Center for Studies in Demography and Ecology

Goodness of Fit of Social Network Models

by

David R. Hunter
Pennsylvania State University

Steven M. Goodreau
University of Washington

Mark S. Handcock
University of Washington

UNIVERSITY OF WASHINGTON

CSDE Working Paper No. 05-05

Goodness of Fit of Social Network Models[1]

David R. Hunter
Pennsylvania State University, University Park

Steven M. Goodreau
University of Washington, Seattle

Mark S. Handcock
University of Washington, Seattle

Working Paper no. 47
Center for Statistics and the Social Sciences
University of Washington

April 28, 2005

[1]David R. Hunter is Assistant Professor of Statistics, Department of Statistics, Pennsylvania State University, University Park, PA 16802 (E-mail: dhunter@stat.psu.edu); Steven M. Goodreau is Assistant Professor of Anthropology, Department of Anthropology, University of Washington, Box 35100, Seattle WA 98195-5100. E-mail: goodreau@u.washington.edu; Web: faculty.washington.edu/goodreau; Mark S. Handcock is Professor of Statistics and Sociology, Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322. E-mail: handcock@stat.washington.edu; Web: www.stat.washington.edu/handcock. The authors are grateful to Martina Morris for numerous helpful suggestions. This research is supported by Grant DA012831 from NIDA and Grant HD041877 from NICHD.

Abstract

We present a systematic examination of real network datasets using maximum likelihood estimation for exponential random graph models as well as new procedures to evaluate how well the models fit the observed graphs. These procedures compare structural statistics of the observed graph with the corresponding statistics on graphs simulated from the fitted model. We apply this approach to the study of friendship relations among high school students from the the National Longitudinal Study of Adolescent Health (AddHealth). The sizes of the networks we fit range from 71 to 2209 nodes. The larger networks represent more than an order of magnitude increase over the size of any network previously fit using maximum likelihood methods for models of this kind. We argue that several well-studied models in the networks literature do not fit these data well, and we demonstrate that the fit improves dramatically when the models include the recently-developed geometrically weighted edgewise shared partner (GWESP) and geometrically weighted degree (GWD) network statistics. We conclude that these models capture aspects of the social structure of adolescent friendship relations not represented by previous models.

Key Words: degeneracy, exponential random graph model, maximum likelihood estimation, Markov chain Monte Carlo, $p-$star model

1 Introduction

Among the many statistical methods for dealing with dependent data of various types developed in recent decades, social network models are especially useful for dealing with the kinds of dependence induced by social relations. Much effort has been focused on inference for social network models (e.g., Holland and Leinhardt 1981; Strauss and Ikeda 1990; Snijders, 2002; Hunter and Handcock, 2004), but comparatively little work tests the goodness of fit of the models. We present an approach within the exponential random graph model (ERGM) framework and illustrate its effectiveness using data from the National Longitudinal Study of Adolescent Health (AddHealth).

Relational data can be described as data whose properties cannot be reduced to the attributes of the individuals involved. They are a particularly common form of data in the social sciences, where relationships among pairs of individual actors represent a central object of inquiry. Such data can be represented as a network, or mathematical graph, consisting of a set of nodes and a set of edges, where an edge is an ordered or unordered pair of nodes. Graphically, it is possible to represent a network as in Figure 1, in which the nodes are of various shapes and the presence of an edge is indicated by a line connecting two nodes. It may be the case that there are measurements associated with each of the actors; we refer to these measurements as *nodal covariates*. The different shapes and labels of the nodes in Figure 1 represent different values of categorical nodal covariates for these network data.

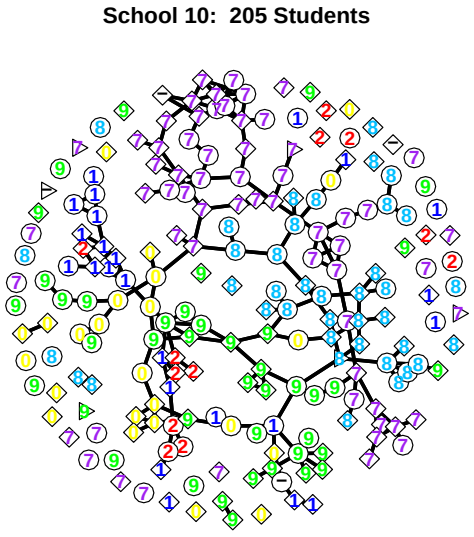

Figure 1: Mutual friendships represented as a network. Shapes of nodes denote sex: circles for female, squares for male, and triangles for unknown. Labels denote the units digit of grade (7 through 12), or "–" for unknown.

In typical applications, the nodes in a graph represent individuals and the edges represent a specified relationship between individuals. Nodes can also be used to represent larger social units, such as groups, families, or organizations; objects, such as physical resources, servers, or locations; or abstract entities, such as concepts, texts, tasks, or random variables. Networks have been applied to a wide variety of situations, including the structure of social networks, the dynamics

of epidemics, the interconnectedness of the World Wide Web, and long-distance telephone calling patterns. This article focuses specifically on network data collected at a nationally representative sample of middle schools and high schools in the United States.

We consider exponential family models, in the traditional statistical sense, for network structure. These models have a long history in the networks literature, and we refer to them here as exponential random graph models (ERGMs). The primary contribution of this article is to demonstrate that it is possible to achieve reasonably good model fit on large networks using ERGMs that include both a small number of covariates measured on the nodes and key network statistics chosen to capture the structural properties of the network. Although such models have been proposed before, no attempts to assess goodness of fit have accompanied them (Wasserman and Pattison, 1996; Snijders et al. 2004). Another contribution of this paper is to demonstrate the use of maximum likelihood to fit reasonable models to network data with hundreds of nodes and obtain results that are scientifically useful. We have developed an `R` package (called `statnet`) to implement the procedures developed in this paper. The package is available at `http://csde.washington.edu/statnet`.

It is possible to simulate random networks from a given ERGM — at least in principle — using well-established Markov chain Monte Carlo techniques. More recently, various researchers have been developing techniques to solve a harder problem: calculating approximate maximum likelihood estimates of the ERGM parameters, given an observed network. While these techniques are conceptually simple (Geyer and Thompson 1992), their practical implementation for large networks has proven elusive. We are now able to apply these techniques to networks encompassing thousands of nodes, problems more than an order of magnitude larger than any previous application of which we are aware.

In problems for which maximum likelihood estimation previously has been possible in ERGMs, a troubling empirical fact has emerged: When ERGM parameters are estimated and a large number of graphs are simulated from the resulting model, these graphs frequently bear no resemblance at all to the observed network. This seemingly paradoxical fact arises because even though the maximum likelihood estimate makes the probability of the observed graph as large as possible, this probability might still be extremely small. In such a case, the ERGM does not fit the data well.

The remainder of this article provides a case study illustrating the application of new model-fitting capabilities and goodness of fit procedures to network datasets from the National Longitudinal Study of Adolescent Health (AddHealth), which is described in Section 2. Section 3 explains the statistical models we will fit to these data along with the techniques we use for doing so, while Section 4 lists potential difficulties encountered along the way. Section 5 illustrates our goodness of fit technique on a couple of simple models that do not fit well. Finally, Section 6 presents a model that fits the mutual friendship data well and explains what we have been able to learn about high school friendship networks as a result.

2 Introduction to the AddHealth Survey

The network data on friendships that we study in this article were collected during the first wave (1994–1995) of the National Longitudinal Study of Adolescent Health (AddHealth). The AddHealth data come from a stratified sample of schools in the US containing students in grades 7 through 12. To collect friendship network data, AddHealth staff constructed a roster of all students

in a school from school administrators. Students were then provided with the roster and asked to select up to five close male friends and five close female friends. Students were allowed to nominate friends who were outside the school or not on the roster, or to stop before nominating five friends of either sex. Complete details of this and subsequent waves of the study can be found in Resnick et al. (1997) and Udry and Bearman (1998) and at `http://www.cpc.unc.edu/projects/addhealth`. In most cases, the individual school does not contain all grades 7–12; instead, data were collected from multiple schools within a single system (e.g. a junior high school and a high school) to obtain the full set of six grades. In these cases, we will use the term "school" to refer to a set of schools from one community. The full dataset contains 86 schools, 90,118 student questionnaires, and 578,594 friendship nominations. Our analysis includes 59 of the schools, ranging in size from 71 to 2209 surveyed students.

The edges in these raw network data are directed, since it is possible A could name B as a friend without B nominating A. However, in this article we will consider the undirected network of *mutual* friendships, those in which both A nominates B and B nominates A.

Each network may be represented by a symmetric $n \times n$ matrix $\mathbf{Y}$ and an $n \times q$ matrix $\mathbf{X}$ of nodal covariates, where n is the number of nodes. The entries of the $\mathbf{Y}$ matrix, termed the *adjacency matrix*, are all zeros and ones, with $Y_{ij} = 1$ indicating the presence of an edge between i and j. Since self-nomination was disallowed, $Y_{ii} = 0$ for all i. The limit on the number of allowed nominations means that the data are not complete, but we will assume for convenience that a lack of nomination in either direction between two individuals means that there is no mutual friendship.

The nodal covariate matrix $\mathbf{X}$ includes many measurements on each of the individuals in these networks. Some such measurements, like sex, are not influenced by network structure in any way, and are termed *exogenous*. Other covariates may exhibit strong non-exogeneity: For example, tobacco use may be influenced through friendships. Exogeneity comes into play, for instance, in claiming that the dyadic independence model of equation (4) truly has the dyadic independence property as advertised. We focus our analysis on only three covariates: sex, grade, and race. Although the latter two may exhibit some endogeneity (e.g., the influence of friends may affect whether a student fails and must repeat a grade, or which race a student of mixed-race heritage chooses to identify with), we assume such effects are minimal and consider the attributes fixed and exogenous. What we term "race" is constructed from two questions on race and Hispanic origin, with Hispanic origin taking precedence. Thus, our categories "Hispanic", "Black", "White", "Asian", "Native American", and "Other" are short-hand names for "Hispanic (all races)", "Black (non-Hispanic)", "White (non-Hispanic)", etc.

Though in this article we focus primarily on a single illustrative school, we analyzed many schools. Schools with large amounts of missing data were excluded from the analysis; this happened, among other reasons, for special education schools and for school districts that required explicit parental consent for student participation. Results for all the schools we analyzed may be found at `http://csde.washington.edu/networks`.

3 Exponential Random Graph Models

Our overall goal in using exponential random graph models (ERGMs) is to model the random behavior of the adjacency matrix $\mathbf{Y}$, conditional on the covariate matrix $\mathbf{X}$. Given a user-defined

p-vector $\mathbf{g}(\mathbf{Y}, \mathbf{X})$ of statistics and letting $\boldsymbol{\eta} \in R^p$ denote the statistical parameter, these models form a canonical exponential family (Lehmann, 1983),

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}|\mathbf{X}) = c^{-1} \exp\{\boldsymbol{\eta}^t \mathbf{g}(\mathbf{y}, \mathbf{X})\}, \tag{1}$$

where the normalizing constant $c \equiv c(\boldsymbol{\eta})$ is defined by

$$c = \sum_{\mathbf{w}} \exp\{\boldsymbol{\eta}^t \mathbf{g}(\mathbf{w}, \mathbf{X})\} \tag{2}$$

and the sum (2) is taken over the whole sample space of allowable graphs. The form of $\mathbf{g}(\mathbf{Y}, \mathbf{X})$ is the essential modelling task and is determined by the researcher. The objective is to choose statistics that represent aspects of the social structure inherent in the network. The range of substantially motivated network statistics that might be included in the $\mathbf{g}(\mathbf{Y}, \mathbf{X})$ vector is vast — see Wasserman and Faust (1994) for the most comprehensive treatment of these statistics — though we will consider only a few key ones in this article. The statistical problem is to estimate $\boldsymbol{\eta}$ given network data and assess the quality of the resulting fit.

3.1 Interpretation and background

Perhaps the simplest way to interpret the probability model (1) is by thinking of the conditional probability that an edge exists between two nodes, conditional on the state of the rest of the graph. Let node indices i and j be fixed, and denote "the rest of the graph" (that is, all of $\mathbf{Y}$ except for Y_{ij}) by $\mathbf{Y}^c_{ij}$. If we let $\text{logit}\,(p) = \log(p) - \log(1-p)$ denote the logit function well known in logistic regression, then equation (1) implies

$$\text{logit}\left\{P(Y_{ij} = 1|\mathbf{Y}^c_{ij} = \mathbf{y}^c_{ij})\right\} = \boldsymbol{\eta}^t \Delta(\mathbf{g}(\mathbf{y}, \mathbf{X}))_{ij}, \tag{3}$$

where $\Delta(\mathbf{g}(\mathbf{y}, \mathbf{X}))_{ij} = \mathbf{g}(\mathbf{y}, \mathbf{X})|_{y_{ij}=1} - \mathbf{g}(\mathbf{y}, \mathbf{X})|_{y_{ij}=0}$ denotes the change in the vector of statistics when y_{ij} is changed from 0 to 1 and the rest of $\mathbf{y}$ (i.e., $\mathbf{y}^c_{ij}$) remains unchanged.

As usual with generalized linear models involving more than one predictor variable, it is dangerous to interpret the individual components of the $\boldsymbol{\eta}$ vector in isolation. For instance, if $g_1(\mathbf{y})$ equals the number t($\mathbf{y}$) of triangles in $\mathbf{y}$ (about which we will say more in Section 3.3), then η_1 represents the increase in log-odds of an edge for each new triangle that edge would create *after all other effects in the model are taken into account*.

Holland and Leinhardt (1981) appear to be the first to propose a specific case of model (1) in the literature. Their model, which they called the p_1 model, resulted in the set of dyads being independent, where a *dyad* is an ordered pair (Y_{ij}, Y_{ji}) for some pair $i < j$ of nodes. Based on developments in spatial statistics (Besag 1974), Frank and Strauss (1986) generalized to the case in which dyads exhibit a kind of Markovian dependence: two dyads are dependent, conditional on the rest of the graph, only when they have a node in common. Frank (1991) mentioned the application of model (1) to social networks in its full generality, a topic pursued in depth by Wasserman and Pattison (1996). In honor of Holland and Leinhardt's p_1 model, Wasserman and Pattison (1996) referred to model (1) as p^* (p-star), a name that has been widely applied to ERGMs in the social networks literature.

Development of estimation methods for ERGMs has not kept pace with development of ERGMs themselves. To understand why, consider the sum of equation (2). A sample space consisting of all

possible undirected graphs on n nodes contains $2^{n(n-1)/2}$ elements, an astronomically large number even for moderate n (e.g., for $n = 20$ there are 1.6×10^{57} graphs). Therefore, direct evaluation of the normalizing constant c in equation (2) is computationally infeasible for all but the smallest networks — except in certain special cases such as the dyadic independence model of equation (5) — and inference using maximum likelihood estimation is extremely difficult. To circumvent this difficulty, we use a technique called Markov chain Monte Carlo maximum likelihood estimation in which a stochastic approximation to the likelihood function is built and then maximized (Geyer and Thompson 1992). This and other methods have been considered by Dahmström and Dahmström (1993), Corander et al. (1998), Crouch et al. (1998), Snijders (2002), and Handcock (2002). Details of the specific technique we use may be found in Hunter and Handcock (2004).

3.2 Dyadic independence models and pseudolikelihood

An important special case of model (1) is the *dyadic independence* model, in which

$$\mathbf{g}(\mathbf{y}, \mathbf{X}) = \sum_{i<j} \sum y_{ij} \mathbf{h}(\mathbf{X}_i, \mathbf{X}_j) \tag{4}$$

for some function $\mathbf{h}$ mapping $\mathbb{R}^q \times \mathbb{R}^q$ into $\mathbb{R}^p$, where the q-dimensional row vectors $\mathbf{X}_i$ and $\mathbf{X}_j$ are the nodal covariate vectors for the ith and jth individuals. In the context of an undirected network, the word *dyad* refers to a single Y_{ij} for some pair (i, j) of nodes (not to be confused with an *edge*, which requires $Y_{ij} = 1$). The ERGM resulting from equation (4) is called the dyadic independence model because equation (1) becomes

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}|\mathbf{X}) = c^{-1} \prod_{i<j} \prod \exp\{y_{ij} \boldsymbol{\eta}^t \mathbf{h}(\mathbf{X}_i, \mathbf{X}_j)\}, \tag{5}$$

and the joint distribution of the Y_{ij} is simply the product of the marginal distributions. In this case, one can obtain the MLE using logistic regression. As the simplest example of a dyadic independence model, we take $p = 1$ and $h(\mathbf{X}_i, \mathbf{X}_j) = 1$, which yields the well-known Bernoulli graph, also known as the Erdős-Rényi graph, in which each dyad is an edge with probability $\exp\{\eta\}/(1 + \exp\{\eta\})$.

Using the notation of equation (3), note that equation (4) implies that $\Delta(\mathbf{g}(\mathbf{y}, \mathbf{X}))_{ij} = \mathbf{h}(\mathbf{X}_i, \mathbf{X}_j)$, so the likelihood (5) may be rewritten

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}|\mathbf{X}) = c^{-1} \prod_{i<j} \prod \exp\{y_{ij} \boldsymbol{\eta}^t \Delta(\mathbf{g}(\mathbf{y}, \mathbf{X}))_{ij}\}. \tag{6}$$

For dyadic dependence models, equation (6) is not generally true, but nonetheless the right hand side of this equation is called the *pseudolikelihood*. Until recently, inference for social network models has relied on maximum pseudolikelihood estimation, or MPLE, which may be implemented using a standard logistic regression algorithm (Besag 1974; Frank and Strauss, 1986; Strauss and Ikeda, 1990; Geyer and Thompson 1992). However, it has been argued that MPLE can perform very badly in practice (Geyer and Thompson, 1992) and that its theoretical properties are poorly understood (Handcock, 2003). Particularly dangerous is the practice of interpreting standard errors from logistic regression output as though they are reasonable estimates of the standard deviations of the pseudolikelihood estimators. The only estimation technique we discuss for the remainder of this article is maximum likelihood estimation.

3.3 Structural properties of networks: Degree, shared partner, and other statistics

Perhaps the simplest ERGMs that are not dyadic independence models are those in which $\mathbf{g}(\mathbf{y}, \mathbf{X})$ consists only of a subset of the degree statistics $\mathrm{d}_k(\mathbf{y})$, $0 \le k \le n-1$. The degree of a node in a network is the number of neighbors it has, where a neighbor is a node with which it shares an edge. We define $\mathrm{d}_k(\mathbf{y})$ to be the number of nodes in the graph $\mathbf{y}$ that have degree k. Note that the $\mathrm{d}_k(\mathbf{y})$ statistics satisfy the constraint $\sum_{i=0}^{n-1} \mathrm{d}_i(\mathbf{y}) = n$, so it is unwise to include all n degree statistics among the components of the vector $\mathbf{g}(\mathbf{y}, \mathbf{X})$; if we did, the coefficients in model (1) would not be identifiable. A common reformulation of the degree statistics is given by the k-star statistics $\mathrm{s}_1(\mathbf{y}), \ldots, \mathrm{s}_{n-1}(\mathbf{y})$, where $\mathrm{s}_k(\mathbf{y})$ is the number of k-stars in the graph $\mathbf{y}$. A k-star is an unordered set of k edges that all share a common node. For instance, "1-star" is synonymous with "edge". Since a node with i neighbors is the center of $\binom{i}{k}$ k-stars (but the "common node" of a 1-star may be considered arbitrarily to be either of two nodes), we see that

$$\mathrm{s}_k(\mathbf{y}) = \sum_{i=k}^{n-1} \binom{i}{k} \mathrm{d}_i(\mathbf{y}), \ 2 \le k \le n-1; \quad \text{and} \quad \mathrm{s}_1(\mathbf{y}) = \frac{1}{2} \sum_{i=1}^{n-1} i \, \mathrm{d}_i(\mathbf{y}). \tag{7}$$

The k-star statistics are highly collinear with one another: For example, any 4-star automatically comprises four 3-stars, six 2-stars, and four 1-stars (or edges).

An additional class of statistics that will be useful later on are the shared partner statistics. We define two distinct sets of shared partner statistics, the *edgewise* shared partner statistics and the *dyadic* shared partner statistics. The edgewise shared partner statistics are denoted $\mathrm{ep}_0(\mathbf{y}), \ldots, \mathrm{ep}_{n-2}(\mathbf{y})$, where $\mathrm{ep}_k(\mathbf{y})$ is defined as the number of unordered pairs $\{i, j\}$ such that $y_{ij} = 1$ and i and j have exactly k common neighbors (Hunter and Handcock, 2004). The requirement that $y_{ij} = 1$ distinguishes the edgewise shared partner statistics from the dyadic shared partner statistics $\mathrm{dp}_0(\mathbf{y}), \ldots, \mathrm{dp}_{n-2}(\mathbf{y})$: We define $\mathrm{dp}_k(\mathbf{y})$ to be the number of pairs $\{i, j\}$ such that i and j have exactly k common neighbors. In particular, it is always true that $\mathrm{dp}_k(\mathbf{y}) \ge \mathrm{ep}_k(\mathbf{y})$, and in fact $\mathrm{dp}_k(\mathbf{y}) - \mathrm{ep}_k(\mathbf{y})$ equals the number of unordered pairs $\{i, j\}$ for which $y_{ij} = 0$ and i and j share exactly k common neighbors.

Since there are $\mathrm{s}_1(\mathbf{y})$ edges and $\binom{n}{2}$ dyads in the entire network, we obtain the identities

$$\mathrm{s}_1(\mathbf{y}) = \sum_{i=0}^{n-2} \mathrm{ep}_i(\mathbf{y}) \tag{8}$$

and

$$\binom{n}{2} = \sum_{i=0}^{n-2} \mathrm{dp}_i(\mathbf{y}). \tag{9}$$

Furthermore, we can obtain the number of triangles in $\mathbf{y}$ by considering the edgewise shared partner statistics: Whenever $y_{ij} = 1$, the number of triangles that include this edge is exactly the number of common neighbors shared by i and j. Therefore, if we count all of the shared partners for all edges, we have counted each triangle three times, once for each of its edges. In other words,

$$\mathrm{t}(\mathbf{y}) = \frac{1}{3} \sum_{i=0}^{n-2} i \, \mathrm{ep}_i(\mathbf{y}). \tag{10}$$

A related formula involving the dyadic shared partner statistics is obtained by noting that each triangle automatically comprises three 2-stars. Therefore, $s_2(\mathbf{y}) - 3t(\mathbf{y})$ is the number of 2-stars for which the third side of the triangle is missing. We conclude that

$$s_2(\mathbf{y}) - 3t(\mathbf{y}) \;=\; \sum_{i=0}^{n-2} i \left[dp_i(\mathbf{y}) - ep_i(\mathbf{y}) \right]. \tag{11}$$

Combining equation (11) with equation (10) produces

$$s_2(\mathbf{y}) = \sum_{i=0}^{n-2} i \, dp_i(\mathbf{y}).$$

Because a 2-star is also a path of length two, $s_2(\mathbf{y})$ is sometimes referred to as the twopath statistic.

Finally, we summarize two additional sets of statistics, due to Snijders et al. (2004), that will be used in Section 6. First, the triangle statistic generalizes to the set of *k*-triangle statistics, where a *k*-triangle is defined to be a set of *k* distinct triangles that share a common edge. In particular, a 1-triangle is the same thing as a triangle. Second, the 2-star statistic (also known as the twopath statistic) generalizes to the set of *k*-twopath statistics, where a *k*-twopath is a set of *k* distinct 2-paths joining the same pair of nodes. In particular, a 1-twopath is the same thing as a 2-star or a 2-path. Snijders et al (2004) actually coined the term "*k*-independent 2-path," but we simplify this to *k*-twopath in this article.

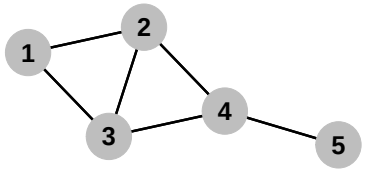

Figure 2: For this simple five-node network, the edgewise and dyadic shared partner distributions are $(ep_0, \ldots, ep_3) = (1, 4, 1, 0)$ and $(dp_0, \ldots, dp_3) = (2, 6, 2, 0)$, respectively; the *k*-triangle and *k*-twopath distributions are $(t_1, t_2, t_3) = (1, 2, 0)$ and $(u_1, u_2, u_3) = (10, 1, 0)$, respectively.

As a concrete example, we note that in the simple network of Figure 2, there are two 1-triangles; one 2-triangle; ten 1-twopaths; and one 2-twopath. (Note that the 2-twopath joining nodes 1 and 4 is the same as the 2-twopath joining nodes 2 and 3, though it is counted only once.) We denote the number of *k*-triangles and *k*-twopaths in the network $\mathbf{y}$ by $t_k(\mathbf{y})$ and $u_k(\mathbf{y})$, respectively. Just as the degree statistics $d_i(\mathbf{y})$ are related to the *k*-star statistics $s_k(\mathbf{y})$ by (7), the edgewise and dyadic shared partner statistics are related to the *k*-triangle and *k*-twopath statistics, respectively, by the equations

$$t_k(\mathbf{y}) = \sum_{i=k}^{n-2} \binom{i}{k} ep_i(\mathbf{y}), \;\; 2 \le k \le n-2$$

and

$$u_k(\mathbf{y}) = \sum_{i=k}^{n-2} \binom{i}{k} dp_i(\mathbf{y}), \;\; 1 \le k \le n-2, k \ne 2.$$

The cases not covered above are that of $t_1(\mathbf{y})$, given in equation (10), and $u_2(\mathbf{y})$, the number of 4-cycles, which includes an extra factor of $1/2$ because any 4-cycle can be considered a 2-path between two distinct pairs of nodes:

$$u_2(\mathbf{y}) = \frac{1}{2}\sum_{i=2}^{n-2}\binom{i}{2}dp_i(\mathbf{y})$$

4 Difficulties in Fitting ERGMs

Although the theory for fitting ERGMs using MCMC to obtain approximate maximum likelihood estimators is well developed, few applications have appeared. This is because these models prove very difficult to fit in practice, and the results are not always useful. There are several interrelated reasons for this.

Because it is impossible to find the MLE or even evaluate the likelihood exactly for graphs of moderate size, as explained in Section 3, approximation methods must be used. Approximations to the loglikelihood based on MCMC sampling can be very bad near the maximizer. Depending on the application, this might be because the Markov chain does not produce a very representative sample, a problem elucidated by Snijders (2002); or because the exponentiation involved in approximating the normalizing constant c^{-1} of equation (1) is unstable in the sense that it tends to magnify small numerical errors. This latter problem is the subject of quite a bit of work over the last couple decades on estimating (ratios of) normalizing constants; see Hunter and Handcock (2004) for a fuller discussion. Even when the loglikelihood function has a reasonably good approximation, there is the numerical challenge of maximizing this approximation, a task that is subject to many of the usual difficulties of high-dimensional numerical optimization. An alternative estimation procedure, due to Snijders (2002), attempts to find a maximum likelihood estimator by solving a moment equation rather than by maximizing a function; yet this procedure is subject to numerical problems of its own.

A related cause of failure of the fitting algorithm occurs when the approximate likelihood function to be maximized does not have a maximizer at all. Indeed, sometimes even the true likelihood function cannot be maximized. This problem is familiar to any practicing statistician who has run a logistic regression that has failed to converge. According to well-known maximum likelihood theory (see, for example, Barndorff-Nielsen, 1978), no maximizer exists whenever the observed vector of statistics $\mathbf{g}(\mathbf{y}_{\text{obs}}, \mathbf{X})$ is not contained in the interior of the convex hull of the set $\mathcal{S}$, where we define $\mathcal{S}$ to be the set of all possible $\mathbf{g}(\mathbf{y}, \mathbf{X})$ as $\mathbf{y}$ ranges over the sample space of all graphs. This problem occurs with ERGMs more frequently than one might suppose; for example, it arises whenever the model contains a degree statistic d_k such that $d_k(\mathbf{y}_{\text{obs}}) = 0$ (that is, the observed graph does not happen to contain any nodes of degree k). In such a case, the coefficient for the d_k term will be driven to $-\infty$ in the estimation procedure. An even more troublesome problem is the fact that we generally cannot know $\mathcal{S}$ and must instead rely on an MCMC-generated subset of $\mathcal{S}$. Call this subset $\mathcal{S}_{\text{MCMC}}$. The same theory cited above implies that there is no maximizer of the approximated likelihood function whenever the interior of the convex hull of $\mathcal{S}_{\text{MCMC}}$ does not contain $\mathbf{g}(\mathbf{y}_{\text{obs}}, \mathbf{X})$.

Finally, and most importantly from the perspective of this article, the maximum likelihood estimator can result in a completely unrealistic model even when it (or its approximation) exists. This

happens when an ERGM is so badly misspecified that even using the vector of parameters most likely under that ERGM to generate the observed graph — by definition, the maximum likelihood estimator — it is still extremely unlikely to do so. In other words, once we obtain an estimator $\hat{\boldsymbol{\eta}}$, the resulting ERGM

$$P(\mathbf{Y} = \mathbf{y}|\mathbf{X}) \propto \exp\{\hat{\boldsymbol{\eta}}^t \mathbf{g}(\mathbf{y}, \mathbf{X})\} \tag{12}$$

places most of the probability mass on a subset of the sample space containing networks that bear no resemblance to the observed network.

Naturally, the next question is: exactly what is meant by "resemblance to the observed network"? To an experienced practitioner, the informal "I'll know it when I see it" might suffice, but clearly a more objective set of criteria would be of benefit. Handcock (2003) addresses some extreme cases, e.g., in which most of the probability mass is placed on the full graph and empty graph. Whenever the ERGM in question is particularly ill-suited to modelling networks, then when networks are repeatedly simulated from model (12), even those statistics that are part of the $\mathbf{g}(\mathbf{y}, \mathbf{X})$ vector will not be near the corresponding values of $\mathbf{g}(\mathbf{y}_{\text{obs}}, \mathbf{X})$. Handcock (2003) terms this "model degeneracy". This is a very serious problem; as explained above, it implies that the observed vector of graph statistics will not be contained in the interior of the convex hull of the sampled statistics. Thus, even when we assume the model is correct, we would be unable to obtain a maximum likelihood estimator in a simulation study.

However, a model does not have to be degenerate in order to fit poorly; it merely has to produce graphs that bear no resemblance to the observed network. In these less extreme cases, determining goodness of fit is more difficult. Our approach begins by selecting a set of graph statistics that are not functions of the statistics in $\mathbf{g}(\mathbf{y}, \mathbf{X})$ and which are believed to represent important structural properties of networks. We then compare these statistics for the original graph and graphs simulated from the fitted model. When the original statistic appears unlikely to have arisen from the statistic's simulated distribution, we call this *lack of fit*. Thus, goodness of fit refers in general to the ability of the fitted ERGM to reproduce certain graph statistics seen in the observed graph. Graphically, it is possible to assess goodness of fit as defined by a chosen graph statistic by plotting the ranges of the values of the statistic(s) in question along with the corresponding value for the observed graph. If the latter is within the range of the sampled statistic values, we say that the model appears to fit well (or, more accurately, we say that we see no evidence of lack of fit). We stress that using this criterion, we are not familiar with any class of ERGMs that has been previously demonstrated in the literature to fit well for social network data.

5 Dyadic independence models for friendship networks

Dyadic independence models (5) for network data are merely logistic regression models, so model-fitting procedures do not suffer from the problems described in Section 4. However, the dyadic independence models we applied to the AddHealth friendship data do not fit well by certain criteria that we introduce below. More promising models for these data are introduced in Section 6.

5.1 Attribute-based models for networks

The first dyadic independence model we consider is perhaps the simplest possible network model, in which $\mathbf{g}(\mathbf{y}, \mathbf{X})$ consists only of $s_1(\mathbf{y})$, the number of edges in $\mathbf{y}$. This is the Bernoulli, or Erdős-

Rényi, graph described in Section 3. For AddHealth school 10, the parameter estimate for the Bernoulli graph is seen in Table 1 to be -4.625. This may be derived exactly: Since school 10 has 205 nodes and 203 edges, the MLE for the probability that any dyad has an edge is $203/\binom{205}{2}$, or 0.00971, and the log-odds of this value is -4.625. If only all estimates were so easy to find!

The second model we consider includes edges and also several statistics based on nodal covariates. Recall that in the dyadic independence model of equation (4), an individual component of the $\mathbf{g}(\mathbf{y}, \mathbf{X})$ vector, say the kth component, may be written as

$$g_k(\mathbf{y}, \mathbf{X}) = \sum_{i<j} \sum y_{ij} h_k(\mathbf{X}_i, \mathbf{X}_j). \tag{13}$$

Because it is not important, we drop the subscript k in equation (13) and simply allow $h(\mathbf{X}_i, \mathbf{X}_j)$ to denote a generic covariate statistic in the following discussion.

For the factors grade, race, and sex, our second model includes two types of statistics. We call the first type a *nodal factor effect*. Given a particular level of a particular factor (categorical variable), the nodal factor effect counts the total number of endpoints with that level for each edge in the graph. In other words, we define

$$h(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} 2 & \text{if both nodes } i \text{ and } j \text{ have the specified factor level;} \\ 1 & \text{if exactly one of } i, j \text{ has the specified factor level;} \\ 0 & \text{if neither } i \text{ nor } j \text{ has the specified factor level.} \end{cases} \tag{14}$$

Interpreted using equation (3), this means that the corresponding parameter is the change in conditional log-odds when we add an edge with one endpoint having this factor level — and this change is doubled when both endpoints of the edge share this level. As an example, consider the grade factor, which has levels 7 through 12 along with one missing-value level *NA*. These seven levels of the grade factor require six separate statistics for the nodal factor effect; one level must be excluded since the sum of all seven equals twice the number of edges in the graph, thus creating a linear dependency among the statistics.

The second type of nodal statistics we employ are *homophily statistics*. A homophily statistic for a particular factor gives each edge in the graph a score or zero or one, depending on whether the two endpoints have matching values of the factor. We distinguish between two kinds of homophily, depending on whether the distinct levels of the factor should exhibit different homophily effects. Thus, for *uniform homophily*, we have a single statistic, defined by

$$h(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ have the same level of the factor;} \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand, for *differential homophily*, we have a set of statistics, one for each level of the factor, where each is defined by

$$h(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ both have the specified factor level;} \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Note that for sex, a two-level factor, we may include a differential homophily effect or a nodal factor effect but not both. This is because in an undirected graph, there are only three types of edges — male-male, female-female, and male-female — so only two statistics are required to completely characterize the sexes of both endpoints of an edge, provided the overall edge effect is

also in the model. A differential homophily effect (two statistics) plus a nodal factor effect (one statistic) would together entail redundant information.

Now that we have defined nodal factor and homophily effects, we are ready to describe our second dyadic independence model. It includes an edge statistic; nodal factor effects and differential homophily for both the race and grade factors; and a nodal factor effect and uniform homophily for the sex factor. Note that all schools have two sexes and six grades, but only some have additional NA categories for these factors. Furthermore, the number of races present varies considerably from school to school. Parameters are excluded from the model when it can be determined in advance that the MLE will be undefined. Such cases occur for node factor effects when only a small number of students possess the factor level and they all have 0 friendships; or for homophily terms, when there are no ties between two students with a given factor level. For example, in AddHealth school 10, grade is a seven-level factor, sex is a three-level factor, and race is a four-level factor; and our dyadic independence model contains 25 parameters: one for edges, six for the grade factor effect, six for differential homophily on grade (excluding the NA category), five for the race factor effect, four for differential homophily on race (excluding the NA and Other categories), two for the sex factor effect, and one for uniform homophily on sex. The fitted values of these 25 parameters are presented as Model I in Table 2.

5.2 Goodness of fit statistics for ERGMs

Our graphical tests of goodness-of-fit require a comparison of certain observed graph statistics with the values of these statistics for a large number of networks simulated according to the fitted ERGM. The choice of these statistics determines which structural aspects of the networks are important in assessing fit. We propose to consider three sets of statistics: the degree distribution, the edgewise shared partner distribution, and the geodesic distance distribution.

The degree distribution for a graph consists of the values $\mathrm{d}_0/n, \ldots, \mathrm{d}_{n-1}/n$. Note that these values sum to unity. Similarly, the edgewise shared partner distribution consists of the values $\mathrm{ep}_0/\mathrm{s}_1, \ldots, \mathrm{ep}_{n-2}/\mathrm{s}_1$. (The statistics d_i, ep_i, and s_i are defined in Section 3.) Finally, the geodesic distance distribution consists of the relative frequencies of the possible values of geodesic distance between two nodes, where the geodesic distance between two nodes equals the length of the shortest path joining those two nodes (or infinity if there is no such path). For instance, because two nodes are at geodesic distance 1 if and only if they are connected by an edge, and because there are $\binom{n}{2}$ possible pairs of nodes, the first value of the geodesic distance distribution equals $\mathrm{s}_1/\binom{n}{2}$. The last term, the fraction of dyads with infinite geodesics, is also called the fraction "unreachable."

We chose to include the degree statistics because of the tremendous amount of attention paid to them in the networks literature — for example, degree statistics are central to the work of Frank and Strauss (1986) on Markov graphs, as explained in Section 6. We included the shared partner statistics based on the work of Snijders et al. (2004) and Hunter and Handcock (2004), and because we will show (in Section 6) that the addition of a parametric formula involving $\mathrm{ep}_0, \ldots, \mathrm{ep}_{n-2}$ improves model fit dramatically. Therefore, these statistics appear to contain a great deal of relevant network information. Furthermore, equation (10) demonstrates that the triangle count, ubiquitous in the networks literature, is a function of the shared partner statistics. Finally, the geodesic distance statistics are the basis for two of the most common measures of centrality, a fundamental concept in social network theory (Wasserman and Faust 1994, page 111), and are clearly relevant to the flow of pathogens, information or other entities among actors. They also represent higher-

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.

Figure 3 depicts the results of 100 simulations for School 10 from the fitted dyadic indepen-

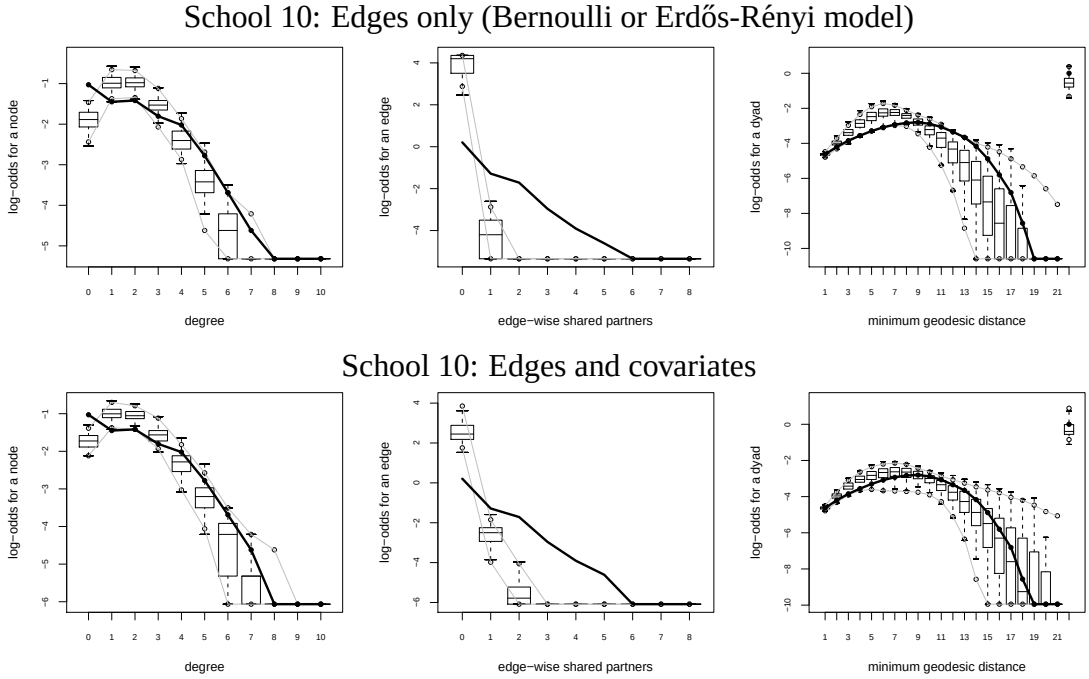

Figure 3: Simulation results for dyadic independence models. In all plots, the vertical axis is the logit of relative frequency; the School 10 statistics are indicated by the solid lines; the boxplots include the median and interquartile range; and the light gray lines represent the range in which 95 percent of simulated observations fall.

dence models given in Tables 1 and 2. The vertical axis in each plot is the logit (log-odds) of the relative frequency, and the solid line represents the statistics for the observed graph. We can immediately see that the models do an extremely poor job of capturing the shared partner distribution. They perform relatively well for the degree distribution and the geodesics distribution, considering their simplicity. Adding the attribute-based statistics improves the fit of the geodesic distribution considerably, but very little for the degree distribution. The large departure in the shared partner plot reflects the fact that the model strongly underestimates the amount of local clustering present in the data. The models predict friends to have no friends in common most of the time, and occasionally one friend in common, whereas in the original data they have up to five. Although we present plots for only one school here, the qualitative results for other schools follow a small number of similar patterns. Plots for other schools can be viewed at `http://csde.washington.edu/networks`.

In the next section, we present some modifications to the models seen here that fit much better as measured both by the graphical criterion we have employed here and by more traditional statistical measures such as Akaike's Information Criterion (AIC). The fact that the simple dyadic independence models do not appear to fit the data well is not surprising; after all, such models are merely logistic regression models in which the responses are the dyads. That we must move beyond dyadic independence in order to construct models that fit social network data well is a result

of the fact that the formation of edges in a network depends upon the existing network structure itself.

6 Dyadic Dependence Models for Friendship Networks

As illustrated in Section 4, many problems can arise when trying to fit ERGMs using maximum likelihood. Only recently has the application of likelihood-based ERGM-fitting methods to network data reliably produced repeatable, interpretable results (Snijders et al., 2004; Hunter and Handcock, 2004). In this section, we describe extensions of the ERGMs of Section 5.1 that improve their fit. We offer graphical and numerical evidence of their improved fit, and we make some generalizations about what these models indicate about high school friendship networks.

We begin by noting some commonly-used social network models that do *not* fit. The well-studied homogeneous Markov ERGM (Frank and Strauss, 1986), in which

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}|\mathbf{X}) = c^{-1} \exp\{\eta_1 \mathrm{d}_1(\mathbf{y}) + \cdots + \eta_{n-1}\mathrm{d}_{n-1} + \eta_n \mathrm{t}(\mathbf{y})\} \tag{16}$$

(where d_k and t are the degree and triangle statistics defined in Section 3.3), did not converge to a finite MLE for any school examined. This is not surprising given that the homogeneity condition, which ignores the nodal covariate information contained in $\mathbf{X}$, is unrealistic for nearly any social network. Deeper reasons for the failure of model (16) are explored by Handcock (2002; 2003), and in fact it was precisely this failure that motivated the work of Snijders et al. (2004) in developing the alternating triangle and alternating k-star statistics that we explain in Sections 6.1 and 6.2.

A Markov model with the additional nodal covariate parameters described in Section 5.1 (main effects and homophily for grade, sex, and race) also did not converge. In each case, the MLE for the triangle parameter headed off to positive infinity and the other parameters to either positive or negative infinity; the resulting probability model places nearly all of its mass on the full or empty graph, in a ratio that results in a mean number of triangles equal to that in the observed graph. Clearly this is not a good model for the data. Nonetheless, the Markov assumption, by allowing for the presence of a triangle statistic in an ERGM, allows us to consider effects such as "triangle closure" — in which $Y_{ij} = 1$ and $Y_{jk} = 1$ increases the chance that $Y_{ik} = 1$ — as arising from an intrinsic property of network formation rather than merely a side effect of homophily. In the remainder of this section, we discuss extensions of the statistics introduced in Section 3.3 that do a better job than Markov models of modelling social network behavior such as triangle closure.

6.1 Geometrically weighted shared partner statistics

Here, we consider the shared partner statistics, both edgewise and dyadic, defined in Section 3.3. Consider first the edgewise shared partner statistics $\mathrm{ep}_0, \ldots, \mathrm{ep}_{n-2}$. It would be possible to add one new term to the model for each of $\mathrm{ep}_1, \ldots, \mathrm{ep}_{n-2}$ — we omit ep_0 to avoid the linear dependence of equation (8) — but this leads to a model with too much flexibility. As Hunter and Handcock (2004) point out, it is often better to restrict the parameter space to avoid problems of degeneracy. To this end, we define the single statistic

$$\mathrm{ep}^G(\mathbf{y}; \tau) = e^{\tau} \sum_{i=1}^{n-2} \left\{1 - \left(1 - e^{-\tau}\right)^i\right\} \mathrm{ep}_i(\mathbf{y}), \tag{17}$$

where τ is an additional parameter. Because the coefficient in equation (17) of the shared partner statistic ep_i includes $(1-e^{-\tau})^i$, we refer to $\mathrm{ep}^G(\mathbf{y};\tau)$ as the *geometrically weighted edgewise shared partner* statistic; thus, the superscript G stands for "geometrically weighted". Hunter and Handcock (2004) introduce this statistic and show that it coincides with the alternating k-triangle statistic proposed by Snijders et al. (2004):

$$\mathrm{ep}^G(\mathbf{y};\tau) = 3\mathrm{t}_1(\mathbf{y}) - \frac{\mathrm{t}_2(\mathbf{y})}{(e^\tau)^1} + \cdots + (-1)^{n-3}\frac{\mathrm{t}_{n-2}(\mathbf{y})}{(e^\tau)^{n-3}}. \tag{18}$$

Similarly, we may define

$$\mathrm{dp}^G(\mathbf{y};\tau) = e^\tau \sum_{i=1}^{n-2}\left\{1-\left(1-e^{-\tau}\right)^i\right\}\mathrm{dp}_i(\mathbf{y}), \tag{19}$$

a statistic equal to the alternating k-twopath statistic of Snijders et al. (2004):

$$\mathrm{dp}^G(\mathbf{y};\tau) = \mathrm{u}_1(\mathbf{y}) - \frac{2\mathrm{u}_2(\mathbf{y})}{(e^\tau)^1} + \cdots + (-1)^{n-3}\frac{\mathrm{u}_{n-2}(\mathbf{y})}{(e^\tau)^{n-3}}. \tag{20}$$

The parameter τ in $\mathrm{ep}^G(\mathbf{y};\tau)$ or $\mathrm{dp}^G(\mathbf{y};\tau)$ is not a canonical exponential family parameter like $\boldsymbol{\eta}$ in equation (1); rather, if τ is considered unknown, so that $(\boldsymbol{\eta},\tau)$ is the full parameter vector, then the ERGM forms a *curved exponential family*, which complicates the estimation procedure. Hunter and Handcock (2004) address this more complicated situation; however, for the purposes of this article, we make the simplifying assumption that τ is fixed and known. In our model-fitting procedure, we tried a range of different values of τ on several schools and found that the estimated likelihood value was generally highest around $\tau = 1.0$ to $\tau = 1.5$. Furthermore, the different likelihood values were very close together, and the goodness-of-fit plots (as in Figure 5) were nearly indistinguishable for different values of τ in the range we tested (0.5-2.0). Values too far outside this range resulted in models that could not be fit for one of the reasons listed in Section 4. Based on these results, we use a fixed value of $\tau = 1.0$ for all the models we discuss below.

As an example, we take $\mathbf{g}(\mathbf{y},\mathbf{X})$ to consist of only two terms, the edge statistic and the geometrically weighted edgewise shared partner (GWESP) statistic. In this case, the ERGM of equation (1) becomes

$$P_{\boldsymbol{\eta}}(\mathbf{Y}=\mathbf{y}|\mathbf{X}) = c^{-1}\exp\{\eta_1\mathrm{s}_1(\mathbf{y}) + \eta_2\mathrm{ep}^G(\mathbf{y};\tau)\}. \tag{21}$$

We fit model (21) and a similar model involving the geometrically weighted dyadic shared partner (GWDSP) statistic $\mathrm{dp}^G(\mathbf{y};\tau)$, to AddHealth school 10 and summarize the results in Table 1.

6.2 Geometrically weighted degree statistic

Similar to the GWESP and GWDSP statistics is the geometrically weighted degree (GWD) statistic

$$\begin{aligned}\mathrm{d}^G(\mathbf{y};\tau) &= e^\tau\left\{2\mathrm{s}_1(\mathbf{y}) - e^\tau\sum_{i=1}^{n-1}\left[1-\left(1-e^{\tau}\right)^i\right]\mathrm{d}_i(\mathbf{y})\right\}\\ &= \left(e^\tau\right)^2\sum_{i=1}^{n-1}\left[\left(1-e^{-\tau}\right)^i - 1 + ie^{-\tau}\right]\mathrm{d}_i(\mathbf{y}).\end{aligned} \tag{22}$$

Coefficient	Model:			
	Edges only	Edges plus GWESP	Edges plus GWDSP	Edges plus GWD
edges	$-3.896(0.12)^{***}$	$-5.314(0.10)^{***}$	$-4.780(0.07)^{***}$	$-4.625(0.07)^{***}$
GWESP	—	$2.404(0.14)^{***}$	—	—
GWDSP	—	—	$0.039(0.009)^{***}$	—
GWD	—	—	—	$1.998(0.31)^{***}$
*** Significant at 0.001 level				

Table 1: Estimated coefficients and standard errors for the parameters of three simple models that consider only network structure but no nodal covariate information. The GWESP statistic $\mathrm{ep}^G(\mathbf{y};\tau)$, the GWDSP statistic $\mathrm{dp}^G(\mathbf{y};\tau)$, and the GWD statistic $\mathrm{d}^G(\mathbf{y};\tau)$ all use $\tau = 1.0$.

The $\mathrm{d}^G(\mathbf{y};\tau)$ statistic coincides with the alternating k-star statistic of Snijders et al (2004):

$$\mathrm{d}^G(\mathbf{y};\tau) = \mathrm{s}_2(\mathbf{y}) - \frac{\mathrm{s}_3(\mathbf{y})}{(e^\tau)^1} + \cdots + (-1)^{n-1}\frac{\mathrm{s}_{n-1}(\mathbf{y})}{(e^\tau)^{n-3}}. \tag{23}$$

Comparing equation (22) with the equation (17) that defines $\mathrm{ep}^G(\mathbf{y};\tau)$, we see that one main difference is the inclusion of $\mathrm{s}_1(\mathbf{y})$ in (22). This difference is superficial, since all models that we consider in this article contain the $\mathrm{s}_1(\mathbf{y})$ term. Therefore, our uses of the terms "geometrically weighted degree statistic" for (22) and "geometrically weighted shared partner statistic" for (17) are completely analogous.

We fit the AddHealth school 10 data to the ERGM containing only edges and GWD:

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}|\mathbf{X}) = c^{-1}\exp\{\eta_1 \mathrm{s}_1(\mathbf{y}) + \eta_2 \mathrm{d}^G(\mathbf{y};\tau)\}. \tag{24}$$

The results are summarized in the second column of Table 1. Based on repeatedly fitting various models using a range of τ values, we settled on $\tau = 1.0$ for the GWD statistic. We also fit models that include the two terms of equation (24), all of the nodal covariate statistics described in Section 5.1, and one or both of the shared partner statistics ep^G and dp^G of Section 6.1. Since the coefficient of dp^G is not significant when both ep^G and dp^G are included in the model, we omit it. The resulting model fit is summarized as Model II in Table 2.

Most dyadic dependence models create such severe numerical difficulties in estimation that we are unable to fit them successfully for a large number of different networks of different sizes. However, both the GWD and GWESP statistics appear to be more robust: Using our MCMC fitting procedure, we were able to be estimate both of these parameters on many of the AddHealth schools, a feat not before seen with any dyadic-dependent model. As a case in point, we fit the full model of Table 2 to the largest school in our sample, with 2209 nodes. Though the fit of the model to data is imperfect, as seen in Figure 4, the model provides reasonable parameter estimates. Plots like Figure 4 can now inform us as to the specific ways in which our model is adequate and in which it needs improvement.

6.3 Graphical goodness-of-fit

As described in Section 5.2, one way to develop an idea of how well a model fits is by comparing a set of observed graph statistics with the range of the same statistics obtained by simulating many

Coefficient	Model I	Model II	Coefficient	Model I	Model II
edges	$-9.233(0.91)^{***}$	$-9.127(0.85)^{***}$			
GWESP	—	$1.586(0.18)^{***}$			
GWD	—	$0.025(0.34)$			
GWDSP	—	$0.019(0.002)^{***}$			
			DH (Gr. 7)	$5.623(0.79)^{***}$	$4.681(0.76)^{***}$
NF (Gr. 8)	$0.536(0.54)$	$0.555(0.52)$	DH (Gr. 8)	$4.520(0.75)^{***}$	$3.722(0.71)^{***}$
NF (Gr. 9)	$1.575(0.48)^{**}$	$1.445(0.46)^{**}$	DH (Gr. 9)	$2.129(0.55)^{***}$	$1.814(0.50)^{***}$
NF (Gr. 10)	$1.896(0.49)^{***}$	$1.712(0.46)^{***}$	DH (Gr. 10)	$1.942(0.62)^{**}$	$1.621(0.55)^{**}$
NF (Gr. 11)	$2.039(0.49)^{***}$	$1.817(0.46)^{***}$	DH (Gr. 11)	$1.953(0.58)^{***}$	$1.486(0.52)^{**}$
NF (Gr. 12)	$2.035(0.52)^{***}$	$1.836(0.49)^{***}$	DH (Gr. 12)	$2.392(0.79)^{**}$	$1.952(0.69)^{**}$
NF (Gr. NA)	$2.270(0.65)^{***}$	$2.099(0.60)^{***}$			
			DH (White)	$1.514(0.61)^{*}$	$1.215(0.53)^{*}$
NF (Black)	$0.438(0.39)$	$0.322(0.32)$	DH (Black)	$1.165(1.26)$	$1.152(1.19)$
NF (Hisp)	$-0.418(0.34)$	$-0.318(0.28)$	DH (Hisp)	$1.107(0.41)^{**}$	$0.935(0.35)^{**}$
NF (Nat Am)	$-0.462(0.30)$	$-0.366(0.25)$	DH (Nat Am)	$1.696(0.42)^{***}$	$1.351(0.36)^{***}$
NF (Other)	$-1.146(0.75)$	$-0.736(0.65)$			
NF (Race NA)	$1.223(0.61)^{*}$	$0.865(0.48)$			
NF (Female)	$0.089(0.09)$	$0.055(0.06)$	UH (Sex)	$0.776(0.15)^{***}$	$0.676(0.14)^{***}$
NF (Sex NA)	$-0.418(0.47)$	$-0.178(0.40)$			
NF stands for Node Factor.			DH stands for Differential Homophily. UH stands for Uniform Homophily.		
* Significant at 0.05 level		** Significant at 0.01 level		*** Significant at 0.001 level	

Table 2: Estimated coefficients (and standard errors) for two models applied to AddHealth school 10. Model I contains terms for edges and the 25 nodal covariate terms described in Section 5.1. Model II contains all of the terms in Model I plus three additional terms, GWESP, GWDSP, and GWD, each with $\tau = 1.0$. Differential homophily terms for Grade NA, Race Other, Race NA, and Sex NA are omitted because there are no edges observed between two actors sharing these attribute values.

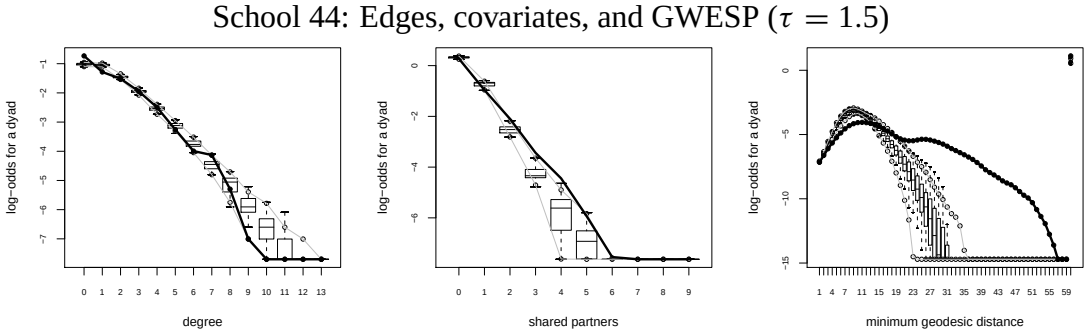

Figure 4: Goodness-of-fit plots for the largest AddHealth school, school 44, with 2209 nodes. The clear lack of fit in the geodesic distribution is typical of this model for the larger AddHealth schools, even though the same model tends to fit well on smaller schools.

graphs from the fitted ERGM. If the observed graph is not typical of the simulated graph for a particular statistic, then the model is either degenerate (if the statistic is among those included in the ERGM vector $\mathbf{g}[\mathbf{y}, \mathbf{X}]$) or poorly-fitting (if the statistic is not included). Figure 5 depicts simulation results for school 10 for the three dyadic-dependent ERGMs in Table 1; Figure 6 depicts Model II from Table 2.

For both School 10 and many of the other smaller AddHealth schools, a simple model containing only individual-level attributes (Figure 3, bottom graph) does a respectable job of recreating the geodesic distribution of the observed data, a global property of the graph. At the same time, it strongly underestimates the amount of local clustering as captured by the shared partner distribution. The former observation is encouraging, since information on attribute matching is far easier to collect than other types of network data in most real-world settings (where only a sample of nodes is available); it requires questions about the attributes of respondents' partners only, not their actual identity. The latter observation tells us that not all features of the network can be ascribed to purely dyadic-level phenomena — yet this fact is not surprising, as it is the very basis for the field of network analysis.

The fact that a simple model is strongly predictive of one higher-order network property (geodesics) and strongly divergent from another (shared partner) is intriguing. This pattern makes clear that a variety of network statistics ought to be tested in order to develop a robust sense of goodness-of-fit. It also provides some evidence that the same macro-level social structure can be built out of multiple distinct social processes. This observation is of potentially great interest to social scientists trying to understand the development of social structure.

Comparing the bottom graph in Figure 3 with the top two graphs in Figure 5, we see that incorporating the heterogeneity of actors through nodal covariates was more important for model fit than either modelling degree or edgewise shared partners alone. This should not be too surprising; high school students are anything but homogeneous with regard to the characteristics of peers with whom they form friendships. In fact, we expect that few if any networks of social relations are likely to demonstrate goodness of fit by the methods we use when all actors are assumed to be homogeneous.

Empirical social relations generally exhibit local clustering, and this one is no exception: the simple Bernoulli model drastically under-predicts the number of shared partners people should have, even though it captures the degree distribution well. Such clustering can come from at

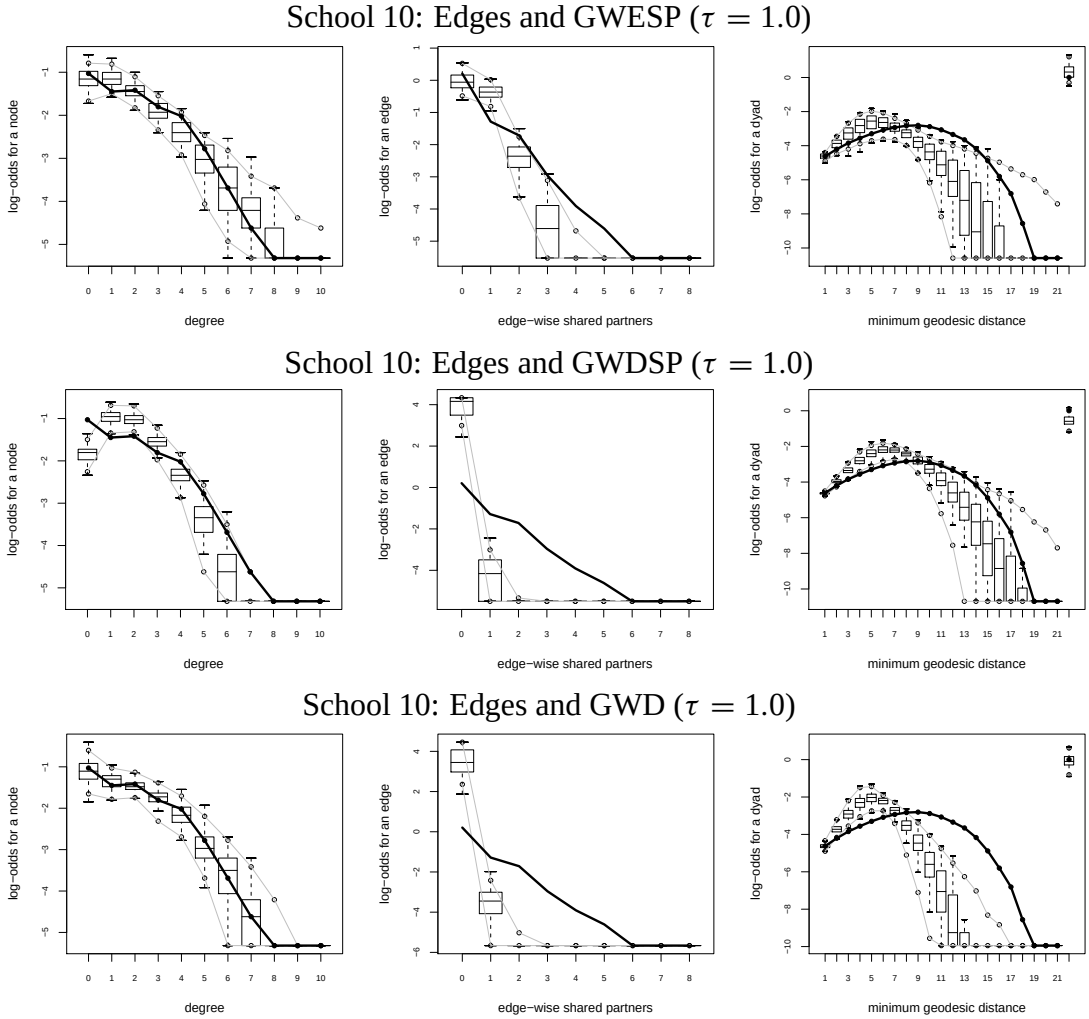

Figure 5: Simulation results for dyadic dependence ERGMs of Table 1

least two different sources: (1) actors matching on exogenous attributes; and (2) actors forming partnerships on the basis of existing shared partners. The two are fundamentally different: the former is dyadic-independent, using factors exogenous to the network structure; while the latter is dyadic-dependent and reflects a tendency known in the social networks literature as "triangle closure" in which friends of my friends are more likely to be my friends. The modelling here shows that neither attributes nor shared partners alone are sufficient to explain the clustering observed in this network (the same is true of other AddHealth schools; see the plots at `http://csde.washington.edu/networks`). Only by including both in tandem did we obtain a good fit to the observed network patterns. One can see the joint effect of these two phenomena by comparing the homophily parameter values in the simple attribute model (Model I, Table 2) with those of the model that also includes shared partner effects explicitly (Model II, Table 2); the magnitudes of the homophily effects are smaller in the larger model, since some of the homophily effects can be explained by triangle closure effects.

In this setting, fitting degree was of tertiary importance. A simple one-term Bernoulli model (Figure 3, top graph) came closer to capturing the degree distribution than it did any other feature of

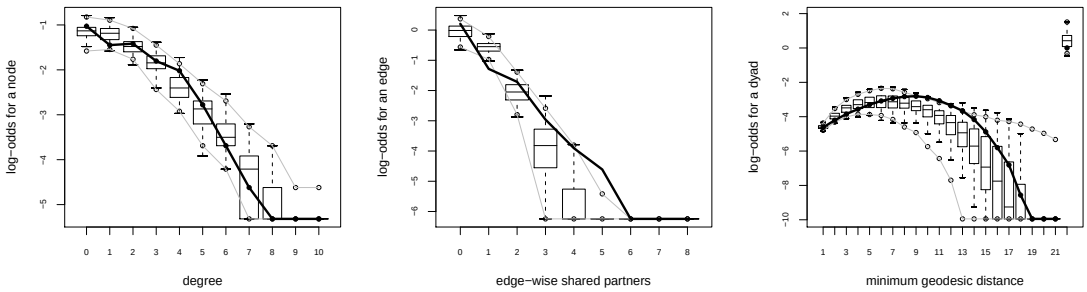

Figure 6: Simulation results for Model II of Table 2.

the network; and a model that included only a term for the degree distribution (Figure 5, top graph) did not capture any of the other structure of the network. This finding seems particularly important given the massive attention that has been placed on degree-only models in some branches of the network literature recently; see Albert and Barabási (2002) for a survey of some of this literature.

6.4 Akaike's information criterion

To see whether the results we observe from the goodness of fit plots are consistent with more traditional goodness-of-fit approaches, we also considered a more regimented approach to model selection based on Akaike's information criterion, or AIC (Akaike, 1973). AIC is among the best-known of the many methods proposed in the literature for solving the problem of balancing the conflicting modelling aims of fidelity to data and parsimony of representation:

$$\text{AIC}(M) = -2(\text{maximized loglikelihood under } M) + 2(\#\text{ of parameters in } M), \tag{25}$$

where M denotes a particular ERGM. The goal is to minimize AIC(M) as a function of M.

Unfortunately, as we pointed out in Section 3, it is not possible to evaluate the likelihood function directly for most ERGMs. Therefore, the value of the loglikelihood used in equation (25) is approximate except in the case of a dyadic independence model, where the pseudolikelihood (6) is equal to the likelihood.

The graphical approach is generally consistent with AIC, in the sense that models that produce large reductions in AIC also seem to yield considerably better fits in the graphical plots; those with smaller reductions in AIC have less pronounced effects on the plots. However, the goodness of fit plots provide a richer picture than AIC alone. From these plots, a number of features of the relationships between these models and the network structure become clear. For instance, both the plots and AIC indicate that incorporating the heterogeneity of actors through nodal covariates is far more important for model fit than modelling either degree or shared partners alone. Yet the picture of this comparison that emerges through the plots is much richer than the information in Table 3. Finally, we note that the approximations of the loglikelihoods, necessary for computing the AIC scores of Table 3, appear to lead to some contradictory results. For example, the AIC score of the largest model, which coincides with Model II in Table 2, is much lower than that of the model that drops only the GWD term — despite the fact that the GWD term is not significant in Table 2.

An interesting question is whether formal model selection criteria other than AIC can be applied to these models. For instance, there is a great deal of statistical literature addressing the

Model, M	# of parameters	AIC(M)
edges only	1	2287.7
edges plus GWESP*	2	2133.0
edges plus GWDSP*	2	2287.4
edges plus GWD*	2	2255.7
edges plus NC	25	1816.3
edges, NC, and GWESP*	26	1753.6
edges, NC, and GWDSP*	26	1818.2
edges, NC, and GWD*	26	1790.0
edges, NC, GWESP, and GWDSP*	27	1756.2
edges, NC, GWESP, and GWD*	27	1738.9
edges, NC, GWESP, GWDSP, and GWD*	28	1727.2

Table 3: Comparison of various ERGMs for school 10 using Akaike's information criterion (AIC). NC stands for the nodal covariates, as described in Section 5.1. For GWD, GWESP, and GWDSP, τ always equals 1.0. Asterisks indicate the models in which approximate loglikelihoods are used.

comparison between AIC and the Bayesian information criterion (BIC); see, for example, Kuha (2004). The definition of BIC is similar to that of AIC:

$$\text{BIC}(M) = -2(\text{maximized loglikelihood under } M) + \log N(\#\text{ of parameters in } M),$$

where N is the sample size. However, for network models, the sample size is not the same as the number of nodes, n. For example, for any dyadic independence model, the sample size is unequivocally $\binom{n}{2}$, the number of dyads. However, when dependence among dyads exists, the *effective* sample size can be smaller than $\binom{n}{2}$. Indeed, in cases of extreme dependence, we may encounter ERGMs in which nearly all of the probability mass is placed on the full graph and the empty graph. In such a case, the effective sample size is roughly one because all dyads nearly always have the same value. Clearly, in order to implement a model selection criterion that relies on the sample size, such as BIC, it is first necessary to establish what "sample size" means. This is a challenging question for network ERGMs, beyond the scope of this article.

7 Discussion

Although the basic idea of exponential random graph models (ERGMs) as a way to model the probabilistic behavior of a network has been around for almost twenty-five years, computing maximum likelihood estimates for these models has proven to be very difficult in the dyadic dependence case. By presenting the first systematic study of a large group of networks using likelihood-based inference for dyadic-dependent ERGMs, this article allows us to consider the goodness of fit of these ERGMs and interpret the parameter estimates obtained. Some of the networks successfully modelled for this article are far larger than for any previously reported dyadic-dependent ERGMs.

Choosing an appropriate set of network statistics on which to compare the observed graph with graphs simulated from the fitted model is an important task in the graphical goodness-of-fit studies we advocate in this article. If possible, these statistics should match the purpose for which one is estimating and simulating networks. It may not be immediately clear what kinds of network

properties are relevant; in fact, that might be precisely the question in which we are interested in the first place. For many social relations, theory may suggest that people do not look beyond more than one or two layers of network neighbors, so adequately modelling statistics such as the edgewise shared partner distribution might be expected to get higher-order statistics correct as well.

When we compare different AddHealth schools, we find that many significant model parameters show remarkably similar qualitative patterns. Even the numerical values of the maximum likelihood estimates are often quite similar across friendship networks. However, it is important when comparing networks with different numbers of nodes that the values of the parameter estimates are not necessarily comparable. The question of how to modify ERGMs so that their coefficients are directly comparable without regard to n, the number of nodes, is a very important issue in network modelling. Furthermore, as we pointed out in Section 6.4, the related question of the effective sample size of a network on n nodes for a particular ERGM is important if we have any hope of applying model selection methods such as BIC that depend on sample size. However, this is a question for the future; for now, the science of likelihood-based methods for fitting ERGMs is still in its early stages.

Although the most complete and best-fitting model presented here appears to come close to capturing the higher-order network statistics examined for School 10, the same is not true for many of the larger schools; for instance, compare Figure 4, based on 2209 nodes. Larger schools depart from the fitted model in a similar way: The model under-predicts the number of long geodesics and over-predicts the number of short ones. In effect, the real social networks are more "stringy" than our best-fitting model predicts. One likely reason for this can be seen in Figure 1: It appears as if (and makes intuitive sense that) students are less likely to be friends as the gap between their grade levels grows. But our models, which include homophily effects, capture this effect only partially; for instance, they treat friendships between a seventh grader and an eighth grader, and between a seventh grader and an twelfth grader, as equally likely assuming that all other nodal covariates are the same. And this is only one of many likely additional processes underlying the structure of some of the larger school groups. We hope that the approach to assessing model fit that we propose in this paper will provide a catalyst for researchers to think about these many different processes and how to test hypotheses about them, not only on these data but a wide variety of social network data generally. We strongly believe that nodal covariate information is vital to any attempt at social network modelling, and the particular covariates of importance will not be the same for all situations.

In the meantime, we believe that the geometrically weighted degree, edgewise shared partner, and dyadic shared partner statistics — equivalent to the alternating k-star, k-triangle, and k-twopath statistics, respectively, of Snijders et al. (2004) — do a credible job of capturing a great deal of dyadic dependence structure of the friendship networks we have studied here. Both the graphical plots and the numerical AIC scores attest to the improvement in goodness of fit attainable by the inclusion of one or both of the these terms. In the interest of parsimony, we note that the GWESP statistic is more valuable than the GWD statistic for the friendship networks we have studied; however, including them both still leads to quite a parsimonious model. On the other hand, inclusion of the GWDSP statistic does not appear to dramatically improve model fit, perhaps because friendship networks tend to be neutral to the formation of twopaths connecting individuals when these individuals are not connected themselves. We hope that such rudimentary observations as these will be continually refined as the tools for fitting ERGMs improve.

References

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in B N Petrov and F Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Budapest: Akademiai Kaidó.

Albert, R. and Barabási, A.-L. (2002), Statistical mechanics of complex networks, *Reviews of Modern Physics*, **74**, 47–97.

Barndorff-Nielsen, O. E. (1978), *Information and Exponential Families in Statistical Theory*, New York: Wiley.

Besag, J. (1974), Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society, series B*, **36**: 192–225.

Corander, J., Dahmström, K., and Dahmström, P. (1998), Maximum likelihood estimation for Markov graphs, Research Report 1998:8, Department of Statistics, University of Stockholm.

Crouch, B. Wasserman, Stanley and Trachtenberg, F. (1998), Markov Chain Monte Carlo Maximum Likelihood Estimation for p^* Social Network Models, Paper presented at the XVIII International Sunbelt Social Network Conference in Sitges, Spain.

Dahmström, K., and Dahmström, P. (1993), ML-estimation of the clustering parameter in a Markov graph model, Stockholm: Research report, Department of Statistics.

Frank, O. (1991), Statistical analysis of change in networks, *Statistica Neerlandica*, **45**: 283–293.

Frank, O. and D. Strauss (1986), Markov graphs, *Journal of the American Statistical Association*, **81**: 832–842.

Geyer, C. J. and E. Thompson (1992), Constrained Monte Carlo maximum likelihood for dependent data, *Journal of the Royal Statistical Society, Series B*, **54**: 657–699.

Handcock, M. S. (2002) Statistical Models for Social Networks: Inference and Degeneracy. Pp. 229 – 240 in *Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers*, edited by Ronald Breiger, Kathleen Carley, and Philippa E. Pattison. National Research Council of the National Academies. Washington, DC: The National Academies Press.

Handcock, M. S. (2003), Assessing degeneracy in statistical models of social networks, Working Paper no. 39, Center for Statistics and the Social Sciences, University of Washington. Available from `http://www.csss.washington.edu/Papers/`

Holland, P. W. and S. Leinhardt (1981), An exponential family of probability distributions for directed graphs, *Journal of the American Statistical Association*, **76**: 33-50.

Hunter, D. R. and M. S. Handcock (2004), Inference in curved exponential family models for networks, Penn State Department of Statistics technical report number 04-02. Available from `http://www.stat.psu.edu/reports/2004/`

Kuha, J. (2004), AIC and BIC: Comparisons of assumptions and performance, *Sociological Methods and Research*, **33**: 188–229.

Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: Wiley.

Resnick, M. D., P. S. Bearman, R. W. Blum, et al. (1997), Protecting adolescents from harm. Findings from the National Longitudinal Study on Adolescent Health, *Journal of the American Medical Association*, **278**: 823–832.

Snijders, T. A. B. (2002), Markov Chain Monte Carlo estimation of exponential random graph models, *Journal of Social Structure*, **3**. Available at
`www.cmu.edu/joss/content/articles/volume3/Snijders.pdf`

Snijders, T. A. B., P. E. Pattison, G. L. Robins, and M. S. Handcock (2004), New specifications for exponential random graph models, Center for Statistics and the Social Sciences working paper no. 42, University of Washington. Available from
`http://www.csss.washington.edu/Papers/`

Strauss, D. and M. Ikeda (1990), Pseudolikelihood estimation for social networks, *Journal of the American Statistical Association*, **85**: 204–212.

Udry, J. R. and P. S. Bearman (1998), New methods for new research on adolescent sexual behavior, in *New Perspectives on Adolescent Risk Behavior*, R. Jessor, ed. New York: Cambridge University Press, pp. 241–269.

Wasserman, S. and K. Faust (1994), *Social Network Analysis: Methods and Applications*, Cambridge, UK: Cambridge University Press.

Wasserman, S. and P. E. Pattison (1996), Logit models and logistic regression for social networks: I. An introduction to Markov graphs and $p*$, *Psychometrika*, **61**: 401–425.