University of New Mexico
UNM Digital Repository

Mathematics & Statistics ETDs | Electronic Theses and Dissertations

2-14-2014

Social Network Analysis of Peer Influence on Adolescent Smoking

Gregory Lambert

Follow this and additional works at: https://digitalrepository.unm.edu/math_etds

Recommended Citation
Lambert, Gregory. "Social Network Analysis of Peer Influence on Adolescent Smoking." (2014). https://digitalrepository.unm.edu/math_etds/79

This Thesis is brought to you for free and open access by the Electronic Theses and Dissertations at UNM Digital Repository. It has been accepted for inclusion in Mathematics & Statistics ETDs by an authorized administrator of UNM Digital Repository. For more information, please contact disc@unm.edu.

Gregory J. Lambert

Candidate

Mathematics and Statistics

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Michael D. Sonksen, Ph.D., Chairperson

Yan Lu, Ph.D.

Nancy Brodsky, Ph.D.

SOCIAL NETWORK ANALYSIS OF PEER INFLUENCE ON ADOLESCENT SMOKING

by

GREGORY J. LAMBERT

B.F.A., PAINTING, SAN FRANCISCO ART INSTITUTE
B.S., MATHEMATICS, PORTLAND STATE UNIVERSITY

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Statistics

The University of New Mexico
Albuquerque, New Mexico

December 2013

© 2013, Gregory J. Lambert

Dedication

I would like to dedicate this paper to my loving wife Jessica, who has been a constant source of support and encouragement during the challenges of graduate school and life. I am truly thankful for having you in my life. I would also like to dedicate this work to my two children Grant and Claire, who I love very dearly and inspire me every day to be a better person.

Acknowledgments

I would first like to thank my advisor Dr. Sonksen, for being tremendously supportive through the difficult process of writing this thesis. I am very grateful for his guidance and the opportunity he has afforded me. Secondly I would like to thank Patrick Finely and Thomas Moore for all there thoughtful comments and help during this endeavor. Without their insight and support this thesis would have never been finished.

I would also like to give gratitude to my thesis committee members, Dr. Brodsky and Dr. Lu for their contributions to this work. I have been very fortunate to have such a superb group of people guiding me through my graduate work. Finally I would like to recognize other members of 6100 at Sandia National Laboratories and the professors at the University of New Mexico's Mathematics and Statistics department, who have made the last two years in graduate school an incredible learning experience. This work was funded in part by the U.S. Food and Drug Administration through a contract with the U.S. Department of Energy/Sandia National Laboratories (funding document 224109011). The information in this thesis is not a formal dissemination of information by the FDA and does not represent agency position or policy.

SOCIAL NETWORK ANALYSIS OF
PEER INFLUENCE ON ADOLESCENT SMOKING

by

GREGORY J. LAMBERT

B.F.A., PAINTING, SAN FRANCISCO ART INSTITUTE
B.S., MATHEMATICS, PORTLAND STATE UNIVERSITY
M.S., STATISTICS, UNIVERSITY OF NEW MEXICO

Abstract

Adolescent cigarette smoking behavior is affected by peer relationships, but how these peer relationships influence the behavior of other adolescent smokers is not well understood. Mixing among cigarette smokers in adolescent friendship networks could indicate peer influence or a homophilous association among likeminded individuals. The goal of this thesis will be to examine a set of adolescent friendship networks to determine if different cigarette smoking behaviors can be predictive of friendship nominations in the network. Examining the structure of social networks requires among other things, inspection of the presence (or absence) of relational ties. Tie formations in social networks are often conditional

on the existence of other ties in the network. This conditional dependency along with purely structural network characteristics, creates a unique set of problems from a statistical modeling perspective. Fitted exponential random graph models for a group of adolescent schools will be examined, to assess how the underlying structure of these social networks is influenced by smoking behaviors.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

According to the American Lung Association, 68 percent of adults who smoke began smoking regularly at the age 18 or younger. Adolescents who reported three or more friends who smoked had a smoking prevalence approximately ten times that of adolescents who reported that none of their friends smoked. Peer smoking relationships are a strong indicator of adolescent cigarette smoking. Despite the large body of work that suggests an association between peer friendships and smoking, an understanding of how peers influence the behavior of other adolescents is still not well understood. Correlation between adolescent smoking and the smoking habits of friends could be an effect of peer influence, as has been implied by many previous studies. Or can this correlation be attributed to what is called assortativity or assortative mixing? Assortative mixing is a preference for a network's nodes (adolescent teens) to attach to others that are similar in some way.

Though we will mainly be interested in assortative mixing of smoking behavior on a dyadic level, one has to wonder if this mixing is due to smoking habits spreading through the network as contagion or epidemic models suggest [1]. It is difficult to ascertain if the mixture of smokers is due to contagion of smoking in the network (peer influence) or if it is merely an effect of adolescent smokers befriending other smokers (assortative mixing). This assumption that peer influence is the only cause of mixing among smokers is the precise point of contention in the much debated Christakis and Fowler's paper on the spread of obesity in social networks [2]. Realistically, both peer influence and assortative mixing are affecting clustering of smokers and nonsmokers in adolescent friendship networks. It is nearly impossible to differentiate between peer influence and assortative mixing if network evolution is ignored completely [3], because assortative mixing at a specific time step in a social network could be the outcome of previous peer influence and not

attributed to assortative mixing. As previously mentioned, both peer influence and assortative mixing can contribute to the homogeneity of peer social networks. This work will focus on the effects that assortative mixing have on adolescent smoking without examining the possible effects of peer pressure.

By modeling the network structure using exponential random graphs models on a cross-section of the longitudinal Add Health data set [4] for individual schools, the hope is that a more detailed description of the network topology can be captured in the model and this will in turn lead to a better understanding of the role that assortative mixing may have on adolescent smoking behaviors. Additionally, by modeling individual schools and the effects that grade level has on mixing of smokers in the network, we hope to account for the differences that schools and grade levels have on smoking patterns. If smoking assortative mixing is prevalent in an adolescent social network then a well executed statistical analysis will show adolescent smokers are more likely to friend smoking peers than non-smoking peers.

Review of Related Literature

Statistical social network analysis has been in development for a few decades. Early work on distributions for graphs were restricted to forcing network modelers to adopt independence assumptions [5]. A significant breakthrough in statistical modeling of networks came in a paper by Frank and Strauss [6] who named their models, Markov random graphs. These Markov random graphs were further developed for estimation of parameters in a paper by Strauss and Ikeda [7]. As stated by Robins and Morris, “A good [statistical network graph] model needs to be both estimable from the data and a reasonable representation of that data, to be theoretically plausible about the type of effects that might have produced the network, and to be amenable to examining which competing effects might be the best ex-

planation of the data." [8]. An exponential random graph model (ERGM) allows us to describe the complex structure of the network, by providing a way to represent both the complex interdependent structure of the network and the individual nodal attributes. An ERGM is not focused on prediction as much as revealing patterns to enable inference on tie formations in complex networks structures.

The possibility of a tie between two actors in a social network can be explained by a combination of factors among the two actors but it can also be described by the presence or absence of other ties in the network. ERGMs compare an observed network to the other possible network configurations. Network structures have a finite number of ways that ties can be arranged, this is called the sample space. When a model includes only terms that represent the composition of node attributes within ties, it is similar to traditional logistic or log-linear models for contingency tables [9]. Such models are said to exhibit dyadic independence because the probability of any tie does not depend on the value of other ties, only on the attributes of the two actors involved. An important early paper that paved the way to modern ERG models was written by Holland and Leinhardt [10]. They introduced a modeling approach that applied to directed graphs (digraphs) that is particularly applicable to social networks. This dyadic independent statistical model is often called the p_1 model. A dyad is a tie (relationship) between two variables (Y_{ij}, Y_{ji}), that can take on many forms from reciprocity to sending and receiving. Even though the p_1 network models did a fair job of representing the dyadic structure of a social network, they didn't go beyond that. In the p_1 network model all parameters are fixed effects, model the four possible dyadic outcomes, and describe the probability of an edge between node i and j.

The Holland-Leinhardt's log-linear equations for the p_1 model can be expressed as follows for the probability of an edge occurring between nodes i and j:

$$\begin{aligned} log\,P(no\ edge) &= P(0,0) = \lambda_{ij} \\ log\,P(i\,to\,j) &= P(1,0) = \lambda_{ij} + \alpha_i + \beta_j + \theta \\ log\,P(j\,to\,i) &= P(0,1) = \lambda_{ij} + \alpha_j + \beta_i + \theta \\ log\,P(bidirected\ edge) &= P(1,1) = \lambda_{ij} + \alpha_i + \beta_i + \alpha_j + \beta_j + 2\theta + \rho_{ij}. \end{aligned} \tag{1}$$

Where the fixed effects are: reciprocity(ρ), sending (α), receiving (β) , normalizing constant (λ), and edge (θ). Although these models could be represented using standard log-linear models the independence assumption between nodes creates a limitation for representation of complex network structures. For example, trying to model reciprocity in a social network using a p_1 model does not allow for taking into account the dependency structure. Additionally, statistical approaches such as logistic regression can be used to model dichotomous outcomes such as the formation of a network tie but require independence among observations. As stated by Hunter et al. [11], ERG models can be used to understand a particular phenomenon or to simulate new random realizations of networks that retain the essential properties of the original. ERGMs are a statistical approach to modeling network data that extend the restrictive dyadic independence assumptions of both p_1 models and logistic regression approaches. This allows ERGMs to describe complex dependency structures of social networks. For example in a direct social network it is not plausible that the tie Y_{ij} is independent of Y_{ji}.

ERGMs have a goal to loosen these restrictions of this independence assumption and to allow for these dependencies among tie-variables. ERGMs are very similar to general linear models specifically the logistic linear models. The main difference between the two modeling techniques is that in ERGMs the observations in the sample are assumed to be dependent. ERGMs do not predict a social tie independent of other social ties as with logistic regression, but rather the conditional probability of a tie given what is observed in the rest of the network. The probability of a set of ties Y given a set of actors and parameters Frank & Strauss [12] and Wasserman & Pattison [13]:

$$P(\mathbf{Y}=\mathbf{y}) = \left(\frac{1}{c(\theta)}\right) exp\left\{\sum_{k} \theta_k S_k(\mathbf{y})\right\} \tag{2}$$

- k is a set of possible edges among a subset of vertices in G
- $\mathbf{Y}$ is a random adjacency matrix for the network and $\mathbf{y}$ is a particular realization of $\mathbf{Y}$
- θ_k vector of parameters for a given configuration
- $S_k(y) = \prod_{y_{ij} \in k} y_{ij}$ vector of sufficient network statistics for a given configuration that is 1 if the configuration is observed and 0 otherwise
- $c(\theta) = \sum_{all\, y} exp\{\theta^T s(y)\}$ is a normalizing constant to ensure that $P(\cdot)$ sums to one over its range of values

Most ERG models are conducted on cross-sectional data but recent work has been done to extend this to longitudinal data Krivitsky [14]. The idea behind Krivitsky's

separable temporal ERGM (STERGM) is to model the transition of a network from Y^t to a network at Y^{t+1}. The formation and dissolution of ties occur independently from each other within the same time step. A key idea is that the process and factors that result in ties being formed are not the same as those for ties being dissolved. Two separate ERG models are created. One ERGM is created for tie formations:

$$P(\mathbf{Y}^+ = \mathbf{y}^+ | \mathbf{Y^t}) = \left(\frac{1}{c(\theta^+)} \right) exp\left\{ (\theta^+)^T s(\mathbf{y}^+) \right\}. \tag{3}$$

And another ERGM is created for tie dissolution:

$$P(\mathbf{Y}^- = \mathbf{y}^- | \mathbf{Y^t}) = \left(\frac{1}{c(\theta^-)} \right) exp\left\{ (\theta^-)^T s(\mathbf{y}^-) \right\}. \tag{4}$$

Then the probability of transitioning from a given network at time t, $\mathbf{y}^t$ to a network at $t+1$, $\mathbf{y}^{t+1}$ can be expressed by the conditional probability statement:

$$\begin{aligned} P(\mathbf{Y^{t+1}} = \mathbf{y^{t+1}} | \mathbf{Y^t}) \quad &= \left(\tfrac{1}{c(\theta^+, \theta^-)} \right) exp\left\{ (\theta)^T s(\mathbf{y^t}, \mathbf{y^{t+1}}, \mathbf{X}) \right\} \\ &= P(\mathbf{Y}^- = \mathbf{y}^- | \mathbf{Y^t}, \mathbf{X}) \times P(\mathbf{Y}^+ = \mathbf{y}^+ | \mathbf{Y^t}, \mathbf{X}). \end{aligned} \tag{5}$$

A visual representation of the how these two models can be formed can be seen

in the following figure.

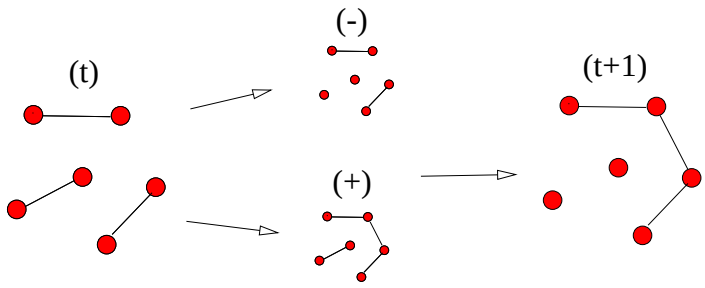


Figure 1: STERGM formation

Two general ways for representing dependence between tie variables have been presented in the literature. One is by postulating latent nodal variables and conditional independence of the observations, given the latent variables, in the classical Lazarsfeld tradition of latent structure models. A discrete latent class approach was proposed in Nowicki and Snijders [15]. The second way is by directly modeling this dependence, as is done in exponential random-graph models. One important advantage of the latent structure models is that they allow missing data in a network to be imputed with likelihood-based inference [16].

Chapter 2: Methods

The National Longitudinal Study of Adolescent Health (Add Health) is a longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States during the 1994-95 school year. The Add Health cohort has been followed into young adulthood with four in-home interviews, the most recent in 2008, when the sample age was 24-32. The study combines longitudinal survey data on respondent's social, economic, psychological and physical well-being. The network data examined in this study is from the first wave (1994-1995) of Add Health. The first wave of Add Health contains data on 86 schools with 90,118 students and 578,594 friendship nominations. Three schools from Add Health will be used in the analysis which we will refer to as: school1, school2, and school3. The three schools will be used to examine the relationship of friendship nominations and smoking habits and were selected for their varying smoking prevalence. To collect the friendship data, each student was asked to nominate five close female and male friends. The students were allowed to nominate friends outside the roster or less than the required five female and male friends. To clean our data for construction of the network, friendship nominations to people outside the school were removed. In addition to including the friendship nominations, both smoking habits and grade level survey data were included into each model. The survey question, "How often did you smoke cigarettes in the past twelve months" was broken into three categories:

- never
- once or twice, once a month or less, 2 or 3 days a month
- once or twice a week, 3 to 5 days a week, nearly everyday.

These three categories will classify a students smoking status as: nonsmoker, light smoker and smoker. The reason for grouping the responses was both for economy during the simulation process of building the model and secondly to look at the relationship between light smokers and regular smokers. Schools with fairly large data sets have been chosen in order to have an ample population of smokers among the different grade levels. The friendship nominations are directed since student A can nominate student B without B nominating student A.

Commonly in statistical social network models, data sets like Add Health will be represented as undirected graphs (mutual friendships). This is often done not only to simplify the estimation of model parameters, but also used to capture only reciprocated friendship nominations. Using an undirected representation of the network helps not only in the Markov chain Monte Carlo (MCMC) calculations but also in the amount of terms that are needed to represent the network structure in the ERG model. In this work we will retain the direction of the friendship nominations and model each individual school as a directed network, to capture more of the smoking assortative mixing in the school networks than just reciprocated friendship nominations. Due to the considerable computational cost of modeling directed graphs from the Add Health data set, parallel processing was employed for the MCMC simulations to help alleviate the computational cost of simulating directed networks. An additional benefit of modeling the network with directed tie formations is to help capture the friendship nominations that are based on admiration. Although reciprocated friendship nominations make up a significant number of the tie formations in the schools analyzed, a lot of detail in the network would be lost by disregarding the unreciprocated friendship nominations.

Exploratory Network Analysis

We can represent a directed social network Y as a set of n actors and m dyads

where $Y_{ij} = 1$ if the actors $(i,\ j)$ are connected and $Y_{ij} = 0$ if they are not connected. Before conducting the formal analysis we will look at some exploratory data analysis of the three social networks. One of the most basic measurements for describing a social network, or any connected network is density. The density of a social network is the proportion of the connections between actors to the total possible number of connections. Density for a $n \times n$ social network can be represented by the following:

$$D = \frac{\sum_{i,j} Y_{ij}}{n(n-1)} \tag{6}$$

There are several common measures for centrality: degree, betweeness, closeness, and eigenvector. Centrality of an actor is a measure of its relative importance in a social network or how influential the person is in a social network. Since centrality is an important indicator of diffusion of smoking in a social network, we will use a form of it to validate our model called geodesic distance. Geodesic distance also represent high-order network statistics not directly related to any of the statistics included in our models, and thus provide a strong independent criterion for goodness of fit [16]. In a graph, a geodesic distance between actors n_i and n_j is the shortest path between them. Therefore the distance between n_i and n_j is equal to the distance between n_j and n_i, $d(i,\ j) = d(j,\ i)$. Betweenness centrality also is a measure of the relative importance of an actor in a social network. It is equal to the number of shortest paths from all actors to all others that pass through that node. The betweenness centrality of an actor n is given by the expression:

$$g(n) = \sum_{s \neq n \neq t} \frac{\sigma_{st}(n)}{\sigma_{st}} \tag{7}$$

where σ_{st} is the total number of shortest paths from actors s to actor t and $\sigma_{st}(n)$ is the number of those paths that pass through n. Histograms of the distribution of betweenness centrality for each school is given in the Appendix: Figure 13, Figure 14, and Figure 15. Another important measure in social network analysis is actor degree, which is relatively easy to calculate. In a directed network we have both in-degree and out-degree statistics. Degree indicates an actors involvement in a social network and will also be examined in the goodness of fit in the model. Figure 2 shows school3 colored from light to dark for actors with in-degree from 1 to 10.

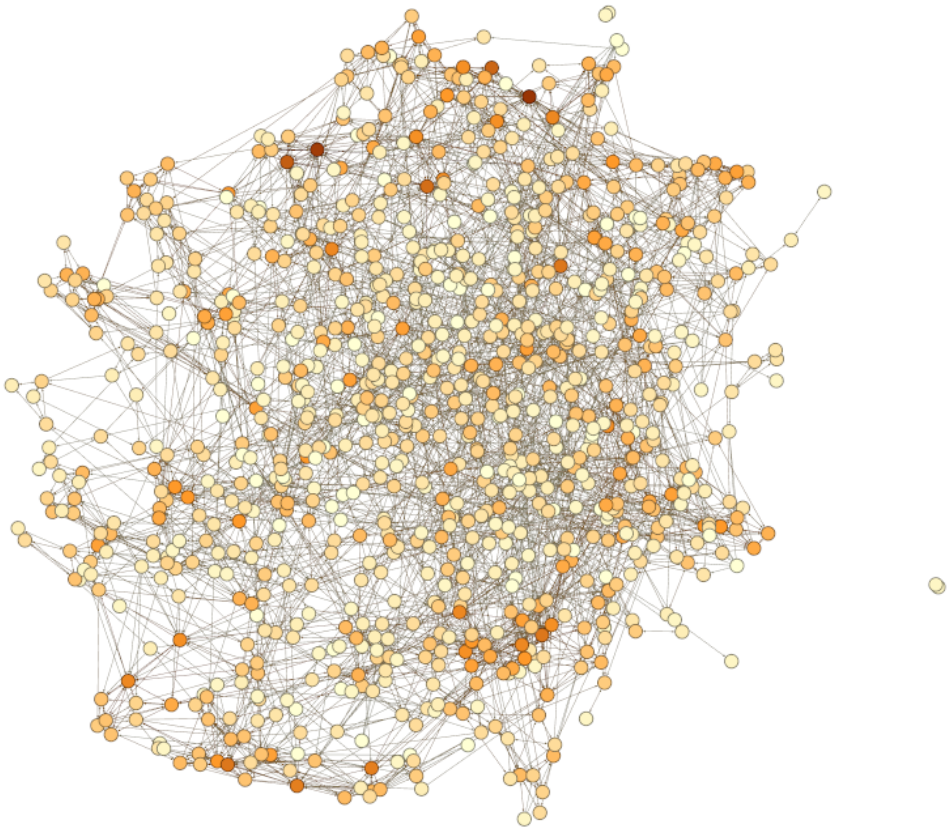


Figure 2: In-degree for school3

Histograms of the degree distributions for each school are given in the Appendix. After density, the next most important statistic in social networks is centrality. Transitivity can be used to measure the local structuring of smoking assortative mixing in a network, defined when there is a tie from actor n_i to n_j and also a tie from n_j to n_h. There is a strong proportion of actors that display transitivity in all three networks. Terms for representing degree and transitivity in the network will be discussed in further detail in the analysis section for modeling smoking assortative mixing in the social networks. Figure 3 represents the social network for school3 colored in red for nonsmokers, green for light smokers and blue for smokers. It is

difficult to capture a good graphical representation for social networks of this size (952 nodes and 4109 directional edges).

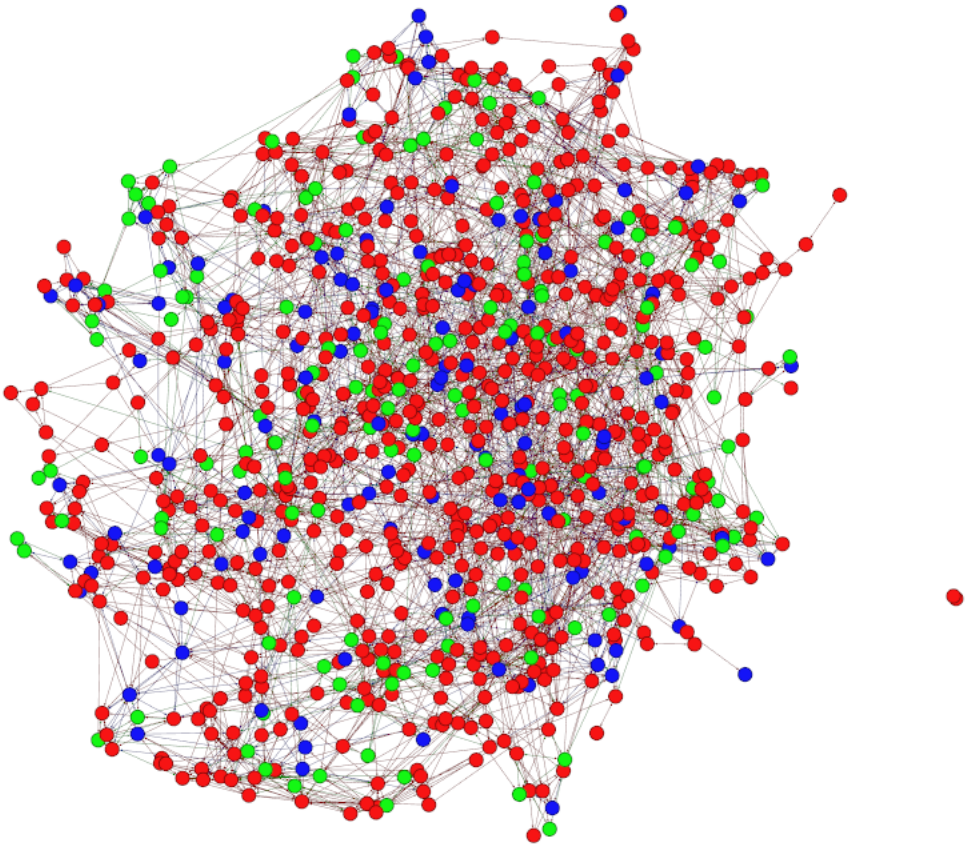


Figure 3: Smoking behaviors for school3

Modeling Adolescent Smoking Behaviors

To model the social behavior of adolescent students in the Add Health network, a class of models called exponential random graph models (ERGMs) described in the literature review section were employed. Some general work with ERGMs has previously been done on the Add Health network such as the paper by Goodreau,

Kitts and Morris [17], which looks at patterns in the adolescent social network through examination of mutual friendship ties and exogenous attributes of sex, race and grade level. Other studies have investigated the effects of cigarette smoking within the Add Health peer network using logistic regression [18]. While some have looked at the dynamics of smoking by way of common descriptive network statistics such as centrality to look at the clustering behavior of the smoking population in the social network [19]. In contrast, this paper attempts to take the mechanics built into the ERG models by looking at the adolescent school network data in a social network framework. The approach will be to quantify the influence smoking behavior has on formation of friendship nominations on a dyadic level to explore the influence smoking habits have on friendship nominations, by creating a formal statistical model to describe adolescents who smoke.

Many of the studies that have been conducted on smoking in social networks have used logistic regression and relied on summary statistics for describing the structure of the network while leaving out a formal statistical model of the network itself. A social network can be thought of as a set of nodes and vertices connected by edges (in this instance friendship nominations). The network can be written as $G = (V,\ Y)$ where V consists of a set of actors $V = \{v_1, \ldots, v_n\}$and Y is an $n \times n$ matrix with binary values for the present or absence of friendship nominations. Then $P(Y_{ij} = y_{ij})$ is the probability of the Y_{ij} edge where it is 1 if the i student nominates j and 0 otherwise. Rather than modeling network ties conditional on a fixed set of attributes, it is possible to model a set of attributes conditional on a fixed network within a ERGM-type framework. Such a model represents a social influence or social contagion process, Robins, Pattison, and Elliot [20]. This is similar to logistic regression but tie probabilities are recursively dependent. This makes them especially valuable for modeling relational data like social networks. This ability to model the dependency structure of the network while retaining the

explanatory nature of the logistic regression model is what makes the ERG model so useful in social network framework.

The conditional log-odds (logit) of an edge occurring from node i to j can be represented by the following:

$$\begin{aligned} logit\ P(Y_{ij} = 1|Y_{ij}^c) &= log \frac{P(Y_{ij} = 1|Y_{ij}^c)}{P(Y_{ij} = 0|Y_{ij}^c)} \\ &= log \left[\frac{\left(\frac{e^{\theta^T S(y_{ij}^+)}}{\theta} \right)}{\left(\frac{e^{\theta^T S(y_{ij}^-)}}{\theta} \right)} \right] \\ &= \theta^T \delta(y)_{ij} \end{aligned} \tag{8}$$

- where y_{ij}^+ is the graph with $Y_{ij} = 1$ and y_{ij}^- is the graph with $Y_{ij} = 0$
- $\delta(\mathbf{y})_{ij} = \left[S(y_{ij}^+) - S(y_{ij}^-) \right]$ is the change statistics if y_{ij} where to change the value of the network statistic from 0 to 1
- $\delta(\mathbf{y})_{ij}$ multiplied by the parameter value is the log-odds of the existence of a tie due to the statistic

Model degeneracy is a common problem that occurs when fitting ERGMs to social networks. When a model does not fit the data well or MLEs don't exist, or they exist but don't fit the data well, then a ERGM is called degenerate. Assessing if a particular MLE fits the data will be determined in a later section using a parametric boostrapping method. Degeneracy is a sign that the model is not constructed properly, which is often caused by overly strong assumptions of homogeneity in one of more descriptive network statistics. The idea is to limit the probability of higher order statistics like k-triangles in network analysis, relaxing the homogeneity assumption. Homogeneity is the idea that all isomorphic graphs have the the

same probability. Parameter specific differences to different vertices are not introduced. This assumption is necessary for many social networks especially for directed graphs due to the large number of parameters associated. One approach to dealing with model degeneracy is to limit the number of configurations or hypothesized parameters [21]. This can be accomplished by using geometrically weighted statistics for describing network structures. Often over estimation of these values are due to lower-order statistics being nested in high-order statistics. One problem is that the model gives large probabilities on graphs with larger degrees. An example of this nesting problem can be seen in Figure 4 showing a 1-triangle nested in a 2-triangle.

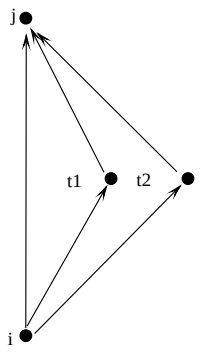


Figure 4: k-triangles

A solution to this is to set declining marginal returns for each additional 3-cycle with decreasing weights on the number of shared partners:

$$S_k(y) = \sum_{k=0}^{n-1} e^{-\alpha} d_k(y). \tag{9}$$

Where $S_k(y)$ is the sufficient network statistic from equation (2), $d_k(y)$ is the number of nodes with degree k, and $\alpha > 0$ controls the geometric rate of decreasing in the weights. All of the models in this analysis (excluding the baseline Bernoulli model) contain a geometrically weighted edgewise shared partner (GWESP) term to capture transitivity in the network and to help with model degeneracy [22]. The GWESP parameter estimates the number of shared friends of two adolescents in the network. This network statistics will help model the clustering in the network by way of triadic structures.

For clarity, smoking mixing will be analyzed for grades 9-12 and for students who responded that they smoked at least once in the past month. The primary interest here is to examine if a parametric model can be created based on network characteristics and the preference of smoking high school students to nominate smoking peers in their school. After creating an ERG model, the creation of ties are probabilistic in nature. Due to the stochastic nature of the ERG model there is some uncertainty when inferring the parameters in the model based on the data. Given an observed network, can we quantify the uncertainty of various structural and parametric characteristics in the model? Goodness-of-fit diagnostics of the estimated model parameters can be assessed by randomly simulating the distribution of graphs to see whether other features of the observed network data (i.e., non-fitted effects) are being properly represented. A large amount of uncertainty will generally produce high $p - values$. Suppose we are interested in inferring reciprocity in a directed network. If the observed network exhibits a high reciprocity statistic, it is unlikely that it was drawn from an ERGM but not impossible. Testing uncertainty about parameter values of an ERG model requires a way to measure their uncertainty. The sampling distribution of an estimator is the distribution of the estimator over multiple samples of data from the same population. The first step is to determine the sampling distribution of the estimators in the ERG model, get

maximum likelihood estimations for parameters in ERG model, simulate networks with the sample probability distributions for MLE estimates, and test individual models for goodness-of-fit for their ability to predict global network properties that are not explicit terms in the ERG model. This is the approach that will be used for assessing uncertainty and goodness-of-fit in the model [14].

Chapter 3: Analysis

Modeling Specification and Verification

The observed adolescent school network is assumed to be the realization of a stochastic process. Parameter estimates of the exponential random graph are estimated using frequentist Markov chain Monte Carlo methods. To assess the goodness-of-fit of the model results sampling from the fitted model is compared to network topology from the observed adolescent friendship network. Once a suitable fitted network is chosen, it can be analyzed for its explanatory ability for describing smoking assortative mixing. Since our ERG model can be rewritten in terms of the log-odds of a tie occurring, we can use logistic regression to represent the dependent variable of a friendship tie. To conduct the model selection process, a set of model configurations for each school will be evaluated and compared. A bottom-up approach will be used by first fitting a simple Bernoulli random graph as our baseline model. This model assumes independence among all friendship ties for each student in the social network. A dependency structure between ties in friendship networks is a more realistic assumption than independence when modeling friendship networks. For instance, if student A nominates student B as a friend and student B nominates student C then it is more likely student C will nominate students A. If our assumption is that there is a strong dependency structure in the network then we should be able to formulate a better model than a baseline Bernoulli random graph. To assist in model selection we will use Akaike information criterion (AIC). When fitting a model, it is possible to increase the likelihood of fit by adding parameters, but doing so may result in overfitting. In addition to using AIC scores for model selection we will conduct goodness-of-fit diagnostics to assess overfitting and overall fit to network statistics from the observed network data [20]. Comparison of a single outcome from the simulation of the fitted model to the

original network is of limited value. Our process will be to generate 200 random networks from our fitted ERG model and then use the distribution of the network topology of these 200 simulations to compare them to the observed Add Health network.

Assessing Model Degeneracy

As stated previously in many MCMC samples of ERG models there is a common problem called model degeneracy, meaning they are very dense or very sparse networks. Often when an ERG model is degenerate the model terms will over estimate or under estimate the observed network even under maximum likelihood coefficients. One possible reason for this occurring is that the Markov chain is too slow at mixing or reaching a stationary distribution. Another possible reason for the model degeneracy is that the parameters are too close to the boundary of the parameter space. To assist in determining if model degeneracy exists in a given model from one of the three schools assessed, both trace and density plots during the final iterations of the MCMC algorithm will be examined for each parameter. The plots should both vary stochastically around the mean, with non-stochastic trends from the mean being indicative of degeneracy. Another way for determining if a given ERG model is degenerate is to simulate from the model parameter estimates and compare the simulated networks to the observed for the statistics in the model as discussed above in section about goodness-of-fit. Both of these methods will be used and discussed to assess model degeneracy in the next chapter.

Chapter 4: Results

As stated previously, we will be comparing three schools from the National Longitudinal Study of Adolescent Health (Add Health) data set. The schools were chosen for their variation in smoking prevalence in the student population. After a baseline Bernoulli model was fit to each school with a single predictor for describing network density, a second model was fit with predictors for density (edges) and reciprocity (mutuality). The third model that was fit to each school had the two terms from the previous model plus an added term to represent network transitivity (geometrically weighted edgewise shared partners). The final model has all the previous terms for describing the topological structure of the network plus $8^2 = 64$ possible terms for all the possible pairings of light smokers and smokers from $9^{th} - 12^{th}$ grade. A number of these pairings should be statistically significant if smoking assortative mixing is a strong predictor for modeling friendship nominations in the observed adolescent social networks.

Model Results

Tables 1, 2 and 3 in the Appendix show the significant parameters for each model chosen to represent the adolescent social network for each school respectively. All three models for each school produced statistically significant predictors for density, reciprocity, and transitivity. Or as show by values for edges, mutual, and GWESP respectively. The last two significant predictors should not come as a surprise, since most social networks show strong tendencies for friendship reciprocity and transitivity. The statistically significant terms that varied were the 64 possible predictors for smoking assortativity based on grade level.

The terms for smoking assortativity selected in the final models for the three schools ranged from thirteen for school2 to twenty-five for school1. For the sake

of brevity we will give the interpretations of the model coefficients only for school2; similar logic can be used to discuss the model results for the other two schools. Models were selected both on AIC scores and goodness-of-fit tests. After choosing a final model, coefficients that were statistically significant at the $\alpha = 0.01$ level were left in the model. Figure 5 shows the goodness-of-fit for the edge-wise shared partners and out degree from 200 random simulations from our final selected model. The boxplots represent the simulations for the model and the dark black lines represent the network statistics for the observed data. Two network statistics will be shown for each of the models, one for edge-wise shared partners and the another for out degree. Edge-wise shared partner statistic is the number of unordered pairs $\{i,\ j\}$ such that $y_{ij} = 1$ and i and j have exactly k common neighbors and out degree is the number of outward friendship nominations for a given actor. As we can see in the following figure, the final model for school2 over estimates slightly for one edge-wise shared partner and under estimates for the observed network for two and three edge-wise shared partners.

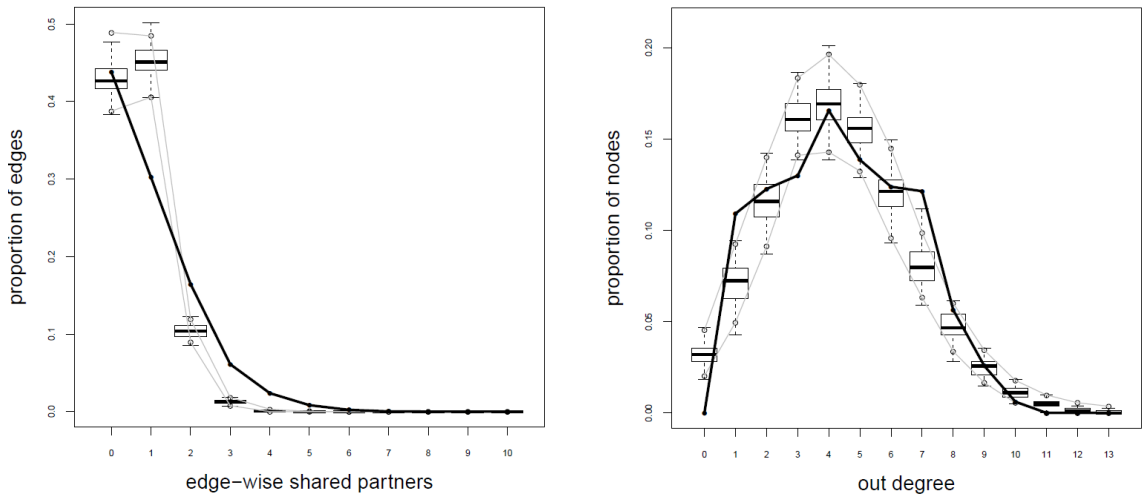


Figure 5: Goodness-of-fit final model school2 with assortative mixing my grade level

Comparing these results to 200 simulations from the model with no predictor variables for smoking assortative mixing by grade level and three terms to represent network characteristics: density(edges), transitivity(GWESP), and reciprocity(mutual). We can see that the model does a fair job representing the two network statistics from the original network, edge-wise shared partners and out degree . Additionally for the other two schools examined in this study, significant improvements in the overall fit to the observed network data can be observed (Appendix).

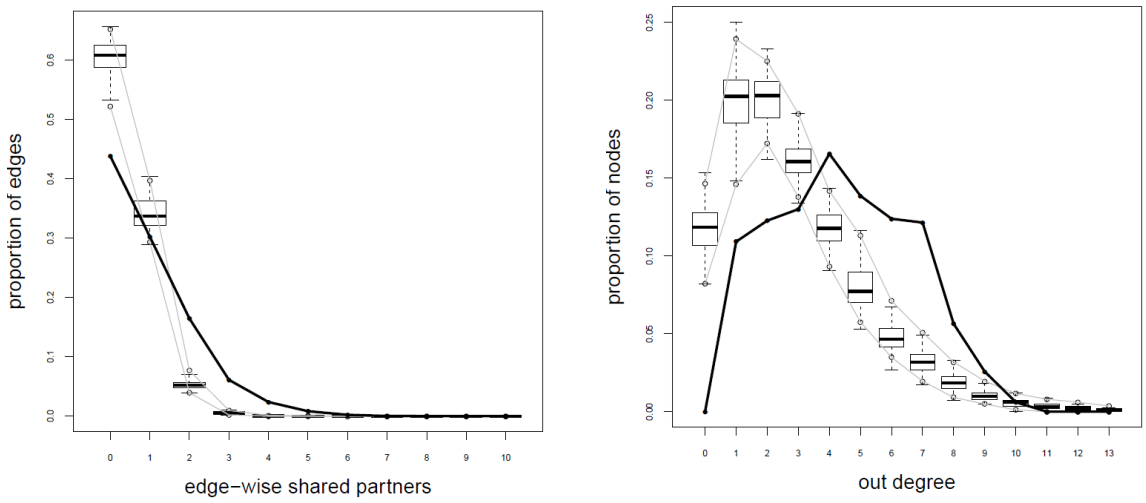


Figure 6: Goodness-of-fit model without smoking assortative mixing by grade level school2

Discussion

Now that we have a final parametric representation for tie formations in the network, we can start to interpret our results. All 3 schools show similar values for edges (-6.2 to -6.6), mutual (4.0 to 4.2) and GWESP (1.7 to 2.0). The model for school2 can be interpreted by using log-odds of a friendship tie occurring, where the log-odds estimate for edges is -6.2453 (Table 2). This estimate can be de-

scribed as the log-odds of two arbitrary students (i, j) having a directed friendship nomination between them when they don't have any friends in common. This log-odds estimate has a corresponding probability of 0.0019. The results for the edge term (significantly negative) can be interpreted as saying the observed network is sparse. Additional the positive term for mutual implies there is a high probability of reciprocating friendship nominations and the positive term for GWESP says there is a high probability for triangle closure in the network. The transitivity effect (GWESP) was generally found to be stronger in all 3 schools than terms for describing smoking habits in predicting the existence of ties in the network.

As shown above, the final model with terms for smoking assortativity by grade level does a much better job of capturing edge-wise shared partnerships and out degree statistics in the observed network data. The positive coefficients for smoking assortativity in the network show that adolescent smokers form friendships more often than two dissimilar adolescents. For example the MLE of the log-odds of a friendship nomination between a 11^{th} grade light smoker and a 12^{th} grade smoker is 1.11859 with a corresponding probability of $\frac{e^{1.11859}}{(1+e^{1.11859})} = 0.75$. This is the largest positive coefficient for smoking assortativity in the model. This indicates that 11^{th} grade light smokers have the strongest preference to nominate 12^{th} grade smokers than any other smoking adolescent pair in the network. From looking at all the estimates for smoking assortativity, it is clear that homophily is not uniform across all grade and smoking levels. For example, in school2 six out of eight positive coefficients for smoking assortativity include 12^{th} grade smokers in the school, indicating that there is a strong tendency in this social network for 12^{th} grade smokers to either be nominated or nominate other smokers. The second largest positive coefficients for smoking assortativity in school2 are 9^{th} grade light smokers nominating 12^{th} grade light smokers with a log-odds estimate of 0.90952 and a corresponding probability of 0.7129.

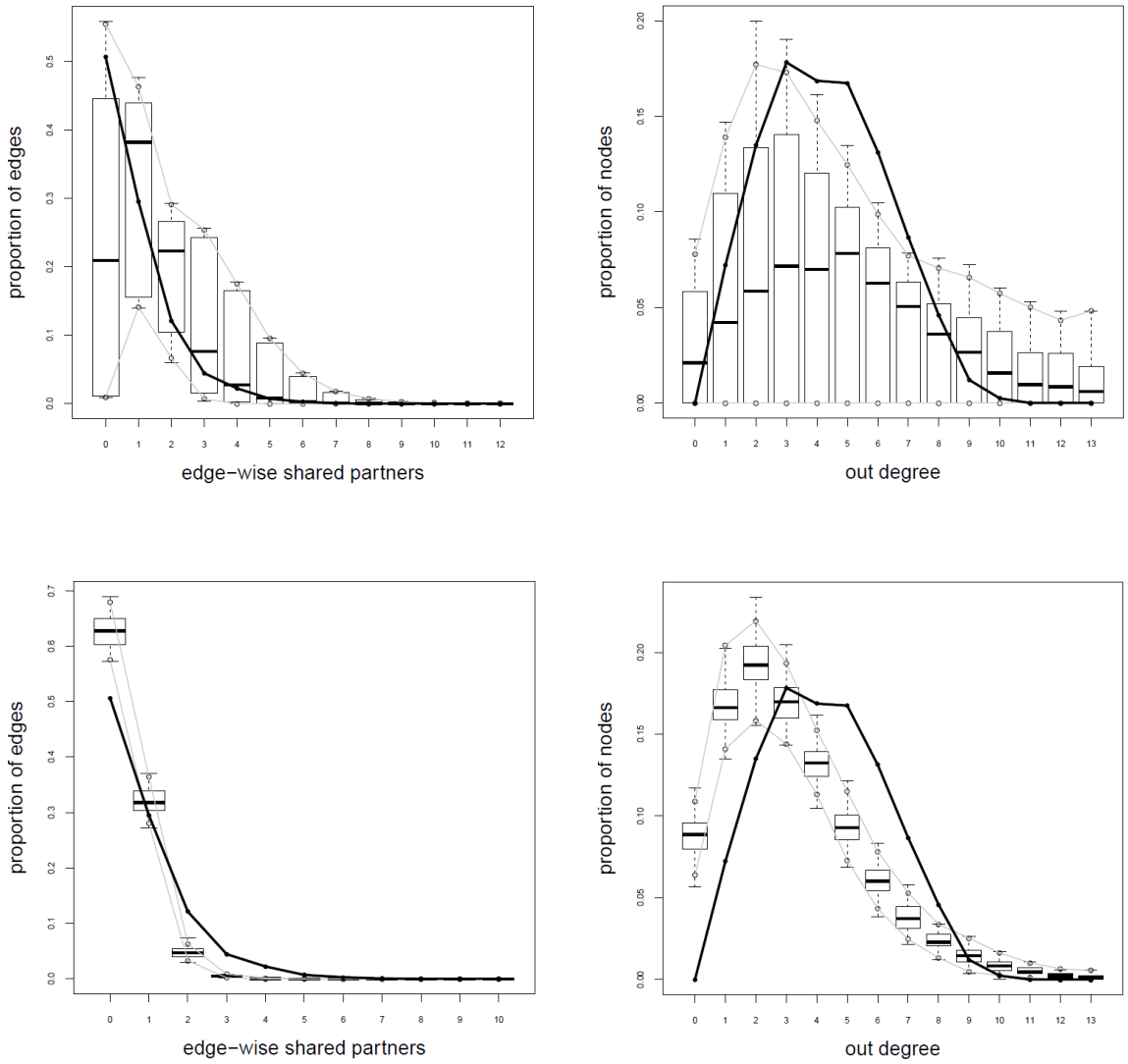
Chapter 5: Conclusion

Exponential random graph models offer a unique approach to modeling both the network and nodal attributes of adolescent friendship networks. The models allow flexible ways of modeling the dependency structure of the data with clear explanatory results of coefficient estimates. Additionally, ERGMs provide a useful technique for characterizing network assortativity and give insight into the detailed assessment of intra-grade and inter-grade smoking relationships. The Add Health data set allows for a reasonable goodness-of-fit to the observed school networks, and an objective way to look at metrics such as probabilities of connections among students of different grades and smoking behaviors. One should be aware that the Add Health data has a fair amount of missing data due to absentee students (affecting out-degree) and nominations to non-unique nodes (students not registered in the same school). These are some of the constraints when working with the limited adolescent friendship network data available. Based on the results from the three final models there are some interesting smoking assortative mixing occurring in the adolescent social networks. The assortative mixing is heterogeneous across smoking habits and grade level, and each school's characteristics are different; however, there are some interesting patterns that emerge across the three schools, like the strong smoking homophily effect for 9^{th} grade smokers nominating smoking peers at all four grade levels examined. Though it has been shown that these models can give adequate descriptions of the network structures examined, there is no reason to assume that these specific model parameters will describe other social networks and it should not be presumed to be a general framework for further analysis.

There are a number of future directions for this research. Further statistically significant model parameters for additional schools with different smoking prevalence could be examined to assess similarities and differences among schools in the Add Health data set. Since much of the dependency in social networks can be observed at the dyadic and triadic network level, investigating other ways of modeling adolescent social networks by taking a more localized approach to parametrization of the dependency structures in the network through the use of multilevel modeling techniques might offer stronger model representations of the observed networks. Finally, another interesting future direction would be to explore dynamic modeling to investigate if the effects of peer influence and smoking homophily in adolescent social networks can be separated by comparing smoking assortative mixing in networks across multiple time steps.

Appendix

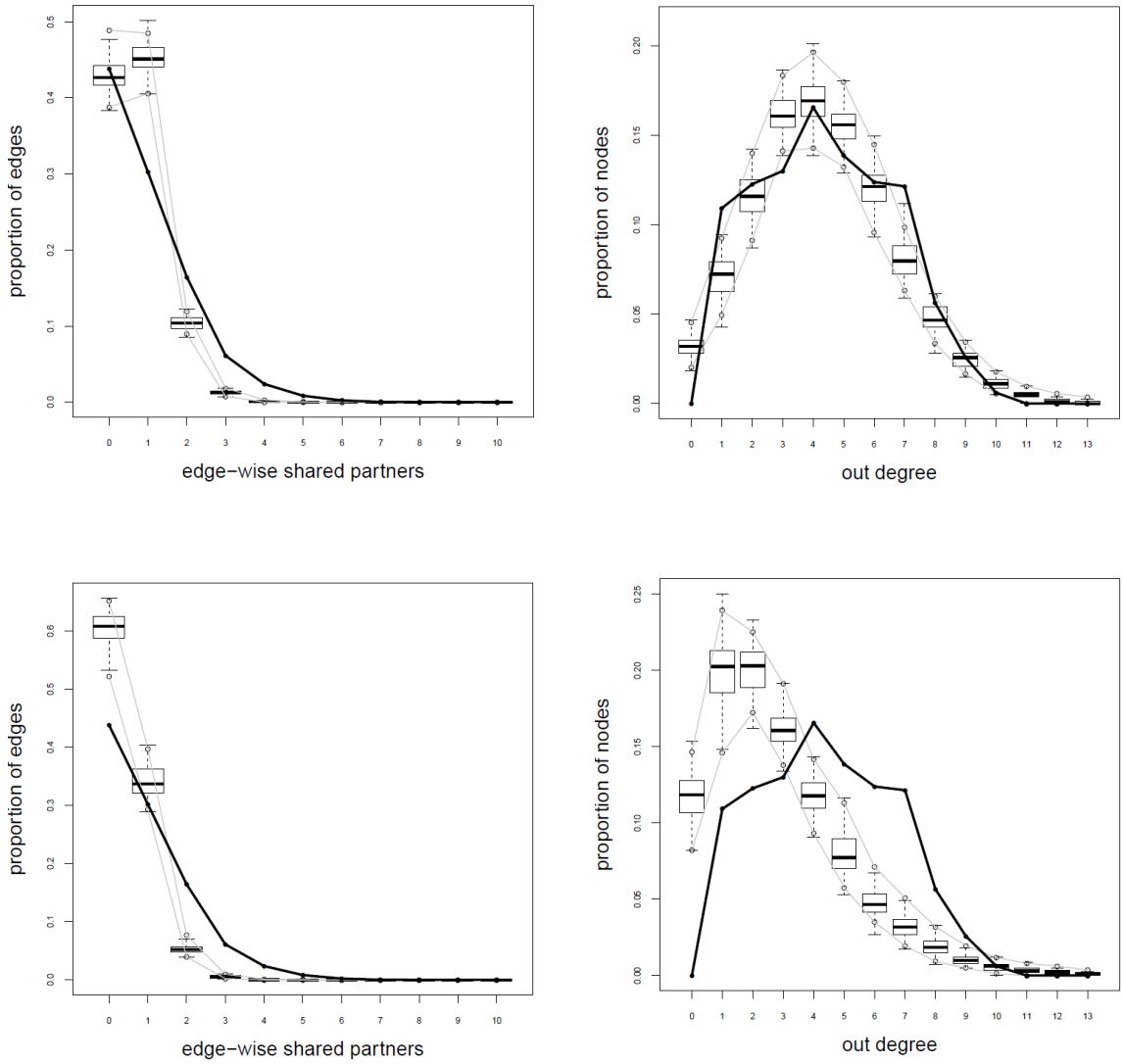
Figure 7: Goodness-of fit comparison school1



1^{st} **row**: 200 simulated networks with all predictor variables.

2^{nd} **row**: 200 simulated networks from first three predictor variables.

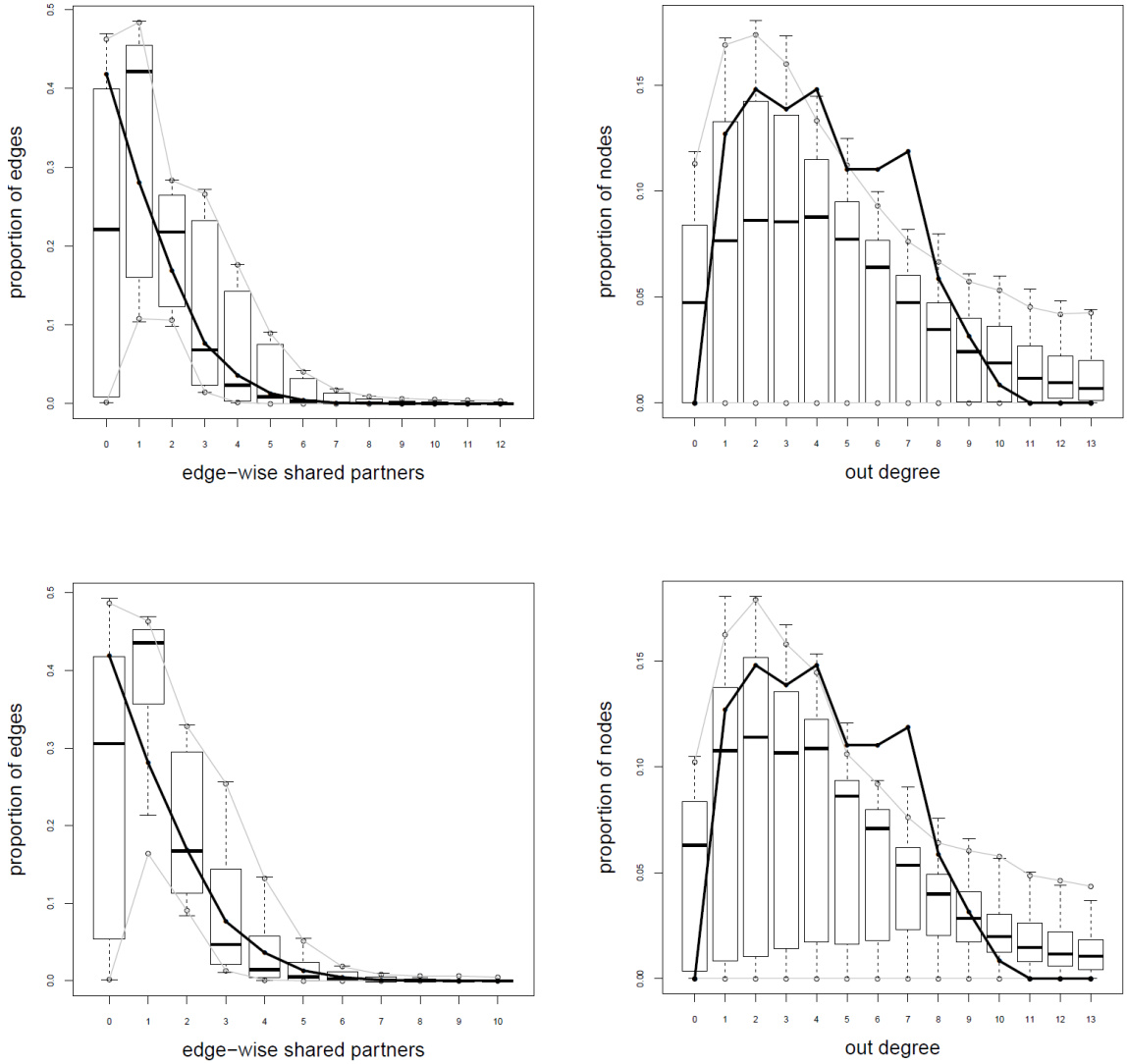
Figure 8: Goodness-of-fit comparison school2



1^{st} **row**: 200 simulated networks with all predictor variables.

2^{nd} **row**: 200 simulated networks from first three predictor variables.

Figure 9: Goodness-of-fit comparison school3



1^{st} **row**: 200 simulated networks with all predictor variables.

2^{nd} **row**: 200 simulated networks from first three predictor variables.

Table 1: Significant terms final model school1

	Estimate	Std. Error	p-value
edges	-6.24949	0.05109	<0.0001
gwesp	1.71636	0.03817	<0.0001
mutual	4.14233	0.02211	<0.0001
9th light smoker<-9th light smoker	-0.68657	0.15040	<0.0001
10th light smoker<-9th light smoker	-0.36167	0.13863	0.00908
10th smoker<-9th light smoker	-1.74732	0.29615	<0.0001
11th light smoker<-9th light smoker	-2.25862	0.36813	<0.0001
9th light smoker<-9th smoker	0.46551	0.11408	<0.0001
11th light smoker<-9th smoker	0.42302	0.16154	0.00883
11th smoker<-9th smoker	-1.16357	0.21811	<0.0001
12th smoker<-9th smoker	0.47075	0.15135	0.00187
9th light smoker<-10th light smoker	0.67523	0.11395	<0.0001
12th smoker<-10th light smoker	0.56347	0.13382	<0.0001
10th smoker<-10th smoker	0.45656	0.08932	<0.0001
11th light smoker<-10th smoker	0.55877	0.16631	0.00078
9th smoker<-11th light smoker	-0.72608	0.21180	0.00061
10th light smoker<-11th light smoker	-0.63264	0.22683	0.00529
12th light smoker<-11th light smoker	-0.76556	0.23330	0.00103
9th light smoker<-11th smoker	0.44770	0.11550	0.00011
10th smoker<-11th smoker	0.51785	0.13664	0.00015
11th light smoker<-11th smoker	0.63327	0.13643	<0.0001
9th smoker<-12th light smoker	0.51027	0.14674	0.00051
11th light smoker<-12th light smoker	0.79414	0.17100	<0.0001
12th smoker<-12th light smoker	0.68685	0.15831	<0.0001
9th smoker<-12th smoker	-1.19371	0.24383	<0.0001
10th smoker<-12th smoker	-0.73044	0.27036	0.00690
11th smoker<-12th smoker	-1.25979	0.23514	<0.0001
12th light smoker<-12th smoker	-0.77377	0.22875	0.00072

Table 2: Significant terms final model school2

	Estimate	Std. Error	p-value
edges	-6.24543	0.07121	<0.0001
gwesp	1.66499	0.05792	<0.0001
mutual	4.03862	0.02877	<0.0001
11th light smoker->12th smoker	1.11859	0.19232	<0.0001
9th light smoker->12th light smoker	0.90952	0.24911	0.000261
12th light smoker->12th light smoker	0.87371	0.15528	<0.0001
12th smoker->12th light smoker	0.76483	0.18838	<0.0001
9th light smoker->10th light smoker	0.70676	0.20574	0.000592
10th light smoker->10th light smoker	0.61774	0.14376	<0.0001
12th smoker->10th light smoker	0.57065	0.17657	0.00123
12th smoker->11th smoker	0.57043	0.15649	0.000267
10th light smoker->12th smoker	-0.73109	0.25743	0.004512
12th smoker ->11th light smoker	-0.95297	0.32592	0.003456
9th light smoker->12th smoker	-1.03374	0.29683	0.000497
10th smoker->10th light smoker	-1.05484	0.37125	0.004493
10th light smoker->11th smoker	-1.33606	0.47025	0.004495

Table 3: Significant terms final model school3

	Estimate	Std. Error	p-value
edges	-6.57566	0.05843	<0.0001
gwesp	1.97628	0.04459	<0.0001
mutual	4.22147	0.02019	<0.0001
9th light smoker<-11th light smoker	1.34554	0.19544	<0.0001
9th smoker<-9th light smoker	1.17585	0.16178	<0.0001
12th light smoker<-10th light smoker	0.88067	0.17193	<0.0001
12th light smoker<-9th light smoker	0.76373	0.23038	0.00092
9th smoker<-12th smoker	0.73466	0.24356	0.00256
10th smoker<-9th light smoker	0.64093	0.19575	0.00106
9th light smoker<-11th smoker	0.62293	0.20521	0.00240
10th light smoker<-9th light smoker	0.61997	0.18357	0.00073
11th smoker<-11th light smoker	0.58426	0.20869	0.00512
11th light smoker<-10th smoker	0.52847	0.18770	0.00487
10th light smoker<-10th smoker	-0.54239	0.20164	0.00715
10th light smoker<-10th light smoker	-0.60316	0.18421	0.00106
11th light smoker<-11th smoker	-0.75593	0.27643	0.00625
10th light smoker<-12 smoker	-0.77772	0.30174	0.00995
9th light smoker<-10th smoker	-0.91676	0.30063	0.00229
11th smoker<-9th light smoker	-0.97169	0.29105	0.00084
9th light smoker<-10th light smoker	-1.23336	0.31734	0.00010
12th light smoker<-11th light smoker	-1.34510	0.50741	0.00803
9th light smoker<-12th light smoker	-1.69170	0.44111	0.00013
11th smoker<-9th smoker	-1.92123	0.46745	<0.0001
11th light smoker<-9th light smoker	-2.01683	0.46059	<0.0001

Figure 10: MCMC density estimate school1

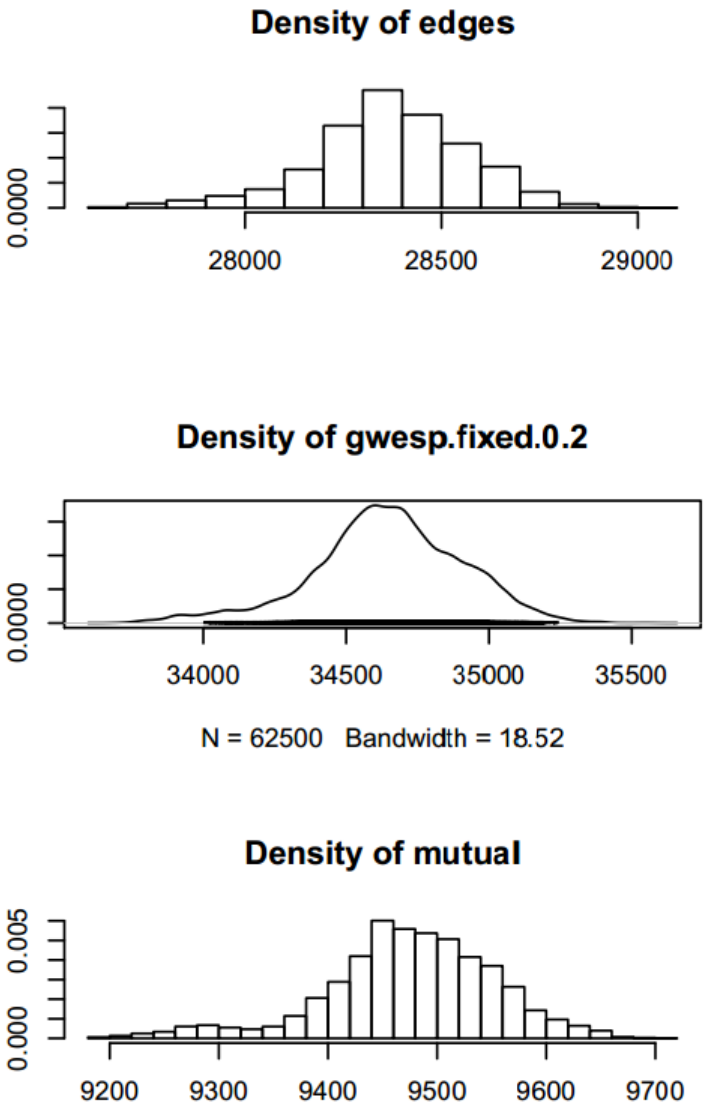


Figure 11: MCMC density estimate school2

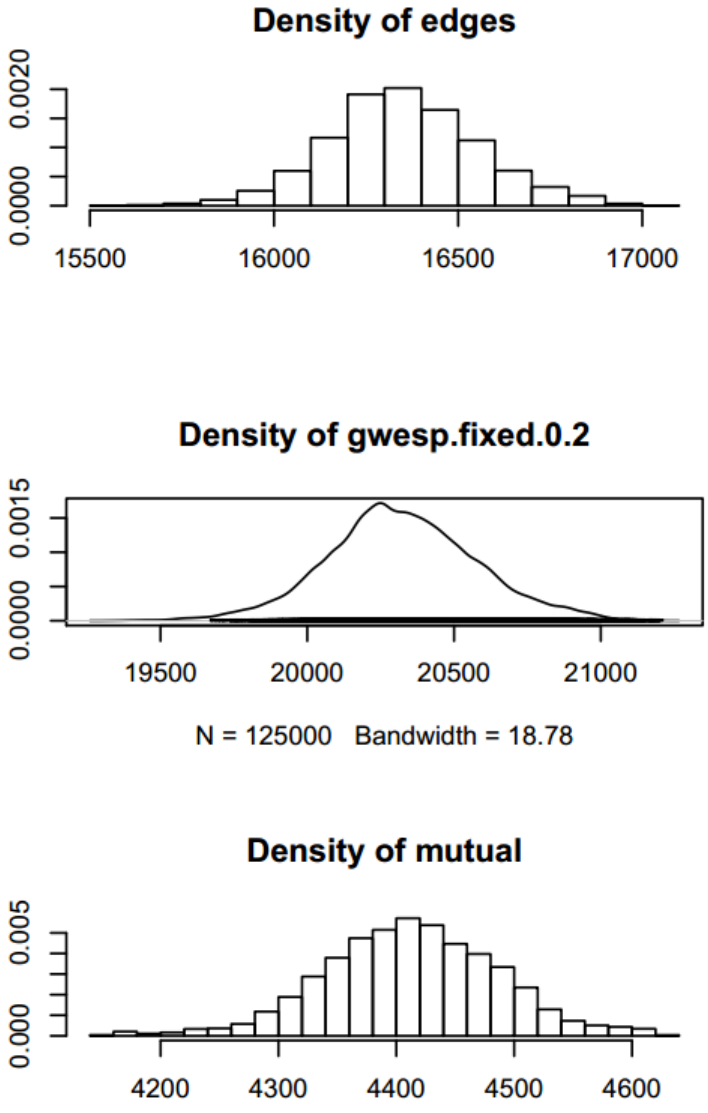
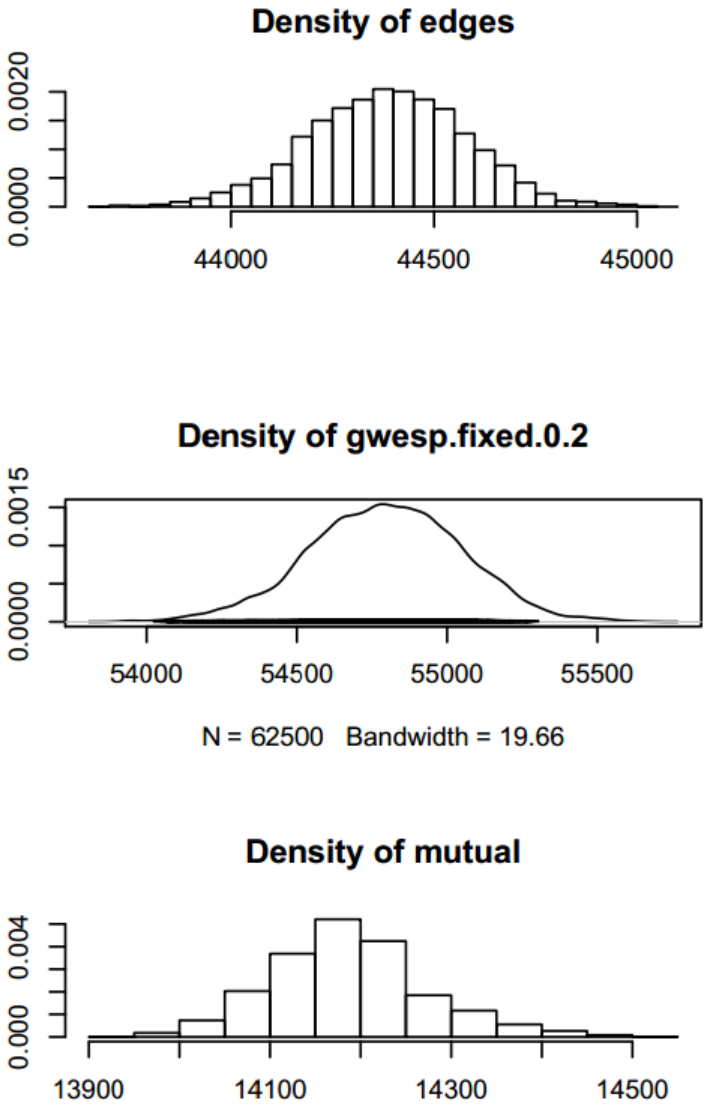


Figure 12: MCMC density estimate school3



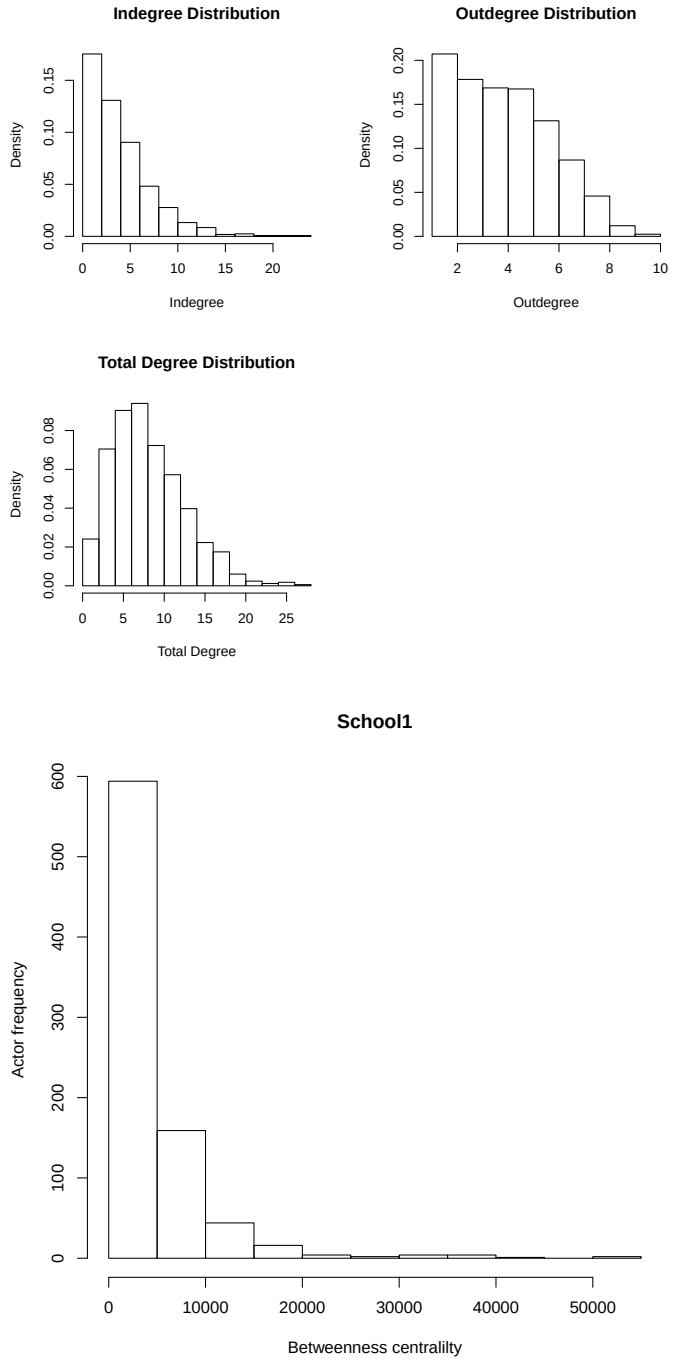


Figure 13: Degree and betweeness centrality school1

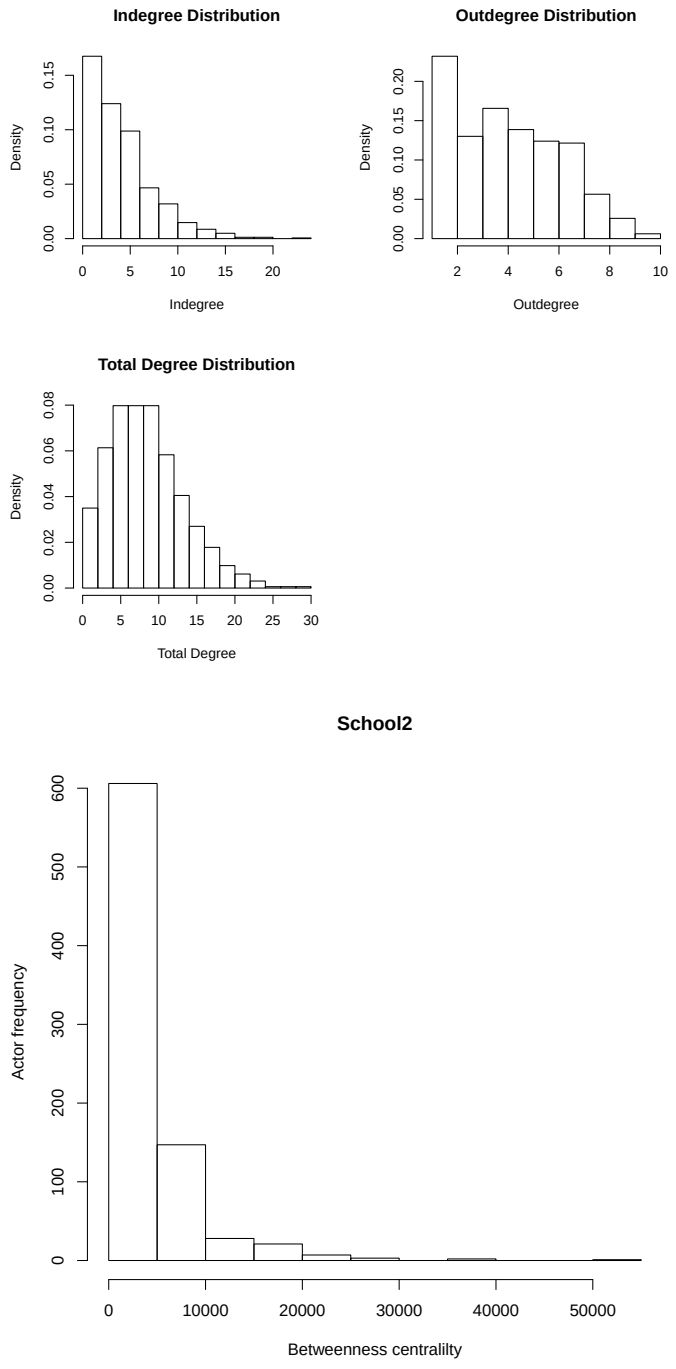


Figure 14: Degree and betweeness centrality school2

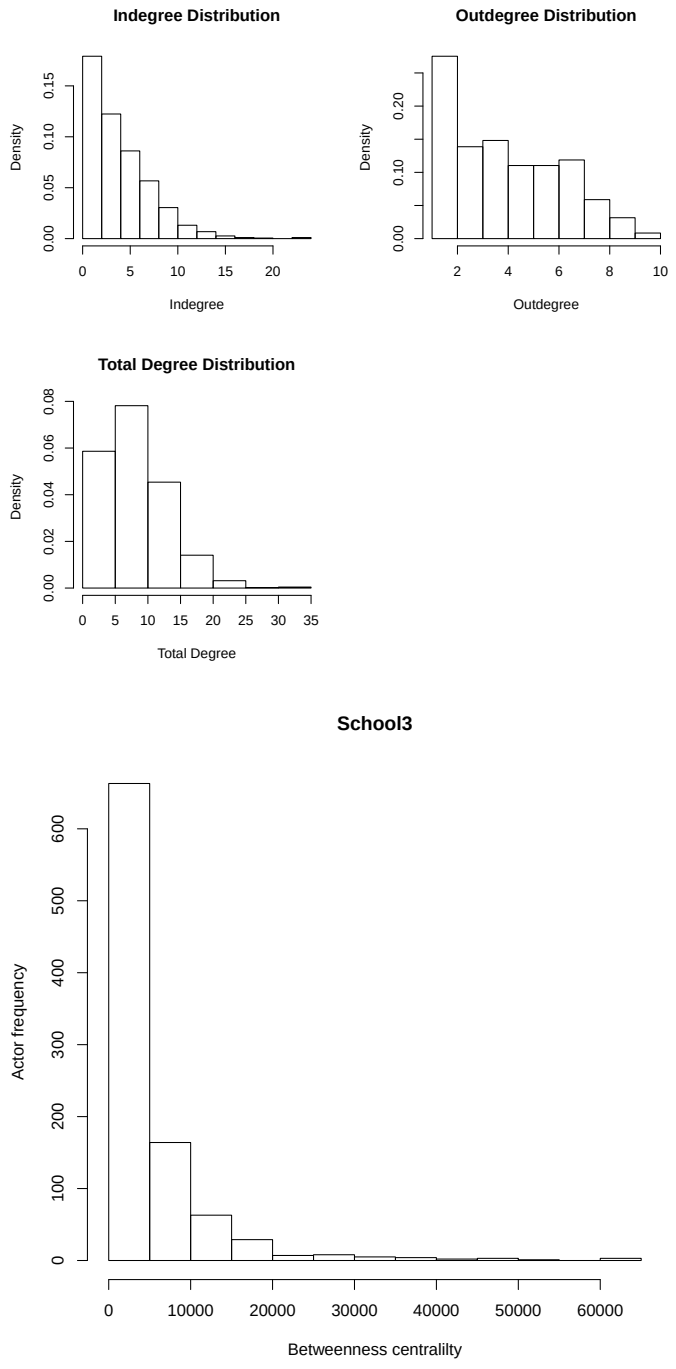


Figure 15: Degree and betweeness centrality school3

References

[1] Rowe DC, Rodgers JL (1991). Adolescent smoking and drinking: Are they epidemics? Journal of Studies on Alcohol. Volume 52.

[2] Nicholas A. Christakis, M.D., Ph.D., M.P.H., and James H. Fowler, Ph.D (2007). The Spread of Obesity in a Large Social Network over 32 Years.

[3] Kobus, K., (2003). Peers and adolescent smoking.

[4] Harris, K.M., C.T. Halpern, E. Whitsel, J. Hussey, J. Tabor, P. Entzel, and J.R. Udry (2009). The National Longitudinal Study of Adolescent Health: Research Design [WWW document].
URL: http://www.cpc.unc.edu/projects/addhealth/design.

[5] Mark S. Handcock and Krista J. Gile (2010). Modeling social networks from sampled data. Annals of Applied Statistics. Volume 4, Issue 1.

[6] Ove Frank; David Strauss (1986). Markov Graphs. Journal of the American Statistical Association, Volume 81, Issue 395.

[7] David Strauss; Michael Ikeda (1990). Pseudolikelihood Estimation for Social Networks. Journal of the American Statistical Association. Volume 85, Issue 409.

[8] G. Robins and M. Morris (2007). “Recent developments in exponential random graph (p^*) models for social networks,” Social Networks. Volume 29, Issue 2.

[9] Koehly L, Goodreau SM, and Morris M. (2004). Exponential family models for sampled and census network data. Sociological Methodology. Volume 34.

[10] Paul W. Holland; Samuel Leinhardt (1981). An Exponential Family of Probability Distributions for Directed Graphs. Journal of the American Statistical Association. Volume 76, Issue 373.

[11] Hunter DR, Handcock MS, Butts CT, Goodreau SM, Morris M (2008). ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. Journal of Statistical Software. Volume 24, Issue 3.

[12] Ove Frank; David Strauss (1986). Markov Graphs. Journal of the American Statistical Association. Volume 81, Issue 395.

[13] Wasserman S, Pattison P (1996). Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p*. Psychometrika. Volume 61, Issue 3.

[14] Pavel N. Krivitsky and Mark S. Handcock (2010). A Separable Model for Dynamic Networks.

[15] Nowicki, K. and T. A. B. Snijders (2001). Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association. Volume 96, Issue 45.

[16] David R. Hunter, Steven M.Goodreau, Mark S. Handcock (2005). Goodness of Fit of Social Networks Models. Center for Statistics and the Social Sciences University of Washington.

[17] Goodreau SM, Kitts JA, and Morris M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. Demography.

[18] Cheryl Alexander, Marina Piazza, Debra Mekos, Thomas Valente (2001). Peers, schools, and adolescent cigarette smoking. Journal of Adolescent Health. Volume 29, Issue 1.

[19] Nicholas A. Christakis, James H. Fowler. (2008). The Collective Dynamics of Smoking in a Large Social Network. New England Journal of Medicine.

[20] Robins, G.L., Elliot, P., & Pattison, P.E. (2001b). Network models for social selection processes. Social Networks. Volume 23 Issue 1.

[21] Snijders TAB, Pattison P, Robins GL, Handcock MS (2006). New Specifications for Exponential Random Graph Models.
Sociological Methodology. Volume 36, Issue 1.
[22] Hunter DR, Handcock MS, Butts CT, Goodreau SM, Morris M (2008). ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. Journal of Statistical Software. Volume 24, Issue 3.