Center for Studies in Demography and Ecology

Inference in curved exponential family models for networks

by

David R. Hunter
Pennsylvania State University, University Park

Mark S. Handcock
University of Washington, Seattle

UNIVERSITY OF WASHINGTON **CSDE Working Paper No. 04-14**

Inference in curved exponential family models for networks [1]

David R. Hunter
Pennsylvania State University, University Park

Mark S. Handcock
University of Washington, Seattle

Working Paper no. 43
Center for Statistics and the Social Sciences
University of Washington

August 3, 2004

[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); 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. This research supported by Grant DA012831 from NIDA and Grant HD041877 from NICHD. The authors are grateful to Tom Snijders for detailed comments and discussions. We thank Steven Goodreau and Martina Morris for comments.

Abstract

Network data arise in a wide variety of applications. Although descriptive statistics for networks abound in the literature, the science of fitting statistical models to complex network data is still in its infancy. The models considered in this article are based on exponential families; therefore, we refer to them as exponential random graph models (ERGMs). Although ERGMs are easy to postulate, maximum likelihood estimation of parameters in these models is very difficult. In this article, we first review the method of maximum likelihood estimation using Markov chain Monte Carlo in the context of fitting linear ERGMs. We then extend this methodology to the situation where the model comes from a curved exponential family. The curved exponential family methodology is applied to new specifications of ERGMs, proposed by Snijders et al. (2004), having non-linear parameters to represent structural properties of networks such as transitivity and heterogeneity of degrees. We review the difficult topic of implementing likelihood ratio tests for these models, then apply all these model-fitting and testing techniques to the estimation of linear and non-linear parameters for a collaboration network between partners in a New England law firm.

KEY WORDS: exponential random graph model; maximum likelihood estimation; Markov chain Monte Carlo; $p-$star model.

1. INTRODUCTION

A network is a way to represent "relational data" — i.e., data whose properties cannot be reduced to the attributes of the individuals involved — in the form of a mathematical graph. For the purposes of this article, a network consists of a set of nodes and a set of edges, where an edge is an ordered or unordered pair of nodes. 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 concerns inference in specific probabilistic models for networks. Throughout, we will represent a generic random network by the matrix Y, an $n \times n$ matrix where n is the number of nodes. Each Y_{ij} can equal zero or one, with one indicating the presence of an edge between i and j and zero indicating the absence of such an edge. More complicated networks may be represented if Y_{ij} is allowed to take on arbitrary values, in which case the edges may be considered to have weights; however, we avoid such complications here. We disallow the possibility of self-edges, so $Y_{ii} = 0$ for all i. Furthermore, for the sake of simplicity we develop arguments using the assumption that Y is undirected — that is, $Y_{ij} = Y_{ji}$ for all i and j so only the lower triangle of Y is relevant. However, none of the theory we present depends essentially on the undirectedness assumption.

The models we consider for the random behavior of Y rely on a p-vector $\mathbf{Z}(Y)$ of statistics and a parameter vector $\boldsymbol{\eta} \in R^p$. The canonical exponential family model is

$$P(Y = y) = \exp\{\boldsymbol{\eta}^t \mathbf{Z}(y) - \psi(\boldsymbol{\eta})\}, \tag{1}$$

where

$$\exp\{\psi(\boldsymbol{\eta})\} = \sum_x \exp\{\boldsymbol{\eta}^t \mathbf{Z}(x)\} \tag{2}$$

is the familiar normalizing constant associated with an exponential family of distributions (Barndorff-Nielsen 1978; Lehmann, 1983). The sum in (2) is taken over the whole sample space, which presents a very important problem in most applications: A sample space consisting of all possible undirected graphs on n nodes contains $\exp\{\binom{n}{2}\log 2\}$ elements, an astronomically large number even for moderately sized n of, say, 20. Thus, for most applications it is impossible even to evaluate the likelihood function for a particular $\boldsymbol{\eta}$, let alone maximize it. We consider ways around this problem in Section 2.

The range of network statistics that might be included in the $\mathbf{Z}(y)$ vector is vast — see Wasserman and Faust (1994) for the most comprehensive treatment of these statistics — though we will consider only a few in this article. We allow the vector $\mathbf{Z}(y)$ to include covariate information about nodes or edges in the graph in addition to information derived directly from the matrix y itself. Thus, $\mathbf{Z}(y)$ should be viewed as a function not only of y, but also potentially of certain *exogenous* covariates, by which we mean covariates on nodes or pairs of nodes whose values are not affected by the presence or absence of edges. For example, if each node is a person, $\mathbf{Z}(y)$ might include the total number of edges between individuals of the same gender, which is a function of both the graph y and the exogenous nodal covariate gender. For notational simplicity, we prefer to allow the dependence of $\mathbf{Z}$ on exogenous covariates to be implicit rather than explicitly indicated by the notation.

There has been a lot of work on models of the form (1), to which we refer as exponential random graph models or ERGMs for short. (We avoid the lengthier EFRGM, for "exponential family random graph models," both for the sake of brevity and because we consider some models in this article that should technically be called *curved* exponential families.) 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 each dyad — by which we mean each pair of nodes — having edges independently of every other dyad. 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 share a node. Frank (1991) mentioned the application of model (1) to social networks in its full generality. This was pursued in depth by Wasserman and Pattison (1996). In honor of Holland and Leinhardt's p_1 model, they referred to model (1) as $p*$ ($p-$star), a name that has been widely applied to ERGMs in the social networks literature.

Inference for this class of models was considered in the seminal paper by Geyer and Thompson (1992), building on the methods of Frank and Strauss (1986) and the above cited papers. Until recently, inference for social networks models has relied on maximum pseudolikelihood estimation (Besag 1974; Frank and Strauss, 1986; Strauss and Ikeda, 1990; Geyer and Thompson 1992). Geyer and Thompson (1992) proposed a stochastic algorithm to approximate maximum likelihood estimates for model (1) among other models; this Markov chain Monte Carlo (MCMC) approach forms the basis of the method described in this article. The development of these methods for social network data has been considered by Dahmström and Dahmström (1993), Corander et al. (1998), Crouch et al. (1998), Snijders (2002), and Handcock (2002).

In this article, we begin with a summary in Section 2 of the basic idea behind the MCMC maximum likelihood approach. Many of the estimation ideas in Section 3 are more or less implicit in the articles of Geyer and Thompson (1992) and Geyer (1994), though their application to fitting

curved exponential family models is new. Section 4 describes several particular ERGMs due to Snijders et al. (2004) and demonstrates how to fit them. Section 5 discusses an approach to the difficult issue of implementing a likelihood ratio test in this context. Finally, Section 6 ties all of the previous sections together, demonstrating the use of these methods to fit an ERGM to a collaboration network among lawyers, a problem considered by Snijders et al. (2004). Whereas Snijders et al. (2004) estimated some of the parameters in their model but assumed others were fixed and known, we apply the curved exponential family machinery to estimating all of the parameters.

2. MARKOV CHAIN MONTE CARLO MAXIMUM LIKELIHOOD ESTIMATION

In Section 1, we pointed out the difficulty of evaluating $\psi(\boldsymbol{\eta})$ in equation (2) due to the fact that it involves a sum with an extremely large number of terms. Here, we discuss a way around this problem in preparation for a discussion in Section 3 about estimating the parameters via maximum likelihood. The method uses Markov chain Monte Carlo to approximate the likelihood function, and then maximizes this approximation.

Let $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^0$ denote two distinct values of the canonical parameter in model (1). We are interested in calculating $\exp\{\psi(\boldsymbol{\eta}) - \psi(\boldsymbol{\eta}^0)\}$ as a function of $\boldsymbol{\eta}$, where $\boldsymbol{\eta}^0$ is fixed and known. Since

$$\begin{aligned} \exp\{\psi(\boldsymbol{\eta}) - \psi(\boldsymbol{\eta}^0)\} &= \sum_x \exp\{(\boldsymbol{\eta} - \boldsymbol{\eta}^0)^t \mathbf{Z}(x)\} \left(\frac{\exp\{(\boldsymbol{\eta}^0)^t \mathbf{Z}(x)\}}{\exp\{\psi(\boldsymbol{\eta}^0)\}} \right) \\ &= \mathrm{E}_{\boldsymbol{\eta}^0} \left[\exp\{(\boldsymbol{\eta} - \boldsymbol{\eta}^0)^t \mathbf{Z}(Y)\} \right], \end{aligned} \tag{3}$$

we may approximate $\exp\{\psi(\boldsymbol{\eta}) - \psi(\boldsymbol{\eta}^0)\}$ by the sample mean

$$\frac{1}{m} \sum_{i=1}^{m} \exp\{(\boldsymbol{\eta} - \boldsymbol{\eta}^0)^t \mathbf{Z}(Y_i)\}, \tag{4}$$

where $Y_1, \ldots, Y_m$ is a sample of random graphs from the distribution defined by $\boldsymbol{\eta}^0$. Such a sample may be obtained using Markov chain Monte Carlo.

Let $\ell(\boldsymbol{\eta})$ be the log-likelihood for model (1) based on observing a single realization y_{obs} of Y. Letting $r(\boldsymbol{\eta}, \boldsymbol{\eta}^0) \stackrel{\text{def}}{=} \ell(\boldsymbol{\eta}) - \ell(\boldsymbol{\eta}^0)$ denote the logarithm of the likelihood ratio, we apply the ideas above and approximate $r(\boldsymbol{\eta}, \boldsymbol{\eta}^0)$ by

$$\hat{r}_m(\boldsymbol{\eta}, \boldsymbol{\eta}^0) \stackrel{\text{def}}{=} (\boldsymbol{\eta} - \boldsymbol{\eta}^0)^t \mathbf{Z}(y_{\text{obs}}) - \log \left[\frac{1}{m} \sum_{i=1}^{m} \exp\{(\boldsymbol{\eta} - \boldsymbol{\eta}^0)^t \mathbf{Z}(Y_i)\} \right]. \tag{5}$$

The strong convergence of $\hat{r}_m(\boldsymbol{\eta}, \boldsymbol{\eta}^0)$ to $r(\boldsymbol{\eta}, \boldsymbol{\eta}^0)$ as $m \to \infty$ is guaranteed by a Markov chain version of the strong law of large numbers (Meyn and Tweedie, 1993). Thus, for a fixed sample size m, maximization of $\hat{r}_m(\boldsymbol{\eta}, \boldsymbol{\eta}^0)$ as a function of $\boldsymbol{\eta}$ gives an approximation to the maximum likelihood

estimator $\hat{\boldsymbol{\eta}}$. This procedure, which may be termed Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE for those who like acronyms), originates in Geyer and Thompson (1992).

Note that $\ell(\boldsymbol{\eta})$ and $r(\boldsymbol{\eta}, \boldsymbol{\eta}^0)$ are unchanged if $\mathbf{Z}(y)$ is replaced by $\mathbf{Z}(y) - \mathbf{a}$ for some constant vector $\mathbf{a}$. For example, we might take $\mathbf{a} = \mathbf{Z}(y_{\text{obs}})$, in which case $\mathbf{Z}(y)$ represents the *change* in the vector of statistics for the graph y relative to the observed graph y_{obs}. This makes $\mathbf{Z}(y_{\text{obs}}) = 0$, which simplifies the definition of $\hat{r}_m(\boldsymbol{\eta}, \boldsymbol{\eta}^0)$ in equation (5). Alternatively, we might take $\mathbf{a} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{Z}(Y_i)$, which has the effect of centering the $\mathbf{Z}(Y_i)$ at zero, leading to more stable numerical calculations.

In some applications, we may want to estimate not merely the likelihood ratio but the actual value of the log-likelihood itself. This may be accomplished by noting that $\ell(\mathbf{0}) = -\log M$, where M is the size of the sample space. For instance, if the sample space includes all undirected graphs on n nodes, then $\log M = \binom{n}{2}\log 2$. By combining $\ell(\mathbf{0})$ with estimates of $\ell(\boldsymbol{\eta}) - \ell(\boldsymbol{\eta}^0)$ and $\ell(\mathbf{0}) - \ell(\boldsymbol{\eta}^0)$, we obtain

$$\hat{\ell}(\boldsymbol{\eta}) \stackrel{\text{def}}{=} \hat{r}_m(\boldsymbol{\eta}, \boldsymbol{\eta}^0) - \hat{r}_m(\mathbf{0}, \boldsymbol{\eta}^0) - \log M. \tag{6}$$

It remains to describe how to generate a Markov chain whose stationary distribution is given by equation (1). The simplest Markov chain proceeds by choosing (by some method, either stochastic or deterministic) a dyad (i, j) and then deciding whether to set $Y_{ij} = 1$ or $Y_{ij} = 0$ at the next step of the chain. One way to do this is using Gibbs sampling, whereby the new value of Y_{ij} is sampled from the conditional distribution of Y_{ij} conditional on the rest of the graph. Denote "the rest of the graph" by Y_{ij}^c. Then $Y_{ij}|Y_{ij}^c = y_{ij}^c$ has a Bernoulli distribution, with odds given by

$$\frac{P(Y_{ij} = 1|Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0|Y_{ij}^c = y_{ij}^c)} = \exp\{\boldsymbol{\eta}^t \Delta(\mathbf{Z}(y))_{ij}\}, \tag{7}$$

where $\Delta(\mathbf{Z}(y))_{ij}$ denotes the difference between $\mathbf{Z}(y)$ when y_{ij} is set to 1 and $\mathbf{Z}(y)$ when y_{ij} is set to 0. A simple variant to the Gibbs sampler (which is an instance of a Metropolis-Hastings algorithm) is a pure Metropolis algorithm in which the proposal is always to change the value of y_{ij}. This proposal is accepted with probability $\min\{1, \pi\}$, where

$$\pi = \frac{P(Y_{ij} = 1 - y_{ij}|Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = y_{ij}|Y_{ij}^c = y_{ij}^c)} = \begin{cases} \exp\left\{\boldsymbol{\eta}^t \Delta(\mathbf{Z}(y))_{ij}\right\} & \text{if } y_{ij} = 0; \\ \exp\left\{-\boldsymbol{\eta}^t \Delta(\mathbf{Z}(y))_{ij}\right\} & \text{if } y_{ij} = 1. \end{cases}$$

The vector $\Delta(\mathbf{Z}(y))_{ij}$ used by these MCMC schemes is often much easier to calculate directly than as the difference of two separate values of $\mathbf{Z}(y)$. For instance, if one of the components of the $\mathbf{Z}(y)$ vector is the total number of edges in the graph, then the corresponding component of $\Delta(\mathbf{Z}(y))_{ij}$ is always equal to 1.

The Metropolis scheme is usually preferred over the Gibbs scheme because it results in a greater probability of changing the value of y_{ij}, a property thought to produce better-mixing chains. However, it is well known that these simple MCMC schemes often fail for various reasons to produce well-mixed chains (Snijders 2002; Handcock 2002, 2003; Snijders et al. 2004). The choice of the model class and more sophisticated MCMC schemes are a topic of ongoing research. We return to the former in Section 4.

3. ESTIMATION FOR CURVED EXPONENTIAL FAMILIES

Suppose that $\boldsymbol{\eta} \in R^p$, the canonical parameter in equation (1), is a function of a lower-dimensional parameter $\boldsymbol{\theta} \in R^q$, $q < p$. If the function is linear, say $\boldsymbol{\eta} = A\boldsymbol{\theta}$ for some $p \times q$ matrix A, then $\boldsymbol{\theta}$ is simply the canonical exponential family parameter for the reduced set of statistics $A^t\mathbf{Z}(y)$. However, if the function mapping $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$ is nonlinear, then in general the situation is more complicated. The family of distributions

$$P(Y = y) = \exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^t\mathbf{Z}(y) - \psi[\boldsymbol{\eta}(\boldsymbol{\theta})]\}, \boldsymbol{\theta} \in R^q$$

is called a *curved exponential family* in the terminology of Efron (1975).

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ satisfies the likelihood equation

$$\nabla\ell(\hat{\boldsymbol{\theta}}) = \nabla\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})^t[\mathbf{Z}(y_{\text{obs}}) - \mathrm{E}_{\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})}\mathbf{Z}(Y)] = \mathbf{0}, \tag{8}$$

where $\nabla\boldsymbol{\eta}(\boldsymbol{\theta})$ is the $p \times q$ matrix of partial derivatives of $\boldsymbol{\eta}$ with respect to $\boldsymbol{\theta}$. We may search for a solution to equation (8) using an iterative technique such as Newton-Raphson; however, the exponential family form of the model makes the Fisher information matrix

$$I(\boldsymbol{\theta}) = \nabla\boldsymbol{\eta}(\boldsymbol{\theta})^t[\mathrm{Var}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\mathbf{Z}(Y)]\nabla\boldsymbol{\eta}(\boldsymbol{\theta}) \tag{9}$$

easier to calculate than the Hessian matrix of second derivatives required for Newton-Raphson. For more about equations (8) and (9), see Efron (1978). The information matrix (9) is the basis for the method of Fisher scoring, which is analogous to Newton-Raphson except that $-I(\boldsymbol{\theta})$ is used in place of the Hessian matrix. Thus, if $\boldsymbol{\theta}^{(k)}$ denotes the estimate of $\boldsymbol{\theta}$ at the kth iteration, Fisher scoring sets

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \left[I(\boldsymbol{\theta}^{(k)})\right]^{-1}\nabla\ell(\boldsymbol{\theta}^{(k)}). \tag{10}$$

The biggest obstacle to overcome in implementing the scoring algorithm (10) is the fact that $\mathrm{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\mathbf{Z}(Y)$ and $\mathrm{Var}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\mathbf{Z}(Y)$ are difficult to calculate directly for ERGMs. One approach to estimating these quantities is to use one of the MCMC methods described in Section 2 to generate a sample $Y_1, \ldots, Y_m$ from the distribution defined by the parameter value $\boldsymbol{\theta}$, then use the sample

mean and covariance of $\mathbf{Z}(Y_1), \ldots, \mathbf{Z}(Y_m)$ to approximate $\mathrm{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\,\mathbf{Z}(Y)$ and $\mathrm{Var}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\,\mathbf{Z}(Y)$. However, such an approach could prove computationally expensive in an optimization routine, since a new sample would have to be generated each time the value of $\boldsymbol{\theta}$ changed. An alternative is to generate a single sample, based on a fixed parameter value $\boldsymbol{\theta}^0$. Let $Y_1, \ldots, Y_m$ denote this sample, and suppose that $\boldsymbol{\theta}^{(k)}$ is the value of the parameter vector at the kth iteration of an iterative algorithm. Then the approximate Fisher scoring method is implemented as

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \left\{\hat{I}(\boldsymbol{\theta}^{(k)})\right\}^{-1} \nabla\boldsymbol{\eta}(\boldsymbol{\theta})^t \left[\mathbf{Z}_{\mathrm{obs}} - \sum_i w_i^{(k)} \mathbf{Z}_i\right], \tag{11}$$

where $\mathbf{Z}_{\mathrm{obs}}$ and $\mathbf{Z}_i$ denote $\mathbf{Z}(y_{\mathrm{obs}})$ and $\mathbf{Z}(Y_i)$, respectively;

$$w_i^{(k)} = \frac{\exp\{[\boldsymbol{\eta}(\boldsymbol{\theta}^{(k)}) - \boldsymbol{\eta}(\boldsymbol{\theta}^0)]^t \mathbf{Z}_i\}}{\sum_{j=1}^n \exp\{[\boldsymbol{\eta}(\boldsymbol{\theta}^{(k)}) - \boldsymbol{\eta}(\boldsymbol{\theta}^0)]^t \mathbf{Z}_j\}};$$

and

$$\hat{I}(\boldsymbol{\theta}^{(k)}) = \nabla\boldsymbol{\eta}(\boldsymbol{\theta}^{(k)})^t \left\{\sum_{i=1}^m w_i^{(k)} \mathbf{Z}_i \mathbf{Z}_i^t - \left(\sum_{i=1}^m w_i^{(k)} \mathbf{Z}_i\right)\left(\sum_{i=1}^m w_i^{(k)} \mathbf{Z}_i\right)^t\right\} \nabla\boldsymbol{\eta}(\boldsymbol{\theta}^{(k)}). \tag{12}$$

Equations (11) and (12) are derived by first writing $\mathrm{E}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\,\mathbf{Z}(Y)$ and $\mathrm{Var}_{\boldsymbol{\eta}(\boldsymbol{\theta})}\,\mathbf{Z}(Y)$ in terms of expectations involving only $\mathrm{E}_{\boldsymbol{\eta}(\boldsymbol{\theta}^0)}$ as in equation (3), then substituting sample means like expression (4) for population means.

The two ideas above for stochastic optimization algorithms, one in which we generate a new sample with every iteration and one in which we generate only a single sample, each have their drawbacks. As pointed out above, the first idea is expensive computationally. However, the second may lead to an estimate $\hat{r}_m$ that is not very close to r (where there is no ambiguity, we write $\hat{r}_m$ and r instead of $\hat{r}_m[\boldsymbol{\eta}(\theta), \boldsymbol{\eta}(\theta^0)]$ and $r[\boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{\eta}(\boldsymbol{\theta}^0)]$, respectively). A compromise is represented by the following scheme, which is similar to the approach used by Geyer and Thompson (1992):

1. Select an initial value $\boldsymbol{\theta}^0$.
2. Generate an MCMC sample $\mathbf{Z}(Y_1), \ldots, \mathbf{Z}(Y_m)$ using the parameter $\boldsymbol{\theta}^0$.
3. Iterate algorithm (11) until convergence, obtaining a maximizer $\tilde{\boldsymbol{\theta}}$ of $\hat{r}_m$.
4. If $\mathrm{Var}_{\mathrm{MC}}\,\hat{r}_m$ of equation (13) is too large compared to $\hat{\ell}(\boldsymbol{\eta}(\tilde{\theta}))$, say $\sqrt{\mathrm{Var}_{\mathrm{MC}}\,\hat{r}_m} > b\hat{\ell}(\boldsymbol{\eta}(\tilde{\theta}))$ for some constant $b < 0$, then set $\boldsymbol{\theta}^0 = \tilde{\boldsymbol{\theta}}$ and return to step 2.
5. Take $\tilde{\boldsymbol{\theta}}$ to be the MCMCMLE.

Since $\tilde{\boldsymbol{\theta}}$ is a maximizer of $\hat{r}_m$ by step 3, the logic of the algorithm is that we take $\tilde{\boldsymbol{\theta}}$ to be the MCMCMLE as long as we're convinced that $\hat{r}_m$ is close to the true r.

Let U_i denote $\exp\{[\boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta}^0)]^t \mathbf{Z}_i\}$ for $i = 1, \ldots, m$ and $\bar{U} = \frac{1}{m}\sum_{i=1}^m U_i$. The variance used in step 4 is

$$\mathrm{Var}_{\mathrm{MC}}\left[\hat{r}_m\right] \stackrel{\mathrm{def}}{=} \frac{1}{m^2\bar{U}^2}\sum_{k=-K}^{K}(m-|k|)\hat{\gamma}_k, \tag{13}$$

where $\hat{\gamma}_k = \hat{\gamma}_{-k}$ denotes the sample lag-k autocovariance of the sequence $U_1, U_2, \ldots$, which we assume to be stationary. Equation (13) is obtained from the Taylor approximation $\log(a/b) \approx (a-b)/b$, whence

$$\mathrm{Var}\left[\log\left(\bar{U}\right)\right] \approx \frac{\mathrm{Var}(\bar{U})}{[E(\bar{U})]^2}.$$

The value of K in equation (13) is chosen large enough so that the lag-k autocovariance is approximately zero for $|k| > K$. If the U_i are approximately uncorrelated (for example, if the Markov chain is sampled only at very large intervals), expression (13) reduces to $[\sum_i U_i^2/(m\bar{U})^2] - 1/m$.

After the algorithm has converged, the question of obtaining standard errors remains. There are two interesting aspects of the error: The MCMC error, which is the error in approximating the true MLE, $\hat{\boldsymbol{\theta}}$, by the MCMCMLE, $\tilde{\boldsymbol{\theta}}$; and the usual error inherent in using the MLE $\hat{\boldsymbol{\theta}}$ to approximate reality. For the latter, we rely on standard asymptotic results and use the estimated Fisher information matrix (12) to obtain an estimate $[\hat{I}(\tilde{\boldsymbol{\theta}})]^{-1}$ of the covariance matrix.

For the former error, incurred by approximating $\hat{\boldsymbol{\theta}}$ by $\tilde{\boldsymbol{\theta}}$, we obtain a separate MCMC covariance matrix. Geyer (1994) gives mild regularity conditions under which $\sqrt{m}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$ is asymptotically normal, conditional on $\hat{\boldsymbol{\theta}}$. The asymptotic covariance matrix of $\sqrt{m}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}})$ forms the basis of our MCMC covariance matrix.

A first-order Taylor expansion gives

$$\sqrt{m}(\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}) \approx -\left[\nabla^2\hat{r}_m(\tilde{\boldsymbol{\theta}})\right]^{-1}\left[\sqrt{m}\nabla\hat{r}_m(\hat{\boldsymbol{\theta}})\right]. \tag{14}$$

(Note that we write $\hat{r}_m(\boldsymbol{\theta})$ instead of $\hat{r}_m[\boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{\eta}(\boldsymbol{\theta}^0)]$ in order to simplify notation.) Suppose that graphs $Y_1, Y_2, \ldots$ arise from a (stationary) Markov chain defined by $\boldsymbol{\theta}^0$. In expression (14), $\sqrt{m}\nabla\hat{r}_m(\hat{\boldsymbol{\theta}})$ converges in distribution as $m \to \infty$ to a q-variate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\left[\frac{c(\boldsymbol{\theta}^0)}{c(\hat{\boldsymbol{\theta}})}\right]^2 \sum_{k=-\infty}^{\infty}\mathrm{Cov}[\mathbf{W}_1(\hat{\boldsymbol{\theta}}), \mathbf{W}_{1+|k|}(\hat{\boldsymbol{\theta}})], \tag{15}$$

where $c(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \exp\{\psi[\boldsymbol{\eta}(\boldsymbol{\theta})]\}$ is the normalizing constant of equation (2) and

$$\mathbf{W}_i(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} \{\mathbf{Z}(y_{\mathrm{obs}}) - \mathbf{Z}(Y_i)\}\exp\left\{[\boldsymbol{\eta}(\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta}^0)]^t\mathbf{Z}(Y_i)\right\}. \tag{16}$$

We do not know the value of $\hat{\boldsymbol{\theta}}$ in expression (15); therefore, we approximate it by $\tilde{\boldsymbol{\theta}}$. Using a sample mean as in equation (4) to approximate the ratio $c(\boldsymbol{\theta}^0)/c(\tilde{\boldsymbol{\theta}})$, expression (15) is approximately

$$\tilde{V} \stackrel{\text{def}}{=} \frac{1}{m^2}\left[\sum_{i=1}^{m}\exp\{[\boldsymbol{\eta}(\boldsymbol{\theta}^0)-\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}})]^t\mathbf{Z}(Y_i)\}\right]^2\sum_{k=-K}^{K}\hat{\xi}_k,$$

where $\hat{\xi}_k = \hat{\xi}_{-k}$ is the sample lag-k autocovariance of the sequence $\mathbf{W}_1(\tilde{\boldsymbol{\theta}}), \mathbf{W}_2(\tilde{\boldsymbol{\theta}}), \ldots$.

As we remarked earlier, the Hessian matrix $\nabla^2\hat{r}_m(\tilde{\boldsymbol{\theta}})$ of equation (14) is difficult to calculate. Therefore, we make one final substitution and use instead the estimated Fisher information matrix $\hat{I}(\tilde{\boldsymbol{\theta}})$, which yields

$$\frac{1}{m}\left[\hat{I}(\tilde{\boldsymbol{\theta}})\right]^{-1}\tilde{V}\left[\hat{I}(\tilde{\boldsymbol{\theta}})\right]^{-1} \tag{17}$$

as our estimated MCMC covariance matrix for $\tilde{\boldsymbol{\theta}}$.

4. ALTERNATING $\boldsymbol{K}$-STARS AND ALTERNATING $\boldsymbol{K}$-TRIANGLES

We illustrate the methods discussed in Sections 2 and 3 by applying them to a class of ERGMs proposed by Snijders et al. (2004). To begin with, we define graph statistics $D_0(y), \ldots, D_{n-1}(y)$, known as the *degree distribution* of y, and $P_0(y), \ldots, P_{n-2}(y)$, known as the *shared partner distribution* of y. The degree distribution statistics are well-known in the networks literature, whereas the shared partner distribution statistics appear to be novel.

For a given i, $1 \le i \le n-1$, $D_i(y)$ is defined to be the number of nodes in y whose degree — the number of edges incident to the node — equals i. For instance, $D_{n-1}(y) = n$ when y is the complete graph and $D_0(y) = n$ when y is the empty graph. Note that $D_0, \ldots, D_{n-1}$ satisfy the linear constraint $D_0 + \cdots + D_{n-1} = n$.

For a given i, $1 \le i \le n-2$, $P_i(y)$ is defined to be the number of dyads (j,k) — where we assume $j < k$ since the graph is assumed undirected — such that j and k are neighbors of each other and they share exactly i neighbors in common. ("Neighbors" are simply nodes connected by an edge.) Unlike the D_i statistics, the P_i statistics do not satisfy a linear constraint; however, note that $P_0 + \cdots + P_{n-2}$ equals the total number of edges in the graph.

Snijders et al. (2004) base some of their ERGMs on graph statistics that may be derived from the D_i and P_i Let $S_k(y)$, $1 \le k \le n-1$, denote the number of k-stars in the graph y. A k-star consists of a node together with a set of k of its neighbors. Like the degree statistics D_i, the k-star statistics are well-known in the networks literature. Since a node with degree i is the center of $\binom{i}{k}$ k-stars,

$$S_k(y) = \sum_{i=1}^{n-1}\binom{i}{k}D_i(y) \quad \text{for } k \ge 2. \tag{18}$$

For $k = 1$, a k-star is simply an edge, and the number of edges is

$$E(y) \stackrel{\text{def}}{=} S_1(y) = \frac{1}{2}\sum_{i=1}^{n-1} i\, D_i(y). \tag{19}$$

In addition to the well-known k-star statistics, Snijders et al. (2004) also introduce a new set of statistics they call k-triangles. They use $T_k(y)$, $1 \leq k \leq n-2$, to denote the number of k-triangles in the graph y. A k-triangle consists of k triangles that share a common edge. Thus, if the endpoints of a particular edge share exactly i neighbors in common, then that edge is the base of exactly $\binom{i}{k}$ k-triangles. The relationship between the k-triangle statistics T_k and the shared partner statistics P_i is very similar to the relationship between the k-star statistics and the degree statistics expressed in equation (18):

$$T_k(y) = \sum_{i=1}^{n-2} \binom{i}{k} P_i(y) \;\text{ for } k \geq 2. \tag{20}$$

For $k = 1$, a k-triangle is simply a triangle, so

$$T_1(y) = \frac{1}{3}\sum_{i=1}^{n-2} i\, P_i(y). \tag{21}$$

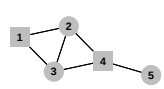

Figure 1: For this undirected, five-node graph, the degree distribution $(D_0, \ldots, D_4)$ is given by $(0, 1, 1, 3, 0)$ and the shared partner distribution $(P_0, \ldots, P_3)$ is given by $(1, 4, 1, 0)$. The edges might represent, say, some social relationship between individuals, and the node shapes might signify some exogenous categorical covariate such as gender.

To make these concepts concrete, consider the simple undirected graph depicted in Figure 1. There are three 3-stars, centered at nodes 2, 3, and 4, and each of these accounts for $\binom{3}{2} = 3$ of the ten 2-stars. There are two 1-triangles (i.e., two triangles), and since these two triangles share an edge there is also one 2-triangle. The degree distribution and the shared partner distribution, given in the caption of Figure 1, may be used to verify equations (18), (19), (20), and (21) along with the fact that $E(y) = \sum_{i=0}^{n-2} P_i(y)$. These relationships may be combined to yield

$$P_0(y) = \frac{1}{2}\sum_{i=1}^{n-1} i\, D_i(y) - \sum_{i=1}^{n-2} P_i(y). \tag{22}$$

Since both D_0 and P_0 may be expressed as linear combinations of the other D_i and P_i statistics, the vector $\mathbf{Z}(y)$ of ERGM (1) based on all degree and shared partner statistics should omit D_0 and P_0:

$$\mathbf{Z}(y) = \left[D_1(y), \ldots, D_{n-1}(y), P_1(y), \ldots, P_{n-2}\right]^t. \tag{23}$$

When $\mathbf{Z}(y)$ of equation (23) is used in model (1) with an unconstrained $\eta \in R^{2n-3}$, the model class is subject to well-known issues of degeneracy (Snijders 2002; Handcock 2002, 2003; Snijders et al. 2004). One type of model degeneracy occurs when the model places most of the probability mass on only a few of the possible graph configurations. The fact that nondegenerate values of η form only a small section of the natural parameter space (Handcock 2003) reduces the value of this model class for describing realistic phenomena. Another problem is the nonexistence of an MLE: Whenever the observed graph statistics fall on the convex hull of the sample space of graph statistics, then the MLE does not exist (Barndorff-Nielsen 1978, Handcock 2003). If the full $\mathbf{Z}(y)$ vector of equation (23) is used, this problem is virtually guaranteed to occur, since typically at least one element of $\mathbf{Z}(y)$ is zero for any realistic network.

To address these problems, we consider constraints on the natural parameter space. In doing so, we hope to limit our attention to subsets of the full parameter space that result in more realistic social network models. Furthermore, the constraints reduce the dimension of the sample space of statistics and make it more probable that an MLE will exist. One way to implement constraints in this case was recommended by Snijders et al. (2004), who introduced an alternating k-star statistic and an alternating k-triangle statistic (in addition, they introduced an alternating independent two-paths statistic that we do not discuss here). In reality, these "statistics" aren't quite statistics because they are based on parameters; however, Snijders et al. (2004) assume that these parameters are fixed and known. In this article, we relax this restriction and estimate these additional parameters.

The alternating k-star and alternating k-triangle "statistics" of Snijders et al. (2004) are defined as

$$u_\lambda(y) = S_2(y) - \frac{S_3(y)}{\lambda} + \cdots + (-1)^n \frac{S_{n-1}(y)}{\lambda^{n-2}}$$

and

$$v_\gamma(y) = 3T_1 - \frac{T_2}{\gamma} + \cdots + (-1)^{n-1} \frac{T_{n-2}}{\gamma^{n-3}},$$

respectively, where λ and γ are additional parameters. Snijders et al. (2004) consider an ERGM that includes statistics E (the number of edges), u_λ, and v_γ:

$$P(Y = y) \propto \exp\{\theta_1 E(y) + \theta_2 u_\lambda(y) + \theta_3 v_\gamma(y)\}. \tag{24}$$

Because we wish both λ and γ to be positive, we reparameterize, letting $\theta_4 = \log\lambda$ and $\theta_5 = \log\gamma$. We may express the canonical parameter $\boldsymbol{\eta}$ of equation (1) in terms of $\theta_1, \ldots, \theta_5$ by replacing S_k and T_k by the expressions in equations (18), (19), (20), and (21): The binomial theorem yields

$$u(y;\theta_4) \stackrel{\text{def}}{=} u_\lambda(y) = e^{2\theta_4} \sum_{i=1}^{n-1} \left\{ \left(1 - e^{-\theta_4}\right)^i - 1 + i e^{-\theta_4} \right\} D_i(y) \tag{25}$$

and

$$v(y;\theta_5) \stackrel{\text{def}}{=} v_\gamma(y) = e^{\theta_5} \sum_{i=1}^{n-2} \left\{ 1 - \left(1 - e^{-\theta_5}\right)^i \right\} P_i(y). \tag{26}$$

Equations (25) and (26) reveal that the coefficients of D_i and P_i are roughly in geometric sequence. For this reason, we refer to θ_4 and θ_5 as the scale parameters of the geometrically weighted degree distribution and geometrically weighted shared partner distribution, respectively. The function $\boldsymbol{\eta}(\boldsymbol{\theta})$ relating the canonical parameter $\boldsymbol{\eta}$ to the parameter $(\theta_1, \ldots, \theta_5)$ of model (24) is required by equations such as (8) and (9); it is summarized by

$$\eta_i = \begin{cases} \theta_1 i + \theta_2 i e^{\theta_4} - \theta_2 e^{2\theta_4} + \theta_2 e^{2\theta_4}(1 - e^{-\theta_4})^i & \text{if } 1 \le i \le n-1; \\ \theta_3 e^{\theta_5}\left[1 - (1 - e^{-\theta_5})^i\right] & \text{if } n \le i \le 2n-3. \end{cases} \tag{27}$$

Model (24) subsumes a number of simpler models. When $\theta_2 = \theta_3 = 0$, the resulting model $P(Y = y) \propto \exp\{\theta_1 E(y)\}$ is the simplistic Bernoulli graph (also known as an Erdős-Rényi graph) in which each edge occurs independently with probability $e^{\theta_1}/(1 + e^{\theta_1})$. When $\theta_3 = \theta_4 = 0$, equation (27) reduces to $\eta_i = i(\theta_1 + \theta_2) - \theta_2$ for $1 \le i \le n-1$, which gives $P(Y = y) \propto \exp\{(\theta_1 + \theta_2)E(y) + \theta_2 D_0(y)\}$. This model contains a "Bernoulli" term and one additional term that governs the propensity for a node to remain unconnected to the rest of the graph. Similarly, when $\theta_2 = \theta_5 = 0$, the model reduces to $P(Y = y) \propto \exp\{(\theta_1 + \theta_3)E(y) - \theta_3 P_0(y)\}$, which contains an additional term that governs how likely neighboring nodes are to resist having any shared neighbors. It is important to note that if $\theta_2 = 0$ (or $\theta_3 = 0$), there is an identifiability problem because in that case the value of θ_4 (or θ_5) is arbitrary. In practical terms, this means that we should not attempt to interpret the value of θ_4 (or θ_5) unless the hypothesis $\theta_2 = 0$ (or $\theta_3 = 0$) can be rejected.

5. LIKELIHOOD RATIO TESTING

Since $2\hat{r}_m[\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}}), \boldsymbol{\eta}(\boldsymbol{\theta}^0)]$ is an estimate of the likelihood ratio statistic $2r[\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}}), \boldsymbol{\eta}(\boldsymbol{\theta}^0)] = 2\ell[\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}})] - 2\ell[\boldsymbol{\eta}(\boldsymbol{\theta}^0)]$ for testing the null hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}^0$, it might seem that likelihood ratio testing is straightforward in this framework. Unfortunately, this is not quite the case: The approximation $2\hat{r}_m[\boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{\eta}(\boldsymbol{\theta}^0)] \approx 2r[\boldsymbol{\eta}(\boldsymbol{\theta}), \boldsymbol{\eta}(\boldsymbol{\theta}^0)]$ becomes worse as $\boldsymbol{\theta}$ gets farther from $\boldsymbol{\theta}^0$. To estimate

$r[\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}}), \boldsymbol{\eta}(\boldsymbol{\theta}^0)]$ accurately necessitates methods to try to lessen the impact of the MCMC error. We do not make any claims here about the distribution of $2r[\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}}), \boldsymbol{\eta}(\boldsymbol{\theta}^0)]$; we concern ourselves in this section only with how best to approximate it using MCMC.

The problem reduces to the problem of estimating the ratio of normalizing constants $c(\tilde{\boldsymbol{\theta}})/c(\boldsymbol{\theta}^0)$, which is a problem that has received quite a bit of attention in the statistics literature in the past decade. Indeed, in presenting some of the history of this problem, Gelman and Meng (1998) point out that it had been studied by physicists before it came to the notice of statisticians, and quite a bit of reinventing the wheel was done by the statistics community. The basic idea of *path sampling* (Gelman and Meng, 1998) is as follows. Define a smooth mapping $\boldsymbol{\theta} : [0, 1] \to R^q$ such that $\boldsymbol{\theta}(0) = \boldsymbol{\theta}^0$ and $\boldsymbol{\theta}(1) = \tilde{\boldsymbol{\theta}}$. Then

$$\mathrm{E}_{\boldsymbol{\theta}(u)} \left\{ \frac{d}{du} \log p[Y|\boldsymbol{\theta}(u)] \right\} = \frac{d}{du} \sum_y p[y|\boldsymbol{\theta}(u)] = 0, \tag{28}$$

where

$$p(y|\boldsymbol{\theta}) \stackrel{\text{def}}{=} \exp\{[\boldsymbol{\eta}(\boldsymbol{\theta})]^t Z(y) - \psi[\boldsymbol{\eta}(\boldsymbol{\theta})]\} \tag{29}$$

is the probability mass function. Combining equations (28) and (29) gives

$$\frac{d}{du}\psi\{\boldsymbol{\eta}[\boldsymbol{\theta}(u)]\} = \mathrm{E}_{\boldsymbol{\theta}(u)} \left\{ \frac{d}{du}\{\boldsymbol{\eta}[\boldsymbol{\theta}(u)]\}^t Z(Y) \right\},$$

which may be integrated to give

$$\psi[\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}})] - \psi[\boldsymbol{\eta}(\boldsymbol{\theta}^0)] = \int_0^1 \mathrm{E}_{\boldsymbol{\theta}(u)} \frac{d}{du}\{\boldsymbol{\eta}[\boldsymbol{\theta}(u)\}^t Z(Y)\, du = \mathrm{E}\, \frac{d}{dU}\{\boldsymbol{\eta}[\boldsymbol{\theta}(U)]\}^t Z(Y). \tag{30}$$

The last expectation in equation (30) is taken with respect to the joint distribution of U and Y, where U is uniform (0,1) and $Y|U$ is distributed according to $\theta(U)$.

Equation (30) suggests that $\ell[\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}})] - \ell[\boldsymbol{\eta}(\boldsymbol{\theta}^0)] = \psi[\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}})] - \psi[\boldsymbol{\eta}(\boldsymbol{\theta}^0)]$ could be estimated by drawing a sample $(U_1, Y_1), \ldots, (U_K, Y_K)$ from the joint distribution of U and Y, then calculating the sample average

$$\frac{1}{K} \sum_{i=1}^{K} [\nabla\boldsymbol{\theta}(U_i)]\{\nabla\boldsymbol{\eta}[\boldsymbol{\theta}(U_i)]\} Z(Y_i),$$

where $\nabla\boldsymbol{\theta}(u)$ is the $1 \times q$ vector of derivatives of $\boldsymbol{\theta}(u)$ with respect to u. We may allow the U_i to be sampled from some density on (0,1), say $q(u)$, other than uniform; each summand in the sample mean above should then be divided by $q(U_i)$. However, the function $q(u)$ may be absorbed into the path map $\boldsymbol{\theta}(u)$, so no generality is lost by assuming that U is uniformly distributed.

On the other hand, it is not hard to generalize the argument leading to equation (30) to allow for the possibility that U has finite support on [0, 1]. In fact, U need not even be random: Suppose

that $0 = u_0 < u_1 < \ldots < u_J = 1$ are given and for each j, $0 \le j \le J$, we draw a random sample $U_{j1}, \ldots, U_{jK_j}$ from the distribution defined by $\boldsymbol{\theta}(u_j)$. The new estimator of $\ell[\boldsymbol{\eta}(\tilde{\boldsymbol{\theta}})] - \ell[\boldsymbol{\eta}(\boldsymbol{\theta}^0)]$ is

$$\sum_{j=1}^{J} \sum_{i=1}^{K_j} \frac{1}{K_j} [\nabla \boldsymbol{\theta}(U_{ji})] \{\nabla \boldsymbol{\eta}[\boldsymbol{\theta}(U_{ji})]\} Z(Y_{ji}). \tag{31}$$

This idea is a simple form of a technique called *bridge sampling* by Meng and Wong (1996). In the implementation of bridge sampling carried out in Section 6, the path between $\boldsymbol{\theta}^0$ and $\tilde{\boldsymbol{\theta}}$ is simply the linear map $\boldsymbol{\theta}(u) = (1-u)\boldsymbol{\theta}^0 + u\tilde{\boldsymbol{\theta}}$.

6. EXAMPLE: COLLABORATION WITHIN A LAW FIRM

As an application of these ideas, we consider the collaborative working relations between 36 partners in a New England law firm. The sociometric relationship is one of many considered by Lazega (2001), Lazega and Pattison (1999) and Snijders et al. (2004) (whom we follow). Specifically, a tie is said to exist between two partners if, and only if, both indicate that they collaborate with the other. As noted in Snijders et al. (2004), the degrees of the nodes range from 0 to 16, with an average of 6.4. The data include covariates collected on each partner. Here we consider seniority (rank number of entry into the firm), gender, office (there were three offices in different cities), and practice (there are two possible values, litigation=0 and corporate law=1).

Our objective is to explain the observed structural pattern of collaborative ties as a function of network statistics, both exogenous and endogenous. The purely endogenous statistics (i.e., those that are true functions of the graph matrix Y) we consider are the number of edges and the alternating k-triangle statistic $v(y;\theta)$ of section 4. We have not included the alternating k-star statistic $u(y;\theta)$, both to simplify the presentation and because our results and those of Snijders et al. (2004) indicate that including that statistic does not appreciably alter the fit of the model.

The statistics involving exogenous data that we consider are all of the form

$$Z(y) = \sum_{1 \le i < j \le n} y_{ij} f(\mathbf{X}_i, \mathbf{X}_j) \tag{32}$$

for some function f of the nodal covariate vectors $\mathbf{X}_i$ and $\mathbf{X}_j$. In expression (32), y_{ij} is the indicator of an edge between nodes i and j, so $f(\mathbf{X}_i, \mathbf{X}_j)$ may be thought of as simply an entry in the change statistic vector $\Delta(\mathbf{Z}(y))_{ij}$ of equation (7). Following Snijders et al. (2004), we first consider the "main effects" of both seniority and practice, for which $f(\mathbf{X}_i, \mathbf{X}_j) = \text{seniority}_i + \text{seniority}_j$ and $f(\mathbf{X}_i, \mathbf{X}_j) = \text{practice}_i + \text{practice}_j$, respectively. We also consider the "similarity effects" of practice, gender, and office. The similarity effect for, say, practice defines $f(\mathbf{X}_i, \mathbf{X}_j)$ to be $I\{\text{practice}_i = \text{practice}_j\}$. Setting $\theta_2 = \theta_4 = 0$ and adding the covariates, model (24) becomes

$$P(Y = y) \propto \exp\{\theta_1 E(y) + \theta_3 v(y;\theta_5) + \boldsymbol{\beta}^T \mathbf{Z}(y)\}, \tag{33}$$

where $\mathbf{Z}(y)$ is the 5-dimensional vector of graph statistics containing the two main effects (seniority and practice) and three similarity effects (practice, gender, and office) described above. Essentially, this model allows us to estimate the effects of the covariates on collaboration while controlling for the network density (as measured by $E(y)$) and a structural transitivity effect (as measured by $v(y;\theta_5)$).

Here we briefly discuss some aspects of implementing the inferential procedures given in Sections 2 and 3. To monitor the statistical properties of the MCMC algorithm, we use the `R` package `coda`. Figure 2 depicts the trace and density plots for a run of sample size 240,000 where only every 1000th step of the Markov chain is sampled (and 50,000 burnin steps were performed). Each row corresponds to a statistic in the model. The values are measured as deviations from the observed value of the statistic. The left column has the trace plots of the sample and the right column has the density plots. Visually the sampler appears to be mixing and the densities are centered about the observed statistics. This visual impression is supported by numerical diagnostics (Raftery and Lewis 1996, Gelman 1996), which indicate that the 240,000 values are more than sufficient. The initial value of $\boldsymbol{\theta}^0$ was the maximum pseudolikelihood estimate. (The pseudolikelihood function is the "likelihood" obtained by considering all edges y_{ij} to be independent, with probabilities given by equation (7); thus, the maximum pseudolikelihood estimate may be obtained by logistic regression.) For the application in this article, only two recalculations of $\boldsymbol{\theta}^0$ as described in Section 3 were necessary.

Table 1 reports the estimates for two models. Model 1 fixes the value of θ_5 at $\log(3) = 1.10$, the value chosen by Snijders et al. (2004). With θ_5 fixed the model is a regular (i.e., non-curved) exponential family. These values replicate those in Snijders et al. (2004), Table 1, Model 2. For compatibility with that paper, we have calculated the estimates conditional on the total number of ties. This conditioning, in which the number of edges is held constant at 115, removes the edges statistic from the model. The unconditional estimates are essentially identical, indicating that the density of collaboration is approximately ancillary to the other statistics.

The $\boldsymbol{\beta}$ coefficients can be interpreted as conditional log-odds ratios (Snijders et al. (2004)). There is also a relative risk interpretation that is often simpler. For example, $\exp(\beta_3)$ is the relative risk of collaboration between two partners from the same practice compared to two partners from different practices with the same values of the other covariates and structural effects. The probabilities involved are conditional on these other covariates and structural effects. The interpretation for non-binary and multiple covariates is similar: $\exp(\beta\Delta)$ is the relative risk of collaboration between two partners compared to two partners with vector of covariates differing by Δ (and with the same values of the structural effects).

The standard errors of Table 1 are obtained from the $\hat{I}$ matrix of equation (12), evaluated at $\tilde{\boldsymbol{\theta}}$ (the MCMC standard errors obtained from equation (17) are much smaller; if they weren't, a

	Model 1		Model 2	
Parameter	est.	s.e.	est.	s.e.
Alternating k-triangles, (θ_3)	0.612	0.091	0.878	0.279
Rate of transitivity (θ_5)	1.099	–	0.814	0.196
Seniority main effect (β_1)	0.024	0.006	0.023	0.006
Practice main effect (β_2)	0.352	0.113	0.390	0.117
Same practice (β_3)	0.708	0.194	0.757	0.194
Same gender (β_4)	0.621	0.257	0.688	0.248
Same office (β_5)	1.151	0.195	1.123	0.194

Table 1: MCMC parameter estimates for the collaboration network. The edge parameter θ_1 has been eliminated from model (33) by conditioning.

larger sample would have been taken). The usual assessments of significance are based on the approximation of the distributions of the t-ratios by standard Gaussian distributions. To assess the accuracy of this approximation, we also applied MCMC p-value tests (Besag and Clifford 1989; Besag 2000). For example, consider evaluating the statistical significance of the main effect of seniority. We use the MCMC procedure to simulate seniority statistics from the model, only allowing steps in the Markov chain that keep all the other statistics fixed, and with $\beta_1 = 0$. This produces a null distribution for the seniority statistic, from which a $p-$value for the observed seniority statistic may be obtained. Using this procedure we were able to validate the Gaussian approximation to the t-ratios. Thus the t-ratios can be used as an informal guide, even though the MCMC p-values are to be preferred for formal testing.

Model 2 fits the curved exponential family model estimating θ_5. The interpretation of the other parameters is similar to Model 1: Collaboration is strongly enhanced by seniority and by working in the same office, and slightly less by having the same practice or gender. Collaboration is also enhanced by practicing corporate law, but at a lower level. The large positive values of θ_3 and θ_5 indicate the presence of complex transitive structure that enhances collaboration beyond the effect that would be expected based on the individual and pairwise partner attributes alone. The scale parameter θ_5 controls the nature of this transitivity: Larger values of θ_5 correspond to increased weight on the higher numbers of shared partners, whereas small positive values correspond to very localized transitive effects (recall the interpretation of the case $\theta_2 = \theta_5 = 0$ following equation (27)).

It is of interest to test if the value of the scaling parameter θ_5 is statistically significantly different from that specified in Snijders et al. (2004). To do this we can conduct likelihood ratio tests using bridge sampling as given in Section 5 with $J = 20$ and $K_j = 200,000$, $j = 0, \ldots, J$, and a sampling interval of 1,000. Table 2 provides the deviance values for a number of models.

The $p-$values in Table 2 are calculated by the MCMC procedure given above; further expla-

Model	Residual Deviance	Deviance	Residual d.f.	$p-$value
NULL	598.78	–	–	–
Covariates only	501.80	96.98	5	0.000
Model 1	457.65	44.15	1	0.000
Model 2	456.21	1.44	1	0.176

Table 2: Deviances for the collaboration network among lawyers.

nation of and justification for this procedure are given by Besag and Clifford (1989) and Besag (2000). The usual χ^2_1 approximation gives a $p-$value for comparing Model 2 to Model 1 of 0.231.

These results indicate that the covariates substantially improve the model fit, as does the inclusion of the transitivity term (Model 1). Allowing the scale of the transitivity to be estimated does not improve the fit significantly from the value specified in Snijders et al. (2004), which is not surprising because that value was chosen by subjective comparison of alternative values. Naturally, however, data-driven estimation of θ_5 is to be preferred unless θ_5 can be pre-set based on theoretical considerations.

7. DISCUSSION

This article gives a fairly comprehensive treatment of maximum likelihood estimation in a particular type of network modeling problem: Beginning from first principles originally set forth by Geyer and Thompson (1992), we discuss estimation and testing based on approximations derived from a Markov chain Monte Carlo scheme. We extend these ideas to curved exponential family models, then discuss particular ERGM specifications due to Snijders et al. (2004) that exploit this extension. Finally, we fit these models to data. Although some of the ideas in this article are about ten years old, the curved exponential family machinery and its application to the particular ERGMs we discuss here are novel.

In our implementation of the Markov chain sampler, we chose to separate our sampled values by a large number of Markov chain iterations, namely 1000. This 1000-step interval is vastly longer than the interval used in several examples mentioned by Geyer and Thompson (1992). The reason we chose such a large separation between sampled values has to do with the tradeoff, mentioned by Geyer and Thompson (1992), between the price paid for more iterations and the price paid for storing and using sampled values. In our implementation, additional iterations are extremely fast. Therefore, we are willing to pay the price (more iterations) for sampled points that are closer to independent than could be expected of points separated by only a few iterations. Additionally, the slow mixing often exhibited by Markov chains of this type makes very long runs (much longer than the sample size we can easily store and use) worthwhile from an exploratory perspective. This

computational tradeoff will vary from application to application.

We have relied in this article on two distinct asymptotic arguments. On one hand, we discussed in depth how the MCMC sample size m contributes to the uncertainty in estimating the true MLE $\hat{\boldsymbol{\theta}}$ by the MCMCMLE $\tilde{\theta}$. On the other hand, we have said relatively little about how the number of nodes n influences the quality of the estimate $\hat{\boldsymbol{\theta}}$, even though we have relied on well-known asymptotic results about the MLE such as the use of Fisher information in approximating its covariance matrix or the implicit assumption that it is approximately normally distributed. However, n is not quite the same as a traditional sample size. What might be given as a sort of "effective sample size" for a graph of size n? Presumably any answer to such a question would have to be model-specific: Note for instance that when edges are independent, the true sample size is $\binom{n}{2}$. There is the further complication that many parameters do not have interpretations that are independent of n; a network might have a totally different MLE from another network twice as large but with qualitatively similar features. Resolving such challenging issues, well beyond the scope of the current article, is of real importance in establishing a cohesive framework of statistical network analysis.

References

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.

Besag, J. (2000), Markov Chain Monte Carlo for Statistical Inference, Working Paper no. 9, Center for Statistics and the Social Sciences, University of Washington. Available from `http://www.csss.washington.edu/Papers/`

Besag, J. and Clifford, P. (1989), Generalized Monte Carlo significance tests, *Biometrika*, **36**, 633–642.

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 Sitga, Spain.

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

Efron, B. (1975), Defining the curvature of a statistical problem (with applications to second order efficiency) (with discussion), *Annals of Statistics*, **3**: 1189–1242.

Efron, B. (1978), The geometry of exponential families, *Annals of Statistics*, **6**: 362–376.

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.

Gelman, A. and X.-L. Meng (1998), Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling, *Statistical Science*, **13**: 163–185.

Geyer, C. J. (1994), On the convergence of Monte Carlo maximum likelihood calculations, *Journal of the Royal Statistical Society, Series B*, **56**: 261–274.

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 Modeling 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.

Lazega, E. (2001), *The Collegial Phenomenon : The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*, Oxford: Oxford University Press.

Lazega, E. and P. E. Pattison (1999), Multiplexity, generalized exchange and cooperation in organizations: a case study. *Social Networks*, **21**: 67–90.

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

Meng, X.-L. and W. H. Wong (1996), Simulating ratios of normalizing constants via a simple identity: A theoretical exploration, *Statistica Sinica*, **6**: 831–860.

Meyn, S. P. and R. L. Tweedie (1993), *Markov Chains and Stochastic Stability*, London: Springer-Verlag.

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.

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.

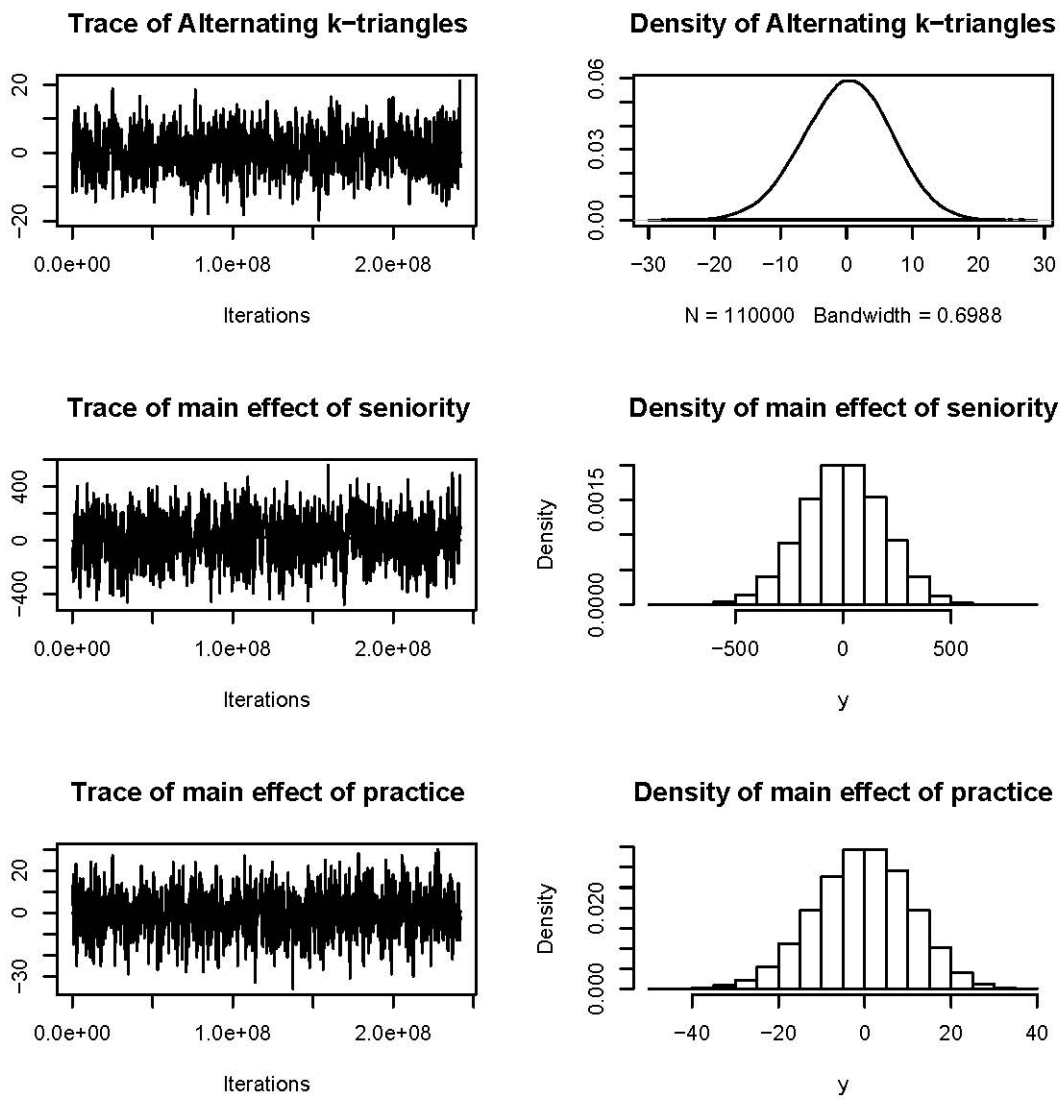

Figure 2: MCMC Diagnostics for the collaboration data. The left-hand side are the trace plots of three statistics, and the right are density estimates (centered on the observed values the actual network).