BAYESIAN ANALYSIS OF EXPONENTIAL RANDOM GRAPHS - ESTIMATION OF PARAMETERS AND MODEL SELECTION

JOHAN KOSKINEN

ABSTRACT. Many probability models for graphs and directed graphs have been proposed and the aim has usually been to reduce the probability of a graph to some function that does not take the entire (graph-) structure into account, e.g. the number of edges (Bernoulli graph), dyadic properties in directed graphs (p_1 Holland and Leinhardt, 1981), subgraph counts (Markov Graphs Frank and Strauss, 1986), etc. Many of these models give you analytically tractable forms for inference about parameters while assuming dependencies that are not always realistic in social science applications, whereas others make up for their increased realism with computational complexity. The Markov Graph of Frank and Strauss (1986), was later developed by Wasserman and Pattison (1996) into the so called p^* model, an exponential model for graphs that comprise arbitrary statistics of graphs and attributes. In this paper we propose a procedure for making Bayesian inference in the exponential graph framework. The aim is to obtain a joint posterior distribution of the parameters in the model, which captures the uncertainty about our parameter values given the observed data. A second objective is to assess how much support different parameterizations of the model are given by data. Typically in Bayesian statistics, the expression for the posterior distribution is not analytically tractable because of the normalizing constant involving a complicated integral or sum. In the case of exponential random graphs we have an additional difficulty, namely that for the p^* model the likelihood is only known up to a constant of proportionality (with respect to data). When the likelihood is easily evaluated, the first problem is easily handled by means of Markov chain Monte Carlo (MCMC) methods. Here, using this fact, the posterior is obtained from a two-step algorithm, which samples from both the sample space and the parameter space. For calculating the marginal likelihood function needed for model comparison, we employ a method suggested by Chib and Jeliazkov (2001). This involves estimating the posterior density evaluated in a suitably chosen point, something which is accomplished using only the key components of the MCMC algorithm, taking averages over the posterior distribution and candidate proposal distribution.

1. INTRODUCTION

Many probability models for graphs and directed graphs have been proposed and the aim has usually been to reduce the probability of a graph to some function that does not take the entire (graph-) structure into account, e.g. the number of edges (Bernoulli graph, Erdös, 1947), dyadic properties in directed graphs (e.g. the p_1 model, Holland and Leinhardt 1981, and its random effects version the p_2 model

Key words and phrases. Bayesian statistical modeling; exponential random graphs; Markov graphs; p-star; normalizing constant family of distributions.

Van Duijn 1995; van Duijn et al. 2004), subgraph counts (Markov Graphs, Frank and Strauss, 1986), etc. (for a comprehensive treatment of random graphs not explicitly adapted for social network analysis see e.g. Janson et al., 2000). Many of these models give you analytically tractable forms for inference about parameters while assuming dependencies that are not always realistic in social science applications, whereas others make up for their increased realism with computational complexity. The Markov Graph of Frank and Strauss (1986), was later developed by Wasserman and Pattison (1996; see also Robins, 1998; and extensions, e.g. Robins et al., 1999) into the so called p^* model, an exponential model for graphs that comprises arbitrary statistics of graphs and attributes.

Since Bayesian analysis like classicist likelihood inference is based upon the likelihood function, both suffers from similar dilemmas in the case of the p^* model. The major obstacle is that the p^* model belongs to the normalizing-constant family of distributions, meaning that the likelihood is only known up to a constant of proportionality. Various simulation techniques have been proposed for finding the maximum likelihood estimates of the parameters for the p^* model and special cases (Frank and Strauss, 1986; Dahmström and Dahmström, 1993; Corander et al., 1998, 2002; Snijders, 2002b; Snijders and van Duijn, 2002) as well as approximations (Strauss and Ikeda, 1990; Frank, 1991; Wasserman and Pattison, 1996). The properties of the approximations are not fully understood and in many situations they compare unfavorably with the maximum likelihood estimators (Besag, 2000; Corander et al., 2002). When it comes to simulation techniques, if we can simulate from a parametric distribution we can often make inference about the parameters with an accuracy that only depends on the amount of computer time we are willing to put down. As pointed out by among others Corander et al. (1998), and Snijders (2002b), sampling from the p^* model is not always straightforward. Recent work (Hancock, 2000; Besag, 2000; Snijders, 2002b) suggests however that difficulties encountered in the estimation procedure are artifacts of certain model deficiencies rather than problems due to estimation strategies. This has led some to model the tie variable as conditionally independent conditional on latent structures. The notion of an underlying social space and/or structural constraints in the context of social networks is thoroughly analysed in (Pattison and Robins, 2002). Much progress in the department of statistical models for investigating these concepts has been made recently. Originating in models for stochastic block models (Fienberg and Wasserman, 1981; Holland et al., 1983), via recent approaches to parameter inference (Snijders and Nowicki, 1997; Nowicki and Snijders, 2001; Tallberg, 2004), the fixed latent blocks have been elaborated to include ultra metrics (Schweinberger and Snijders, 2003) and arbitrary metric spaces (Hoff et al., 2002).

The methods presented in here do not deal specifically with the problems that may occur when simulating exponential random graphs. To fully investigate questions of model deficiency is beyond the scope of this paper. What should count as a model deficiency and what should be regarded merely as a technical issue is perhaps not clear-cut. Here follows some comments in brief on the technical aspects.

When sampling from the model, certain regions of the parameter space may give rise to dominant regimes in the sample space. The existence of regimes is defined in Snijders (2002b) in relation to the sampling scheme. Specifically, regimes are subsets of the state space with high probability mass that are separated from each other by areas with low probability mass making transitions from one regime to the other extremely unlikely. Since the existence of regimes is an artifact of the state space, constructing a different sampling scheme could solve this problem. One solution could be using an over-dispersed Metropolis-coupled Markov chain Monte Carlo strategy. Metropolis-coupled Markov chain Monte Carlo, proposed by Geyer (1991), runs several parallel MCMC chains with different stationary distributions and at each iteration a swap between two of the chains is proposed and accepted with a certain probability. By starting in by some definition extreme states, over dispersing, and letting the chains implement varying degrees of incremental heating (Geyer and Thompson, 1995), the chain converging to the desired distribution is, at least heuristically speaking, relatively freely moving between different regimes.

Another strategy that can be implemented in the Metropolis algorithm (Metropolis et al., 1953), is to randomly select symmetric updating distributions. For example, with a certain probability you change exactly one element of the current adjacency matrix into its complement, and with another probability you change the whole graph into its complement, or you sample a graph from the Gibbs' distribution which has the current graph as its central graph. Snijders (2002b) has suggested a few additional move types.

A problem that some of these and many methods for improving mixing has in common is that they require some degree of fine tuning and in the present context we can not allow for to much of the fine tuning being dependent on the parameter values. In other words, the sampling scheme should work equally well for all parameter values in the range of "likely" values.

Model selection in the class of exponential random graphs has henceforth (as far as we are concerned) been limited to performing t-test for individual parameters. When maximizing the pseudo likelihood (Strauss and Ikeda, 1990) rather than the likelihood function, you may obtain various goodness of fit statistics. The dangers involved in relying on these for model selection were however pointed out by Besag (2000).

Knowledge of how the exponential random graphs work continually accumulate. Recent advances include new specifications limiting the risk of degenerate models (Snijders et al., 2004; "degenerate" is here used very casually, an not in the strict sense of Strauss, 1986, regarding the large sample behaviour). New insights into the performance of exponential random graphs can easily be incorporated into the Bayesian analysis through prior specifications for the model parameters. It is for example know that certain combinations of parameter values are prone to create explosive models (Snijders 2002a; see also Strauss 1986). This should, from a Bayesian perspective, be dealt with by assigning as little prior probability mass as possible to these regions.

The paper is structured as follows. After providing the bare essentials in terms of model specification, we proceed to present the proposed algorithm for exploring the posterior distributions of the parameters. In the latter section, we discuss various issues relating to convergence and convergence control, and present the procedure for calculating the quantities needed for model selection. We close by fitting a p^* model to Krackhardt's (1987) high-tech managers. Two different models are also compared for different prior assumptions.

2. The model

For a fixed set of actors represented by vertices in $V = \{1, \ldots, n\}$, let $\mathbf{x}$ with range space $\mathscr{X}$ be the ordinary adjacency matrix of a (di-) graph G on V, and let

$$\mathbf{z}(\mathbf{x}) = (z_1(\mathbf{x}), \ldots, z_p(\mathbf{x}))'$$

be a collection of functions of $\mathbf{x}$. Introducing a $p \times 1$ vector $\theta = (\theta_1, \ldots, \theta_p)'$ with parameter space $\Theta \subseteq \mathbb{R}^p$, and the so called p^*, or exponential random graph (ERG), model can be expressed as

$$p(\mathbf{x}|\theta) = \exp\{\theta'\mathbf{z}(\mathbf{x}) - \psi(\theta)\}, \tag{2.1}$$

where

$$\psi(\theta) = \log \sum_{\mathbf{u} \in \mathscr{X}} \exp\{\theta'\mathbf{z}(\mathbf{u})\}. \tag{2.2}$$

To distinguish (2.2) from the normalizing constant in the posterior distribution, we call the former (or, rather $e^{\psi(\theta)}$, to be precise) the partition function in accordance with its use in statistical mechanics (Strauss, 1986). The directed Bernoulli(n, p) graph on V, is equivalent to the exponential random graph for

$$p = \frac{e^{\theta_1}}{1 + e^{\theta_1}}, \tag{2.3}$$

and the the degree statistic $z_1(\mathbf{x}) = \sum_{i \neq j} x_{ij}$. The reciprocity model assumes that $\mathbf{z}(\mathbf{x}) = (z_1(\mathbf{x}), z_2(\mathbf{x}))'$, where the first element is the degree statistic and the second the number of mutual dyads $z_2(\mathbf{x}) = \sum_{i \neq j} x_{ij}x_{ji}$. The probability distribution function of $\mathbf{x}$ under the reciprocity model with parameters θ_1 and θ_2 can be written

$$\exp\{\theta_1 z_1(\mathbf{x}) + \theta_2 z_2(\mathbf{x}) - \psi(\theta)\} = e^{\theta_1 z_1(\mathbf{x}) + \theta_2 z_2(\mathbf{x})}(1 + 2e^{\theta_1} + e^{2\theta_1 + \theta_2})^{-n^*},$$

where $n^* = \binom{n}{2}$. The dependence graph on E, the set of all possible (arcs) edges of G, is a graph D with vertex set E, and an edge between two elements $(i, j), (k, \ell) \in E$ if $x_{i,j}$ and $x_{k,\ell}$ are dependent conditional on the rest of $\mathbf{x}$. Any random graph G on V with dependence structure D has probability proportional to

$$\exp \sum_{A \subseteq G} \alpha_A,$$

where α_A is an arbitrary constant if A is a clique of D, and zero otherwise. A graph G is said to be a Markov graph if D only has edges between (i, j) and (k, ℓ), if $\{i, j\} \cap \{k, \ell\} \neq \emptyset$ (Frank and Strauss, 1986; Wasserman and Robins, 2004). Some important progress in generalizing dependence graphs and dependence structures is reported in for example Robins and Pattison (2004).

Typically $\mathbf{z}(\mathbf{x})$ includes functions of observable covariates as well as functions of the graph structure. Exhaustive lists of theoretically motivated models can be found in the literature (Frank and Strauss, 1986; Frank and Nowicki, 1993; Wasserman and Pattison, 1996; Snijders, 2002b; Snijders et al., 2004).

2.1. **Prior distributions.** With prior information about the parameters represented by a distribution with density function π defined on Θ, the posterior distribution of θ given data $\mathbf{x}$ is given by Bayes theorem as

$$\pi(\theta|\,\mathbf{x}) = \frac{\exp\{\theta'\mathbf{z}(\mathbf{x}) - \psi(\theta)\}\,\pi(\theta)}{\int_\Theta \exp\{\theta'\mathbf{z}(\mathbf{x}) - \psi(\theta)\}\,\pi(\theta)\,d\theta}.$$

Now, it is clear that neither the denominator nor the numerator are analytically tractable other than in a few relatively trivial cases.

In the application in the empirical section we use a vague prior (prior proportional to a constant) when obtaining the posterior distribution. For the model selection example we have used independent normal priors. To check that the posterior is proper is, apart from in a few trivial cases (for example $\mathbf{z}(\mathbf{x}) = \mathbf{0}$), in general laborious (or so we conjecture). In the mentioned illustration, we have resorted to the ad-hoc method of assessing whether the sample from the posterior converges. Naturally, practical problems may occur when it is difficult to distinguish potential inappropriateness from general symptoms of bad mixing in the MCMC scheme.

For all proper priors the posterior is also proper, regardless of data. A proper prior, however, should ideally not be used only for convenience, rather the proper prior should reflect our prior knowledge and belief regarding the parameters. In addition, not all values on the parameters make sense since the model is known to be explosive for certain parts of $\mathbb{R}^p$ (Snijders, 2002a). Consequently some subsets of the parameter space should be excluded from the analysis. For most exponential random graphs for example, with $p > 1$ it is usually the case that not all parameters may be strictly positive. A sensibly constructed prior (and a properly subjective one) should reflect this by dependencies between the parameters a priori. One way of getting a working handle on the relation between the parameters and the model is by starting with the parameter corresponding to the degree statistic. The priors for the rest of the parameters are then constructed conditional on the degree parameter, "as if" the (marginal expected) number of edges (arcs) were fixed. A convenient prior distribution for studying the influence of the degree parameter on the density of the graph is a logistic distribution with mean μ and scale parameter τ. The study of the graph density can then be couched in terms of the transformation (2.3). The induced prior on p (given that all other parameters are excluded at this stage) also has a convenient form and it seems easier to think in terms of edge probabilities than about the parameter in its exponential form. That this induced prior should have most of its probability mass to the left of .5 is reasonable to assume in most applied cases, which suggests that μ should be chosen to be negative. To determine more specifically what hyper parameters to be used, different values of μ can be tested and by manipulating the spread with τ, the shape of the distribution of p and its quantiles can be investigated.

We might also mention that a compromise is to have a uniform prior on a bounded subset $\mathbf{\Theta} \subset\mathbb{R}^p$. This could, however, seriously "bias" model selection (in some cases the Bayes factor can be seen as direct function of the volume of the parameter space as is shown in e.g. Lindley 1957; some authors make the dependence on these bounds explicit in the analysis, as does for example Mitchell and Beauchamp, 1988, using a "spike and slab" prior; see Kass and Raftery, 1995, for a general discussion and review on Bayes factors).

3. MCMC sampling scheme

Using a Metropolis sampler (Metropolis et al., 1953), where a proposed move from $\theta^{(t)}$ to θ^*, sampled from a candidate generating distribution $q(\cdot|\theta^{(t)})$, is accepted with probability

$$\alpha\left(\theta_t, \theta^*\right) = \min\left\{1, \frac{\pi\left(\theta^*|\,\mathbf{x}\right)}{\pi\left(\theta^{(t)}|\,\mathbf{x}\right)}\frac{q(\theta^{(t)}|\theta^*)}{q(\theta^*|\theta^{(t)})}\right\}, \tag{3.1}$$

$\left\{\theta^{(t)}\right\}$ converges to a sample from $\pi\left(\theta|\,\mathbf{x}\right)$. Although one important feature of (3.1) is that the normalizing constant in the posterior cancels out, it does however still depend on the partition function ψ. Even though ψ in theory can be simulation consistently estimated with a suitable importance function, the estimator typically has a non negligible variance. Assuming for now that we have a simple form for the kernel of the prior or that we have a vague prior, the aim is to somehow evaluate

$$p\left(\mathbf{x}|\,\theta^*\right)/p\left(\mathbf{x}|\,\theta\right), \tag{3.2}$$

where p is of the form (2.1).

For two fixed parameter vectors θ and θ^*, and a fixed choice of statistics, consider the function

$$f(\mathbf{y}) = \exp\left\{(\theta - \theta^*)'\,\mathbf{z}\left(\mathbf{y}\right)\right\}, \tag{3.3}$$

defined for $\mathbf{y} \in\mathscr{X}$.

Proposition 1. *For a fixed θ^* and θ in a bounded subset of $\mathbb{R}^p$ and f defined as in (3.3)*

$$E\left(f(\mathbf{Y})\right) = \exp\left\{\psi\left(\theta\right) - \psi\left(\theta^*\right)\right\},$$

where expectancy is with respect to the random variable $\mathbf{Y}$ with probability function $p\left(\cdot|\,\theta^*\right)$ and the same choice of statistics as f.

Proof. The proof follows simply by noting that the expectation is the sum of terms

$$f(\mathbf{y})p\left(\mathbf{y}|\,\theta^*\right) = \exp\{\theta'\mathbf{z}\left(\mathbf{y}\right) - \psi\left(\theta^*\right)\} = p\left(\mathbf{y}|\,\theta\right)\exp\{\psi(\theta) - \psi\left(\theta^*\right)\}.$$

□

That the cumulant generating function of the statistics $\mathbf{z}$ has a simple expression has been used by Frank and Strauss (1986), Corander et al. (1998), and Corander et al. (2002) in the context of maximum likelihood estimation. Here we are however interested in the actual ratio of partition functions and not the cumulants. Using the result above, for a fixed θ^*, θ and data $\mathbf{x}$ an estimator of (3.2) is given by

$$\exp\left\{(\theta^* - \theta)' \mathbf{z}(\mathbf{x})\right\} N^{-1} \sum_{k=1}^{N} f(\mathbf{y}_k), \tag{3.4}$$

where $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are generated from $p(\cdot|\,\theta^*)$.

This suggests the following algorithm.

Algorithm 1. *For positive integers T and N*

```
    Step 0: Initialize by setting $\theta_0 := \phi$, where $\phi$ is drawn from distribution $P_0$
    Step 1: if $t < T$, proceed to step 1a, else terminate
    Step 1a: draw $\theta^*$ from $q(\cdot|\theta^{(t)})$
    Step 1b: draw a number $u$ uniformly at random on the interval $(0,1)$
    Step 2: draw a sample $\{\mathbf{y}_k\}_{k=1}^N$ from $p(\cdot|\,\theta^*)$
    Step 3: from $\{\mathbf{y}_k\}$ calculate $p(\mathbf{x}|\,\theta^*)/p(\mathbf{x}|\,\theta)$ using (3.4), if
            $u < \frac{p(\mathbf{x}|\,\theta^*)\,\pi(\theta^*)}{p(\mathbf{x}|\,\theta)\,\pi(\theta^{(t)})} \frac{q(\theta^{(t)}|\theta^*)}{q(\theta^*|\theta^{(t)})}$
      set $\theta^{(t+1)} := \theta^*$, other wise $\theta^{(t+1)} := \theta^{(t)}$
    Step 4: return to Step 1.
```

The distribution P_0 from which the initial values are drawn is a mere formality (and the initial values can be set to some arbitrary vector, or drawn from the prior distribution is this is proper; however see e.g. Fishman, 1996, ch. 6). For step 2 and 3, we need a procedure for drawing samples $\{\mathbf{y}_k\}_{k=1}^N$ from the exponential random graph model. The sample points do not necessarily have to be independent. How to draw a sample $\{\mathbf{y}_k\}_{k=1}^N$ using MCMC is described in for example Strauss (1986), Corander et al. (1998) and Snijders (2002b). When MCMC is used for generating the sample, the function f, inherits the properties of sampled points and the ergodic average (3.4) is a simulation consistent estimator of the ratio of partition functions. It follows from the ergodic theorem (Tierney, 1994) that this quantity converges to its mean almost surely as $N \to \infty$. Now, considering that we need to compute the acceptance ratio not once but in every iteration of the algorithm, at least a few remarks regarding the generation of exponential graphs and of the behaviour of (3.4) are required.

3.1. **Some issues relating to the second step.** As mentioned above, simulating graphs from the exponential graph distribution using MCMC has been suggested in Strauss (1986), Corander et al. (1998) and Snijders (2002b), among others. Whereas the former two relied principally on Metropolis up-dating steps for single edges at a time, the latter also suggested various strategies for improving mixing by combining different up-dating rules. In our experience, a combination of three type of moves yields fully satisfactory convergence properties. The first move type is the Metropolis up-dating step for a single edge. To cater for the need to sometimes propose moves that takes larger jumps in the state space we have also included a Gibbs distribution type up-dating step. Let $\chi \in \mathscr{X}$ be a central (modal) graph in the sense of Banks and Constantine (1998), and define a distribution on $\mathscr{X}$ defined by the probability mass function

$$q(\mathbf{y}|\chi, \tau) = \kappa^{-1} e^{-\tau d(\mathbf{y}, \chi)},$$

for a suitable metric d, partition function κ, and a scaling factor $\tau \geq 0$. To use this in the MCMC procedure for drawing exponential random graphs, let d be the Hamming metric, τ a suitably chosen constant and use the graph from the last iteration as the central graph. For $\mathbf{y}$ following this distribution, the partition function becomes $\kappa = (1 + e^{-\tau})^{-n^*}$, where $n^* = n(n-1)$ for directed graphs and $n^* = \binom{n}{2}$ for un-directed graphs. The proposal distribution reduces to a Bernoulli model where the edge probabilities depend upon whether the edge is present in the modal graph or not. This naturally requires that some thought is put into assessing what values τ are appropriate. Too small a τ will tend to propose graphs that have a very low probability of being accepted, and if τ is too large the proposed moves will not be very different from the ones proposed by the single edge up-dating scheme (and in extreme cases, only the modal graph will be proposed).

For the eventuality that the model has dominant regimes (Snijders, 2002b), a third move type is included in the MCMC scheme. It consists of proposing to move to the complement of the graph in the last iteration.

Given a sequence $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N$ sampled as described, where the sample points from a suitably chosen burn-in period have been discarded, the acceptance probability is estimated with $\hat{\alpha}_3 = (N_1\hat{\alpha}_1 + (N - N_1)\hat{\alpha}_2)/N$, where $\hat{\alpha}_1$ is calculated using (3.4) for the sample points $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{N_1}$, whereas $\hat{\alpha}_2$ is calculated based on $\mathbf{y}_{N_1+1}, \mathbf{y}_{N_1+2}, \ldots, \mathbf{y}_N$, for a $1 \leq N_1 < N$. The sizes of these sub-samples do not necessarily have to be chosen equal, but it is convenient to do so. A measure of the stability of the estimator is thus given by the difference between the estimator based on the first fraction of the sample and the estimator based on the second part of the sample. This carries some of the logic behind the Geweke diagnostics (1992), in that we would expect different parts of a function of an MCMC output to behave similarly if "stationarity" is indeed achieved. The first half and the second half are however correlated to a certain extent. For example, should the proposed parameter values θ^* result in a degenerate model, the sample $\{\mathbf{y}_k\}$ on which the computation of $\hat{\alpha}$ is based could be stuck in a single point. For well behaved parameter values and less extreme cases on the other hand, inspection of the differences $\hat{\alpha}_2 - \hat{\alpha}_1$ gives an indication of how the MCMC sample in the second step is performing.

The variance of (3.3) independent sample points is straightforward to obtain. The expression is however not very useful since it involves the partition function. Additionally, the sample points are not independent in the sample $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N$. As a consequence of this the numerical standard error of (3.4) is not equal to the square root of the variance of (3.3) times $N^{-1/2}$. An idea would otherwise be to set the number of iterations N such that squared error is less than some predetermined level of tolerance with high probability. There are several standard tools in the literature for estimating the numerical standard errors of function of variables drawn using MCMC . These typically involve extensive calculations to be performed on the sample and considerations on a case to case basis, e.g. inspecting the sample autocorrelation functions, running several chains, etc. (see for example Geyer, 1992, and the adjoint discussion). Possibly the most promising approach would be a more sophisticated monitoring of the batched means than that we have proposed, which could flag for unduly high discrepancies. Note also that we have not taking any measures for reducing the variance of the estimator. It would be worthwhile to adapt the variance reducing inversion step procedure suggested in Snijders (2002b), especially as it comes at virtually no extra computational cost.

A related problem is deciding the burn-in time. There are a number if different methodologies and test being developed (see e.g. Cowles and Carlin, 1996) but we daresay, no method is suitable for this particular application. The intuition behind for example the Geweke diagnostics (1992), which is mimicked in the proposed monitoring of differences, is appealing. These diagnostic tools however require an estimate of the standard error. It would be optimal if the required burn-in time was given automatically. The so called perfect sampling (or Propp-Wilson algorithm) of Propp and Wilson (1996) and (Wilson, 2000) is a particularly attractive strategy in this respect since the algorithm itself tells you when it has produced a sample point from the stationary distribution. The exponential random graph model however suffers from the fact that there is no obvious way of ordering the state space in the manner prescribed for defining maximal and minimal elements. Recently though, Corcoran and Tweedie (2002) suggested a way of ordering the state space such that one can sample perfectly using the Metropolis-Hastings algorithm under fairly weak conditions. Their modification however requires that the proposal distribution draws values independently of the previous values (sometimes called the independence sampler). Our preliminary findings for the case of exponential random graphs are that it is in general difficult to construct a proposal distribution that proposes moves that have a high probability of being accepted, and (not unexpectedly) that it is not always straightforward to identify the minimal element.

For determining the appropriate burn-in period, we have largely relied on trial and error. With the type of moves proposed and always initiating the MCMC sample using data $\mathbf{x}$ as the sample point, the length of the burn-in seems to be of a minor importance compared to the overall number of iterations. A heuristic argument for starting in $\mathbf{x}$ is that (provided priors are not highly informative) we expect the posterior to put most of its mass in regions of the parameter space for which the likelihood of generating data is relatively high. If this is the case, most proposed values of θ^* will lie in this region, provided the spread in the proposal distribution is not to large, and the distribution $p(\cdot|\theta^*)$ will put most of its mass in the vicinity of $\mathbf{x}$ (with respect to the move types).

One undesirable aspect of (3.4) is that the importance distribution only is constructed as conditional on a parameter vector. The information in data is hardly used at all, except perhaps as the starting point for the MCMC sample of $\mathbf{y}$. Introducing latent variables that augment the observed data (a method usually attributed to Tanner and Wong, 1987) is a commonly used tool in Bayesian inference that introduces a variable which often has the sample space as its state space but whose distribution incorporates a dependence on both parameters and data. In our experience, designing latent variables for the exponential random graph models often fails in that it is hard to avoid having to deal with the partition function in the full conditional posterior of the parameters. In theory the connection between the stochastic actor-oriented model (SAM) of Snijders (2001) and exponential graph models can be used. More specifically, the exponential random graph model is the limiting distribution of a certain specification of the SAM. If data from the exponential random graph distribution is seen as an observation from the SAM, we can define a latent variable as being an unobserved network that preceded the observed network in time. Because of the time reversibility of the continuous-time Markov chain of the SAM, we can easily simulate "backwards in time". Conditional on a realization of the latent variable and data, inference can be carried out in a manner described in Koskinen (2004). This conditioning is however not legitimate since the marginal distribution of the latent variable has to be given explicitly, the form of which would be the exponential random graph distribution unless the backward process is run for long enough so as to lose its memory.

3.2. **Model selection.** Let $\mathscr{M}$ be a collection of models reflecting different hypothesis regarding the data generation process. Each model $M \in \mathscr{M}$ is characterised by a sampling probability mass function $p(\mathbf{x}|\theta, M)$, and a model specific set of parameters $\theta \in \boldsymbol{\Theta}_M \subseteq \mathbb{R}^{p_M}$, where p_M may vary over models. Here we consider models with sampling probability mass function of the form (2.1). To obtain the marginal likelihood of a model M,

$$m\left(\mathbf{x}\right| M) = \int_{\boldsymbol{\Theta}_M} p\left(\mathbf{x}\right| M, \theta) \, \pi\left(\theta\right| M) \, d\theta,$$

we can use the basic marginal likelihood identity

$$m\left(\mathbf{x}\right| M) = \frac{p\left(\mathbf{x}\right| M, \theta) \, \pi\left(\theta\right| M)}{\pi\left(\theta\right| \mathbf{x}, M)}, \tag{3.5}$$

obtained from Bayes theorem by simply solving for the marginal likelihood. Following Chib and Jeliazkov (2001), the posterior ordinate $\pi\left(\theta^*\right| \mathbf{x}, M)$ for an arbitrary point θ^* is estimated by

$$\widehat{\pi}\left(\theta^*\right| \mathbf{x}, M) = \frac{T^{-1} \sum_{t=1}^{T} \alpha\left(\theta^{(t)}, \theta^*\right) q\left(\theta^{(t)} | \theta^*\right)}{J^{-1} \sum_{j=1}^{J} \alpha\left(\theta^*, \theta^{*(j)}\right)} \tag{3.6}$$

where $\left\{\theta^{(t)}\right\}$ are sampled draws from the posterior distribution and $\{\theta^{*(j)}\}$ are draws from the proposal distribution $q\left(\cdot|\theta^*\right)$. The problem is that we in general do not have the numerator in (3.5) since the likelihood function includes the partition function.

To be able to evaluate the likelihood in θ^*, note that we can often find θ^* such that the likelihood simplifies to a known distribution setting certain coordinates in θ^* to 0. More specifically, if either the degree statistic or the reciprocity statistic, or both, are included in the model setting all coordinates not corresponding to these statistics to zero will yield a known likelihood. Should this not be, although the degree statistic should always be included, one can resort to the naive option $\theta^* = \mathbf{0}$, giving the likelihood $p\left(x \mid \theta^*\right) = \left|\mathscr{X}\right|^{-1}$, which does not depend on data and hence the numerator in (3.5) for θ^* is given by $\left|\mathscr{X}\right|^{-1} \pi\left(\theta^* \mid M\right)$. When comparing ratios of marginal likelihoods for nested models, i.e. Bayes factors, θ^* can be chosen such that the likelihood function does not have to be evaluated. Take for example the case when two model specifications only differ in that one includes a parameter γ, whereas the other does not. For the smaller model the parameter vector can for example be $\theta = (\beta_1, \ldots, \beta_p)'$, and for the larger model by $\widetilde{\theta} = (\beta_1, \ldots, \beta_p, \gamma)'$, where the first p parameters correspond to the same statistics for both models. If the marginal likelihood functions are estimated using the form of (3.5), for θ^*, and $\widetilde{\theta}^* = (\theta^{*\prime}, 0)'$, the likelihood functions vanishes in the ratio.

The acceptance probability in the numerator of (3.6) can be consistently estimated using (3.4), possibly with a few more sample points from the exponential random graph distribution than in the main algorithm. For the denominator in (3.6), (3.4) is calculated for each $\{\theta^{*(j)}\}$. Although θ^* can be chosen arbitrarily, it is preferable, in addition to the requirement that the likelihood function can be evaluated (or nested models be treated in the way described above), that θ^* belongs to a region with relatively hight posterior probability mass. The reason being that the estimate based on (3.6) becomes increasingly unstable with, say, the distance from the posterior mode (Chib and Jeliazkov, 2001).

4. Example: Krackhardt's high-tech managers

A model with the effects choice, mutuality, transitivity, cyclicity, 2-in-stars, 2-out-stars, and 2-mixed-stars is fitted to the Krackhardt's (1987) high-tech manager friendship data, with $n = 21$ actors. A sample $\{\theta^{(t)}\}$ of $T = 10,000$ observations is obtained from the two-step procedure using a multivariate normal proposal centered over the previous iteration value. A provisional variance covariance matrix for the proposal distribution is obtained from a test run. For Figures 1 through 3 , the total length of the MCMC sample in the second step was $N = 10,000$, for which the first 4000 iterations were discarded as burn-in. To study the stability of the estimator (3.4), the samples in the second step of the algorithm were divided in two equal parts of 3,000 iterations each, i.e. $N_1 = 7000$, and the differences between the estimator of the partition function (3.4) for both halves of the sample are shown in Figures 1 and 2. Judging by Figure 1 there seems to be no particular relation between the differences and the distance between the parameter vectors for which the ratio of partition functions is estimated. Conclusions such as these should however be complemented by conditioning on the magnitudes of the acceptance ratios, which is done in Figure 2. We note that the tight concentration around zero in Figure 1 in part is due to the fact that many proposed moves have a very low probability of being accepted (top panel of Figure 2). Had this pattern been more extreme it could have indicated a somewhat low acceptance rate, due

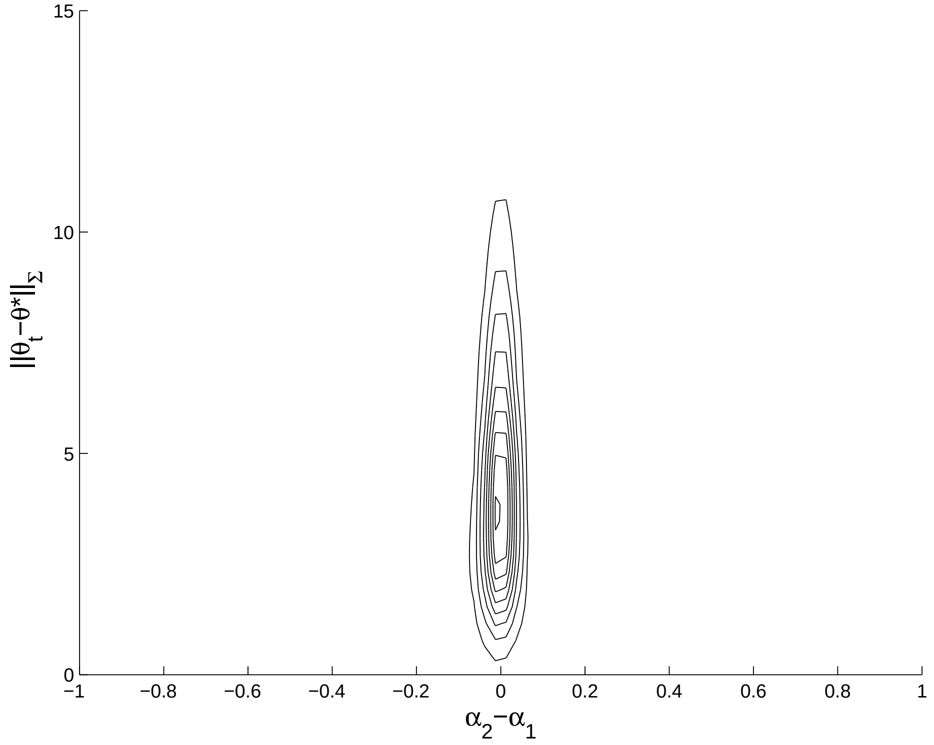

FIGURE 1. Contour plot of estimated density of difference $\hat{\alpha}_2 - \hat{\alpha}_1$ against jump-distance in covariance norm conditional on $\max(\hat{\alpha}_2, \hat{\alpha}_1) < 1$

to too large variance in the candidate generating distribution, and thereby bad mixing.

The marginal posterior distributions are given in Figure 3 with summaries in Table 1 (the figures for the MLE:s and standard errors are reproduced from Snijders, 2002b). The credibility intervals in Figure 3 are conclusive except for β_7, corresponding to 2-mixed stars. What is striking in Table 1 is how well the maximum likelihood analysis approximates the posteriors, not only in terms of the point estimates (something which is expected with vague priors and symmetric distributions), but also in terms of the correspondence between the (approximate) standard errors of the maximum likelihood estimates and the posterior standard deviations of parameters.

As an example of model selection, consider testing a model with all effects in Table 1 included against a model where the effect of cycles is omitted. By evaluating the RHS of the basic marginal likelihood identity in a vector β^* with $\beta_4^* = 0$, there is no need to evaluate the likelihood function since this vanishes in the expression for the Bayes factor for the reduced model against the full model (as noted above).

In this example we choose normal shrinkage priors, i.e. priors of the form $N_p(\mathbf{0}, \lambda\mathbf{I})$, for convenience. There are various arguments for using these type of

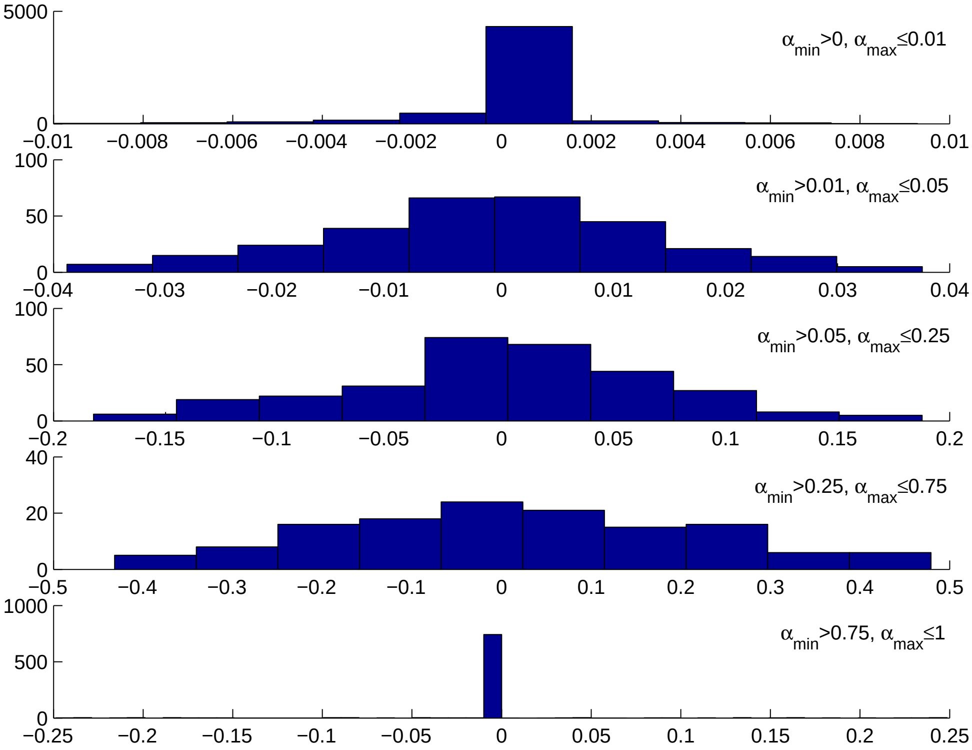

FIGURE 2. The difference $\hat{\alpha}_2 - \hat{\alpha}_1$, where $\alpha_{\min} = \min(\hat{\alpha}_2, \hat{\alpha}_1)$, and $\alpha_{\max} = \max(\hat{\alpha}_2, \hat{\alpha}_1)$

		MLE		Bayes	
statistic	par.	estimate	s.e.	mean	std
number of ties	β_1	−2.066	.656	−2.296	.666
mutuality	β_2	2.035	.437	1.976	.630
transitivity	β_3	.070	.087	.044	.120
cyclicity	β_4	−.004	.225	−.201	.266
2-in stars	β_5	−.025	.110	−.049	.143
2-out stars	β_6	.219	.049	.231	.077
2-mixed stars	β_7	−.104	.0666	−.096	.0612

TABLE 1. Maximum likelihood estimates and posterior means and standard deviations of parameters for Markov graph model fitted to Krackhardt's high-tech managers

priors. Raftery (1996) investigated the Bayes factors as a function of the prior spread in the context of generalised linear models. Using a multivariate normal prior distributions of a similar form to ones used here. Whereas their aim (partly) was to find the scale factor that minimized the influence of the prior distribution on the Bayes factor, we merely mean to illustrate the potential of Bayesian model selection. We prefer also this fairly "neutral" prior to their data-dependent prior

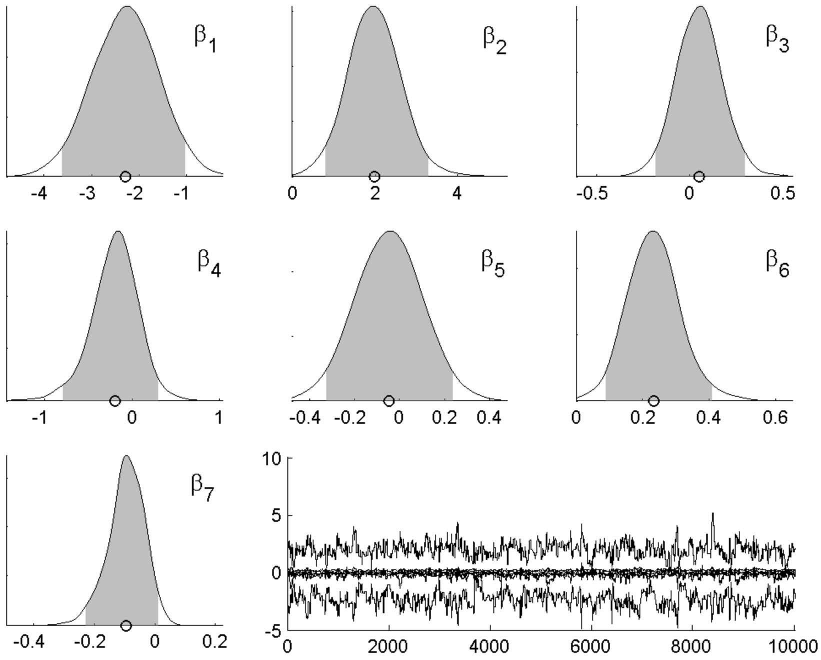

Figure 3. Marginal posterior distributions (panels 1 through 7) with 95% Credibility intervals, and a trace plot for all parameters (panel 8)

for reporting the results. Although it is sensible in a way to adapt the scale to data, it tends to obscure the progression from prior to posterior distribution.

Bayes factors are known to be sensitive to the a priori spread of the parameters. In particular, a sharp null-hypothesis, for example stating that a parameter is 0, is favoured by large spread in the prior distribution. This is sometime referred to as ”Bartlett's paradox” (see Kass and Raftery, 1995). For the choices $\lambda = .1, 1, 5, 10, 20$ in the prior distributions of the parameters the Bayes factors of the full against the reduced model are given in Figure 4. The interpretation of the Bayes factor for a given λ can be expressed in terms of posterior probabilities. If we, for example, considered both models equally probable a priori, and our prior belief regarding the parameters was modeled by independent $N(0, \lambda = .1)$ distributions, the posterior probability of the full model given observed data $\mathbf{x}$ would have been almost .5 judging by Figure 4. That is to say, with two models, M_1 and M_0, the Bayes factor B_{10} for M_1 against M_0 is related to the posterior probability through

$$\Pr(M_0|\mathbf{x}) = \frac{1}{1 + B_{10}}.$$

Hence with extremely strong belief that all of the parameters were 0, the two models would have been equally probable a posteriori. This would have been

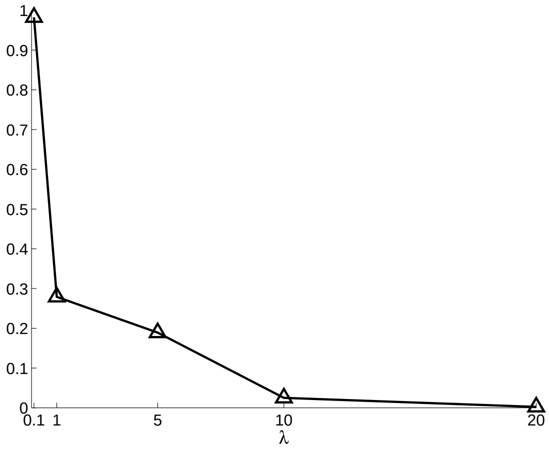

FIGURE 4. Bayes factors for including cyclicity against excluding cyclicity for different shrinkage factors $\lambda = .1, 1, 5, 10, 20$

something of a truism (noting that the two models are equivalent when all parameters are nearly nought) had we known beforehand what extremely strong belief was. The interpretation of Figure 4 can be said to be that regardless of how we modeled "no" or "little" information a priori regarding the parameter values, we can conclude that we do not believe in including cyclicity in the random graph model (this of course depends on what measuring stick we apply and whether we are willing to give the reduced model, with out cyclicity, the status of a null hypothesis, etc. Kass and Raftery, 1995). Considering the marginal credibility intervals in Figure 3, we would not expect the inclusion of cyclicity to be supported by data.

We emphasize the relative simplicity of the algorithm for calculating the marginal likelihood. The computing time as compared to the computing time for obtaining the posteriors is small, since the only additional sampling to be made is drawing random vectors from the proposal distribution. The last step, for the denominator of (3.6), involves estimating the acceptance probabilities, which is the time consuming part. However since the average over the posteriors is the main source of variation in estimating the posterior ordinate, we do not need as many draws from the proposal as from the posterior. For the examples reported above, the lengths of samples from the posteriors were between 10,000 and 20,000, which proved to provide satisfactory accuracy (Chib and Jeliazkov, 2001, give suggestions as to how to calculate the numerical standard errors when the numerator and denominator of (3.6) are based on an equal number of sample points).

Now, we might be interested in knowing how well the reduced model does in comparison with the reciprocity model. Assume that we gave all three of the above models (i.e. including the reciprocity model) equal prior probabilities and modeled our uncertainty regarding the parameters with independent normal priors with variance $\lambda = 10$, centered over the origin. Calculations are performed as above with the exception that the second step in the algorithm can be skipped since the partition function of the reciprocity model is known. The posterior probability of the full model being the right one given data would be 0.0016. The posterior probability of the reduced model would be 0.0629 and hence the posterior probability of the reciprocity model would be 0.9356.

References

Banks, D. and Constantine, G. M. (1998). "Metric models for random graphs." *J. Classification*, 15, 2, 199–223.

Besag, J. (2000). "Markov chain Monte Carlo for statistical inference." working paper 9, University of Washington, Center for Statistics and the Social Sciences.

Chib, S. and Jeliazkov, I. (2001). "Marginal likelihood from the Metropolis-Hastings output." *J. Amer. Statist. Assoc.*, 96, 453, 270–281.

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

— (2002). "Maximum likelihood estimation for exponential random graph model." In *Contributions to Social Network Analysis, Information Theory, and Other Topics in Statistics; A Festschrift in honour of Ove Frank*, ed. J. Hagberg, 1–17. Stockholm: Dept. of Statistics, Stockholm University.

Corcoran, J. N. and Tweedie, R. L. (2002). "Perfect sampling from independent Metropolis-Hastings chains." *J. Statist. Plann. Inference*, 104, 2, 297–314.

Cowles, M. K. and Carlin, B. P. (1996). "Markov chain Monte Carlo convergence diagnostics: a comparative review." *J. Amer. Statist. Assoc.*, 91, 434, 883–904.

Dahmström, K. and Dahmström, P. (1993). "ML-estimation of the clustering parameter in a Markov Graph model." Research report 1993:4, Department of Statistics, Stockholm University.

Erdös, P. (1947). "Some remarks on the theory of graphs." *Bull. Amer. Math. Soc.*, 53, 292–294.

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

Fishman, G. S. (1996). *Monte Carlo – Concepts, Algorithms, and Applications*. New York: Springer–Verlag. Corrected third printing, 1999.

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

Frank, O. and Nowicki, K. (1993). "Exploratory statistical analysis of networks." In *Quo vadis, graph theory?*, vol. 55 of *Ann. Discrete Math.*, 349–365. Amsterdam: North-Holland.

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

Geweke, J. (1992). "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments." In *Bayesian statistics, 4 (Peñíscola, 1991)*, 169–193. New York: Oxford Univ. Press.

Geyer, C. J. (1991). "Markov chain Monte Carlo maximum likelihood." In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, ed. E. M. Keramidas, 156–163. Fairfax Station: Interface Foundation.

— (1992). "Practical Markov chain Monte Carlo." *Statistical Science*, 7, 473–483. With discussion.

Geyer, C. J. and Thompson, E. A. (1995). "Annealing Markov chain Monte Carlo with applications to ancestral inference." *Journal of the American Statistical Association*, 90, 431, 909–920.

Hancock, M. (2000). "Progress in statistical modeling of drug user and sexual networks." Unpublished manuscript, University of Washington, Center for Statistics and the Social Sciences.

Hoff, P., Raftery, A. E., and Handcock, M. (2002). "Latent space approaches to social network analysis." *Journal of the American Statistical Association*, 97, 1090–1098.

Holland, P., Laskey, K. B., and Leinhardt, S. (1983). "Stochastic block-models: Some first steps." *Social Networks*, 5, 109–137.

Holland, P. and Leinhardt, S. (1981). "An exponential family of probability distributions for directed graphs (with discussion)." *Journal of the American Statistical Association*, 76, 33–65.

Janson, S., Luczak, T., and Rucinsnki, A. (2000). *Random graphs*. New York: Wiley.

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90, 430, 773–795.

Koskinen, J. (2004). "Bayesian inference for longitudinal social network data." Research Report 2004:4, Department of Statistics, Stockholm University.

Krackhardt, D. (1987). "Cognitive social structures." *Social Networks*, 9, 109–134.

Lindley, D. V. (1957). "A statistical paradox." *Biometrika*, 44, 187–192.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equations of state calculations by fast computing machine." *Journal of Chemical Physics*, 21, 1087–1091.

Mitchell, T. J. and Beauchamp, J. J. (1988). *Journal of the American Statistical Association*, 83, 1023–1032.

Nowicki, K. and Snijders, T. A. B. (2001). "Estimation and prediction for stochastic blockstructures." *Journal of the American Statistical Association*, 96, 1077–1087.

Pattison, P. and Robins, G. (2002). "Neighbourhood-based models for social networks." *Sociological Methodology*, 32, 301–337.

Propp, J. G. and Wilson, D. B. (1996). "Exact sampling with coupled Markov chains and applications to statistical mechanics." In *Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)*, vol. 9, 223–252.

Raftery, A. E. (1996). "Approximate Bayes factors and accounting for model uncertainty in generalised Linear models." *Biometrika*, 83, 2, 251–266.

Robins, G. (1998). "Personal Attributes in Interpersonal Contexts: Statistical Models for Individual Characteristics and Social Relationships." Ph.d. dissertation, University of Melbourne, Department of Psychology.

Robins, G. and Pattison, P. (2004). "Interdependencies and social processes: dependence graphs and generalized dependence structures." Forthcoming in Models and Methods in Social Network Analysis, P. J. Carrington and J. Scott and S. Wasserman (eds), New York: Cambridge University Press.

Robins, G., Pattison, P., and Wasserman, S. (1999). "Logit models and logistic regression for social networks: III. Values relations." *Psychometrika*, 64, 371–394.

Schweinberger, M. and Snijders, T. A. B. (2003). "Settings in social networks: a measurement model." *Sociological Methodology*, 33, 307–341.

Snijders, T. A. B. (2001). "The statistical evaluation of social network dynamics." *Sociological Methodology*, 30, 361–395.

— (2002a). Personal communication.

— (2002b). "Markov chain Monte Carlo estimation of exponential random graph models." *Journal of Social Structure*, 3, 2.

Snijders, T. A. B. and Nowicki, K. (1997). "Estimation and prediction for stochastic blockmodels for graphs with latent block structure." *Journal of Classification*, 14, 75–100.

Snijders, T. A. B., Pattison, P. E., and Robins, G. (2004). "New specifications for exponential random graph models." Manuscript in preparation.

Snijders, T. A. B. and van Duijn, M. A. J. (2002). "Conditional maximum likelihood estimation under various specifications of exponential random graph models." In *Contributions to Social Network Analysis, Information Theory, and Other Topics in Statistics; A Festschrift in honour of Ove Frank*, ed. J. Hagberg, 117–134. Stockholm: Dept. of Statistics, Stockholm University.

Strauss, D. (1986). "On a general class of models for interaction." *SIAM Rev.*, 28, 4, 513–527.

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

Tallberg, C. (2004). "A Bayesian approach to modeling stochastic blockstructures with covariates." To appear in Journal of Mathematical Sociology.

Tanner, M. A. and Wong, W. H. (1987). "The calculation of posterior distributions by data augmentation (with discussion)." *Journal of the American Statistical Association*, 82, 528–550.

Tierney, L. (1994). "Markov chains for exploring posterior distributions." *Ann. Statist.*, 22, 4, 1701–1762. With discussion and a rejoinder by the author.

Van Duijn, M. A. J. (1995). "Estimation of a random effect model for directed graphs." In *Symposium Statistische Software 1995*.

van Duijn, M. A. J., Snijders, T. A. B., and Zijlstra, B. H. (2004). "p2: a random effects model with covariates for directed graphs." *Statistica Neerlandica*. In press.

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

Wasserman, S. and Robins, G. (2004). "An introduction to random graphs, dependence graphs, and p^*." Forthcoming in Models and Methods in Social Network

Analysis, P. J. Carrington and J. Scott and S. Wasserman (eds), New York: Cambridge University Press.

Wilson, D. B. (2000). "How to couple from the past using a read-once source of randomness." *Random Structures & Algorithms*, 16, 1, 85–113.

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden
E-mail address: johan.koskinen@stat.su.se