing some aspects of the degree distribution (Morris et al., 2008a). The `nodematch` term on smoking behavior measures the number of edges with same smoking behaviors on two ends. In other words, it counts the edge $(i - j)$ when i and j are both smokers or non-smokers. Note that this model is saturated on smoking based mixing behavior (Handcock et al., 2021). The node count of smokers counts the number of smokers and this is specifically for ERNM because the number of smokers is invariant in ERGM.

5.3 Computational aspects

All models in this chapter are fit with the open-source user-friendly R packages `ergm` (Handcock et al., 2021) or `ernm` (Fellows, 2014; R Development Core Team, 2022). The easy availability of powerful, sophisticated community supported software allows broad accessibility of both these modeling classes for researchers. In particular, the `ergm` package is a part of the `statnet` community of packages (Krivitsky et al., 2003-2020). Together these allow robust MCMC based maximum likelihood estimation of ERGM and ERNM model parameters. In addition, they offer powerful models and computational diagnostic tools that we applied here and are available to all. We do not focus on these computational aspects here, but refer the reader to the extensive material in the references.

As we discussed earlier in Section 3.2, due to the intricate dependence feature of ERGM and ERNM, computation of approximate maximum likelihood estimates of the parameters may be complicated by model degeneracy. MCMC diagnostics are needed to check the appropriation of the model, in other words, whether the model converges. From the results of MCMC diagnostics (Appendix D), the trace plots of simulated statistics from the fitexttted model indicate low dependency and Markov chains convergent to the stationary distribution for both ERGM and ERNM. The MCMC samplers mix well.

The code to reproduce this analysis is provided in `https://github.com/Andrea-ZW/ERNM_Paper.git`.

5.4 ERGM and ERNM Fits

We show the results of ERGM fit under the suggested model of four networks (corresponding to grades 9, 10, 11, and 12) in Table 5.2. We can interpret the coefficients using the log-odds definition in Section 6.1.3. The combination of Edges, Diff-homophily-smoke and Homophily-non-smoker represents the propensity for forming a tie between all the possible combinations of smoking attributes between paring of nodes. The baseline (Edges) corresponds to a heterogeneous pairing. The Homophily-non-smoke term represents the homophily for non-smokers. All networks exhibit a positive estimated coefficient on the homophily of non-smokers (although Grade 11 is not significant). To interpret this result, taking Grade 9 as an example, the positive coefficient estimate of homophily on non-smoking (0.40) suggests that students who have not smoked are more likely to nominate as friends others who have not smoked (all else held constant). The Diff-homophily-smoke coefficients give us the differential homophily for smokers. In other words, it represents the excess (or differential) homophily for smokers over that for non-smokers. Summing the Homophily-non-smoker and Diff-homophily-smoke coefficients give us the homophily for smokers. All networks exhibit stronger homophily for smokers (although Grade 12 is only marginally significant). To interpret this result, taking Grade 9 as an example, this indicates that the homophily for smokers is more than that for non-smokers by about 0.10 on the log-odds scale and all else held constant. In other words, friendship formed between two smokers is more likely than friendship formed between two non-smokers. The result of the two terms is evidence that the homophilies between smokers and between non-smokers are both stronger than the heterogamy of smokers. In other words, non-smokers are less likely to make friends with smokers (or vice versa), all other aspects held constant. Moreover, the model indicates that the homophily of smokers is generally higher than the homophily of non-smokers.

The GWESP and ESP terms together model the edgewise shared partner distribution and represent the transitivity in the network (not represented by the other terms, especially the homophily terms). The three ESP terms account for the number of pairs of nodes having 0, 1, and 2 alter in common. The adjustment for any deviation in the lower end of the edgewise shared partner distribution for the additional transitivity implied by the GWESP term. We see that the total effects of these terms are positive and significant in all but the Grade 11 network. This indicates that there is generally transitivity above and beyond that implied by the level of homophily in the network.

From the ERGM fit, we conclude that the social connections of 9 to 12-grade adolescents generally show a tendency for transitivity. The homophily of non-smokers and smokers are both positive, with that of smokers higher than non-smokers. Adolescents with different smoking behaviors are less likely to make friends with each other.

5.4.1 ERNM Fit for the same terms as the ERGM

We fitted the same networks with ERNM with the same terms, and the results are shown in Table 5.3. As we discussed earlier, the qualitative interpretation of graph statistics of the ERNM is comparable to that of the comparable ERGM. The dyadic variables of ERNM can be interpreted under the conditional logistic regression as we show in section 6.1.3. The parameter estimations of the basic terms (Edges, ESP, GWESP, and GWDegree) of ERNM are similar to those of the ERGM fit for each of the four networks. The Homophily-non-smoker term is also very similar to the ERGM coefficient. The estimated coefficients of homophily on non-smoker are positive and significant in all four networks (except Grade 11). The standard errors are on the same scale as the ERGM (around 0.1). To interpret the coefficients, take Grade 9 as an example: the log-odds of a homogeneous tie of a non-smoker (that is a tie between non-smoker and non-smoker) versus a heterogeneous tie (that is a tie between a non-smoker and a smoker) is 0.4 higher, holding all else fixed. This suggests that there is a higher statistically significant probability of a tie between two smokers than a tie between one smoker and one non-smoker (holding all else constant). This coincides with the conclusion of ERGM. What stands out is the difference between the homophilies for smokers and non-smokers (that is, the Diff-homophily-smoke term). Unlike ERGM which has statistically significantly higher homophily for smokers compared to non-smokers, ERNM has coefficients of homophily of smokers very close to homophily of non-smokers, which suggests uniform homophily. The difference in homophily on smoking is close to zero for each grade. This suggests that there is no big difference in the tendency to make friends between non-smokers and non-smokers compared to smokers and smokers. We will discuss this result in detail in the next chapter as it sheds light on the difference between the two models.

5.4.2 ERNM fit when a count of the number of smokers is included

Another ERNM (ERNM-Count) is fitted by adding the node count of the smoke term to the first model, and the results are shown in Table 5.4. Note that this is equivalent to the model counting the number of non-smokers as the total number of nodes/students is fixed for each grade/network.

Note that the coefficients of all terms except the added term are very close to those in the ERGM of Table 5.2 (which excludes the smoker count term). The reason for this is a geometric feature of exponential family models. Consider conditioning on the number of smokers in the ERNM-Count model:

$$P_\eta(Y = y, X = x) = \frac{1}{c(\eta)} \exp\left\{\eta \cdot \mathrm{g}(y, x) + \eta_c n_{smokers}(x)\right\}$$

$$P_\eta(n_{smokers}(X) = n_{smokers}) = \frac{c(\eta; n_{smokers})}{c(\eta)}$$

$$\begin{aligned} &P_\eta(Y = y, X = x | n_{smokers}(X) = n_{smokers}) \\ =&\frac{1}{c(\eta; n_{smokers})} \exp\left\{\eta \cdot \mathrm{g}(y, x; n_{smokers})\right\} \end{aligned} \tag{5.1}$$

The equation (5.1) represents the exponential family form of the ERNM-Count model. By specifying the number of smokers, it acts as if conditioning on the nodal attributes X representing the smoking behavior. Recall the functional form of ERGM (2.2), and we would expect the parameter estimates η of ERGM and ERNM-Count relating to homophilous smoking to be close.

As Table 5.4 shows, the majority of the results from four networks are consistent with the previous two models, except for the homophily terms. In particular, we observe differential homophily on smoking in ERNM-Count (that is the statistically significant positive estimated coefficients of Diff-homophily-smoker term). A positive coefficient suggests that the homophily of smokers is more than the homophily of non-smokers. The Nodecount-smoker term is significant (or marginally significant in Grade 12) with negative coefficient estimates. Taking Grade 9 as an example, we see that the log-odds of a student being a smoker is -0.62, holding the rest constant. This suggests that there are fewer smokers than expected based on the social structure of the friendship ties.

Comparing the two models of ERNM, the discrepancy in homophily measures gives an illuminating finding. Without the node count on smokers, we find uniform homophily between smokers and non-smokers. However, after taking into account the node count of smokers, the homophily of smokers is driven up, for example, 0.38 to 0.49 in Grade 9, which indicates that the homophily of smokers is more than the homophily of non-smokers. And the conditional log-odds of a tie for a smoker who chooses a smoker is 0.09 higher after adding the node count of the smoker.

5.5 Assessing Goodness of Fit

It is necessary to check the goodness of fit (GOF) of fitted models to verify the rationality of the model. Hunter et al. (2008) introduced a procedure for goodness of fit, which generates simulations on target network statistics and compares them to the observed graph statistics.

The construction of the tests is as follows: for both ERGM and ERNM, we generate 1000 simulations from the fitted models and compare their simulated distributions to the observed statistics. The GOF plot (Appendix D) consists of statistics that are included in the model and statistics (Degree(0:20), ESP(3:10)) that are not included in the model. We provide side-by-side box plots, including the statistics mentioned above, in order to compare the models without the node count term fitted by ERNM and ERGM. We also compare the model with the node count of the smoker fitted by ERNM and without the node count model fitted by ERGM. Both ERGM and ERNM simulate distribution aligning closely to the observed statistics, which is under expectation. Although there are some deviations in the degree distribution, it does an adequate job, as no degree terms are included in the model (except for the GWDegree). The comparison of the count of smokers and non-smokers between ERNM and ERNM-Count is shown in Figure C.1 in Appendix C. The model with full freedom of varying smoking covariates (ERNM) thinks there are more smokers than observed, whereas ERNM-Count and observed statistics suggest that we actually have fewer smokers than expected due to chance.

Table 5.2: Summary of the fit of the ERGM in Section 5.2 on four grades

	ERGM			
Coefficient	Grade 9	Grade 10	Grade 11	Grade 12
Edges	-8.97(2.99) **	-19.55(3.06)***	9.04(3.16) **	-6.73(0.28) ***
ESP-0	3.35(2.99)	14.01(3.05) ***	3.87(3.15)	1.21(0.23) ***
ESP-1	0.67(1.11)	4.77(1.17) ***	0.81(1.18)	NA
ESP-2	-0.23(0.40)	1.38(0.45) **	-0.05(0.43)	NA
GWESP	3.85(1.88) *	10.44(1.89) ***	4.06(1.97) .	2.40(0.18) ***
GWDegree	2.98(0.48) ***	1.58(0.32) ***	1.66(0.35) ***	1.73(0.33) ***
Diff-homophily-smoke	0.10(0.02) ***	0.06(0.01) ***	0.10(0.03) ***	0.06(0.03) .
Homophily-non-smoker	0.39(0.06) ***	0.40(0.06) ***	0.12(0.08)	0.33(0.08) ***

Homophily-non-smoker is the number of ties between nodes with the same smoker activity;

Diff-homophily-smoke is the number of ties incident on a non-smoking node;

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$

Table 5.3: Summary of the fit of the ERNM in Section 5.2 on four grades

	ERNM			
	Grade 9	Grade 10	Grade 11	Grade 12
Edges	-7.95(2.95) ***	-19.36(3.12) ***	-9.54(3.19) ***	-6.82(0.27) ***
ESP-0	2.23(2.94)	13.79(3.12) ***	4.24(3.18)	1.24(0.24) ***
ESP-1	0.28(1.10)	4.69(1.20) ***	0.94(1.20)	NA
ESP-2	-0.33(0.40)	1.36(0.47) ***	-0.01(0.45)	NA
GWESP	3.12(1.85) *	10.30(1.93) ***	4.30(1.99) **	2.42(0.18) ***
GWDegree	3.05(0.48) ***	1.59(0.33) ***	1.63(0.35) ***	1.68(0.33) ***
Diff-homophily-smoke	0.01(0.03)	0.02(0.03)	-0.01(0.04)	0.00(0.04)
Homophily-non-smoker	0.37(0.08) ***	0.40(0.08) ***	0.14(0.11)	0.35(0.09) ***

Homophily-non-smoker is the number of ties between nodes with the same smoker activity;

Diff-homophily-smoke is the number of ties incident on a non-smoking node;

$*** \quad p < 0.001, \; ** \quad p < 0.01, \; * \quad p < 0.05, . \;\; p < 0.1$

Table 5.4: Summary of the fit of the ERNM including the count of the number of smokers on four grades (Section 5.2)

	ERNM-Count			
	Grade 9	Grade 10	Grade 11	Grade 12
Edges	-8.80(2.95) ***	-19.56(3.12) ***	-8.97(3.22) ***	-6.73(0.28) ***
ESP-0	3.18(2.94)	14.02(3.11) ***	3.80(3.20)	1.21(0.24) ***
ESP-1	0.62(1.10)	4.77(1.19) ***	0.78(1.21)	NA
ESP-2	-0.24(0.40)	1.38(0.47) ***	-0.05(0.45)	NA
GWESP	3.73(1.85) **	10.45(1.93) ***	4.02(2.01) **	2.40(0.18) ***
GWDegree	2.97(0.47) ***	1.58(0.32) ***	1.66(0.35) ***	1.72(0.34) ***
Diff-homophily-smoke	0.10(0.03) ***	0.06(0.03) **	0.10(0.05) **	0.06(0.05)
Homophily-non-smoker	0.39(0.08) ***	0.40(0.08) ***	0.11(0.11)	0.34(0.09) ***
Nodecount-smoker	-0.62(0.20) ***	-0.45(0.18) **	-0.85(0.25) ***	-0.42(0.24) *

Homophily-non-smoker is the number of ties between nodes with the same smoker activity;
Diff-homophily-smoke is the number of ties incident on a non-smoking node;
Nodecount-smoker is the number of students that report smoker activity;
$*** \quad p < 0.001, \; ** \quad p < 0.01, \; * \quad p < 0.05, . \; p < 0.1$

CHAPTER 6

Latent Class Exponential-family Random Network Model

The random graph describes the probability distribution on the graphs. For a social network (X, Y), with n nodes, $Y \subset \mathbb{R}^{n\times n} \in \mathcal{Y}$ is the graph with $X \subset \mathbb{R}^{n\times n\times q} \in \mathcal{X}$ as dyadic attributes. The space of tie variables, $\mathcal{Y}$, can be arbitrary. Although here we explicitly consider binary tie variables. Here

$$Y_{ij} = \begin{cases} 1 & \text{if actor } i \text{ is connected to actor } j \\ 0 & \text{otherwise} \end{cases}, \tag{6.1}$$

$i, j = 1, \cdots, n$. For undirected network, $Y_{ij} = Y_{ji}$. Y is often called an adjacency matrix. The dyadic covariates, X_{ijk} are measures on the $(i, j)^{th}$ pair. Examples include a homophily term:

$$X_{ijk} = \begin{cases} 1 & \text{if actor } i \text{ and actor } j \text{ have identical values} \\ & \text{on the } k^{th} \text{ nodal characteristic} \\ 0 & \text{otherwise} \end{cases}, \tag{6.2}$$

$k = 1, \cdots, q$, for each of q separate nodal characteristics.

6.1 Latent Class ERNM

6.1.1 ERNM Specification

Exponential-family random network models (ERNMs) generalize ERGMs by treating the nodal attributes as endogenous variables (Fellows and Handcock, 2012). This was inspired by Leenders (1997), which argued that the process of social selection and nodal attributes influence are simultaneous. ERNMs model the joint relationship between edges and nodal variates. The ERNM distribution for Y is

$$\begin{aligned} &P_\eta(Y=y, X=x) \\ =&\frac{1}{c(\eta,\mathcal{N})}\exp\{\eta\cdot g(y,x)\} \quad (y,x)\in\mathcal{N}, \end{aligned} \tag{6.3}$$

where $\mathcal{N}$ is the sample space of Y and X, $\eta \in \Lambda$ is a q-vector of parameters, $g(y,x)$ is a q-vector of network statistics, with $g(Y,X)$ jointly sufficient for the model, and $c(\eta,\mathcal{N})$ is the normalization constant. The formal definition of $c(\eta,\mathcal{N})$ is given in Fellows and Handcock (2012): Let $(N,\mathcal{N},P_0)$ be a σ-finite measure space with reference measure P_0, then, a probability measure to this space is an ERNM if it is dominated by P_0. The normalization constant is defined as

$$c(\eta,\mathcal{N}) = \int_{y,x\in\mathcal{N}} \exp\{\eta\cdot g(y,x)\} dP_0(y,x), \tag{6.4}$$

where $\Lambda \subset \{\eta\in\mathbb{R}^q : c(\eta,\mathcal{N})<\infty\}$.

6.1.2 Latent Variables

Many network characteristics are unobservable; for example, the class of each node belongs to. When the group of each node in the network is latent, the stochastic block models (SBMs) are widely used tools for detecting community structure in networks. The most recent degree-corrected block models (DCBMs) include heterogeneity in the degrees of vertices (Karrer and Newman, 2011). Unlike ordinary SBMs, which assume a homogeneous edge distribution with the probability of relationship only depending on their group memberships, DCBMs incorporate the degree-corrected parameters into the edge distribution. Hence, under DCBMs, high-degree nodes are more likely to be connected than those low-degree nodes, holding everything else constant.

6.1.3 Model Interpretation

To interpret the coefficient of ERNM, consider the logit form of exponential family models (6.3):

$$\begin{aligned}
&\text{logit}\left(P_\eta(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c, X = x)\right) \\
&= \eta \cdot \left(g(y_{ij}^+, x) - g(y_{ij}^-, x)\right),
\end{aligned} \tag{6.5}$$

where y_{ij}^c is the set of tie values $y \backslash y_{ij}$, y_{ij}^+ and y_{ij}^- correspond to the graphs $(y_{ij}^c, y_{ij} = 1)$ and $(y_{ij}^c, y_{ij} = 0)$, respectively, Y_{ij}^c is the random variable $Y \backslash Y_{ij}$. This is often referred to as the conditional log-odds of a tie Y_{ij}. We see that η has the interpretation of the change in conditional log-odds of a tie Y_{ij} per unit change in the graph statistics were y_{ij} toggled from zero to one.

Fellows and Handcock (2012) discusses the way to interpret coefficients of dyadic variables of ERNM with logistic regression. The dyadic attributes X in equation (6.3) can be partitioned into two parts, Z and T by defining $Z \in 0, 1$ as a binary dyadic variate of interest (that is, an outcome variable) and T as a matrix of regressors, where $Z, T \subset X$. We can rewrite the equation (6.3) with the new definition:

$$\begin{aligned}
&P_\eta(Y = y, Z = z, T = t) \\
=&\frac{1}{c(\eta, \mathcal{N})} \exp\left\{\alpha \cdot g(y, t) + \lambda \cdot h(y, z) + z \cdot t\beta\right\} \\
&(y, x) \in \mathcal{N},
\end{aligned} \tag{6.6}$$

where $\eta = (\alpha, \beta, \lambda)$ are parameters, $g(y, x)$ and $h(y, z)$ are network statistics, $z \cdot t\beta$ is the relationship of T to Z. We can then derive the logit form of the distribution of z_{ij} from equation (6.6) condition upon the rest of the network:

$$\begin{aligned}&\text{logit}\left(P_\eta(z_{ij}=1|z_{ij}^c, t_{ij}, Y=y)\right)\\&=(t_{ij}\beta)-\left(\lambda\cdot(h(y,z_{ij}^-)+h(y,z_{ij}^+))\right),\end{aligned}\tag{6.7}$$

where z$_{ij}$ and t_{ij} are the measures on the $(i,j)^{\text{th}}$ pair of Z and T, z_{ij}^c is the set of variants $z\setminus z_{ij}$, z_{ij}^+ and z_{ij}^- correspond to the the variant of z_{ij} where $z_{ij}=1$ and $z_{ij}=0$, respectively. Suppose the matrix of regressors t_{ij} changes to $t_{ij}^{'}$, with all other variables and networks remaining fixed, then the logarithm of odds ratio (R) is

$$\ln R=\beta\cdot(t_{ij}-t_{ij}^{'}).\tag{6.8}$$

Therefore, the coefficients of the outcome variable z may be interpreted as a conditional logistic regression model. For one unit change in t_{ij}, the log-odds changes by β, keeping all other variables constant.

6.1.4 Inference on Groups

An advantage of latent ERNM is that we can investigate the probability of cluster group, which is well defined through the marginal distribution $P(X=x|Y=y_{obs},\eta)$. To compute $P(X=x|Y=y_{obs},\eta)$, we can simulate a large number of samples from $P(X=x|Y=y_{obs},\hat{\eta})$ using MCMC to show the probability of each node being in each cluster.

To better visualize the resulting cluster distribution, we create a network plot framework where each node is shown as a pie chart representing the probability of being classified into each group computed from the simulations, and the network layout consists of the original network layout (Handcock et al., 2007).

6.2 SBM

The Stochastic Block Model (SBM) introduced by Nowicki and Snijders (2001) is an important statistical model for clustering problems. It models the probability of the change of an edge using their group memberships only. Given the social network (X, Y) defined in (6.1) and (6.2), we also assume $Q \in \mathbb{N}$ classes on node ,and define the membership matrix Z, where

$$Z_{iq} = \begin{cases} 1 & \text{if actor } i \text{ is of class } q \\ 0 & \text{otherwise} \end{cases}, \tag{6.9}$$

$i = 1, \cdots, n; q = 1, 2, \cdots, Q.$

Stochastic block models are extensively developed for clustering problems. However, we did not find an applicable Degree-Corrected Stochastic Block Model to perform the clustering analysis. An alternative model we used is simple SBM in R package `SBM`. The mathematical background is provided in Leger (2016). It constructs the simple SBM with a latent layer which represents the distribution of the class membership, and an observed layer for general SBM. The latent class membership is independent identically distributed under multinomial distribution:

$$Z_i \sim \mathcal{M}(1, \alpha),$$

where $i = 1, \cdots, n$ and α is a parameter satisfying the constraint $\sum_{q=1}^{Q} = 1$. The observed layer describes the distribution of each edge conditional upon the membership assignment of two end nodes independently:

$$Y_{ij} | Z_{iq} Z_{jl} = 1 \sim \mathcal{F}_{ql}(X_{ij}),$$

where $i, j = 1, \cdots, n; i \neq j; q, l = 1, \cdots, Q$. $\mathcal{F}$ depends on the choice of model distributions, such as Bernoulli, Poisson, and Gaussian, without or with covariates.

It presents an implementation of a Variational EM algorithm for the Stochastic Block Model for some common probability functions, Bernoulli, Gaussian, and Poisson, without or with covariates (that is, the model for the relationship between nodes). This package implements group number exploration and selection via the Integrated Completed Likelihood (ICL) (Biernacki et al., 2000) automatically. The ICL criterion is essentially the ordinary BIC penalized by the subtraction of the estimated mean entropy. During the estimation, models with different numbers of blocks are explored, and the best model with the lowest ICL will be displayed.

6.3 Computational aspects

All models in this chapter are fit with the open-source user-friendly R packages `ernm` (Fellows, 2014; R Development Core Team, 2022) or `SBM`. The easy availability of powerful, sophisticated community supported software allows broad accessibility of both these modeling classes for researchers. For certain latent ERNM models (Model 2 in Section 6.4.3), the tapered version ERNM is used.

6.4 Case Study–Sampson's Monastery Data

The case study is performed on the sampson monastery dataset among 18 monks in a New England monastery. Sampson (1968) collected the social interactions among the monks by asking about each monk's positive relations to other monks and recorded their top three choices. A direct edge from monk A to monk B is formed if A assigns B as one of his top choices. The vertex names are unique to the 18 monks, and they are classified into three groups:

1. Outcasts: Elias, Simplicius, Basil, Amand;

2. Turks: John Bosco, Gregory, Winfrid, Albert, Boniface, Hugh, and Mark;

3. Loyal: Peter, Bonaventure, Berthold, Victor, Ambrose, Louis, Romuald.

The data was gathered across three time points, and we use a time-aggregated version, called `samplike` in R, to implement the study.

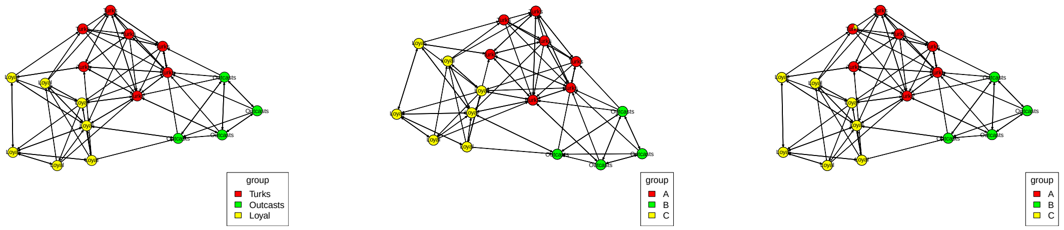

Figure 6.1: Sampson Network: Left: Original; Middle: Model 1; Right: Model 2

6.4.1 Latent ERNM Fits

The latent ERNM model is fitted on the sampson data, where the dyadic variable (`latent`) is set to be latent with three levels:

- Latent ERNM 1: `Network` $\sim$ `edges` + `reciprocity` + `homophily(latent)` + `nodecount(latent)` | `latent`
- Latent ERNM 2: `Network` $\sim$ `edges` + `reciprocity` +`nodematch(latent)` +`nodefactor(latent)` + `nodecount(latent)` | `latent`

The statistics consist of the definition in Chapter 5. Moreover, reciprocity counts for the number of reciprocated ties, and it can only be applied to a directed network. The homophily term is a regularised version nodematch term defined in Fellows and Handcock (2012). Model 2 is fitted with tapered version ERNM with tapering parameter 2.

The summary statistics are shown in Table 6.1. The coefficient estimates, and standard errors of edges and reciprocity terms are very close between the two models. For model 1, the regularised homophily term is significant with a positive coefficient estimate, which suggests a high tendency of connection between nodes with the same latent attributes. The Nodematch term in ERNM 1 is also the uniform homophily with a significant positive coefficient estimate. The Nodefactors on the latent attribute in Model 2 account for the differential homophily for latent attribute level 2 over that for level 1 and level 3 over that for level 1, respectively. Both terms are not statistically significant at 1% significance level, mainly due to the low variation in the degree distribution. The Nodecount terms on the latent attribute are both negative and insignificant in model 1, whereas they are highly significant with positive coefficients in model 2.

Table 6.1: Summary of the Latent Class ERNM fit to Sampons Monk's Data

	Latent Class ERNM	
Coefficient	Model 1	Model 2
Edges	-1.15(0.24) $***$	-1.50(0.37) $***$
Reciprocity	1.37(0.48) $**$	1.77(0.51) $***$
Homophily-latent	5.67(0.92) $***$	NA
Nodematch-latent	NA	1.43(0.30) $***$
Nodefactor-latent.1	NA	-1.10(0.44) $*$
Nodefactor-latent.2	NA	-1.09(0.59) .
Nodecount-latent.1	-0.02(1.14)	8.04(2.55) $**$
Nodecount-latent.2	-2.12(1.31)	9.16(2.51) $***$

$***$ $p < 0.001$, $**$ $p < 0.01$, $*$ $p < 0.05$, . $p < 0.1$

To assess the Goodness of Fit, we refer to the method in Section 5.5. We have generated 1000 simulations from the fitted latent ERNM. Then we plot the original network layout with the cluster assignment probabilities for each actor shown as pie charts illustrated in Section 6.1.4. Model 1 achieves almost 100% accuracy in classifying the correct groups shown in

Model	Model 1	Model 2
logLLH	15.43864	54.07796
BIC	-8.490592	-76.81457

Table 6.2: Sampson: Latent Class ENRM BIC Comparison

Figure 6.1 - Middle. From Figure 6.1 - Right, the correct classification of Model 2 is almost 100% except for one node, which is in group A with 80% chance and in group C for the rest of time. It initially is Turks, but among 1000 simulations, Model 2 classifies it as a Loyal 180 times. One explanation is it has some in-edges from Loyal group nodes. The log-likelihood and BIC are calculated for each model (Table 6.2). Model 2 gives a smaller BIC.

6.4.2 Simple SBM Fits

We fit the simple SBM with Bernoulli density to Sampson's data. The automated ICL suggests the SBM with two groups (SBM-2) shown in the left plot of Figure 6.2. The SBM with three blocks (SBM-3) (the actual number of groups) yields a larger ICL than the SBM-2. Moreover, if we sample from the fitted SBM-2 and SBM-3, the resulting blocks suggest that SBM-2 classifies original Loyal and Outcasts into one group and original Turks into the other group (Figure 6.2, Middle), and SBM-3 gives exactly same memberships distribution as the original network (Figure 6.2, Right).

6.4.3 Comparison Between the Two Methods

Sampson Monastery data has a well-classified group structure. The latent ERNM did a nice job of classifying the group membership, which generates three groups with correct membership in all simulations from the fitted model. Unfortunately, the simple SBM with the Bernoulli model suggests the SBM-2 with two groups according to the ICL. Moreover, as we mentioned in Section 6.1.4, we can calculate the probability of monks being in the true

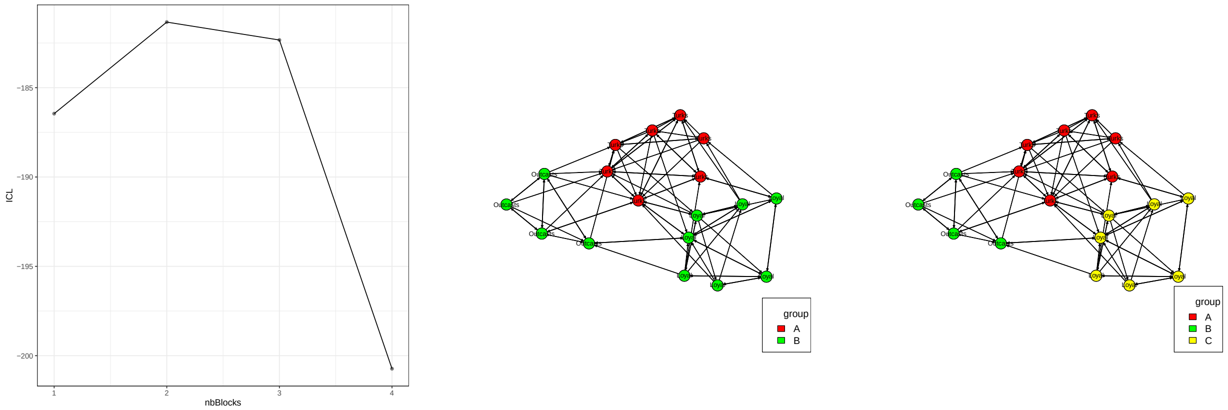

Figure 6.2: Sampson: Left: SBM ICL; Middle: SBM with 2 groups; Right: SBM with 3 groups

group, $P(X = x|Y = y_{obs}, \eta)$, and the accuracy is almost 100%.

6.5 Case Study–Addhealth Data

The Addhealth data is from the National Longitudinal Study of Adolescent Health (Harris et al., 2007) (Add Health). This nationally representative study includes a longitudinal sample of more than 20,000 adolescents in grades 7 to 12 who were surveyed with in-school questionnaires in the US in 1994 and 1995.

In this paper, we analyse a single school of 71 adolescents from grades 7–12. The network connections are defined as the nomination of friendships. The self-nomination is disallowed, and the isolated nodes are not considered.

6.5.1 Simple SBM Fits

We fit the Addhealth data with the simple SBM. The SBM with the Bernoulli model suggests the best SBM with four levels (Figure 6.3) according to the ICL criterion (Figure 6.6). The SBM with four groups (SBM-4) classifies most of the nodes from the original grade 7 as one group, the original grade 8 and some of grade 9 as another group, the rest of the nodes from the original grade 9 as a 3rd group, and the original 10, 11, and 12 as the 4th group. The SBM with five groups (SBM-5) further splits the second group in SBM-4, where grade 8 and grade 9 are previously grouped together into two groups (Figure 6.4). In SBM-6, some of the nodes from the original grade 11 are classified as a new group based on SBM-5 (Figure 6.5).

From the three simulated network plots, a drawback of SBM is that it fails to account for the uncertainty in the group memberships. For instance, in SBM-4, the original grade 10, 11, and 12 are classified as one group without any uncertainty. However, if we turn to SBM-6, some nodes originally from grade 11 are classified into a new group. In future works, we shall implement the latent ERNM on the Addhealth data and compare it to the simple SBM.

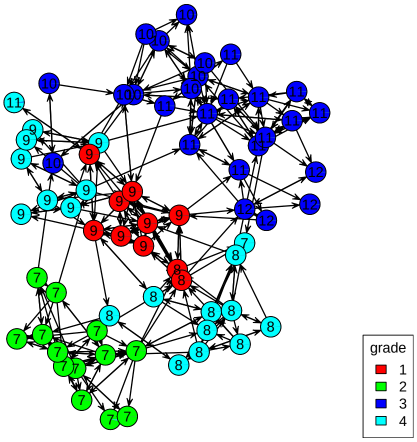

Figure 6.3: Addhealth: SBM 4 groups

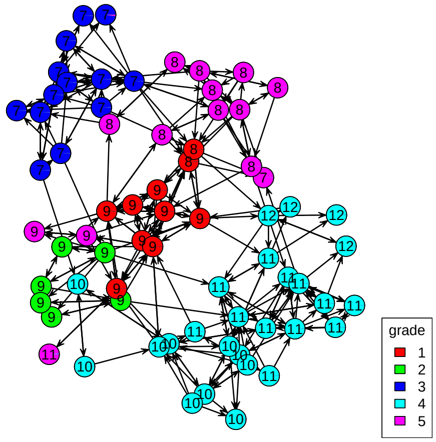

Figure 6.4: Addhealth: SBM 5 groups

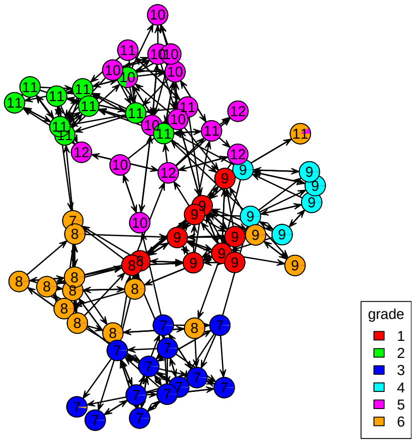

Figure 6.5: Addhealth: SBM 6 groups

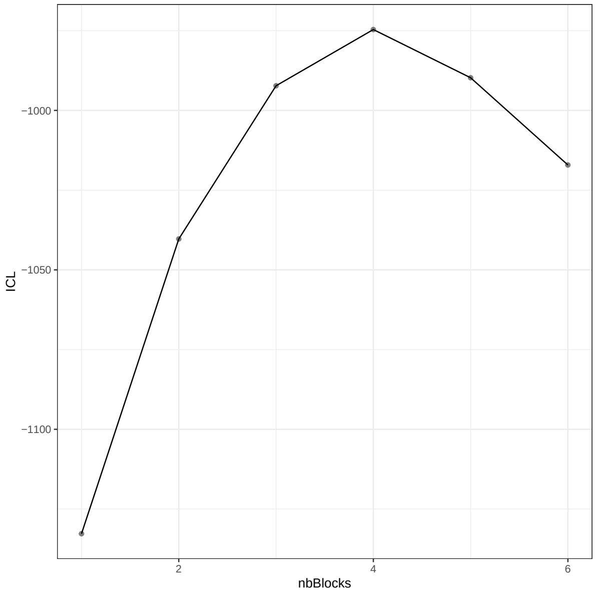

Figure 6.6: Addhealth: SBM ICL

CHAPTER 7

Summary

7.1 Discussion on ERNM Classes

Despite the wide success of exponential-family random graph models in representing complex network data, they treat the nodal and dyadic covariates as exogenous. This is not true for many realistic social processes. In this paper, we show that this treatment misspecifies the social structure of network processes. We also provide evidence that Exponential-family random network models represent a much better class of models for representing processes with nodal and dyadic covariates that are endogenous to the tie variables. ERNMs have many advantages. First, they are also in the exponential-family class of models, which have been shown to be able to represent complex social structures. Exponential-family classes of models have been extensively studied, and their properties have been explored. Because of this, the extensive knowledge and software platforms that have been developed for ERGM can, and have been, extended to ERNM.

In this paper, we compare ERGM and ERNM, with a special interest in situations where at least some of the covariates are stochastic. We use as a case-study: a friendship network among students within a school from the National Longitudinal Study of Adolescent Health. Within this network, the student's smoking behavior is likely endogenous to their friendship ties. Both ERNM and ERGM models represent the four friendship networks well, as evidenced by the goodness-of-fit and MCMC diagnostics. The coefficient estimates and interpretations of the ERGM and ERNM are very similar after adding the node count term into ERNM (Table 5.2 and Table 5.4). Although both models show significant differential homophilies on smoke, the node count term on smoke is a notable addition to the ERNM fitting. Combining the result of ERNM and ERNM-Count model, we find there are fewer smokers in the network than expected due to chance: the simulation results of the ERNM model gives a higher number of smokers compared to the observed statistics and the ERNM-Count (Figure C.1) simulated statistics; The negative coefficient of the node count variable in the ERNM-Count model also suggests this. It illustrates the importance of treating smoking status as endogenous, rather than exogenous as in ERGM.

The impact of the endogeneity is throughout the model. As the case study shows, primary properties such as the presence of differential homophily can be misspecified by ERGM. Coefficients can be both under and overestimated by ERGM and standard errors can be affected both ways. In our case-study, this can be seen by comparing the coefficients and standard errors in Table 5.2 (ERGM) to those in Table 5.3 (ERNM).

The findings from the ERGM and ERNM are fundamentally different: Either the smokers have similar homophily to non-smokers (Table 5.4), or the number of smokers is lower than we would expect (Table 5.4). This is missed by the ERGM and clear (and statistically significant) in the ERNM. Note that the ERGM model is conceptually wrong in this case as it is a pure social selection model where ERNM allows both social selection and social influence.

On the one hand, based on the result of ERNM and ERGM, this is consistent with a process of social selection. Smokers tend to connect with smokers, and non-smokers tend to connect with non-smokers. If this argument holds empirically, then this would be a selection causal mechanism that leads to it. However, there might be other tie formation processes, such as different social contexts (Feld, 1981). Take a simple example: since smoker A and smoker B go to the same tobacco shop, they meet there a lot of times and become friends. This cannot be seen as a causal relationship of the homophily, rather, it is the social context that leads to the formation of friendship. Hence, we can conclude an association between nodal attributes and social networks, instead of causality. On the other hand, AddHealth networks can be explained by the process of social influence. Given the result of ERNM-Count, smokers who are connected with non-smokers may choose to quit smoking and become non-smokers. If this holds, then this would be an influence causal mechanism that leads to it, which is a mechanism that only involves influence. Steglich et al. (2010) has verified the existence of such a mechanism. However, it may also be possible that smokers quit smoking because they want to make connections with non-smokers. Hence, the tie formation is a homophily or social selection process. Although we cannot disentangle the social selection and influence mechanisms, based on our conceptual knowledge, we tend to believe that the social influence process is less credible under this context, and it is much more likely that it is the selection mechanism that dominates.

The availability of powerful user-friendly open-source software allows broad accessibility and use of both ERGM and ERNM (Krivitsky et al., 2003-2020; Fellows, 2014). The analysis in this paper supports the notion that ERNM is preferred when networks have stochastic covariates.

Finally, we note the ERNM provides a way to specify the complex dependency structures that would empower autologistic actor attribute models (ALAAM) (Robins et al., 2001b). This connection to ALAAM will be the topic of future investigation.

7.2 Discussion on Latent Class Models

One key question in network data analysis is to identify the clusters, which are subsets of the nodes in the network and are generally unknown. This can be generalized to questions when network exhibits unobservable quantities (ie. nodal attributes).

Stochastic Block Model are well-known models for clustering problems (Fienberg and Wasserman, 1981; Holland et al., 1983). With known cluster groups, classic models the distribution of the vertices' relationship Y_{ij} depends only on their cluster groups, and are irrelevant to their degree distribution. In other words, the nodes from same group have the same probabilities of relationships with all other nodes. Wasserman and Anderson (1987) and Snijders and Nowicki (1997) introduced a posteri blockmodelling which enables the cluster groups to be unknown in SBMs. However, SBMs perform poorly in many applications to real world social and biological networks. The model assumption of homogeneous degree distribution is the most factor to be blamed. Degree-Corrected Stochastic Block Models (DCSBMs) introduces a set of degree-correction parameters, so nodes with larger parameter values are expected to have higher connections than the nodes with smaller parameter values within the same group (Dasgupta et al., 2004; Lei and Rinaldo, 2015). However, the development of DCSBMs stays more on theoretical and algorithm levels, and we did not find a compatible package to perform.

Classic SBMs fails to represent degree distribution of the network, moreover, they struggle to capture the transitivity and homophily by attributes (Handcock et al., 2007). Holland and Leinhardt (1981) introduced the p1 model, which models the whole adjacency matrix, Y, by considering jointly the reciprocity of the network and the differential connections of each node. The ties between each pair of nodes are independently modelled from every other dyads. However, p1 model not only fails to capture clusters, but also the transitivity or homophily of networks. The popular Exponential-family Random Graph Models (ERGMs) (Frank and Strauss, 1986; Snijders et al., 2006) though generalize the p1 model to allow dyads exhibit Markovian dependence, it can only represent the transitivity of the graph. Although the subsequent development of ERGMs builds the homophily statistic by attributes, it requires the attributes to be observable. Since the clustering structure of many social networks are influenced by the homophily on some unobserved attributes (Wasserman and Faust, 1994), this limits the usage of ERGM in analysing the clustering structure of network. The Latent position cluster models (Handcock et al., 2007) posit the existence of unobservable continuous nodal quantities that provide a spatial geometry for the network structure.

Exponential-family Random Network Models (ERNMs) Fellows and Handcock (2012) models the social connections and nodal attributes jointly. It is exponential-family model, hence inherits the some good properties: models with the dyadic dependency, includes dyadic level and attributes level statistics. Moreover, since ERNMs treat the nodal attributes as endogenous variables, it can make inference on the nodal attributes, that is cluster groups in this paper. The random nodal attributes can be observed or latent in ERNM. When the attributes is treated as latent, the homophily of unobserved nodal attributes can be analysed.

Under latent ERNM framework, we can investigate the probability of cluster group, which is well defined through the marginal distribution $P(X = x|Y = y_{obs}, \eta)$. We have also built a pie chart network visualisation to achieve a clear understanding of group classification. In Sampson Monk's case study, we build the latent ERNM and the simple SBM with Bernoulli model. Latent ERNM suggests the model with 3 groups with high accuracy, where as the default SBM suggests the model with 2 groups. SBMs are also inefficient in measuring the uncertainty in the group memberships. For future work, we will fit the latent ERNM on Addhealth data.

APPENDIX A

ERNM Logit Form for Edge Connections

$$
\begin{aligned}
\mathbb{P} &\equiv P_\eta(y_{ij} = 1 | x, y_{ij}^c) \\
&= \frac{P_\eta(y_{ij} = 1, x, y_{ij}^c)}{P_\eta(x, y_{ij}^c)} \\
&= \frac{P_\eta(y_{ij} = 1, x, y_{ij}^c)}{P_\eta(y_{ij} = 1, x, y_{ij}^c) + P_\eta(y_{ij} = 0, x, y_{ij}^c)} \\
&= \frac{\exp \eta \cdot g(y_{ij}^-, y_{ij} = 1, x)}{\exp\{\eta \cdot g(y_{ij}^-, y_{ij} = 1, x)\} + \exp\{\eta \cdot g(y_{ij}^-, y_{ij} = 0, x)\}} \\
&= \frac{1}{1 + \exp\left\{\eta \cdot \left(g(y_{ij}^-, x) - g(y_{ij}^+, x)\right)\right\}},
\end{aligned}
\tag{A.1}
$$

$$
\begin{aligned}
\text{logit}\mathbb{P} &= \log \frac{\mathbb{P}}{1 - \mathbb{P}} \\
&= \log \frac{\frac{1}{1+\exp\left\{\eta\cdot\left(g(y_{ij}^-,x)-g(y_{ij}^+,x)\right)\right\}}}{1 - \frac{1}{1+\exp\left\{\eta\cdot\left(g(y_{ij}^-,x)-g(y_{ij}^+,x)\right)\right\}}} \\
&= \log \exp\left\{-\eta \cdot \left(g(y_{ij}^-, x) - g(y_{ij}^+, x)\right)\right\} \\
&= \eta \cdot \left(g(y_{ij}^+, x) - g(y_{ij}^-, x)\right).
\end{aligned}
\tag{A.2}
$$

APPENDIX B

ERNM Logit Form for Dyadic Attributes

$$
\begin{aligned}
\mathbb{P} &\equiv P_\eta(z_{ij} = 1 | z_{ij}^c, t_{ij}, Y = y) \\
&= \frac{P_\eta(z_{ij} = 1, z_{ij}^c, t_{ij}, Y = y)}{P_\eta(z_{ij}^c, t_{ij}, Y = y)} \\
&= \frac{P_\eta(z_{ij} = 1, z_{ij}^c, t_{ij}, Y = y)}{P_\eta(z_{ij} = 0, z_{ij}^c, t_{ij}, Y = y) + P_\eta(z_{ij} = 1, z_{ij}^c, t_{ij}, Y = y)} \\
&= \frac{\exp\left\{t_{ij}\beta + \alpha g(y, t_{ij}) + \lambda(h(y, z_{ij} = 1) + h(y, z_{ij}^c))\right\}}{\exp\left\{t_{ij}\beta + \alpha g(y, t_{ij}) + \lambda(h(y, z_{ij} = 0) + h(y, z_{ij}^c))\right\} + \exp\left\{\alpha g(y, t_{ij}) + \lambda(h(y, z_{ij} = 1) + h(y, z_{ij}^c))\right\}} \\
&= \frac{\exp\{t_{ij}\beta\}}{\exp\{t_{ij}\beta\} + \exp\{\lambda(h(y, z_{ij} = 0) + h(y, z_{ij} = 1)\}} \\
&= \frac{1}{1 + \exp\left\{\lambda(h(y, z_{ij}^-) + h(y, z_{ij}^+)) - t_{ij}\beta\right\}},
\end{aligned}
\tag{B.1}
$$

$$
\begin{aligned}
\text{logit}\mathbb{P} &= \log \frac{\mathbb{P}}{1 - \mathbb{P}} \\
&= \log \frac{\frac{1}{1+\exp\left\{\lambda(h(y,z_{ij}^-)+h(y,z_{ij}^+))-t_{ij}\beta\right\}}}{1 - \frac{1}{1+\exp\left\{\lambda(h(y,z_{ij}^-)+h(y,z_{ij}^+))-t_{ij}\beta\right\}}} \\
&= \log\exp\left\{-\left(\lambda(h(y, z_{ij}^-) + h(y, z_{ij}^+)) - t_{ij}\beta\right)\right\} \\
&= (t_{ij}\beta) - \left(\lambda(h(y, z_{ij}^-) + h(y, z_{ij}^+))\right).
\end{aligned}
\tag{B.2}
$$

APPENDIX C

Smoke Label Comparison

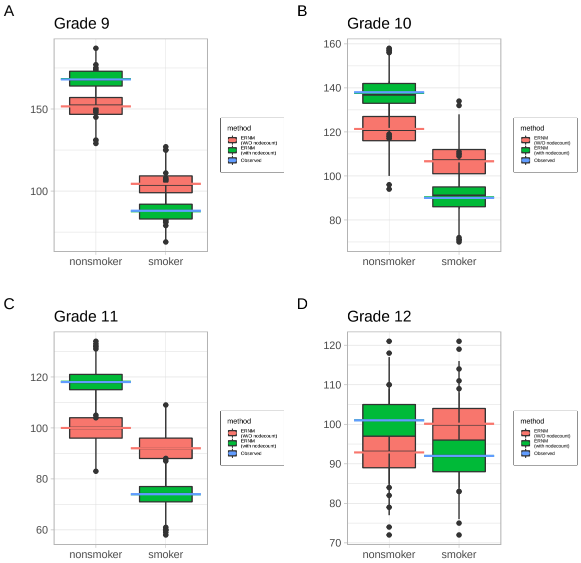

Figure C.1: Comparison between two ERNMs of count of smokers and non-smokers

APPENDIX D

Goodness of Fit

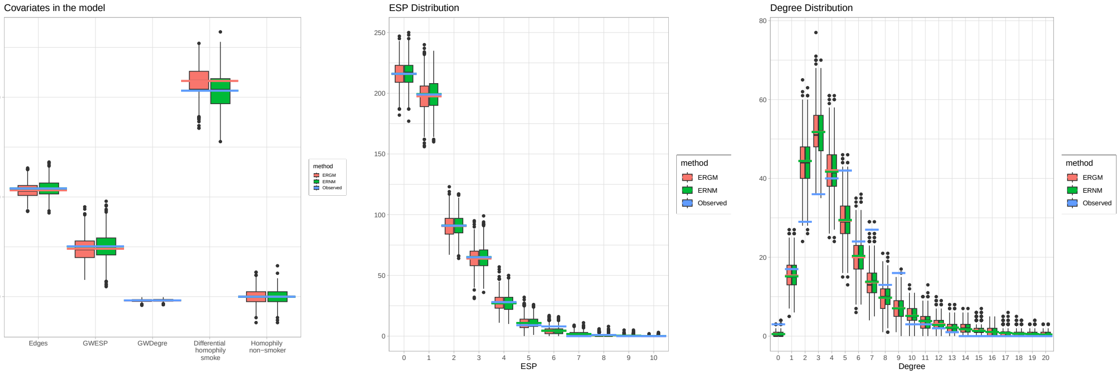

Figure D.1: GOF Comparison of ERGM and ERNM (Model 1): Grade 9

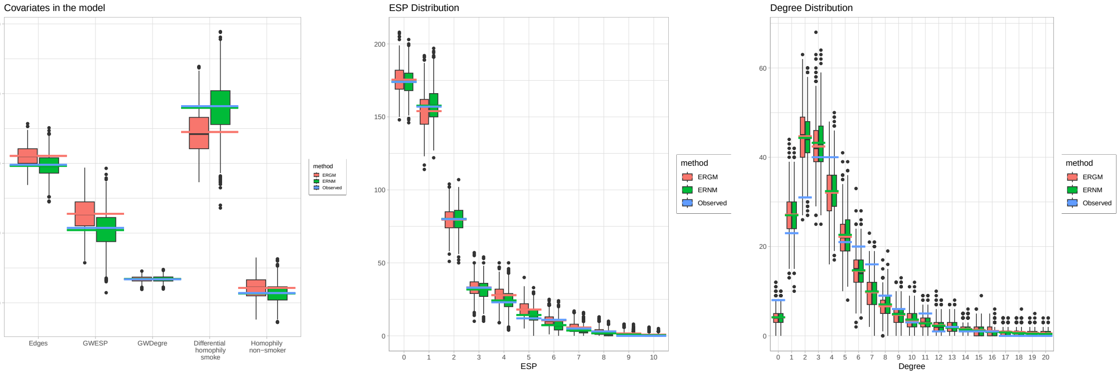

Figure D.2: GOF Comparison of ERGM and ERNM (Model 1): Grade 10

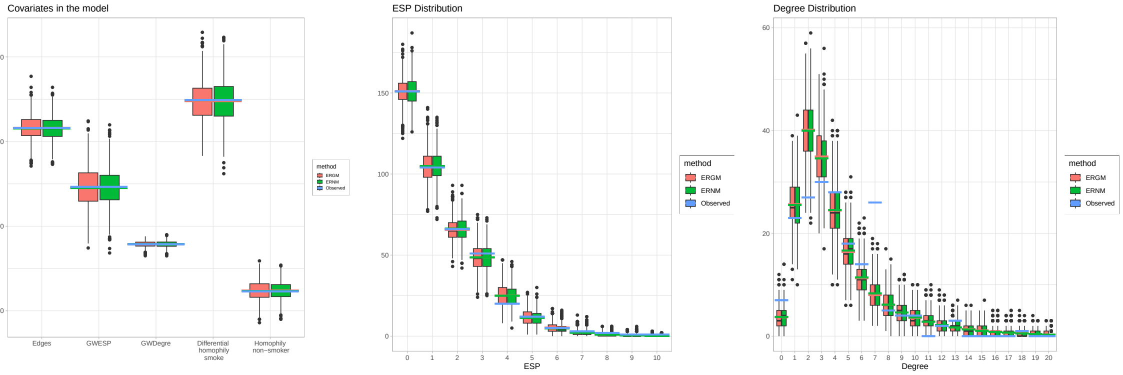

Figure D.3: GOF Comparison of ERGM and ERNM (Model 1): Grade 11

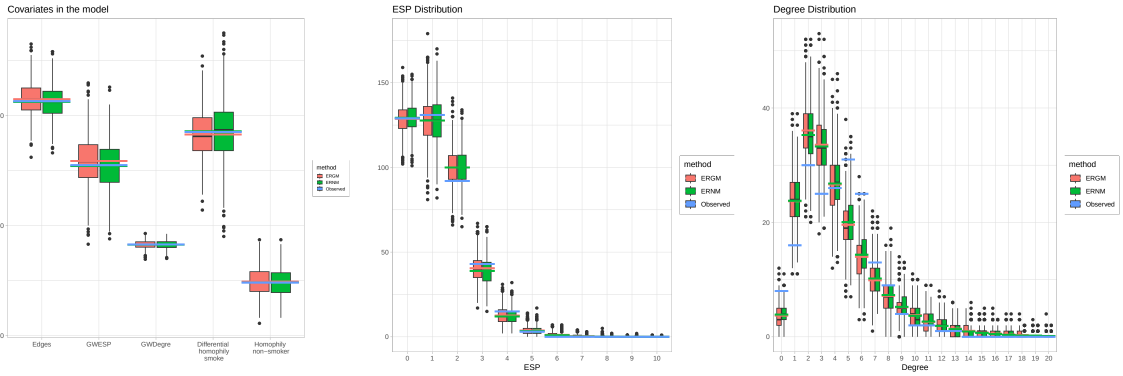

Figure D.4: GOF Comparison of ERGM and ERNM (Model 1): Grade 12

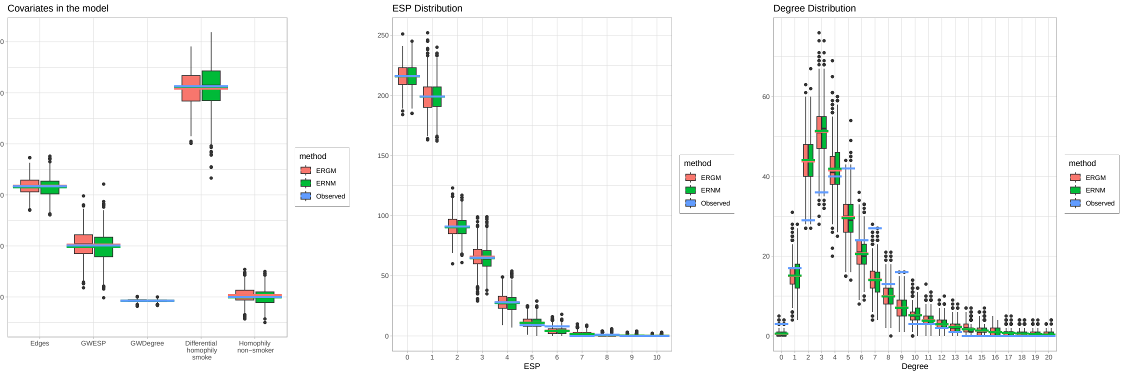

Figure D.5: GOF Comparison of ERGM and ERNM-Count (Model 2): Grade 9

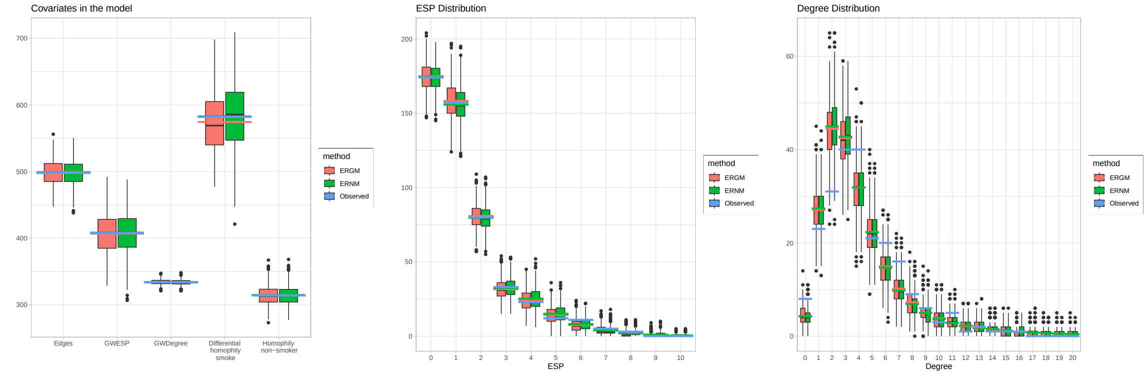

Figure D.6: GOF Comparison of ERGM and ERNM-Count (Model 2): Grade 10

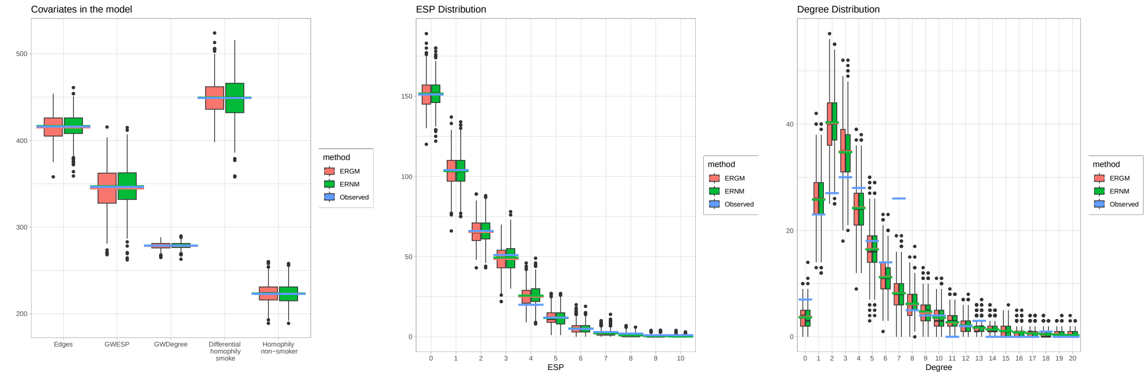

Figure D.7: GOF Comparison of ERGM and ERNM-Count (Model 2): Grade 11

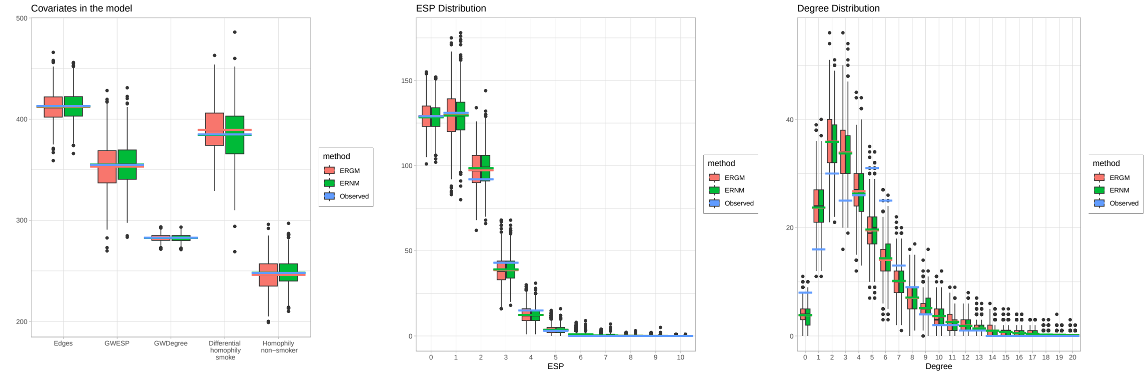

Figure D.8: GOF Comparison of ERGM and ERNM-Count (Model 2): Grade 12

APPENDIX E

MCMC Diagnostics

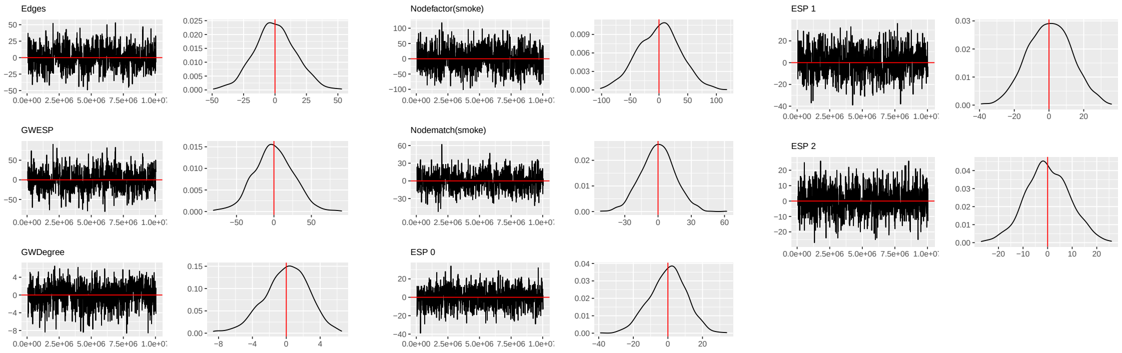

Figure E.1: MCMC Diagnostics of ERNM (Model 1) for grade 9: Left Column: Trace plot; Right Column: Density plot

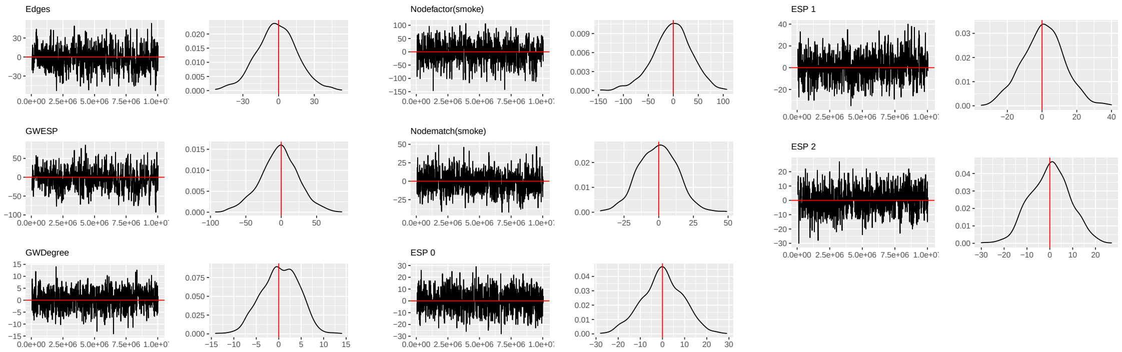

Figure E.2: MCMC Diagnostics of ERNM (Model 1) for grade 10: Left Column: Trace plot; Right Column: Density plot

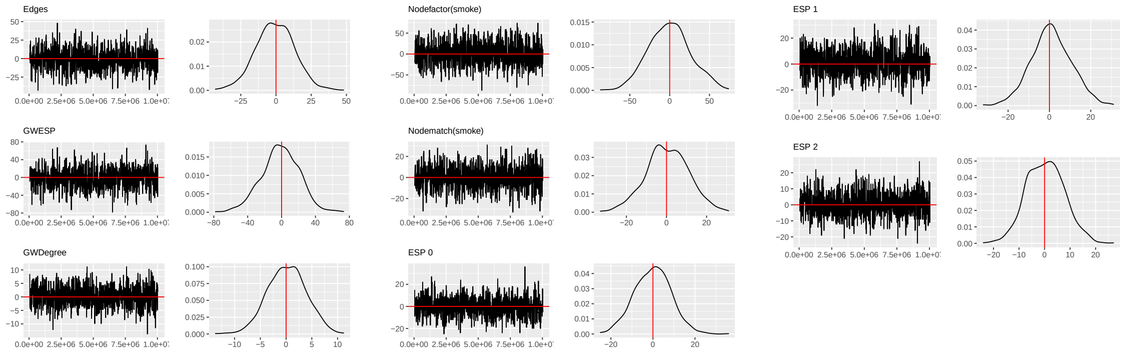

Figure E.3: MCMC Diagnostics of ERNM (Model 1) for grade 11: Left Column: Trace plot; Right Column: Density plot

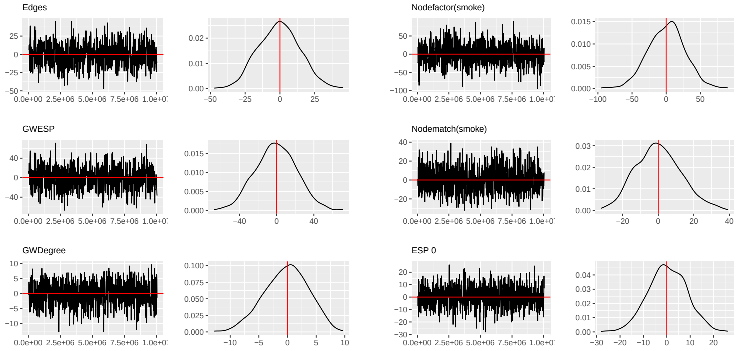

Figure E.4: MCMC Diagnostics of ERNM (Model 1) for grade 12: Left Column: Trace plot; Right Column: Density plot

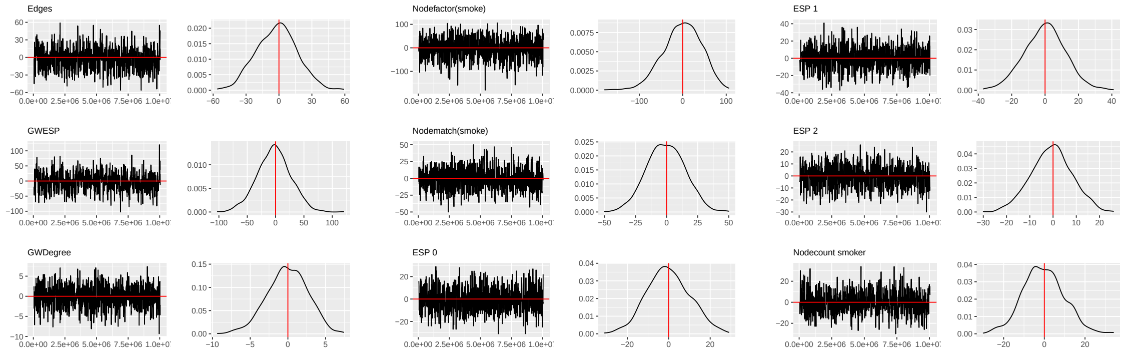

Figure E.5: MCMC Diagnostics of ERNM-Count (Model 2) for grade 9: Left Column: Trace plot; Right Column: Density plot

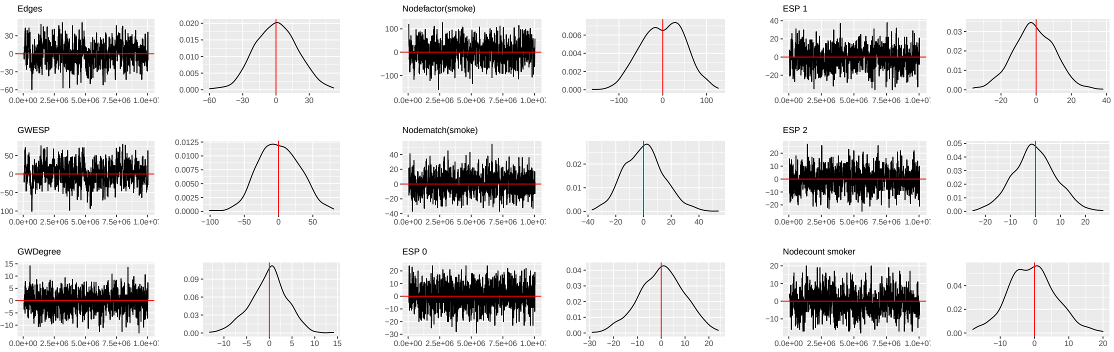

Figure E.6: MCMC Diagnostics of ERNM-Count (Model 2) for grade 10: Left Column: Trace plot; Right Column: Density plot

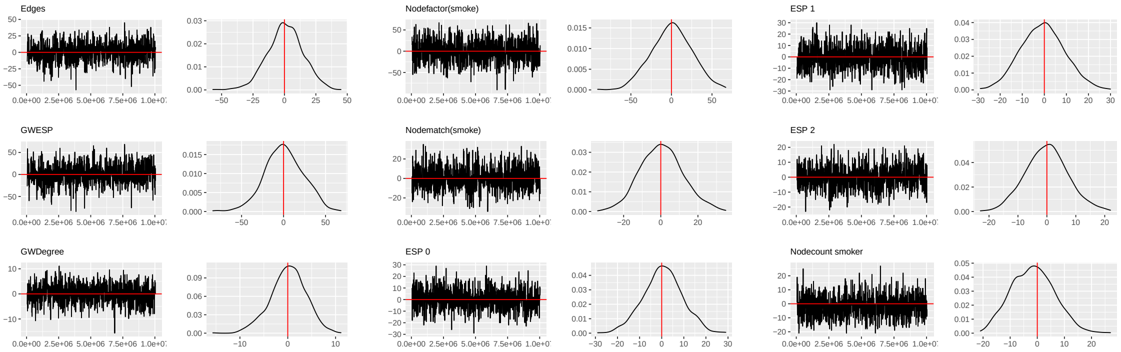

Figure E.7: MCMC Diagnostics of ERNM-Count (Model 2) for grade 11: Left Column: Trace plot; Right Column: Density plot

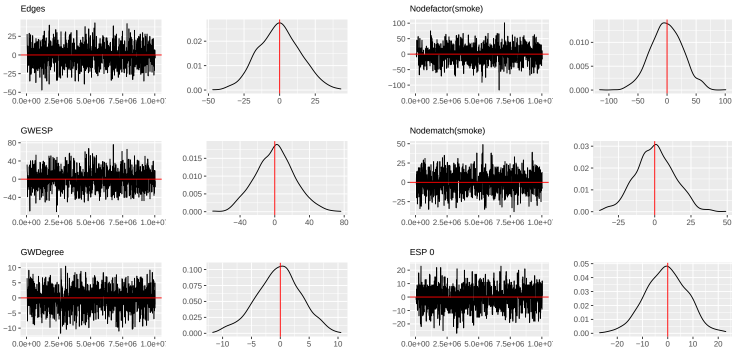

Figure E.8: MCMC Diagnostics of ERNM-Count (Model 2) for grade 12: Left Column: Trace plot; Right Column: Density plot

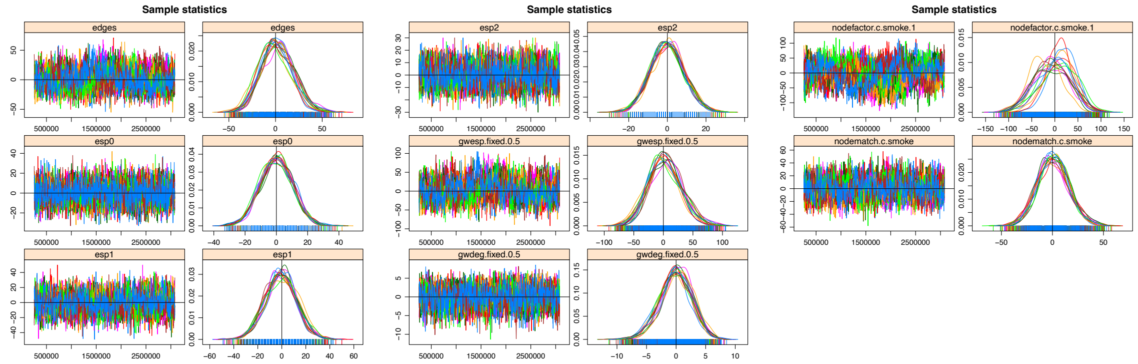

Figure E.9: MCMC Diagnostics of ERGM for Grade 9: Left Column: Trace plot; Right Column: Density plot. (The line with different color represents each parallelled Markov chain)

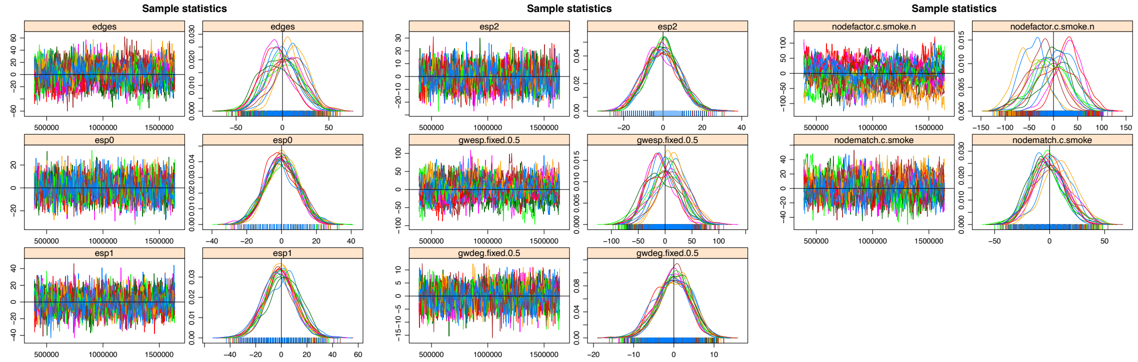

Figure E.10: MCMC Diagnostics of ERGM for Grade 10: Left Column: Trace plot; Right Column: Density plot. (The line with different color represents each parallelled Markov chain)

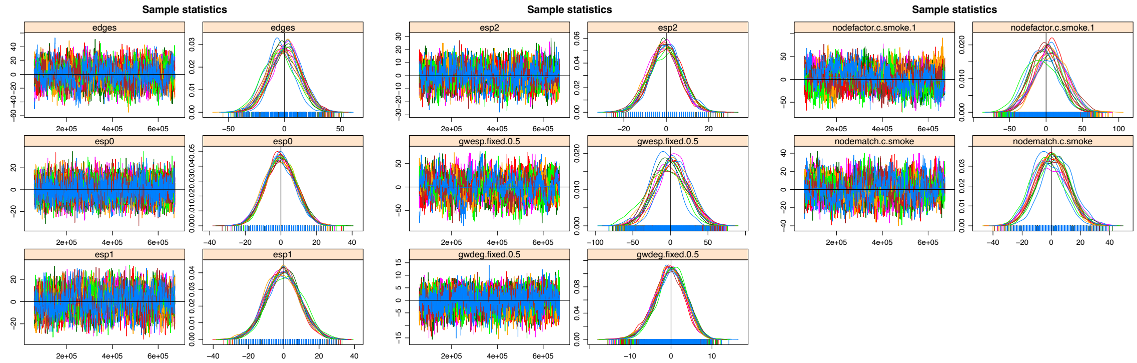

Figure E.11: MCMC Diagnostics of ERGM for Grade 11: Left Column: Trace plot; Right Column: Density plot. (The line with different color represents each parallelled Markov chain)

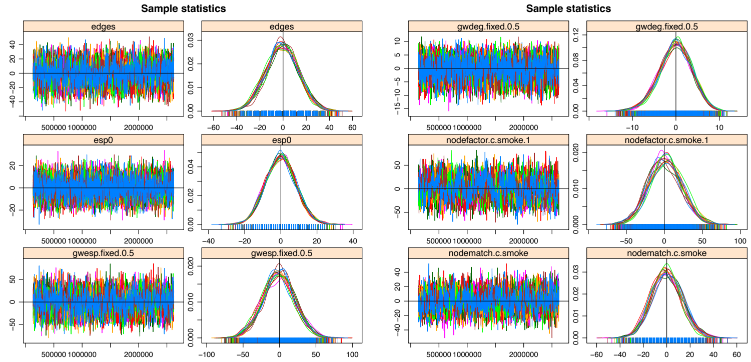

Figure E.12: MCMC Diagnostics of ERGM for Grade 12: Left Column: Trace plot; Right Column: Density plot. (The line with different color represents each parallelled Markov chain)

Bibliography

Almquist, Z. W. and Butts, C. T. (2014) Logistic network regression for scalable analysis of networks with joint edge/vertex dynamics. *Sociological Methodology*, **44**, 273–321. URL: https://doi.org/10.1177/0081175013520159. PMID: 26120218.

Biernacki, C., Celeux, G. and Govaert, G. (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22**, 719 – 725.

Clark, D. A. and Handcock, M. S. (2022) An approach to causal inference over stochastic networks.

Dasgupta, A., Hopcroft, J. E. and McSherry, F. (2004) *Spectral analysis of random graphs with skewed degree distributions.*

Erickson, B. H. (1988) The relational basis of attitudes. In *Social Structures: A Network Approach* (eds. B. Wellman and S. D. Berkowitz), 99–121. Cambridge: Cambridge University Press.

Feld, S. L. (1981) The focused organization of social ties. *American journal of sociology.*, **86**.

Fellows, I. and Handcock, M. S. (2012) Exponential-family random network models. URL: https://arxiv.org/abs/1208.0121.

Fellows, I. E. (2012) *Exponential Family Random Network Models.* PhD in Statistics, University of California, Los Angeles. Advisor: Mark S. Handcock.

— (2014) *ernm: Exponential-Family Random Network Models.* URL: https://github.com/fellstat/ernm. R package version 1.1.

Fienberg, S. E. and Wasserman, S. S. (1981) Categorical data analysis of single sociometric relations. *Sociological methodology*, **12**, 156–192.

Frank, O. and Strauss, D. (1986) Markov graphs. *Journal of the American Statistical Association*, **81**, 832–842. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478342.

Friemel, T. N. (2015) Influence versus selection: A network perspective on opinion leadership. *International Journal of Communication*, **9**, 1002–1022. URL: https://ijoc.org/index.php/ijoc/article/view/2806.

Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T. and Morris, M. (2008) A statnet tutorial. *Journal of Statistical Software*, **24**, 1–26. URL: https://www.jstatsoft.org/index.php/jss/article/view/v024i09.

Handcock, M. S. (2003a) Assessing degeneracy in statistical models of social networks. *Working paper #39*, Center for Statistics and the Social Sciences, University of Washington. URL: https://csss.uw.edu/Papers/wp39.pdf.

— (2003b) Statistical models for social networks: Inference and degeneracy. In *Dynamic social network modeling and analysis* (eds. R. Breiger, K. Carley and P. Pattison), 229–252. National Academies, Washington, DC.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N. and Morris, M. (2021) *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (https://statnet.org). URL: https://CRAN.R-project.org/package=ergm. R package version 4.0-6406.

Handcock, M. S., Raftery, A. E. and Tantrum, J. M. (2007) Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **170**, 301–354.

Harris, K., Halpern, C., Smolen, A. and Haberstick, B. (2007) The national longitudinal study of adolescent health (add health) twin data. *Twin research and human genetics : the official journal of the International Society for Twin Studies*, **9**, 988–97.

Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983) Stochastic blockmodels: First steps. *Social Networks*, **5**, 109–137. URL: https://www.sciencedirect.com/science/article/pii/0378873383900217.

Holland, P. W. and Leinhardt, S. (1981) An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, **76**, 33–50. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1981.10477598.

Hunter, D. R. (2007) Curved exponential family models for social networks. *Social networks*, **29**, 216–230.

Hunter, D. R. and Handcock, M. S. (2006) Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, **15**, 565–583. URL: https://doi.org/10.1198/106186006X133069.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. and Morris, M. (2008) ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, **24**, 1–29.

Karrer, B. and Newman, M. E. J. (2011) Stochastic blockmodels and community structure in networks. *Physical Review E*, **83**. URL: https://doi.org/10.1103%2Fphysreve.83.016107.

Krivitsky, P. N., Handcock, M. S., Hunter, D. R., Butts, C. T., Klumb, C., Goodreau, S. M. and Morris, M. (2003-2020) *statnet: Software tools for the Statistical Modeling of Network Data*. Statnet Development Team. URL: http://statnet.org.

Leenders, R. (1997) Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection. *Evolution of social networks*, **1**, 165–184.

Leger, J.-B. (2016) Blockmodels: A r-package for estimating in latent block model and stochastic block model, with various probability functions, with or without covariates.

Lei, J. and Rinaldo, A. (2015) Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, **43**, 215 – 237. URL: https://doi.org/10.1214/14-AOS1274.

Lusher, D., Koskinen, J. and Robins, G. (2013) *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

Morris, M., Handcock, M. S. and Hunter, D. R. (2008a) Specification of exponential-family random graph models: Terms and computational aspects. *Journal of Statistical Software*, **24**. URL: http://www.jstatsoft.org/v24/i04/.

— (2008b) Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects. *Journal of Statistical Software*, **24**. URL: https://ideas.repec.org/a/jss/jstsof/v024i04.html.

Nowicki, K. and Snijders, T. A. B. (2001) Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, **96**, 1077–1087. URL: https://doi.org/10.1198/016214501753208735.

R Development Core Team (2022) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/. ISBN 3-900051-07-0.

Robins, G., Elliott, P. and Pattison, P. (2001a) Network models for social selection processes. *Social Networks*, **23**, 1–30.

Robins, G., Pattison, P. and Elliott, P. (2001b) Network models for social influence processes. *Psychometrika*, **66**, 161–189.

Sampson, S. F. (1968) *A novitiate in a period of change: An experimental and case study of social relationships*. Cornell University.

Schweinberger, M. (2011) Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, **106**, 1361–1370. URL: https://doi.org/10.1198/jasa.2011.tm10747. PMID: 22844170.

Schweinberger, M., Krivitsky, P. N., Butts, C. T. and Stewart, J. R. (2020) Exponential-Family Models of Random Graphs: Inference in Finite, Super and Infinite Population Scenarios. *Statistical Science*, **35**, 627 – 662. URL: https://doi.org/10.1214/19-STS743.

Snijders, T. A. B. and Nowicki, K. (1997) Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, **14**, 75–100. URL: https://doi.org/10.1007/s003579900004.

Snijders, T. A. B., Pattison, P. E., Robins, G. L. and Handcock, M. S. (2006) New specifications for exponential random graph models. *Sociological Methodology*, **36**, 99–153. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9531.2006.00176.x.

Steglich, C., Snijders, T. A. and Pearson, M. (2010) Dynamic networks and behavior: Separating selection from influence. *Sociological methodology*, **40**, 329–393.

Wasserman, S. and Anderson, C. (1987) Stochastic a posteriori blockmodels: Construction and assessment. *Social networks*, **9**, 1–36.

Wasserman, S. and Faust, K. (1994) Social network analysis: Methods and applications.

Wasserman, S. and Pattison, P. (1996) Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp. *Psychometrika*, **61**, 401–425.

Weng, H. (2020) *A Social Interaction Model with Endogenous Network Formation*. Ph.D. thesis, University of Cincinnati, University of Cincinnati. URL: http://rave.ohiolink.edu/etdc/view?acc_num=ucin159317152899108.