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 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), geometrically weighted dyadic shared partner (GWDSP), 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 developed in recent decades for analyzing dependent data, network models are especially useful for dealing with the kinds of dependence induced by social relations. Applications of social network models are becoming important in a number of fields, such as epidemiology, with the emergence of new infections diseases like AIDS and SARS; business, with the study of "viral marketing"; and political science, with the study of coalition formation dynamics. 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.

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 models have been applied to a wide variety of phenomena spanning many disciplines, including the structure of social networks, the dynamics of epidemics, the interconnectedness of the World Wide Web, and protein-protein interactions within a cell. This article focuses specifically on network data collected at a nationally representative sample of 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 propose a systematic approach to the assessment of network ERGMs. The models we examine here achieve a good fit to key structural properties of the network with a small number of covariates. The approach and the findings address a central question in the network literature: Can the global structure features observed in a network be generated by a modest number of local rules?

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 meaningful and interesting. 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

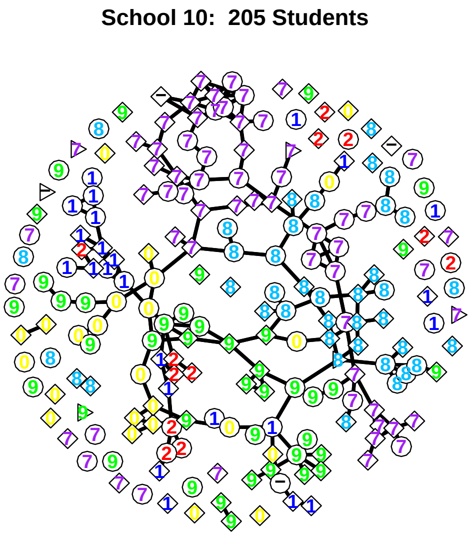
School 10: 205 Students

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.

likelihood estimates of the ERGM parameters, given an observed network. While these techniques are conceptually simple (Geyer and Thompson 1992), their practical implementation for relatively large social 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 (for instance, see Figure 6).

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 little resemblance at all to the observed network (Handcock, 2003). 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 relative to other graphs. 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 fit to these data. Section 4 illustrates our goodness of fit technique on a couple of simple models that do not fit well. Section 5 explains some network statistics that are used to build good-fitting models in Section 6. Finally, Section 7 discusses Akaike's information criterion and related model selection criteria.

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. Schools with large amounts of missing data were excluded from our analysis; this happened, among other reasons, for special education schools and for school districts that required explicit parental consent for student participation. Thus, our analysis includes 59 of the schools, ranging in size from 71 to 2209 surveyed students. Though in this article we focus primarily on a single illustrative school, School 10, results for all the schools we analyzed may be found at `http://csde.washington.edu/networks`.

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. This feature of reciprocation of nomination is common to many conceptualizations of friendship.

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 (3) 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. This coding follows standard practice in the social science literature.

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 objective in defining $\mathbf{g}(\mathbf{Y}, \mathbf{X})$ is to choose statistics that summarize the social structure of 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. We will consider only a few key statistics here, chosen to represent friendship selection rules that operate at a local level. The goal is to test whether these local rules can reproduce the global network patterns of clustering and geodesic distances (Morris, 2003).

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. 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 (3) — 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), while a discussion of the background of ERGMs in the networks literature may be found in Snijders (2002) or Hunter and Handcock (2004).

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

$$\mathbf{g}(\mathbf{y}, \mathbf{X}) = \sum\sum_{i<j} y_{ij} \mathbf{h}(\mathbf{X}_i, \mathbf{X}_j) \tag{3}$$

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 $Y_{ij} = 1$). In the ERGM resulting from equation (3), equation (1) becomes

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}|\mathbf{X}) = c^{-1} \prod\prod_{i<j} \exp\{y_{ij} \boldsymbol{\eta}^t \Delta(\mathbf{g}(\mathbf{y}, \mathbf{X}))_{ij}\}, \tag{4}$$

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 $\mathbf{y}$ remains unchanged. In equation (4), the joint distribution of the Y_{ij} is simply the product of the marginal distributions — hence the name "dyadic independence model". The MLE in such a model may be obtained 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 $\exp\{\eta\}/(1 + \exp\{\eta\})$.

For dyadic dependence models, equation (4) 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.

4 Goodness of fit for dyadic independence models

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 $\mathrm{s}_1(\mathbf{y})$, 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.

The second model we consider includes edges and also several statistics based on nodal covariates. Recall that in the dyadic independence model of equation (3), an individual component of the $\mathbf{g}(\mathbf{y},\mathbf{X})$ vector, say the kth component, may be written as

$$g_k(\mathbf{y},\mathbf{X}) = \sum_{i<j}\sum y_{ij} h_k(\mathbf{X}_i, \mathbf{X}_j). \tag{5}$$

Because it is not important, we drop the subscript k in equation (5) 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

$$h(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} 2 & \text{if both nodes } i \text{ and } j \text{ have the specified factor level;} \\ 1 & \text{if exactly one of } i, j \text{ has the specified factor level;} \\ 0 & \text{if neither } i \text{ nor } j \text{ has the specified factor level.} \end{cases} \tag{6}$$

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

$$h(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ both have the specified factor level;} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

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.

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

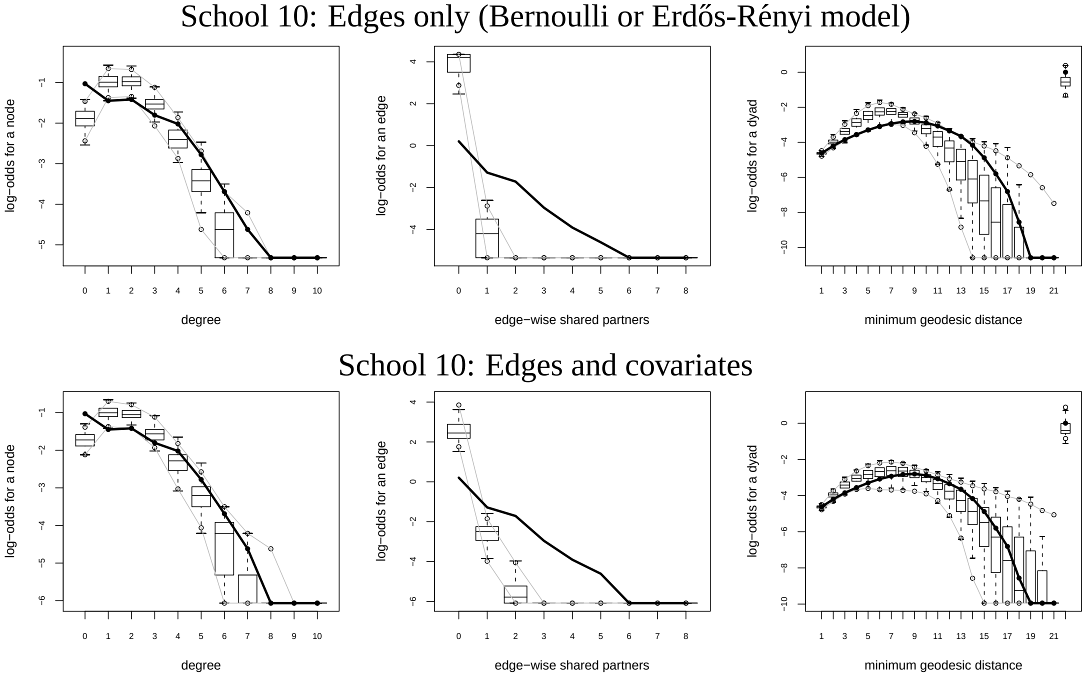

Figure 2: Simulation results for dyadic independence models. In all plots, the vertical axis is the logit of relative frequency; the School 10 statistics are indicated by the solid lines; the boxplots include the median and interquartile range; and the light gray lines represent the range in which 95 percent of simulated observations fall.

$\mathrm{ep}_0/\mathrm{s}_1, \ldots, \mathrm{ep}_{n-2}/\mathrm{s}_1$. (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 (11) 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 speed and robustness of diffusion across networks. 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 2 depicts the results of 100 simulations for School 10 from the fitted dyadic independence 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. The lack of fit 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.

5 Degree, shared partner, and other network 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 $\mathrm{d}_k(\mathbf{y})$, $0 \leq k \leq n-1$. 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 (Frank and Strauss, 1986) 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

$$\mathrm{s}_k(\mathbf{y}) = \sum_{i=k}^{n-1} \binom{i}{k} \mathrm{d}_i(\mathbf{y}),\ 2 \leq k \leq n-1; \quad \text{and} \quad \mathrm{s}_1(\mathbf{y}) = \frac{1}{2} \sum_{i=1}^{n-1} i\, \mathrm{d}_i(\mathbf{y}). \tag{8}$$

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 $\mathrm{dp}_k(\mathbf{y}) \geq \mathrm{ep}_k(\mathbf{y})$, 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

$$\mathrm{s}_1(\mathbf{y}) = \sum_{i=0}^{n-2} \mathrm{ep}_i(\mathbf{y}) \tag{9}$$

and

$$\binom{n}{2} = \sum_{i=0}^{n-2} \mathrm{dp}_i(\mathbf{y}). \tag{10}$$

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,

$$\mathrm{t}(\mathbf{y}) = \frac{1}{3} \sum_{i=0}^{n-2} i \, \mathrm{ep}_i(\mathbf{y}). \tag{11}$$

A related formula involving the dyadic shared partner statistics is obtained by noting that each triangle automatically comprises three 2-stars. Therefore, $\mathrm{s}_2(\mathbf{y}) - 3\mathrm{t}(\mathbf{y})$ is the number of 2-stars for which the third side of the triangle is missing. We conclude that

$$\mathrm{s}_2(\mathbf{y}) - 3\mathrm{t}(\mathbf{y}) \;\; = \;\; \sum_{i=0}^{n-2} i \left[\mathrm{dp}_i(\mathbf{y}) - \mathrm{ep}_i(\mathbf{y}) \right]. \tag{12}$$

Combining equation (12) with equation (11) produces

$$\mathrm{s}_2(\mathbf{y}) = \sum_{i=0}^{n-2} i \, \mathrm{dp}_i(\mathbf{y}).$$

Because a 2-star is also a path of length two, $\mathrm{s}_2(\mathbf{y})$ 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.

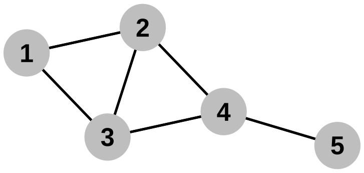

Figure 3: For this simple five-node network, the edgewise and dyadic shared partner distributions are $(\mathrm{ep}_0, \ldots, \mathrm{ep}_3) = (1, 4, 1, 0)$ and $(\mathrm{dp}_0, \ldots, \mathrm{dp}_3) = (2, 6, 2, 0)$, respectively; the k-triangle and k-twopath distributions are $(\mathrm{t}_1, \mathrm{t}_2, \mathrm{t}_3) = (1, 2, 0)$ and $(\mathrm{u}_1, \mathrm{u}_2, \mathrm{u}_3) = (10, 1, 0)$, respectively.

As a concrete example, we note that in the simple network of Figure 3, 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 $\mathrm{t}_k(\mathbf{y})$ and $\mathrm{u}_k(\mathbf{y})$, respectively. Just as the degree statistics $\mathrm{d}_i(\mathbf{y})$ are related to the k-star statistics $\mathrm{s}_k(\mathbf{y})$ by (8), the edgewise and dyadic shared partner statistics are related to the k-triangle and k-twopath statistics, respectively, by the equations

$$\mathrm{t}_k(\mathbf{y}) = \sum_{i=k}^{n-2} \binom{i}{k} \mathrm{ep}_i(\mathbf{y}), \;\; 2 \le k \le n-2$$

and

$$\mathrm{u}_k(\mathbf{y}) = \sum_{i=k}^{n-2} \binom{i}{k} \mathrm{dp}_i(\mathbf{y}), \;\; 1 \le k \le n-2, k \ne 2.$$

The cases not covered above are that of $\mathrm{t}_1(\mathbf{y})$, given in equation (11), and $\mathrm{u}_2(\mathbf{y})$, 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:

$$\mathrm{u}_2(\mathbf{y}) = \frac{1}{2} \sum_{i=2}^{n-2} \binom{i}{2} \mathrm{dp}_i(\mathbf{y}).$$

6 Goodness of fit for dyadic dependence models

A fundamental principle of social network analysis is that dependence among edges is a guiding force in the formation of networks — i.e., the pattern of 0's and 1's in an adjacency matrix $\mathbf{Y}$ cannot be described merely by a logistic regression model. A major step in the development of dyadic dependence ERGMs was taken by Frank and Strauss (1986), who proposed Markov random graphs. As originally proposed, these *homogeneous* Markov random graphs treated all nodes as equivalent, ignoring any covariate information. Since covariates are clearly important in social networks, homogeneous Markov random graph models fail to describe such networks adequately; yet even when covariate information of the sort discussed in Section 4 is included, these models generally do a poor job of representing the local clustering (edgewise shared partner distribution) in friendship networks. Nonetheless, the Markov assumption, by allowing for the presence of a triangle statistic in an ERGM, allows us to consider effects such as transitivity — 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. Reasons for the failure of Markov random graph models are explored by Handcock (2002; 2003). This failure motivated the work of Snijders et al. (2004) in developing the alternating k-triangle, k-twopath, and k-star statistics that we explain presently.

Consider using the shared partner statistics and the degree statistics defined in Section 5 to build an ERGM. For instance, it is possible to add one new term to the model for each of the edgewise shared partner statistics $\mathrm{ep}_1, \ldots, \mathrm{ep}_{n-2}$ — we omit ep_0 to avoid the linear dependence of equation (9) — 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 statistics

$$\begin{aligned}
\mathrm{ep}^G(\mathbf{y}; \tau) &= e^\tau \sum_{i=1}^{n-2} \left\{1 - \left(1 - e^{-\tau}\right)^i\right\} \mathrm{ep}_i(\mathbf{y}), & (13)\\
\mathrm{dp}^G(\mathbf{y}; \tau) &= e^\tau \sum_{i=1}^{n-2} \left\{1 - \left(1 - e^{-\tau}\right)^i\right\} \mathrm{dp}_i(\mathbf{y}), & (14)\\
\text{and } \mathrm{d}^G(\mathbf{y}; \tau) &= e^\tau \left\{2\mathrm{s}_1(\mathbf{y}) - e^\tau \sum_{i=1}^{n-1} \left[1 - \left(1 - e^{\tau}\right)^i\right] \mathrm{d}_i(\mathbf{y})\right\} &\\
&= \left(e^\tau\right)^2 \sum_{i=1}^{n-1} \left[\left(1 - e^{-\tau}\right)^i - 1 + ie^{-\tau}\right] \mathrm{d}_i(\mathbf{y}). & (15)
\end{aligned}$$

where τ in each case is an additional parameter. The superscript G stands for "geometrically weighted" in each case; we refer to these three statistics as *geometrically weighted* edgewise shared partner, dyadic shared partner, and degree statistics, respectively.

Although the definitions of ep^G, dp^G, and d^G may appear somewhat unusual, they are chosen to coincide with the alternating k-triangle, alternating k-twopath, and alternating k-star statistics, respectively, of Snijders et al. (2004):

$$\begin{aligned}
\mathrm{ep}^G(\mathbf{y}; \tau) &= 3\mathrm{t}_1(\mathbf{y}) - \frac{\mathrm{t}_2(\mathbf{y})}{(e^\tau)^1} + \cdots + (-1)^{n-3}\frac{\mathrm{t}_{n-2}(\mathbf{y})}{(e^\tau)^{n-3}}, & (16)\\
\mathrm{dp}^G(\mathbf{y}; \tau) &= \mathrm{u}_1(\mathbf{y}) - \frac{2\mathrm{u}_2(\mathbf{y})}{(e^\tau)^1} + \cdots + (-1)^{n-3}\frac{\mathrm{u}_{n-2}(\mathbf{y})}{(e^\tau)^{n-3}}, & (17)\\
\text{and } \mathrm{d}^G(\mathbf{y}; \tau) &= \mathrm{s}_2(\mathbf{y}) - \frac{\mathrm{s}_3(\mathbf{y})}{(e^\tau)^1} + \cdots + (-1)^{n-1}\frac{\mathrm{s}_{n-1}(\mathbf{y})}{(e^\tau)^{n-3}}. & (18)
\end{aligned}$$

As Snijders et al. (2004) explain, these three statistics appear to capture high-order dependency structure in networks in a parsimonious fashion while avoiding the problems of degeneracy described by Handcock (2002; 2003).

The τ parameters in equations (13), (14), and (15) are not canonical exponential family parameters like $\boldsymbol{\eta}$ in equation (1); rather, if $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$ is considered unknown and $(\boldsymbol{\eta}, \boldsymbol{\tau})$ is the full parameter vector, then the ERGM forms a *curved exponential family*, which complicates

	Model:			
Coefficient:	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 estimation procedure. Hunter and Handcock (2004) address this more complicated situation; however, for the purposes of this article, we make the simplifying assumption that each 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 for each statistic, 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 4) were nearly indistinguishable for different values of τ in the range we tested (0.5 to 2.0). Values far outside this range resulted in models that could not be fit. 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

$$P_{\boldsymbol{\eta}}(\mathbf{Y} = \mathbf{y}|\mathbf{X}) = c^{-1}\exp\{\eta_1 \mathrm{s}_1(\mathbf{y}) + \eta_2 \mathrm{ep}^G(\mathbf{y};\tau)\}. \tag{19}$$

We fit model (19), as well as similar models using the geometrically weighted dyadic shared partner (GWDSP) and geometrically weighted degree (GWD) statistics, to AddHealth school 10. The results are found in Table 1.

Many dyadic dependence models create such severe numerical difficulties in estimation (Handcock, 2002; 2003) that we are unable to fit them successfully for a large number of different networks of different sizes. However, models with the GWESP, GWDSP, and GWD statistics appear to be more robust: Using our MCMC fitting procedure, we were able to be estimate their parameters on many of the AddHealth schools, the first such application of maximum likelihood estimation to a dyadic dependence model for a range of different-sized networks with hundreds of nodes. As a case in point, consider Figure 6, in which we successfully fit a dyadic dependence model to the largest school in the sample, with 2209 nodes, and obtained reasonable parameter estimates. (We discuss this school further in Section 8.)

As described in Section 4, 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 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 4 depicts simulation results for school 10 for the three dyadic-dependent ERGMs in Table 1; Figure 5 depicts Model II from Table 2.

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 4. 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.

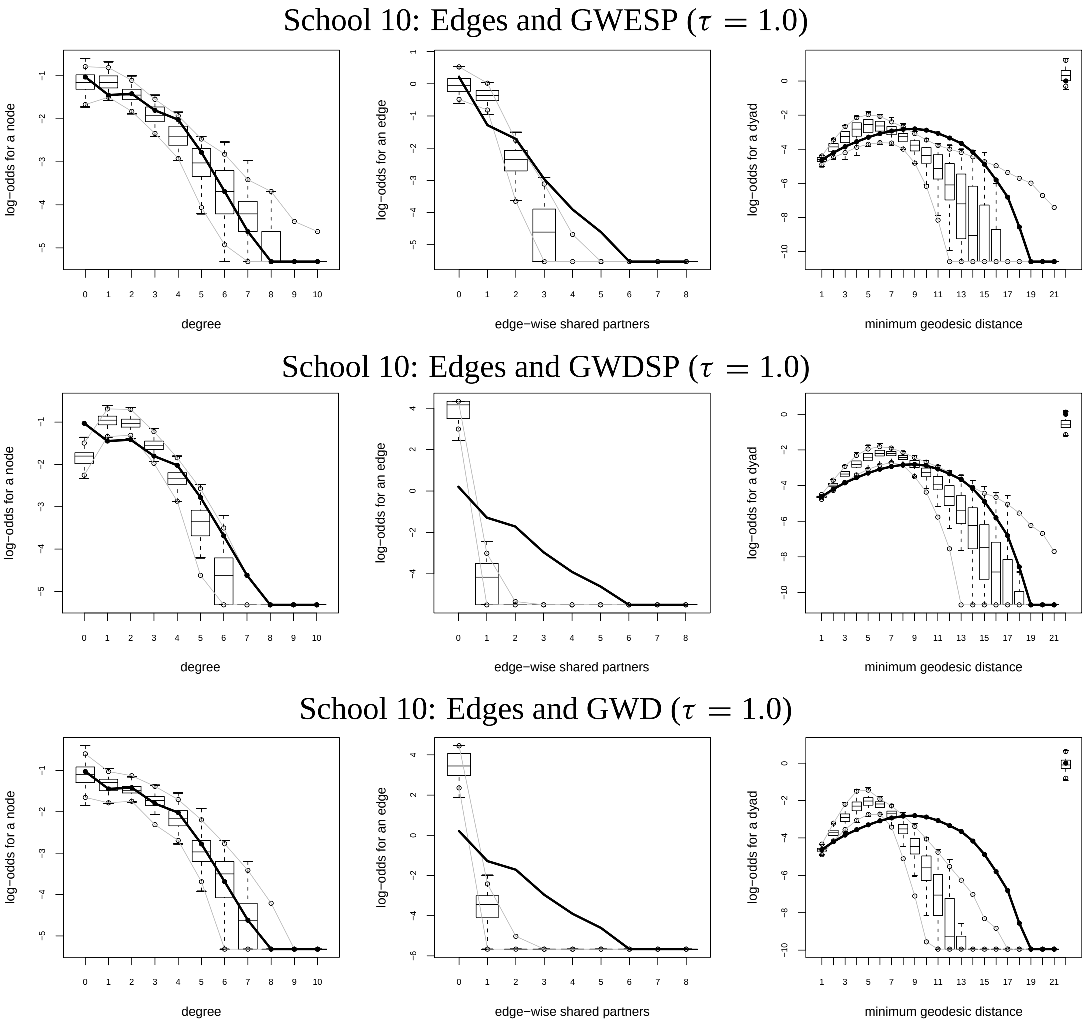

Figure 4: Simulation results for dyadic dependence ERGMs of Table 1

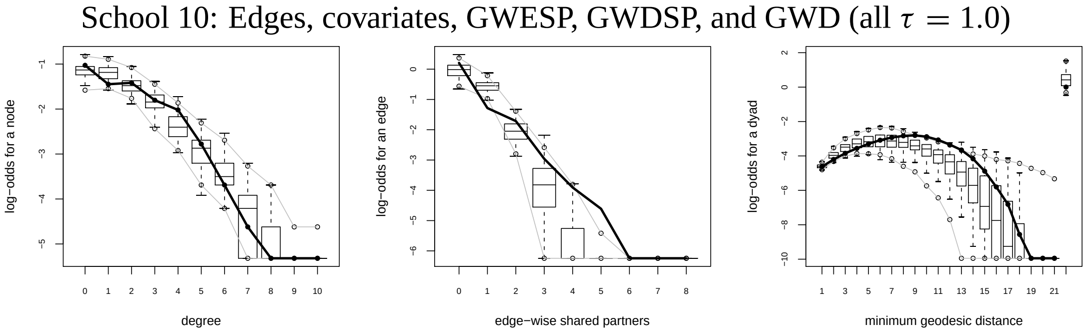

Figure 5: Simulation results for Model II of Table 2.

For both School 10 and many of the other smaller AddHealth schools, a simple model containing only individual-level attributes (Figure 2, 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: Gathering such information only requires questions about the attributes of respondents' partners, not their actual identities. 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. Finally, the fact that a simple model is strongly predictive of one higher-order network property (geodesics) and strongly divergent from another (shared partner) suggests that a variety of network statistics ought to be tested in order to develop a robust sense of goodness-of-fit.

Comparing the bottom graph in Figure 2 with the top two graphs in Figure 4, 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; we expect that nodal covariates are very important in predicting most types of social relationships, and certainly high school friendships are no exception.

Social relations generally exhibit local clustering, and in this case we observe that 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 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 the transitivity property that friends of my friends are more likely to be my friends. The modelling here shows that neither homophily nor shared partners alone is sufficient to explain the clustering observed in this friendship network (the same is true of other AddHealth schools; see the plots at `http://csde.washington.edu/networks`). Indeed, in Table 2 we see that the homophily effects are smaller in magnitude in Model II, which includes the shared partner statistics, than in Model I.

In this setting, a simple one-term Bernoulli model (Figure 2, top graph) turned out to fit the degree distribution fairly well, and adding the GWD statistic improved the fit to the degree distribution. But neither of these models reproduced the clustering and geodesics observed in this network. This finding seems particularly important given the great attention that has been placed on degree-only models in some branches of the networks literature recently; see Albert and Barabási (2002) for a survey of some of this literature.

7 Other model selection criteria

To see whether the results we observe from the goodness of fit plots are consistent with traditional criteria, 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

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 4. For GWD, GWESP, and GWDSP, τ always equals 1.0. Asterisks indicate the models in which approximate loglikelihoods are used.

fidelity to data and parsimony of representation:

$$\text{AIC}(M) = -2(\text{maximized loglikelihood under } M) + 2(\text{\# of parameters in } M), \qquad (20)$$

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 (20) is approximate except in the case of a dyadic independence model, where the pseudolikelihood (4) is equal to the likelihood.

The graphical approach agrees 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 plots are more informative than the AIC results in the sense that they tell which structural features are fit well and which are not. 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 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 the value of a single dyad essentially determines the values of all others, making the effective sample size about one. 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.

8 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 7, 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 and many of the smaller schools, the same is not true for many of the larger schools. For instance, consider Figure 6, based on the largest school in our sample, with 2209 nodes. This and other large 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 students are less likely to be friends as the gap between their grade levels widens, a hypothesis supported by earlier research on assortative mixing on quantitative traits (Morris, 1991). This is

School 44: Edges, covariates, and GWESP ($\tau = 1.5$)

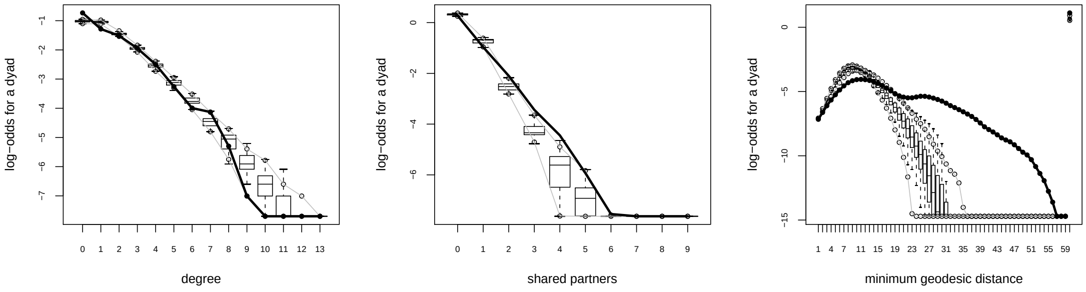

Figure 6: Goodness-of-fit plots for the largest AddHealth school, school 44, with 2209 nodes. The clear lack of fit in the geodesic distribution is typical of this model for the larger AddHealth schools, even though the same model tends to fit well on smaller schools.

only one of many additional processes underlying the structure of some of the larger school groups that could be incorporated into a more detailed analysis.

As this empirical application has shown, both exogenous nodal covariates and endogenous network effects can play an important role in the generative processes that give rise to network structure. There is no *a priori* reason to assume that all networks will have the same structure, and the methods here provide a systematic framework for evaluation of models that can be adapted to test a wide range of hypotheses. In the context of mutual friendships among high-school adolescents, 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 the aggregate network structures of interest.

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. Washington, DC: National Academy 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.

Morris, M. (1991), A log-linear modeling framework for selective mixing, *Mathematical Biosciences*, **107**: 349–377.

Morris, M. (2003), Local rules and global properties: Modeling the emergence of network structure, pp. 174 – 186 in *Dynamic Social Network Modelling and Analysis: Workshop Summary and Papers*, edited by Ronald Breiger, Kathleen Carley, and Philippa E. Pattison. Washington, DC: National Academy Press.

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.