

# Computational Statistical Methods for Social Network Models

David R. Hunter, Pavel N. Krivitsky, and Michael Schweinberger\*

September 11, 2012

## Abstract

This article reviews the broad range of recent statistical work in social network models, with particular emphasis on computational aspects of these methods.

*Keywords:* degeneracy, ERGM, latent variables, MCMC MLE, variational methods

## 1 Introduction

A typical statistical data frame includes sampling units, which may be considered individuals, and analysis often focuses on some property of these units. Loosely speaking, social networks arise whenever the “property” of interest involves interactions between multiple sampling units, rather than the units themselves. We do not limit ourselves to the case in which the sampling units are actually human beings, though this is by far the most common application that has appeared in the literature on social network models.

There is a long history of work that may be characterized as related to social networks—as Carrington and Scott (2011) point out, it is difficult to pinpoint the genesis of this field but its roots may be traced at least as far back as the 1930s—though we do not focus on this development here, both because there already exist numerous treatises on networks in general and social networks in particular, and because for the audience of *JCGS*, we wish to focus on computational questions. However, we can at least give a partial list of survey-type references for readers interested in delving into the subject of social networks more deeply: Though almost two decades old, the

---

\*All three authors contributed equally to this article.

classic book by Wasserman and Faust (1994) is still considered a comprehensive introduction to the important quantitative concepts of social network analysis. As to statistical analysis for social networks, more recent works include the survey article by Goldenberg et al. (2009) and the book-length treatment of various network-related statistical topics by Kolaczyk (2009), both of which give numerous references. Finally, we recommend the other network-related articles in this issue.

The current article highlights some current topics in social network modeling, with special emphasis on computational aspects. In Section 2, we introduce a classic dataset, which, though extremely small especially by modern standards, serves to illustrate some of these computational techniques even if it does not demonstrate the state-of-the-art in computational techniques designed for massive social networks. To keep the article to a manageable length, in Section 3 we merely highlight many important topics illustrating the range of statistical work on social network applications, citing recent references to enable interested readers to learn more. The topics that we describe in more detail in Sections 4 and 5 share a common feature: they focus on complete networks that are cross-sectional, which means that they observed at only one point in time. The distinction between the techniques in Section 4 and those in Section 5 is that the former covers network models whose dependence structure does not have a clear hierarchy, typically expressed as a joint distribution of edge variables, and often focused on modeling global network features and social forces; whereas the latter covers network models that are hierarchical in nature, where the edge variable distributions are parameterized in terms of latent variables, and are focused on identifying individual nodes' roles and positions. We conclude in Section 6 with a discussion of some future challenges. Our hope throughout is to stimulate interest among the readership of *JCGS* to delve into the rapidly expanding field of statistical modeling of social networks.

## 2 Data and notation

We introduce some notation used throughout the article and then discuss a classic social network data set, which we use as an illustrative example throughout the paper.

## 2.1 Notation

Let  $N$  be the set of nodes in the network of interest, indexed  $\{1, \dots, n\}$ . The relationships in the network may be directed (e.g., friendship nominations, messages) or undirected (e.g., sexual partnerships, conversations). In the former case, we define the set of *dyads* (here used to refer to potential relationships)  $\mathbb{Y}$  to be a subset of  $N \times N$ , the set of ordered pairs of nodes; in the latter case, it is a subset of  $\{\{i, j\} : (i, j) \in N \times N\}$ , the unordered pairs of nodes. (We will also use  $u(\mathbb{Y})$  to refer to an “unordered” version of  $\mathbb{Y}$ , i.e.,  $u(\mathbb{Y}) \equiv \{\{i, j\} : (i, j) \in \mathbb{Y}\}$ .) Usually,  $\mathbb{Y}$  is further constrained in that in most social networks studied, a node cannot have a relationship of interest with itself, excluding pairs of the form  $(i, i)$ .

For binary networks, in which the relationship of interest must be either present or absent, we use  $\mathcal{Y} \subseteq 2^{\mathbb{Y}}$ , the set of subsets of  $\mathbb{Y}$ , to refer to the set of possible networks of interest, which may be further constrained (that is,  $\mathcal{Y}$  may be a proper subset of  $2^{\mathbb{Y}}$ ). We will use  $\mathbf{Y}$  to refer to network random variables and  $\mathbf{y} \in \mathcal{Y}$  to refer to their realizations, and  $\mathbf{y}_{i,j}$  be an 0–1 indicator of whether a relationship of interest is present between  $i$  and  $j$  in a binary network context.

## 2.2 Data

The data set collected by Sampson (1968) and described by Batagelj and Mrvar (2003) is a classic data set in social network analysis. The data set summarizes relationships, observed at three distinct time points, among 18 monks who were about to enter a monastery when a conflict erupted. We use here the directed network where  $\mathbf{y}_{i,j} = 1$  denotes that monk  $i$  liked monk  $j$  at any of the three time points and  $\mathbf{y}_{i,j} = 0$  otherwise. The directed network is shown in Figure 1, where circles represent monks and directed edges are oriented from  $i$  to  $j$  whenever  $\mathbf{y}_{i,j} = 1$ . The monks are divided by Sampson into three groups: Loyal Opposition, Turks, and Outcasts.

## 3 Range of social network models

The range of statistical modeling techniques for social networks is too broad to address adequately in a single article. We have chosen to focus in some depth in Sections 4 and 5 on cross-sectional

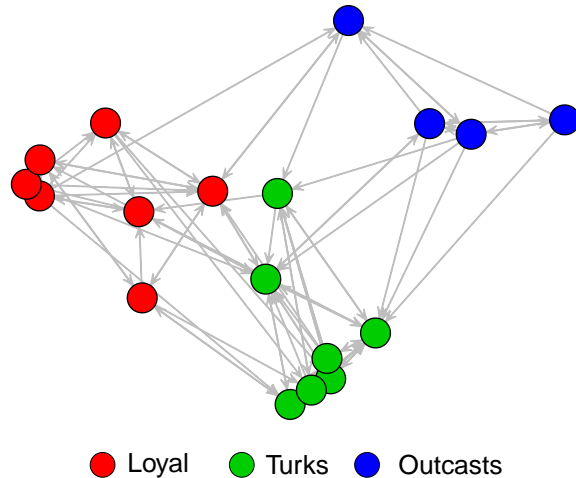


Figure 1: Monk social network dataset of Sampson (1968), where circles represent monks and directed edges represent liking relationships.

(i.e., observed once only), completely observed networks. The current section, by contrast, seeks to illuminate the myriad other applications by describing several additional important recent trends in social network modeling for which we do not have space for a lengthier exposition.

### 3.1 Dynamic Markovian models of networks

The modeling of social network dynamics—i.e., changes over time—has attracted much attention, starting with the groundbreaking work of Holland and Leinhardt (1977a,b) and Wasserman (1977, 1979, 1980). Holland and Leinhardt argued that continuous-time Markov processes with state space  $\mathcal{Y}$  are natural models of social network dynamics. Wasserman’s work, later sharpened by Leenders (1995), studied maximum likelihood estimation for continuous-time Markov process models based on the assumptions that dyad processes are independent and stationary.

This early work was expanded by Snijders (2001), who introduced parameterizations of continuous-time Markov processes that allow dyad processes to be dependent and relaxed the restrictive stationarity assumption provided two or more discrete-time observations of the process are available. Motivated by the work of McFadden (1974) on random utility models, Snijders’ parameterizations have the advantage that the Markov process can be interpreted as “actor-driven”, i.e., driven

by nodes who maximize random utility functions of the network. More importantly from the standpoint of statistical computing, Snijders (2001) also adapted the method of simulated moments (McFadden, 1989) to method of moments estimation of continuous-time Markov models, implemented by stochastic approximation (Robbins and Monro, 1951; Pflug, 1996). Some computational improvements were discussed by Schweinberger and Snijders (2007), and maximum likelihood and Bayesian estimation were proposed by Snijders et al. (2010) and Koskinen and Snijders (2007), respectively. These computational methods are based on non-standard (Markov chain) Monte Carlo data-augmentation methods and are implemented in the Windows-based program *Siena* (Snijders et al., 2010a) and the platform-independent R package *RSiena* (Ripley and Snijders, 2011).

More recently, discrete-time Markov models have been explored as alternatives to continuous-time Markov models. Hanneke et al. (2010) explored discrete-time Markov models in which the transition probabilities are expressed by ERGMs. Krivitsky and Handcock (2012a) proposed separable parameterizations of discrete-time Markov models, where one process governs the addition of edges and the other process governs the deletion of edges at each time step; their estimation methods, which extend the (Markov chain) Monte Carlo maximum likelihood methods of Geyer and Thompson (1992), are implemented in the R package *ergm* package (Handcock et al., 2012).

### **3.2 Dynamic non-Markovian models of networks**

Several non-Markovian models for changing networks, including loglinear models and models in which changes are selected uniformly conditional on the degree structure of the network, are discussed briefly by Frank (1991), who cites multiple references. A more recent method is the random-effects model suggested by Westveld and Hoff (2011), which may be adapted to network data that are either binary or in which edges between nodes have normally distributed weights.

One recent trend in the non-Markovian vein in which computation plays an increasingly important role is the application of survival analysis to continuously collected network data (Butts, 2008; Brandes et al., 2009), an increasingly common paradigm with internet-based and other computer-generated network datasets. If an “event” is the formation of a new edge, then we

attach a counting process either to every node or to every pair of nodes. Vu et al. (2011) refer to these as the “egocentric” and “relational” models, respectively. The analysis of large datasets (with thousands of nodes and tens of thousands of edges) based on these counting processes is demonstrated in the egocentric case by Vu et al. (2011) and in the relational case by Vu et al. (2011) and Perry and Wolfe (2011).

### **3.3 Joint models of networks and other outcome variables**

There is a growing literature in which statistical models for the edges in a network are only one part of a larger joint model. Computing plays a huge role in this area, as simulations of these joint models are vital; yet statistical inference is a relatively recent addition to the computational mix. Snijders et al. (2007), for instance, model jointly a dynamically evolving network along with behavior measured on the nodes in that network; parameter estimation for such models given longitudinally observed data is implemented in SIENA (Snijders et al., 2010a; Ripley and Snijders, 2011). Fellows and Handcock (2012) demonstrate the computational viability of models in which covariates on the nodes in a network are considered random variables. An important subclass of these joint models are models in which a social network is modeled together with some other process of interest that takes place among the nodes in the network, as in the study of infectious diseases. For instance, Britton and O’Neill (2002), whose models were later extended and applied by Groendyke et al. (2011, 2012), use an MCMC-based Bayesian approach to infer model parameters for a contact network underlying an outbreak of an infectious disease when certain information is available about the outbreak, even in cases where none of the contact network edges is actually observed.

### **3.4 Measurement error models and partially sampled networks**

Accuracy of network data may be suspect for a number of reasons. A vast sociological literature on the question of how reliable respondents’ reports might be has failed to resolve the question; yet even if it is not quite true that “[self-report] data about communication cannot be used as a proxy for behavioral data” (Bernard et al., 1979), the statistical community will certainly not

be surprised to learn that noise in data collection, along with missing data, can pose significant problems in data analysis. Butts (2003) explores this theme from a statistical point of view, advancing a hierarchical Bayesian framework for simultaneously analyzing network structure along with respondent accuracy. Wyatt et al. (2008) report promising results inferring parameters for a network model from noisy data using a stochastic gradient ascent algorithm to optimize the computationally intractable likelihood function.

An important subcategory of models for which data are imperfectly observed are those in which some (known) portion of the network observations is missing. Both maximum likelihood (Gile and Handcock, 2010) and Bayesian (Koskinen et al., 2010) approaches to this missing-data problem have been proposed and demonstrated to be viable, though each relies heavily on computational techniques: Both use MCMC to help approximate an intractable likelihood, while the latter also uses Bayesian data augmentation.

## 4 Exponential-family random graph models: global network characteristics

In this section, we discuss models represented in terms of the joint distribution of all edges, as opposed to a series of conditional distributions of a hierarchical model as described in Section 5. The most popular framework for these joint models is the class of exponential-family random graph models (ERGMs). Originally proposed by Holland and Leinhardt (1981) to model individual heterogeneity of nodes and reciprocity of their edges (called the  $p_1$  model), the framework was generalized by Frank and Strauss (1986), Wasserman and Pattison (1996) (who called it the  $p^*$  model), and Snijders et al. (2006). It takes the form of a (curved) exponential family on the sample space  $\mathcal{Y}$ :

$$P_{\boldsymbol{\theta}}(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \mathbf{g}(\mathbf{y})\}}{\kappa(\boldsymbol{\theta})}, \quad \mathbf{y} \in \mathcal{Y}, \quad (1)$$

where  $\boldsymbol{\theta}$  is a  $q$ -vector of model parameters, which are mapped to a  $p$ -vector of natural parameters by  $\boldsymbol{\eta}(\cdot)$ ; and  $\mathbf{g}(\cdot)$  is a  $p$ -vector of sufficient statistics, which capture network features of interest, its postulated dependence structure, or both. We present some examples of ERGM statistics in


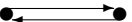
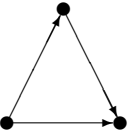
	Edges	Mutual dyads	Transitive triads
			
$\mathbf{g}(\mathbf{y})$	$\sum_{(i,j) \in \mathbb{Y}} \mathbf{y}_{i,j}$	$\sum_{(i,j) \in \mathbf{u}(\mathbb{Y})} \mathbf{y}_{i,j} \mathbf{y}_{j,i}$	$\sum_{(i,j) \in \mathbb{Y}} \mathbf{y}_{i,j} \sum_{k \in N \setminus \{i,j\}} \mathbf{y}_{i,k} \mathbf{y}_{k,j}$
$\Delta_{i,j} \mathbf{g}(\mathbf{y})$	1	$\mathbf{y}_{j,i}$	$\sum_{k \in N \setminus \{i,j\}} (\mathbf{y}_{i,k} \mathbf{y}_{k,j} + \mathbf{y}_{i,k} \mathbf{y}_{j,k} + \mathbf{y}_{k,i} \mathbf{y}_{k,j})$

Table 1: Well-known network statistics, where circles represent nodes and directed lines represent directed edges between nodes. The first two of these statistics are used in the model of Section 4.4, whereas the last is replaced in the models by a statistic that is less prone to the degeneracy problem discussed in Section 4.3.

Table 1. Lastly, to make all probabilities sum to one,

$$\kappa(\boldsymbol{\theta}) = \sum_{\mathbf{y}' \in \mathcal{Y}} \exp \{ \boldsymbol{\eta}(\boldsymbol{\theta})^\top \mathbf{g}(\mathbf{y}') \}.$$

Notably, this differs from the “textbook” exponential family formula, in that it omits an “ $h(\mathbf{y})$ ” factor, which, together with  $\mathcal{Y}$ , controls the “reference measure” for the model, the distribution of networks when  $\boldsymbol{\eta}(\boldsymbol{\theta}) = \mathbf{0}$ . While usually omitted in binary ERGMs, it gains a great deal of importance when using exponential families to model valued networks ties (Krivitsky, 2012).

Exponential families for networks can be derived by postulating a substantively relevant conditional dependence structure, and, based on this structure, approaches like the Hammersley-Clifford Theorem (Besag, 1974) can be used to derive sufficient statistics associated with the model. Frank and Strauss (1986) derive a set of statistics under the “Markovian” assumption that the states of two relationships are, conditional on the rest of the network, stochastically dependent only if they have at least one node in common. Other examples of this approach include the work of Pattison and Wasserman (1999) on multivariate relations, Robins et al. (1999) on polytomous relations, and the realization-dependent conditional independence of Snijders et al. (2006). At the same



time, choice of statistics can be driven by beliefs about the factors that influence the underlying social process, without regard for dependence structure (Morris et al., 2008; Krivitsky, 2012).

For each dyad, one can derive its conditional distribution given the rest of the network, i.e.,  $\text{Odds}_{\boldsymbol{\theta}}(\mathbf{Y}_{i,j} = 1 \mid \mathbf{Y} - (i,j) = \mathbf{y} - (i,j)) = \frac{\exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \mathbf{g}(\mathbf{y} + (i,j))\} / \kappa(\boldsymbol{\theta})}{\exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \mathbf{g}(\mathbf{y} - (i,j))\} / \kappa(\boldsymbol{\theta})} = \exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{\Delta}_{i,j} \mathbf{g}(\mathbf{y})\}$ , with  $\boldsymbol{\Delta}_{i,j} \mathbf{g}(\mathbf{y}) \equiv \mathbf{g}(\mathbf{y} \cup \{(i,j)\}) - \mathbf{g}(\mathbf{y} \setminus \{(i,j)\})$ , the change in the sufficient statistic vector associated with adding an edge at  $(i,j)$ . These “change statistics” (Hunter and Handcock, 2006) or “change scores” (Snijders et al., 2006) facilitate “local” interpretation of ERGMs. For example, consider the mutual dyads statistic in Table 1:  $\boldsymbol{\theta}_{\leftrightarrow}$  associated with this statistic can be said to increase the conditional odds of an edge  $(i,j)$  by  $\exp\{\boldsymbol{\theta}_{\leftrightarrow}\}$  if there is already a tie from  $j$  to  $i$ .

When  $\boldsymbol{\Delta}_{i,j} \mathbf{g}(\mathbf{y})$  does not depend on  $\mathbf{y}$ , the model is dyad-independent, and can be decomposed into logistic regression for each edge.

## 4.1 Simulation methods

Many applications of these models, including all of the Monte Carlo-based methods for finding a maximum likelihood estimator (MLE) as well as Bayesian methods, require making draws from the ERGM.

Because evaluating the conditional probability of an edge given the rest of the network is often relatively inexpensive, it is straightforward to simulate network realizations even from intractable ERGMs using a Metropolis-Hastings sampling procedure: given a proposal  $\hat{\mathbf{y}}$  from density  $q(\hat{\mathbf{y}} \mid \mathbf{y})$ , accept with probability

$$\frac{q(\mathbf{y} \mid \hat{\mathbf{y}}) \exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \mathbf{g}(\hat{\mathbf{y}})\} / \kappa(\boldsymbol{\theta})}{q(\hat{\mathbf{y}} \mid \mathbf{y}) \exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \mathbf{g}(\mathbf{y})\} / \kappa(\boldsymbol{\theta})} = \frac{q(\mathbf{y} \mid \hat{\mathbf{y}})}{q(\hat{\mathbf{y}} \mid \mathbf{y})} \exp\{\boldsymbol{\eta}(\boldsymbol{\theta})^\top (\mathbf{g}(\hat{\mathbf{y}}) - \mathbf{g}(\mathbf{y}))\}$$

or 1, whichever is smaller. Supposing that  $q(\hat{\mathbf{y}} \mid \mathbf{y}) > 0$  only if  $\sum_{(i,j) \in \mathbb{Y}} \mathbb{I}(\hat{\mathbf{y}}_{i,j} \neq \mathbf{y}_{i,j})$  equals one—i.e., the proposed network is constructed by toggling a single edge, say  $(i,j)$ —then  $\mathbf{g}(\hat{\mathbf{y}}) - \mathbf{g}(\mathbf{y})$  reduces to  $\pm \boldsymbol{\Delta}_{i,j} \mathbf{g}(\mathbf{y})$ , with the sign depending on the direction of the toggle, facilitating a fast Gibbs sampling algorithm. Morris et al. (2008) further optimize it for large, sparse networks by implementing an asymmetric proposal.

The rate of convergence for this Gibbs procedure was studied by Bhamidi et al. (2008). There is also some work in progress on exact sampling (Butts, 2012).

## 4.2 Inference methods

The major challenge associated with applying these models is the intractability of the normalizing constant  $\kappa(\boldsymbol{\theta})$  in the likelihood. Here, we describe the methods used to address this challenge.

### 4.2.1 Pseudo-likelihood estimation

An approximate approach to maximum likelihood estimation is based on the so-called pseudo-likelihood function (Strauss and Ikeda, 1990), defined as

$$\prod_{(i,j) \in \mathbb{Y}} P_{\boldsymbol{\theta}}(\mathbf{Y}_{i,j} = \mathbf{y}_{i,j} \mid \mathbf{Y} \setminus \{(i,j)\} = \mathbf{y} \setminus \{(i,j)\}) = \prod_{(i,j) \in \mathbb{Y}} \frac{1}{1 + \exp\{-\boldsymbol{\theta}^\top \boldsymbol{\Delta}_{i,j} \mathbf{g}(\mathbf{y})\}} \approx P_{\boldsymbol{\theta}}(\mathbf{Y} = \mathbf{y}).$$

In other words, for each potential edge in the network, its conditional probability given the state of the rest of the network is evaluated, and the product of those probabilities is used to approximate the likelihood. Maximizing the pseudo-likelihood results in a maximum pseudo-likelihood estimator (MPLE), which, computationally, reduces to logistic regression (Strauss and Ikeda, 1990). van Duijn et al. (2009) and others show that the MPLE is often biased and far less efficient than the MLE, particularly when the social process modeled has strong dyadic dependence, and is particularly vulnerable to the “degeneracy” issues discussed below. van Duijn et al. (2009) also propose a maximum bias-corrected pseudolikelihood estimator (MBLE).

### 4.2.2 Maximum likelihood estimation: stochastic approximation

In the case of linear ERGMs, the maximum likelihood estimator of  $\boldsymbol{\theta}$  solves

$$\nabla_{\boldsymbol{\theta}} \log P_{\boldsymbol{\theta}}(\mathbf{Y} = \mathbf{y}) = \mathbf{g}(\mathbf{y}) - \mathbb{E}_{\boldsymbol{\theta}}(\mathbf{g}(\mathbf{Y})) = 0, \quad (2)$$

where  $\nabla_{\boldsymbol{\theta}} \log P_{\boldsymbol{\theta}}(\mathbf{Y} = \mathbf{y})$  denotes the gradient of the log likelihood function  $\log P_{\boldsymbol{\theta}}(\mathbf{Y} = \mathbf{y})$  with respect to  $\boldsymbol{\theta}$ . Snijders (2002) proposes solving equation (2) using stochastic approximation (Robbins and Monro, 1951; Pflug, 1996). Starting with an initial guess, the stochastic approximation method updates  $\boldsymbol{\theta}^t$  to  $\boldsymbol{\theta}^{t+1}$  at iteration  $t + 1$  as follows:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - a_t \hat{D}_t^{-1} (\mathbf{g}(\mathbf{Y}_{\boldsymbol{\theta}^t}) - \mathbf{g}(\mathbf{y})),$$

where  $D_t^{-1}$  is an approximation of the gradient  $\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}}(\mathbf{g}(\mathbf{Y}))$  in the neighborhood of  $\boldsymbol{\theta}^t$ ,  $a_t$  is a sequence of numbers which tends sufficiently slowly to 0 as  $t$  increases, and  $\mathbf{Y}_{\boldsymbol{\theta}^t}$  is a network sample from the ERGM with parameter  $\boldsymbol{\theta}^t$  by Markov chain Monte Carlo (MCMC) methods. Okabayashi and Geyer (2012) propose a linear search algorithm along similar lines, while Jin and Liang (2012) develop yet another stochastic approximation algorithm.

### 4.2.3 Maximum likelihood estimation: Monte Carlo maximization

An alternative approach to maximum likelihood estimation is based on Monte Carlo approximations of the likelihood function (1). The MCMC MLE approach of Geyer and Thompson (1992) was first adapted by Handcock (2003) to ERGMs and extended by Hunter and Handcock (2006). Both the stochastic approximation algorithm of Snijders (2002) described in Section 4.2.2 and the MCMC MLE approach of Hunter and Handcock (2006) rely on MCMC simulations of networks, but the MCMC MLE approach has the advantage that it makes more efficient use of MCMC samples, as pointed out by Geyer and Thompson (1992, Section 1.3).

Starting with an initial guess  $\boldsymbol{\theta}$ , which is often taken to be the easy-to-calculate MPLE, the ratio of intractable normalizing constants is approximated as

$$\begin{aligned} \frac{\kappa(\boldsymbol{\theta}')}{\kappa(\boldsymbol{\theta})} &= \sum_{\mathbf{y} \in \mathcal{Y}} \exp \left\{ (\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta}))^\top \mathbf{g}(\mathbf{y}) \right\} \frac{\exp \{ \boldsymbol{\eta}(\boldsymbol{\theta})^\top \mathbf{g}(\mathbf{y}) \}}{\kappa(\boldsymbol{\theta})} \\ &= \mathbb{E}_{\boldsymbol{\theta}} \left( \exp \left\{ (\boldsymbol{\eta}(\boldsymbol{\theta}') - \boldsymbol{\eta}(\boldsymbol{\theta}))^\top \mathbf{g}(\mathbf{Y}) \right\} \right), \end{aligned}$$

with the last expectation approximated by a sample of sufficient statistics under  $\boldsymbol{\theta}$ , allowing the likelihood to be maximized as a function of  $\boldsymbol{\theta}'$ .

In practice, the accuracy of the estimate decreases as  $\boldsymbol{\theta}'$  moves farther from  $\boldsymbol{\theta}$ ; in particular, if a guess  $\boldsymbol{\theta}$  is so far from the MLE that the interior of the convex hull of a sample under  $\boldsymbol{\theta}$  does not contain the observed sufficient statistic, the maximized  $\boldsymbol{\theta}'$  does not exist (Hummel et al., 2012). Thus, practical implementation such as that of Handcock et al. (2012) involves several iterations of refining the guess and sampling from it. The MPLE is often used for the purpose, and Okabayashi and Geyer (2012) suggest using their linear search algorithm for the purpose, while Hummel et al. (2012) propose an adaptive method to attenuate the change in  $\boldsymbol{\theta}'$ .

Publicly available implementations of ERGM MLE and MPLE include the `ergm` package (Hunter et al., 2008; Handcock et al., 2012) from the Statnet suite of R packages and `PNet` (Wang et al., 2009).

#### 4.2.4 Bayesian methods

Bayesian inference is based on the posterior distribution of  $\boldsymbol{\theta}$  given  $\mathbf{y}$ :

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}),$$

where  $p(\mathbf{y}) = \int p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$  is the marginal probability of  $\mathbf{y}$ ,  $p(\mathbf{y} \mid \boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\mathbf{Y} = \mathbf{y})$  is the conditional probability of  $\mathbf{y}$  given  $\boldsymbol{\theta}$ ,  $p(\boldsymbol{\theta})$  is the prior probability density of  $\boldsymbol{\theta}$ , and  $p(\boldsymbol{\theta} \mid \mathbf{y})$  is the posterior density of  $\boldsymbol{\theta}$  given  $\mathbf{y}$ . The posterior density  $p(\boldsymbol{\theta} \mid \mathbf{y})$  is typically intractable, because its normalizing constant  $p(\mathbf{y})$  is intractable.

Standard MCMC methods, e.g., the Metropolis-Hastings algorithm, can deal with intractable normalizing constants of a posterior density as long as the posterior density in question is known up to a constant. The problem is that the posterior density is not known up to a constant, because the normalizing constant  $\kappa(\boldsymbol{\theta})$  of the likelihood function  $p(\mathbf{y} \mid \boldsymbol{\theta})$  is intractable. Since the posterior density  $p(\boldsymbol{\theta} \mid \mathbf{y})$  of complex ERGMs includes two intractable normalizing constants,  $p(\mathbf{y})$  on the one hand and  $\kappa(\boldsymbol{\theta})$  on the other hand, the posterior density is *doubly intractable* (Murray et al., 2006).

To demonstrate that standard MCMC methods cannot sample from doubly intractable posterior densities, suppose that we want to construct a Markov chain with stationary distribution  $p(\boldsymbol{\theta} \mid \mathbf{y})$  by a Metropolis-Hastings algorithm. If  $\boldsymbol{\theta}$  denotes the current value of the parameter and  $\boldsymbol{\theta}'$  denotes a proposal of the parameter generated from a proposal density  $q(\cdot \mid \boldsymbol{\theta})$ , then  $\boldsymbol{\theta}'$  is accepted with probability  $\min(1, a)$ , where

$$a = \frac{p(\mathbf{y} \mid \boldsymbol{\theta}') p(\boldsymbol{\theta}')}{p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}. \quad (3)$$

The problem is that acceptance probability (3) is intractable, because  $p(\mathbf{y} \mid \boldsymbol{\theta})$  and  $p(\mathbf{y} \mid \boldsymbol{\theta}')$  involve the normalizing constants  $\kappa(\boldsymbol{\theta})$  and  $\kappa(\boldsymbol{\theta}')$ , respectively.

A naive approach would be to replace the intractable likelihood ratio by an approximation, either deterministic or stochastic. A deterministic approximation could be based on variational approximations of the log-normalizing constants, while stochastic approximations could be based on MCMC estimators of the ratio of normalizing constants such as the importance sampling estimator used in maximum likelihood estimation (e.g., Murray, 2007). However, the stationary distribution of the resulting Markov chains may not be the desired target distribution, i.e., the posterior density of interest (Murray, 2007).

The problematic nature of the naive approach has led to the development of a body of MCMC methods that by design generate samples from doubly intractable posterior density's. Most of them are based on augmenting the posterior density so that the augmented posterior probability distribution is easy to sample from. We discuss here one of the simplest and most appealing auxiliary-variable MCMC methods, following Murray et al. (2006) and Caimo and Friel (2011). Other auxiliary-variable approaches are discussed by Møller et al. (2006) and Koskinen et al. (2010); an alternative approach, inspired by the Monte Carlo maximum likelihood algorithm of Geyer and Thompson (1992), was proposed by Atchadé et al. (2012).

The basic idea can be described as follows. The data  $\mathbf{y}$  are augmented by an auxiliary random graph  $\mathbf{Y}'$  and an auxiliary parameter vector  $\theta'$ . Suppose the joint density of  $\theta, Y, \theta', \mathbf{Y}'$  is of the form

$$p(\theta, \mathbf{y}, \theta', \mathbf{y}') = p(\theta) p(\mathbf{y} | \theta) q(\theta' | \theta, \mathbf{y}) p(\mathbf{y}' | \theta'), \quad (4)$$

where  $q(\theta' | \theta, \mathbf{y})$  is an auxiliary density and  $p(\mathbf{y}' | \theta')$  is the conditional probability of  $\mathbf{y}'$  given  $\theta'$ , which is of the same exponential-family form as  $\mathbf{Y}$ , implying the same reference measure and the same sufficient statistics. The augmented posterior density is of the form

$$p(\theta, \theta', \mathbf{y}' | \mathbf{y}) \propto p(\theta, \mathbf{y}, \theta', \mathbf{y}') \propto p(\theta) p(\mathbf{y} | \theta) q(\theta' | \theta, \mathbf{y}) p(\mathbf{y}' | \theta'). \quad (5)$$

The posterior density of interest,  $p(\theta | \mathbf{y})$ , is the marginal distribution of the augmented posterior density,  $p(\theta, \theta', \mathbf{y}' | \mathbf{y})$ . A simple Metropolis-Hastings algorithm to sample from the augmented posterior density, which has been called the *exchange algorithm* (Murray et al., 2006; Caimo and Friel, 2011), operates as follows:

- (1) Sample  $\theta' | \theta, \mathbf{y} \sim q(\cdot | \theta, \mathbf{y})$  and then sample  $\mathbf{Y}' | \theta' \sim p(\cdot | \theta')$ .

(2) Propose to swap the values of  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}'$  and accept the proposal with probability  $\min(1, a)$ , where

$$a = \frac{p(\boldsymbol{\theta}') p(\mathbf{y} | \boldsymbol{\theta}') q(\boldsymbol{\theta} | \boldsymbol{\theta}', \mathbf{y}) p(\mathbf{y}' | \boldsymbol{\theta})}{p(\boldsymbol{\theta}) p(\mathbf{y} | \boldsymbol{\theta}) q(\boldsymbol{\theta}' | \boldsymbol{\theta}, \mathbf{y}) p(\mathbf{y}' | \boldsymbol{\theta}')}. \quad (6)$$

Simple calculation shows that the intractable normalizing constants  $\kappa(\boldsymbol{\theta})$  and  $\kappa(\boldsymbol{\theta}')$  cancel from the acceptance probability (6), so the Metropolis-Hastings algorithm operating on the augmented state space is tractable. Sampling  $\mathbf{Y}'$  requires exact sampling, which is typically infeasible. Caimo and Friel (2011) proposed to sample  $\mathbf{Y}'$  instead by MCMC methods, as described in Section 4.1. The work of Liang (2010) on double Metropolis-Hastings algorithms provides some justification for doing so. Last, to reduce model degeneracy of conventional ERGMs, Schweinberger and Handcock (2011) proposed hierarchical ERGMs with local dependence and extended the exchange algorithm to hierarchical ERGMs.

### 4.3 ERGM degeneracy

In addition to the intractable normalizing constant, a major difficulty in modeling complex social processes using ERGMs is a phenomenon referred to as “degeneracy,” which arises for example when attempting to model processes exhibiting individual heterogeneity in activity level using 2-star statistics and triad closure using triangle counts as suggested in Frank and Strauss (1986). In a contemporaneous publication, Strauss (1986) noted that the Markov graph models Frank and Strauss (1986) had proposed tend to, as the network size increases, concentrate an increasingly large probability mass on an increasingly small fraction of possible networks. Jonasson (1999) and Häggström and Jonasson (1999) studied the asymptotics of the so-called triangle model, corresponding to the ERGM with the number of edges and triangles as sufficient statistics, and showed that as the network size increases, the parameter space tends to be divided between configurations that exhibit little triadic closure—typically less than observed—and configurations that are degenerate, in that almost all of the probability mass is concentrated on a single (complete) network. Handcock (2003) and Rinaldo et al. (2009), among others, mapped the shape of the region of the parameter space in which this degeneracy occurred, and Schweinberger (2011) showed that some statistics, including triangle and  $k$ -star counts, induce this behavior asymptotically.

It can be argued that this degeneracy is merely a symptom of a broader problem of model misspecification. In their development of curved ERGMs, Snijders et al. (2006) used change statistics to show that models with positive coefficients on triangles and/or  $k$ -stars induced strong positive dependence among dyad values: edges beget more edges, leading to what the authors called an “avalanche” toward a complete graph. In another realization, the avalanche might not take place, leading to a sparse network. Thus, the model with such positive dependence can induce a bimodal distribution of networks, and, at the MLE, however it is found, the observed network would be between the modes. A bimodal distribution, with little mass between the modes, is difficult for MCMC sampling to explore, which can cause any simulation-based method to fail, whether maximum likelihood or Bayesian.

Indeed, Krivitsky (2012), in developing ERGM statistics for networks with weighted edges that are unbounded counts, reports that a statistic analogous to the count of 2-stars induced a bimodal distribution of networks, with neither mode being degenerate (i.e., concentrated) itself. Thus, a fruitful approach to degeneracy may be to model network phenomena using statistics less prone to result in model degeneracy, such as those of Snijders et al. (2006) and Krivitsky (2012).

## 4.4 Application

We demonstrate the three methods of estimation by applying them to Sampson’s network dataset and the ERGM with the three statistics of Table 1 except that the transitive triad statistic is replaced by the less degeneracy-prone transitive edges statistic given by

$$\sum_{(i,j) \in \mathbb{Y}} \mathbf{y}_{i,j} \max_{k \in N \setminus \{i,j\}} \mathbf{y}_{i,k} \mathbf{y}_{k,j}$$

(Snijders et al., 2010b). This statistic resembles the transitive triad count in Table 1 except the summation over  $k$  is replaced by a maximum over  $k$ .

Table 2 shows various point estimates of the edges, mutual edges, and transitive edges parameters. Although there is some question as to the validity of ERGM standard errors due to questionable asymptotics (Hunter and Handcock, 2006, for example), the standard errors in this case seem to agree closely with the posterior standard deviations, suggesting that they may be valid in practice. The marginal posterior densities are shown in Figure 2, along with the MCMC

	MPLE (S.E.)	MCMC MLE (S.E.)	Posterior mean (S.D.)
edges	-1.6981 (0.2502)	-1.9468 (0.3672)	-1.9216 (0.3535)
mutual edges	2.3277 (0.2940)	2.3093 (0.4195)	2.2732 (0.4273)
transitive edges	-0.0503 (0.1335)	0.1328 (0.2210)	0.1309 (0.2135)

Table 2: Point estimates of edges, mutual edges, and transitive edges parameters, including posterior mean and standard deviation.

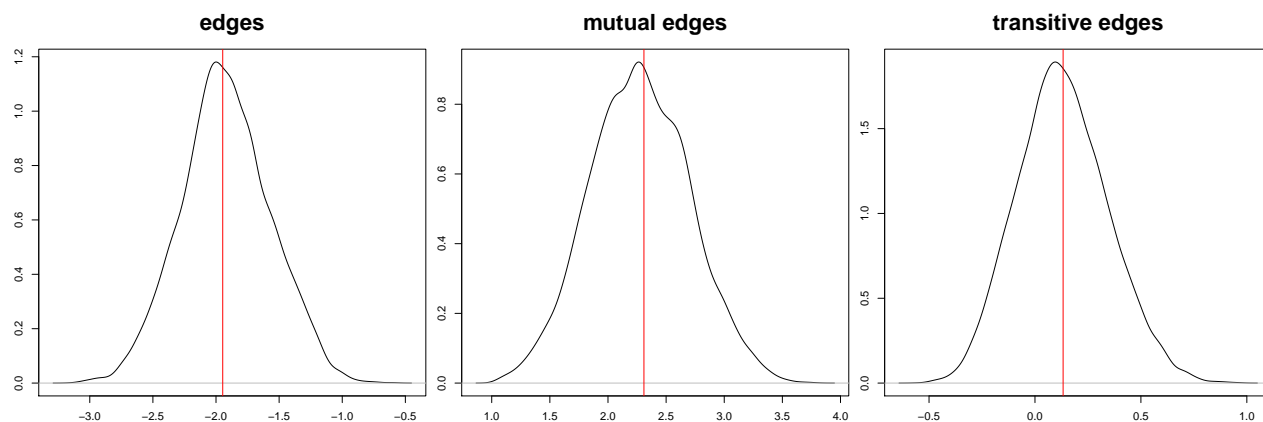


Figure 2: Marginal posterior densities of canonical edges, mutual edges, and transitive edges parameter; the red lines indicate the MCMC MLE. Transitive edges should not be confused with the transitive triads of Table 1.

MLE. It is evident that the MCMC MLE is close to the posterior mode. Example code to obtain the results shown here using the R packages `ergm` and `Bergm` can be found in the appendix (in both R packages, the transitive edges statistic is called `transitiveties`).

Hunter et al. (2008) argue that assessing the goodness-of-fit of models is important, not least due to the model degeneracy problem discussed in Section 6. We show posterior predictions of the sufficient statistics in Figure 3. The posterior predictions are centered at the observed data, which is encouraging.



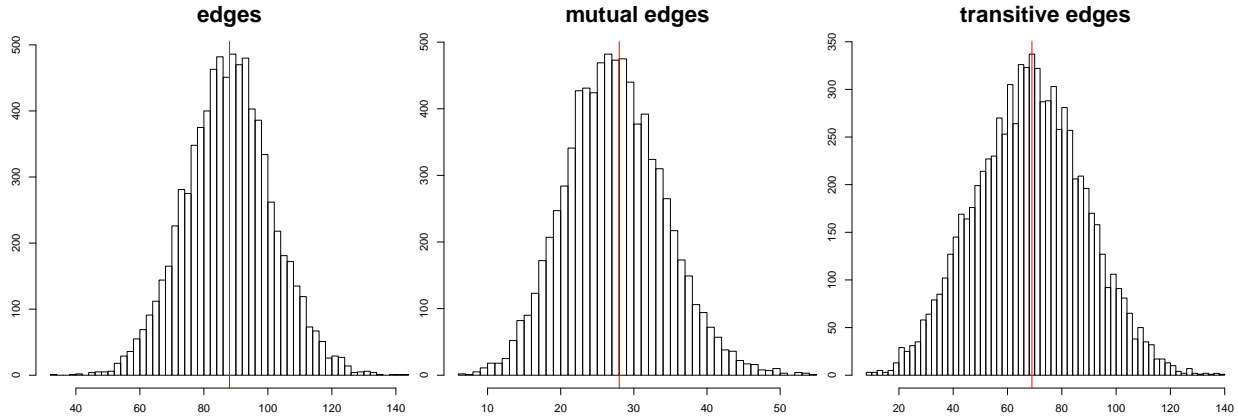


Figure 3: Posterior predictions of number of edges, mutual edges, and transitive edges; red line indicates observed value. Transitive edges should not be confused with the transitive triads of Table 1.

## 5 Models based on independence conditional on latent variables

While ERGMs are useful for modeling global network characteristics, models based on conditional independence (given latent variables) are useful for multiple reasons. First, ERGMs are not well-understood and sometimes possess undesirable properties, e.g., model degeneracy. Second, the likelihood function of ERGMs is intractable, complicating statistical computing. Third, there may be unobserved heterogeneity or unobserved structure.

For these reasons, models building on conditional independence (given latent variables) have attracted much attention. At least three streams of latent variables can be distinguished: random effects models and mixed effects models and extensions; stochastic block models and extensions, including mixed membership models; and latent space models and extensions. All of these latent variable models, with the exception of Wyatt et al. (2008), Koskinen (2009), and Schweinberger and Handcock (2011), assume conditional independence of dyads: if  $\mathbf{Z}$  denotes generic latent variables, which may be node-bound or dyad-bound and discrete or continuous, then the models assume that dyads (distinct unordered pairs of nodes)  $(\mathbf{Y}_{i,j}, \mathbf{Y}_{j,i})$  are independent conditional

on  $\mathbf{Z}$ :

$$P_{\theta}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}) = \prod_{(i,j) \in \text{eu}(\mathbb{Y})} P_{\theta}(\mathbf{Y}_{i,j} = \mathbf{y}_{i,j}, \mathbf{Y}_{j,i} = \mathbf{y}_{j,i} \mid \mathbf{Z} = \mathbf{z}).$$

In the case of directed networks, conditional independence of dyads  $(\mathbf{Y}_{i,j}, \mathbf{Y}_{j,i})$  does not imply conditional independence of edges  $\mathbf{Y}_{i,j}$  and  $\mathbf{Y}_{j,i}$ : even conditional on  $\mathbf{Z}$ , edges  $\mathbf{Y}_{i,j}$  and  $\mathbf{Y}_{j,i}$  may be dependent due to reciprocity. However, with the notable exception of the  $p_2$  model (van Duijn, 1995; van Duijn et al., 2004), most latent variable models make the more restrictive assumption that edges  $\mathbf{Y}_{i,j}$  are independent conditional on  $\mathbf{Z}$ :

$$P_{\theta}(\mathbf{Y} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}) = \prod_{(i,j) \in \mathbb{Y}} P_{\theta}(\mathbf{Y}_{i,j} = \mathbf{y}_{i,j} \mid \mathbf{Z} = \mathbf{z}). \quad (7)$$

Three aspects of these models are worth noting. First, conditional independence does not imply that latent variable models cannot capture network dependencies of interest. Indeed, some of the more advanced latent variables (notably latent space models) make clever use of latent variables to capture such dependence structure, including mutuality and transitivity. Second, the conditional independence of edges implies that model degeneracy is not an issue, which facilitates the construction of models. Third, the conditional independence of edges has computational advantages in that standard MCMC methods can be used.

We discuss random effects and mixed effects models very briefly in Section 5.1 and then describe two classic latent variable models in more detail: stochastic block models in Section 5.2 and latent space models in Section 5.3.

## 5.1 Random effects and mixed effects models

The earliest latent variable model, the random effects model of van Duijn (1995), was motivated by the lack of model parsimony of the so-called  $p_1$  model of Holland and Leinhardt (1981)—which is an ERGM of the form (1) with the in-degrees, out-degrees, and mutual edges as sufficient statistics—and the fact that the  $p_1$  model ignores covariates. Since it can be considered a random effects version of  $p_1$ , the model of van Duijn (1995) is known as the  $p_2$  model. To estimate its parameters, van Duijn (1995) and van Duijn et al. (2004) exploit the fact that the  $p_1$  model can be represented as a generalized linear model and the  $p_2$  model as a generalized linear mixed model and

develop an iterative generalized least squares procedure. Zijlstra et al. (2009) develops Bayesian MCMC methods for the  $p_2$  model.

Hoff (2003, 2005, 2009) introduces multiple generalizations of generalized linear mixed models, along with Bayesian MCMC methods. Some of these computational methods are implemented in the R package `eigenmodel` (Hoff, 2012). Krivitsky et al. (2009) extend the latent cluster models of Section 5.3 to include random effects.

## 5.2 Stochastic block models

Stochastic block models of networks, which are related to finite mixture models, were explored by Snijders and Nowicki (1997) and extended by Nowicki and Snijders (2001). These models partition the set of nodes is partitioned into subsets, called blocks, where conditional on block memberships, edges are independent and the probability of an edge between two blocks depends on the blocks to which the nodes belong. Tallberg (2005) describes an extension that incorporates covariates to predict block memberships. Airoldi et al. (2008) propose more advanced stochastic block models, called mixed membership models, that allow the block memberships of nodes to depend on the pair of nodes, i.e., the block to which you belong depends with whom you interact.

Yet we focus here on the basic stochastic block model of Nowicki and Snijders (2001), which is based on two fundamental assumptions. First, the set of nodes is partitioned into  $K$  blocks, where  $K$  is fixed and known. Let

$$\mathbf{Z}_i \mid \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_K \stackrel{\text{i.i.d.}}{\sim} \text{Multinomial}(1; \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_K). \quad (8)$$

Second, equation (7) holds, with  $P_{\boldsymbol{\theta}}(\mathbf{Y}_{i,j} = 1 \mid \mathbf{Z}_i = \mathbf{z}_i, \mathbf{Z}_j = \mathbf{z}_j) = \boldsymbol{\theta}_{\mathbf{z}_i, \mathbf{z}_j}$ , where  $\boldsymbol{\theta}_{k,l}$  is the probability that a given actor in group  $k$  has a tie to a given actor in group  $l$ .

The Bayesian MCMC algorithm is simple: given conjugate priors, the posterior can be sampled by Gibbs sampling. In particular, if the priors for  $\boldsymbol{\pi}$  and  $\boldsymbol{\beta}$  are independent and given by Dirichlet and beta, respectively, then the full conditional distributions of  $\boldsymbol{\pi}$  and  $\boldsymbol{\beta}$  are also Dirichlet and beta. The parameters of stochastic block models are not identifiable in that the likelihood function is invariant to the labeling of the blocks and Bayesian MCMC samples may therefore show evidence of label-switching. Snijders and Nowicki (1997) and Nowicki and Snijders (2001) deal with the

label-switching problem either by imposing order restrictions on parameters or restricting attention to functions that are invariant to the labeling of the blocks, e.g., an indicator of whether two nodes belong to the same block. An alternative is given by the Bayesian decision-theoretic approach of Stephens (2000), which relabels the Bayesian MCMC sample. The first approach is implemented in the Windows-based program BLOCKS (Snijders and Nowicki, 2007) and the second approach is implemented in the R package `hergm` (Schweinberger, 2012).

The variational methods of Section 5.4 provide an alternative to MCMC-based estimation. They are approximate but fast and feasible, and they can be applied to networks with more than 100,000 nodes and 10 billion dyads.

### 5.3 Latent space models

Hoff et al. (2002) proposed latent space models based on the assumption that nodes are embedded in a latent, metric space. Latent space models come in two basic flavors: One that assumes the metric space is Euclidean (Hoff et al., 2002), and another that assumes the metric space is ultrametric (Schweinberger and Snijders, 2003). In either case, equation (7) is assumed to hold with

$$\text{logit}(P_{\beta}(\mathbf{Y}_{i,j} = 1 \mid \mathbf{Z} = \mathbf{z}_i, \mathbf{Z} = \mathbf{z}_j)) = \beta_0 + \mathbf{x}_{i,j}^{\top} \boldsymbol{\beta} + d(\mathbf{z}_i, \mathbf{z}_j),$$

where  $\mathbf{x}_{i,j}$  is a  $q$ -vector of covariates for dyad  $(i, j)$ ,  $\beta_0$  is a parameter which controls the density of the network, and  $d(\cdot, \cdot)$  is a distance function. Two examples of distance functions are given by  $d(\mathbf{z}_i, \mathbf{z}_j) = -\|\mathbf{z}_i - \mathbf{z}_j\|$  for a Euclidean latent effect and  $d(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^{\top} \mathbf{z}_j$  for an inner-product effect.

The likelihood function, which is equivalent to the likelihood function of nonlinear logistic regression models, can be maximized to obtain maximum likelihood estimates. An alternative is a Bayesian approach, specifying a prior distribution for  $\boldsymbol{\beta}$  and  $\mathbf{Z}$  and using the Gibbs sampler to sample from the posterior distribution of  $\boldsymbol{\beta}$  and  $\mathbf{Z}$ . If the posterior distribution is approximated by Bayesian MCMC methods, then a nonidentifiability problem arises, in that the likelihood is invariant to rotation and reflection of latent positions about the origin and, for Euclidean latent space models, it is also invariant to translation. Hoff et al. (2002) address this problem

by using a Procrustes transformation to rotate and reflect these posterior draws to be as close as possible to a reference configuration. Shortreed et al. (2006) further address this problem via a Minimum Kullback-Leibler divergence (MKL) estimate to summarize the posterior positions, finding a configuration  $(\boldsymbol{\beta}_{\text{MKL}}, \mathbf{Z}_{\text{MKL}})$  that induces a distribution of dyad values closest to their posterior predictive distribution. That is, given two parameter configurations  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{z})$  and  $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \mathbf{z}')$ , and letting  $\mathbf{Y}_{\boldsymbol{\theta}}$  be a random network variable induced by  $\boldsymbol{\theta}$ , so that  $P(\mathbf{Y}_{\boldsymbol{\theta}} = \mathbf{y}) \equiv P_{\boldsymbol{\theta}}(\mathbf{Y} = \mathbf{y})$ , the divergence is given by

$$\text{KL}(\mathbf{Y}_{\boldsymbol{\theta}'}, \mathbf{Y}_{\boldsymbol{\theta}}) = \mathbb{E}_{\mathbf{Y}'|\boldsymbol{\theta}'} \left( \log \frac{P_{\boldsymbol{\theta}'}(\mathbf{Y} = \mathbf{Y}')}{P_{\boldsymbol{\theta}}(\mathbf{Y} = \mathbf{Y}')} \right),$$

with the estimate minimizing its expectation under the posterior  $\boldsymbol{\theta} \mid \mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}$ :

$$\boldsymbol{\theta}_{\text{MKL}} = \arg \min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\theta}' \mid \mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}} (\text{KL}(\mathbf{Y}_{\boldsymbol{\theta}'}, \mathbf{Y}_{\boldsymbol{\theta}})).$$

Shortreed et al. (2006) show that the estimate could be obtained easily for an exponential family, and for binary networks in particular, by evaluating the posterior predictive expectation  $\mathbb{E}_{\mathbf{Y}' \mid \mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}}(\mathbf{Y}')$  and maximizing

$$\arg \min_{\boldsymbol{\theta}} \prod_{(i,j) \in \mathbb{Y}} P_{\boldsymbol{\theta}}(\mathbf{Y}_{i,j} = 1)^{\mathbb{E}_{\mathbf{Y}' \mid \mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}}(\mathbf{Y}'_{i,j})} P_{\boldsymbol{\theta}}(\mathbf{Y}_{i,j} = 0)^{1 - \mathbb{E}_{\mathbf{Y}' \mid \mathbf{Y}_{\text{obs}} = \mathbf{y}_{\text{obs}}}(\mathbf{Y}'_{i,j})},$$

the binomial PMF evaluated up to proportion.

This model is further extended by Hoff (2005) to add node-specific random sender ( $\boldsymbol{\delta}$ ) and receiver ( $\boldsymbol{\gamma}$ ) effects, i.e.,

$$\text{logit}(P_{\boldsymbol{\beta}}(\mathbf{Y}_{i,j} = 1 \mid \mathbf{Z} = \mathbf{z}_i, \mathbf{Z} = \mathbf{z}_j)) = \boldsymbol{\beta}_0 + \mathbf{x}_{i,j}^{\top} \boldsymbol{\beta} + d(\mathbf{z}_i, \mathbf{z}_j) + \boldsymbol{\delta}_i + \boldsymbol{\gamma}_j,$$

with  $\boldsymbol{\delta}$  and  $\boldsymbol{\gamma}$  having a normal distribution, and by Handcock et al. (2007) to model the latent positions as a mixture of  $K$  spherical Gaussian clusters, i.e.,

$$\mathbf{Z}_i \stackrel{\text{i.i.d.}}{\sim} \sum_{k=1}^K \pi_k \text{MVN}(\boldsymbol{\mu}, \sigma_{\mathbf{Z}}^2 \mathbf{I}),$$

allowing a soft clustering of the nodes in the network to be produced. This leads to another nonidentifiability problem, in that the prior for the latent positions is invariant to permutation of cluster labels, which is resolved using the relabeling algorithm of Stephens (2000).

Alternative computational approaches to approximating the likelihood function include the case-control sampling idea of Raftery et al. (2011) and the variational approach of Salter-Townshend and Murphy (2010). The latter is described in Section 5.4.

## 5.4 Variational methods

Bayesian MCMC methods, which represent the gold standard and dominate the practice of latent variable modeling, tend to be slow and cannot be applied to large networks with hundreds of nodes. Variational methods have emerged as a fast and feasible though approximate alternative to Bayesian MCMC methods. Approximate maximum likelihood estimation of stochastic block models based on variational methods was introduced by Daudin et al. (2008), who presented a variational EM algorithm and applied it to more than 600 nodes. Online versions of Daudin et al. (2008) were considered by Zanghi et al. (2010) and applied to more than 12,000 nodes. Airolidi et al. (2008) applied a variational EM algorithm to mixed membership models. A generalized variational EM algorithm, which exploits an MM algorithm (Hunter and Lange, 2004) to implement the E-step, was proposed by Vu et al. (2012) and applied to more than 131,000 nodes. The consistency of approximate maximum likelihood estimators was established by Celisse et al. (2011), who considered the stochastic block models of Nowicki and Snijders (2001). Approximate Bayesian estimation of stochastic block models can be found using methods in Schweinberger et al. (2011) and Vu et al. (2012). Approximate Bayesian estimation of latent space models was discussed by Salter-Townshend and Murphy (2010) and applied to more than 80 nodes.

We focus here on approximate Bayesian estimation of latent variable models based on variational methods. To keep the discussion manageable, we let  $\mathbf{u}$  represent the vector of unknowns—which includes all latent variables and all parameters.

The key obstacle to Bayesian model estimation and model selection is the marginal likelihood  $p(\mathbf{y})$ , which can be expressed as

$$p(\mathbf{y}) = \int \frac{p(\mathbf{y} | \mathbf{u}) p(\mathbf{u})}{q(\mathbf{u} | \boldsymbol{\vartheta})} q(\mathbf{u} | \boldsymbol{\vartheta}) d\mathbf{u}, \quad (9)$$

where  $q(\mathbf{u} | \boldsymbol{\vartheta})$  is a distribution with support  $\mathcal{U}$ , the space of  $\mathbf{u}$ , parameterized by a vector of auxiliary parameters  $\boldsymbol{\vartheta}$ . The idea of variational methods is to translate the intractable integration

problem (9) into a tractable optimization problem by bounding  $\log p(\mathbf{y})$  below using a bound that is as tight as possible. To this end, we maximize the right-hand side of

$$\log p(\mathbf{y}) \geq \int \left[ \log \frac{p(\mathbf{y} | \mathbf{u}) p(\mathbf{u})}{q(\mathbf{u} | \boldsymbol{\vartheta})} \right] q(\mathbf{u} | \boldsymbol{\vartheta}) d\mathbf{u}. \quad (10)$$

with respect to  $\boldsymbol{\vartheta}$ . It can be shown that the difference between the left- and right-hand side of (10) is equal to the Kullback-Leibler divergence from the auxiliary distribution  $q(\mathbf{u} | \boldsymbol{\vartheta})$  parameterized by  $\boldsymbol{\vartheta}$  to the posterior distribution  $p(\mathbf{u} | \mathbf{y})$ :

$$\text{KL}(q(\mathbf{u} | \boldsymbol{\vartheta}); p(\mathbf{u} | \mathbf{y})) = \int \left[ \log \frac{q(\mathbf{u} | \boldsymbol{\vartheta})}{p(\mathbf{u} | \mathbf{y})} \right] q(\mathbf{u} | \boldsymbol{\vartheta}) d\mathbf{u}. \quad (11)$$

Thus, maximizing the lower bound (10) is equivalent to minimizing the Kullback-Leibler divergence (11).

To make the optimization problem tractable, it is common to choose fully factorized auxiliary distributions of the form

$$q(\mathbf{u} | \boldsymbol{\vartheta}) = \prod_i q(u_i | \boldsymbol{\vartheta}_i), \quad (12)$$

where  $u_i$  denotes the  $i$ th element of the vector of unknowns  $\mathbf{u}$  and  $q(u_i | \boldsymbol{\vartheta}_i)$  denotes the marginal auxiliary distribution of  $u_i$ , parameterized by a vector of auxiliary parameters  $\boldsymbol{\vartheta}_i$ .

A variational EM algorithm to maximize the lower bound can be sketched as follows. Let  $\text{LB}(\boldsymbol{\vartheta}_{\mathbf{Z}}^{(t)}, \boldsymbol{\vartheta}_{\boldsymbol{\theta}}^{(t)})$  be the lower bound (10), where  $\boldsymbol{\vartheta}_{\mathbf{Z}}$  denotes the vector of auxiliary parameters corresponding to the latent variables, with value  $\boldsymbol{\vartheta}_{\mathbf{Z}}^{(t)}$  at iteration  $t$ , and  $\boldsymbol{\vartheta}_{\boldsymbol{\theta}}$  denotes the vector of auxiliary parameters corresponding to the parameters  $\boldsymbol{\theta}$ , with value  $\boldsymbol{\vartheta}_{\boldsymbol{\theta}}^{(t)}$  at iteration  $t$ .

E-STEP: Let  $\boldsymbol{\vartheta}_{\mathbf{Z}}^{(t+1)}$  be the maximizer of  $\text{LB}(\boldsymbol{\vartheta}_{\mathbf{Z}}, \boldsymbol{\vartheta}_{\boldsymbol{\theta}}^{(t)})$  with respect to  $\boldsymbol{\vartheta}_{\mathbf{Z}}$ .

M-STEP: Let  $\boldsymbol{\vartheta}_{\boldsymbol{\theta}}^{(t+1)}$  be the maximizer of  $\text{LB}(\boldsymbol{\vartheta}_{\mathbf{Z}}^{(t+1)}, \boldsymbol{\vartheta}_{\boldsymbol{\theta}})$  with respect to  $\boldsymbol{\vartheta}_{\boldsymbol{\theta}}$ .

By construction, iteration  $t + 1$  increases the lower bound  $\text{LB}(\boldsymbol{\vartheta}_{\mathbf{Z}}, \boldsymbol{\vartheta}_{\boldsymbol{\theta}})$ :

$$\text{LB}(\boldsymbol{\vartheta}_{\mathbf{Z}}^{(t)}, \boldsymbol{\vartheta}_{\boldsymbol{\theta}}^{(t)}) \leq \text{LB}(\boldsymbol{\vartheta}_{\mathbf{Z}}^{(t+1)}, \boldsymbol{\vartheta}_{\boldsymbol{\theta}}^{(t)}) \quad (13)$$

$$\leq \text{LB}(\boldsymbol{\vartheta}_{\mathbf{Z}}^{(t+1)}, \boldsymbol{\vartheta}_{\boldsymbol{\theta}}^{(t+1)}). \quad (14)$$

Implementation details are discussed in the cited literature.

A variational approach to Bayesian inference approximates the marginal likelihood  $p(\mathbf{y})$  and the posterior distribution  $p(\mathbf{u} \mid \mathbf{y})$  and thus tackles both Bayesian model estimation and model selection.

In practice, the most important problem with variational methods is that it is impossible to assess how tight the lower bound (10) is. In the case of fully factorized auxiliary distributions of the form (12), the Kullback-Leibler divergence (11), and thus the tightness of the lower bound (10), is determined by the posterior dependence of the unknowns. In the absence of posterior dependence, equality can be achieved in (10), though in general this is not possible. However, as long as the posterior dependence is weak, the variational approximation of the posterior distribution and the log marginal likelihood may be useful, at least as a starting point of Bayesian MCMC algorithms and when Bayesian MCMC algorithms are too slow.

## 5.5 Application

To illustrate latent variable models, we apply the stochastic block model of Nowicki and Snijders (2001) and the latent space models of Handcock et al. (2007) and Krivitsky et al. (2009) to Sampson’s monk data.

We obtained a Bayesian posterior MCMC sample from the Nowicki and Snijders (2001) model using the Windows-based program BLOCKS (Snijders and Nowicki, 2007). Because BLOCKS returns summaries of the posterior distribution rather than a sample from it, we modified the `Delphi` source code of BLOCKS to obtain an MCMC sample. We handled the label-switching in the posterior sample using a stochastic label-switching algorithm described by Schweinberger and Handcock (2011) and implemented in the R package `hergm` (Schweinberger, 2012). MCMC samples from the posterior distributions of the latent space models of Handcock et al. (2007) and Krivitsky et al. (2009) were obtained by the R package `latentnet` (Krivitsky and Handcock, 2008), for which the code can be found in the appendix.

Figure 4 shows the clustering of Sampson’s monks by the three models, where we use both two and three dimensions for the Krivitsky et al. (2009) model. It is evident that the clustering is almost the same with all three approaches and agrees with the ground truth—the three-group



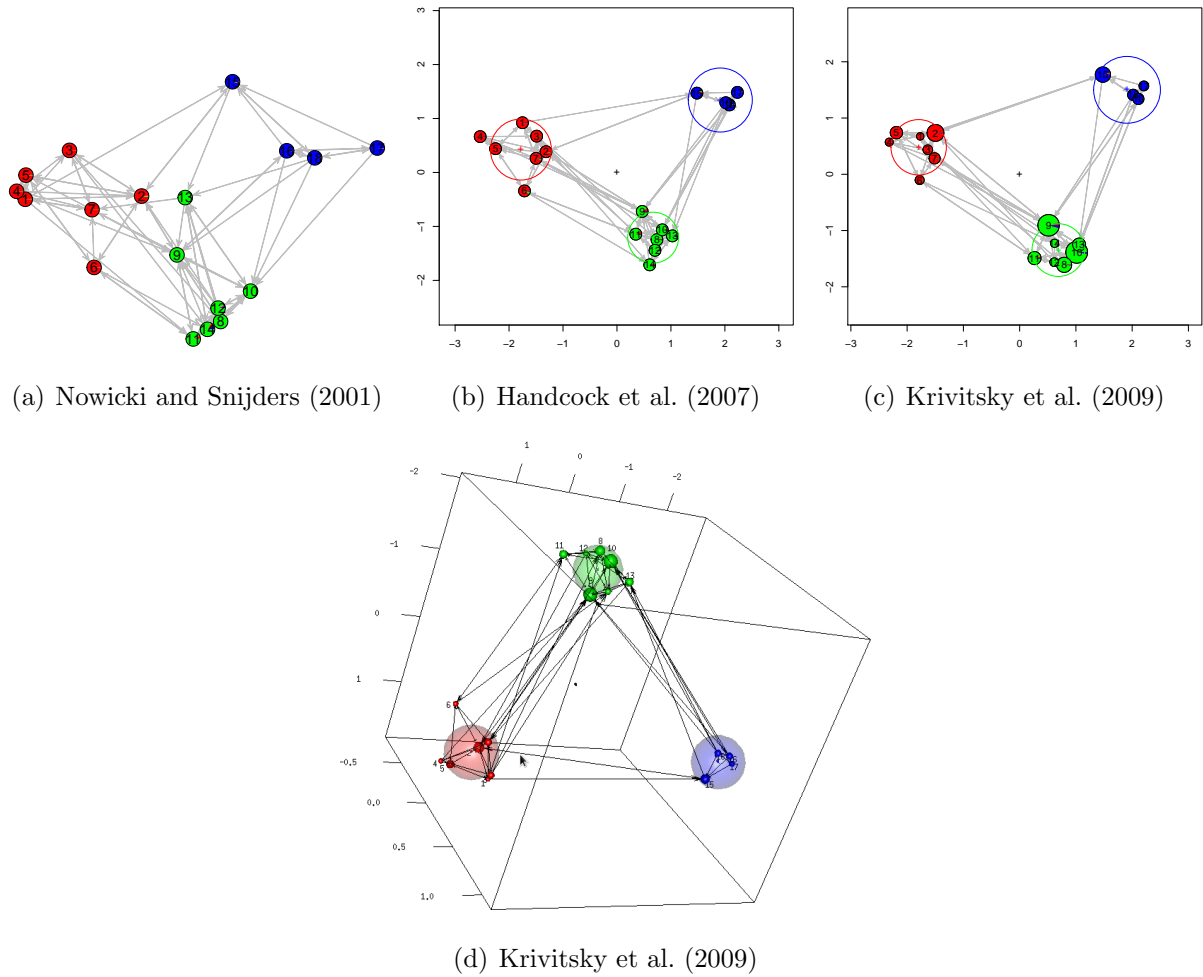


Figure 4: Clustering of Sampson’s monks by (a) stochastic block model of Nowicki and Snijders (2001); (b) latent space model of Handcock et al. (2007); and latent space model of Krivitsky et al. (2009) in (c) two dimensions and (d) three dimensions.

partition of the monks into Loyal Opposition, Turks, and Outcasts. In fact, there is hardly any uncertainty about the clustering.

## 6 Discussion

As we have tried to show, applications that combine computational statistics and social network modeling are already very broad, but research combining is still quite young relative to these two separate fields. As such, numerous challenges remain.

First, there are myriad open questions regarding theoretical statistics in network modeling. Research in this area is too broad to summarize adequately in a single paragraph, though we may highlight some computational issues related to this theory. There are several recent articles that extend exponential family theory to network models, with particular emphasis on cases in which maximum likelihood estimators may not exist. Geyer (2009) shows how exact linear programming algorithms, implemented in the R package `rcdd` (Geyer, 2012), may be used to analyze such cases. There is very little asymptotic theory about the use of ERGMs—indeed, Shalizi and Rinaldo (2011) point out that maximum likelihood estimators for certain such models may be shown not even to be consistent as the number of nodes grows without bound—which means that computational techniques for finite-sample statistical inference and for Bayesian inference become important. Monte Carlo-based testing (Krivitsky, 2012, for example) is an obvious possibility here, though there exists little work to date on practical implementation of valid inferential techniques in general.

Beyond the theoretical questions, however, perhaps the most important computational issue looming for statistical analysis of social network data is the high-dimensional nature of more and more network datasets. Since the number of edge variables in a network of  $n$  nodes scales like  $O(n^2)$ , mere storage of large networks becomes problematic. Often, social networks are sparse in the sense that many more edge variables are zero (indicating absence of an edge) than one, which means that computational techniques for sparse matrices become relevant. Though software implementing such techniques already exists for many common matrix operations, novel statistical analyses might require novel matrix operations and therefore new algorithms. Furthermore, it is increasingly the case that  $n$  itself is quite large—when data are collected automatically by

electronic devices from a large population, for example—so algorithms that scale no better than  $O(n^2)$  are problematic.

We feel that the interplay of statistical theory and computing in social network modeling applications will also prove to be fertile ground for researchers in statistics to make advances. As a single example, consider testing a null hypothesis against an alternative hypothesis. Standard tests based on the Wald or likelihood ratio test statistics require maximizing the likelihood function under the alternative hypothesis, which may be infeasible or, at the very least, may be more time-consuming than maximizing the likelihood function under the null hypothesis. Thus, score tests, which only require to maximizing the likelihood function under the null hypothesis, may present an attractive alternative provided clever methods are used to evaluate the score function under the null hypothesis. In the case of dynamic network models of the form Snijders (2001); Snijders et al. (2007), frequentist score-type tests (Schweinberger, 2012) have become popular and are widely used. For exponentially-parameterized stochastic block models, Bayesian score-type tests are explored by Schweinberger et al. (2011).

With regard to degeneracy, as discussed in Section 4.3, dependence creates multimodality for degenerate models, which in turn creates issues for computational algorithms that depend on MCMC. Solving these computational issues will not make degeneracy go away; it can help to cure the symptom, but fails to address the root of the problem. In any case, goodness-of-fit measures are indispensable to diagnose degeneracy where it occurs (Hunter et al., 2008).

## Acknowledgments

The authors gratefully acknowledge support by the Office of Naval Research under the MURI program, Award Number N00014-08-1-1015, and the National Institutes of Health, Award Numbers R01 HD068395 and R01 GM083603.

## Example code

### ERGM MCMC MLE

```
library(ergm)
data(sampson)
monks.ergm.mle <- ergm(samplike~edges+mutual+transitiveties)
summary(monks.ergm.mle)
```

### ERGM MPLE

```
library(ergm)
data(sampson)
monks.ergm.mple <- ergm(samplike~edges+mutual+transitiveties,estimate="MPLE")
summary(monks.ergm.mple)
```

### Two-dimensional Euclidean latent space model with three clusters and random receiver effects

```
library(latentnet)
data(sampson)
monks.d2G3r <- ergmm(samplike~euclidean(d=2,G=3)+rreceiver)
Z <- plot(monks.d2G3r, rand.eff="receiver", pie=TRUE, vertex.cex=2)
text(Z, label=1:nrow(Z))
```

### Three-dimensional Euclidean latent space model with three clusters and random receiver effects

```
library(latentnet)
data(sampson)
monks.d3G3r <- ergmm(samplike~euclidean(d=3,G=3)+rreceiver)
```

```
plot(monks.d3G3r, rand.eff="receiver", use.rgl=TRUE, labels=TRUE)
```

## Bayesian ERGM

```
library(Bergm)
data(sampson)
mcmc<-bergm(samplike ~ edges + mutual + transitivities,
  burn.in=20000, #Number of burn-in iterations
  main.iter=100000, #Number of post-burn-in iterations
  aux.iter=10000, #Auxiliary network: number of iterations
  mprior=rep.int(0.0,3), #Prior: means of Gaussian
  sdprior=rep.int(5.0,3), #Prior: standard deviations of Gaussian
  nchains=1 #Number of Markov chains
)
```

## References

- Airoldi, E., Blei, D., Fienberg, S., and Xing, E. (2008), “Mixed Membership Stochastic Block-models,” *Journal of Machine Learning Research*, 9, 1981–2014.
- Atchadé, Y., Lartillot, N., and Robert, C. P. (2012), “Bayesian computation for intractable normalizing constants,” *Brazilian Journal of Statistics*, to appear.
- Batagelj, V., and Mrvar, A. (2003), *Pajek. Program for Large Network Analysis*, University of Ljubljana, Slovenia, Ljubljana.
- Bera, A. K., and Biliyas, Y. (2001a), “On some optimality properties of Fisher-Rao score function in testing and estimation,” *Communications in Statistics—Theory and Methods*, 30, 1533–1559.
- (2001b), “Rao’s score, Neyman’s  $C(\alpha)$  and Silvey’s LM tests: an essay on historical developments and some new results,” *Journal of Statistical Planning and Inference*, 97, 9–44.

- Bernard, H. R., Killworth, P. D., and Sailer, L. (1979), “Informant accuracy in social network data IV,” *Social Networks*, 2, 191–218.
- Besag, J. (1974), “Spatial interaction and the statistical analysis of lattice systems,” *Journal of the Royal Statistical Society, Series B*, 36, 192–225.
- Bhamidi, S., Bresler, G., and Sly, A. (2008), “Mixing time of exponential random graphs,” in *2008 IEEE 49th Annual IEEE Symposium on Foundations of Computer Science*, pp. 803–812.
- Brandes, U., Lerner, J., and Snijders, T. A. B. (2009), “Networks evolving step by step: Statistical analysis of dyadic event data,” in *Advances in Social Network Analysis and Mining*, IEEE, pp. 200–205.
- Britton, T., and O’Neill, P. D. (2002), “Statistical Inference for Stochastic Epidemics in Populations With Network Structure,” *Scandinavian Journal of Statistics*, 29, 375–390.
- Butts, C. T. (2003), “Network inference, error, and informant (in)accuracy: a Bayesian approach,” *Social Networks*, 25, 103–140.
- (2008), “A relational event framework for social action,” *Soc. Meth.*, 38, 155–200.
- (2012), “A Perfect Sampling Method for Exponential Random Graph Models,” Department of Sociology, University of California, Irvine.
- Caimo, A., and Friel, N. (2011), “Bayesian inference for exponential random graph models,” *Social Networks*, 33, 41–55.
- Carrington, P. J., and Scott, J. (2011), “Introduction,” in *The SAGE Handbook of Social Network Analysis*, London: SAGE Publications Inc., Chap. 1, pp. 1–8.
- Celisse, A., Daudin, J.-J., and Pierre, L. (2011), “Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model,” Preprint available at <http://arxiv.org/pdf/1105.3288.pdf>.
- Daudin, J. J., Picard, F., and Robin, S. (2008), “A mixture model for random graphs,” *Statistics and Computing*, 18, 173–183.

- Fellows, I., and Handcock, M. S. (2012), “Exponential-family Random Network Models,” *Technical report, Department of Statistics, University of California, Los Angeles*.
- Frank, O. (1991), “Statistical analysis of change in networks,” *Statistica Neerlandica*, 45, 283–293.
- Frank, O., and Strauss, D. (1986), “Markov graphs,” *Journal of the American Statistical Association*, 81, 832–842.
- Geyer, C. J. (2009), “Likelihood inference in exponential families and directions of recession,” *Electronic Journal of Statistics*, 3, 259–289.
- (2012), *R package rcdd version 1.1-7*.
- Geyer, C. J., and Thompson, E. A. (1992), “Constrained Monte Carlo maximum likelihood for dependent data,” *Journal of the Royal Statistical Society, Series B*, 54, 657–699.
- Gile, K., and Handcock, M. (2010), “Respondent-Driven Sampling: An Assessment of Current Methodology,” *Sociological Methodology*, 40, 285–327.
- Goldenberg, A., Zheng, A., Fienberg, S., and Airoldi, E. (2009), “A survey of statistical network models,” *Foundations and Trends® in Machine Learning*, 2, 129–233.
- Groendyke, C., Welch, D., and Hunter, D. R. (2011), “Bayesian inference for contact networks given epidemic data,” *Scandinavian Journal of Statistics*, to appear.
- (2012), “A Network-based Analysis of the 1861 Hagelloch Measles Data,” *Biometrics*.
- Häggström, O., and Jonasson, J. (1999), “Phase transition in the the random triangle model,” *Journal of Applied Probability*, 36, 1101–1115.
- Handcock, M. (2003), “Assessing degeneracy in statistical models of social networks,” Tech. rep., Center for Statistics and the Social Sciences, University of Washington, <http://www.csss.washington.edu/Papers>.

- Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., and Morris, M. (2012), *ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks*, Seattle, WA, version 3.0-3. Project home page at [statnet.org](http://statnet.org).
- Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007), “Model-Based Clustering for Social Networks,” *Journal of the Royal Statistical Society, Series A*, 170, 301–354, with discussion.
- Hanneke, S., Fu, W., and Xing, E. P. (2010), “Discrete temporal models of social networks,” *Electronic Journal of Statistics*, 4, 585–605.
- Hoff, P. (2003), “Random Effects Models for Network Data,” in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, eds. Breiger, R., Carley, K., and Pattison, P., Washington, D.C.: National Academies Press, pp. 303–312.
- (2005), “Bilinear Mixed-Effects Models for Dyadic Data,” *Journal of the American Statistical Association*, 100, 286–295.
- (2009), “A Hierarchical Eigenmodel for Pooled Covariance Estimation,” *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 71, 971–992.
- (2012), *eigenmodel: Semiparametric factor and regression models for symmetric relational data*, r package version 1.01.
- Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), “Latent Space Approaches to Social Network Analysis,” *Journal of the American Statistical Association*, 97, 1090–1098.
- Holland, P. W., and Leinhardt, S. (1977a), “A Dynamic Model for Social Networks,” *Journal of Mathematical Sociology*, 5, 5–20.
- (1977b), “Social structure as a network process,” *Zeitschrift für Soziologie*, 6, 386–402.
- (1981), “An Exponential Family Of Probability Distributions For Directed Graphs,” *Journal of the American Statistical Association*, 76, 33–65.
- Hummel, R. M., Hunter, D. R., and Handcock, M. S. (2012), “Improving simulation-based algorithms for fitting ERGMs,” *Journal of Computational and Graphical Statistics*, in press.



- Hunter, D., and Lange, K. (2004), “A Tutorial on MM Algorithms.” *The American Statistician*, 58, 30–38.
- Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008), “Goodness of fit of social network models,” *Journal of the American Statistical Association*, 103, 248–258.
- Hunter, D. R., and Handcock, M. S. (2006), “Inference in curved exponential family models for networks,” *Journal of Computational and Graphical Statistics*, 15, 565–583.
- Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M. (2008), “ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks,” *Journal of Statistical Software*, 24, 1–29.
- Jin, I. H., and Liang, F. (2012), “Fitting exponential random graph models using stochastic approximation MCMC,” *Department of Biostatistics, University of Texas*.
- Jonasson, J. (1999), “The random triangle model,” *Journal of Applied Probability*, 36, 852–876.
- Kolaczyk, E. D. (2009), *Statistical Analysis of Network Data: Methods and Models*, Springer.
- Koskinen, J. (2009), “The Linked Importance Sampler Auxiliary Variable Metropolis Hastings Algorithm for Distributions with Intractable Normalising Constants,” Working paper.
- Koskinen, J. H., Robins, G. L., and Pattison, P. E. (2010), “Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation,” *Statistical Methodology*, 7, 366–384.
- Koskinen, J. H., and Snijders, T. A. B. (2007), “Bayesian Inference for Dynamic Social Network Data,” *Journal of Statistical Planning and Inference*, 137, 3930–3938.
- Krivitsky, P., Handcock, M. S., Raftery, A. E., and Hoff, P. (2009), “Representing Degree Distributions, Clustering, and Homophily in Social Networks With Latent Cluster Random Effects Models,” *Social Networks*, 31, 204–213.
- Krivitsky, P. N. (2012), “Exponential-Family Models for Valued Networks,” *Electronic Journal of Statistics*, 6, 1100–1128.

- Krivitsky, P. N., and Handcock, M. (2012a), “A Separable Model for Dynamic Networks,” *Technical report, Department of Statistics, Pennsylvania State University*.
- Krivitsky, P. N., and Handcock, M. S. (2008), “Fitting position latent cluster models for social networks with latentnet,” *Journal of Statistical Software*, 24.
- (2012b), *latentnet: Latent position and cluster models for statistical networks*, r package version 2.4-2.
- Leenders, R. T. A. J. (1995), “Models for Network Dynamics: A Markovian Framework,” *Journal of Mathematical Sociology*, 20, 1–21.
- Liang, F. (2010), “A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants,” *Journal of Statistical Computing and Simulation*, 80, 1007–1022.
- McFadden, D. (1974), “Conditional Logit Analysis of Qualitative Choice Behavior,” in *Frontiers in Econometrics*, ed. Zarembka, P., New York: Academic Press, pp. 105–142.
- (1989), “A method of simulated moments for estimation of discrete response models without numerical integration,” *Econometrica*, 57, 995–1026.
- Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006), “An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants,” *Biometrika*, 93, 451–458.
- Morris, M., Handcock, M. S., and Hunter, D. R. (2008), “Specification of Exponential-family Random Graph Models: Terms and Computational Aspects,” *Journal of Statistical Software*, 24.
- Murray, I. (2007), “Advances in Markov chain Monte Carlo methods,” Ph.D. thesis, University College London, <http://www.cs.toronto.edu/~murray/pub/>.
- Murray, I., Ghahramani, Z., and MacKay, D. J. (2006), “MCMC for doubly-intractable distributions,” in *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, AUAI Press, pp. 359–366.

- Nowicki, K., and Snijders, T. A. B. (2001), “Estimation and Prediction for Stochastic Blockstructures,” *Journal of the American Statistical Association*, 96, 1077–1087.
- Okabayashi, S., and Geyer, C. J. (2012), “Long range search for maximum likelihood in exponential families,” *Electronic Journal of Statistics*, 6, 123–147.
- Pattison, P., and Wasserman, S. (1999), “Logit models and logistic regressions for social networks: II. Multivariate relations,” *British Journal of Mathematical and Statistical Psychology*, 52, 169–193.
- Perry, P. O., and Wolfe, P. J. (2011), “Point process modeling for directed interaction networks,” *Under review*, <http://arxiv.org/abs/1011.1703>.
- Pflug, G. C. (1996), *Optimization of stochastic models. The interface between simulation and optimization*, Boston: Kluwer Academic.
- Raftery, A. E., Niu, X., Hoff, P. D., and Yeung, K. Y. (2011), “Fast Inference for the Latent Space Network Model Using a Case-Control Approximate Likelihood,” *Journal of Computational and Graphical Statistics*, in press.
- Rao, C. R. (1948), “Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation,” *Proceedings of the Cambridge Philosophical Society*, 44, 50–57.
- Rinaldo, A., Fienberg, S. E., and Zhou, Y. (2009), “On the geometry of discrete exponential families with application to exponential random graph models,” *Electronic Journal of Statistics*, 3, 446–484.
- Ripley, R., and Snijders, T. A. B. (2011), *Manual for RSiena*.
- Robbins, H., and Monro, S. (1951), “A Stochastic Approximation Method,” *Annals of Mathematical Statistics*, 22, 400–407.
- Robins, G., Pattison, P., and Wasserman, S. (1999), “Logit models and logistic regressions for social networks, III. Valued relations,” *Psychometrika*, 64, 371–394.

- Salter-Townshend, M., and Murphy, T. (2010), “Variational Bayesian Inference for the Latent Position Cluster Model,” in *NIPS Networks Workshop*.
- Sampson, S. (1968), “A novitiate in a period of change: An experimental and case study of relationships,” Ph.D. thesis, Department of Sociology, Cornell University.
- Schweinberger, M. (2011), “Instability, Sensitivity, and Degeneracy of Discrete Exponential Families,” *Journal of the American Statistical Association*, 106, 1361–1370.
- (2012), *R package hergm version 1.3-6. Hierarchical Exponential-Family Models for Networks*, <http://CRAN.R-project.org/package=hergm>.
- Schweinberger, M. (2012), “Statistical modeling of digraph panel data: goodness-of-fit,” *British Journal of Mathematical and Statistical Psychology*, 65, 263–281.
- Schweinberger, M., and Handcock, M. S. (2011), “Hierarchical exponential-family random graph models: local versus dependence,” Tech. rep., Department of Statistics, Pennsylvania State University.
- Schweinberger, M., Petrescu-Prahova, M., and Vu, D. Q. (2011), “Disaster Response on September 11, 2001 Through the Lens of Statistical Network Analysis,” Tech. rep., Technical report, Department of Statistics, Pennsylvania State University.
- Schweinberger, M., and Snijders, T. A. B. (2003), “Settings in Social Networks: A Measurement Model,” in *Sociological Methodology*, ed. Stolzenberg, R. M., Boston & Oxford: Basil Blackwell, Vol. 33, Chap. 10, pp. 307–341.
- (2007), “Markov models for digraph panel data: Monte Carlo-based derivative estimation,” *Computational Statistics & Data Analysis*, 51, 4465—4483.
- Shalizi, C. R., and Rinaldo, A. (2011), “Consistency under Sampling of Exponential Random Graph Models,” ArXiv:1111.3054v1.
- Shortreed, S., Handcock, M., and Hoff, P. (2006), “Positional estimation within the latent space model for networks,” *Methodology*, 2, 24–33.

- Snijders, T. A. B. (2001), “The Statistical Evaluation of Social Network Dynamics,” in *Sociological Methodology*, eds. Sobel, M., and Becker, M., Boston and London: Basil Blackwell, pp. 361–395.
- (2002), “Markov chain Monte Carlo Estimation of exponential random graph models,” *Journal of Social Structure*, 3, 1–40.
- Snijders, T. A. B., Koskinen, J., and Schweinberger, M. (2010), “Maximum Likelihood Estimation for Social Network Dynamics,” *Annals of Applied Statistics*, 4, 567–588.
- 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., and Nowicki, K. (2007), *Manual for BLOCKS version 1.8*, <http://www.gmw.rug.nl/stocnet/>.
- Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), “New specifications for exponential random graph models,” *Sociological Methodology*, 36, 99–153.
- Snijders, T. A. B., Steglich, C. E. G., and Schweinberger, M. (2007), “Modeling the co-evolution of networks and behavior,” in *Longitudinal models in the behavioral and related sciences*, eds. van Montfort, K., Oud, H., and Satorra, A., Lawrence Erlbaum, pp. 41–71.
- Snijders, T. A. B., Steglich, C. E. G., Schweinberger, M., and Huisman, M. (2010a), *Manual for Siena 3.0*.
- Snijders, T. A. B., van de Bunt, G., and Steglich, C. E. G. (2010b), “Introduction to stochastic actor-based models for network dynamics,” *Social Networks*, 32, 44–60.
- Stephens, M. (2000), “Dealing with label-switching in mixture models,” *Journal of the Royal Statistical Society, Series B*, 62, 795–809.
- Strauss, D. (1986), “On a general class of models for interaction,” *SIAM Review*, 28, 513–527.
- Strauss, D., and Ikeda, M. (1990), “Pseudolikelihood estimation for social networks,” *Journal of the American Statistical Association*, 85, 204–212.

- Tallberg, C. (2005), “A Bayesian approach to modeling stochastic blockstructures with covariates,” *Journal of Mathematical Sociology*, 29, 1–23.
- van Duijn, M., Gile, K., and Handcock, M. (2009), “A Framework for the Comparison of Maximum Pseudo Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models,” *Social Networks*, 31, 52–62.
- van Duijn, M. A. J. (1995), “Estimation of a Random Effects Model for Directed Graphs,” in *Toeval zit overal: programmatuur voor random-coëfficiënt modellen*, eds. Snijders, T. A. B., Engel, B., Van Houwelingen, J. C., Keen, A., Stemerink, G. J., and Verbeek, M., Groningen: IEC ProGAMMA, pp. 113–131.
- van Duijn, M. A. J., Snijders, T. A. B., and Zijlstra, B. J. H. (2004), “P2: a random effects model with covariates for directed graphs,” *Statistica Neerlandica*, 58, 234–254.
- Vu, D. Q., Asuncion, A., Hunter, D. R., and Smyth, P. (2011), “Dynamic egocentric models for citation networks,” in *Proceedings of the 28th International Conference on Machine Learning*.
- Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2012), “Model-based clustering of large networks,” Tech. rep., Department of Statistics, Pennsylvania State University.
- Wang, P., Robins, G., and Pattison, P. (2009), *PNet. Program for the Simulation and Estimation of Exponential Random Graph ( $p^*$ ) Models*.
- Wasserman, S. (1977), “Stochastic models for directed graphs,” Ph.D. thesis, Department of Statistics, University of Harvard.
- (1979), “A Stochastic Model for Directed Graphs with Transition Rates Determined by Reciprocity,” in *Sociological Methodology*, ed. Schuessler, K. F., San Francisco, CA: Jossey-Bass, pp. 392–412.
- (1980), “Analyzing Social Networks as Stochastic Processes,” *Journal of the American Statistical Association*, 75, 280–294.

- Wasserman, S., and Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University 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.
- Westveld, A., and Hoff, P. (2011), “A Mixed Effects Model for Longitudinal Relational and Network Data, with Applications to International Trade and Conflict,” *Annals of Applied Statistics*, 5, 843–872.
- Wyatt, D., Choudhury, T., and Bilmes, J. (2008), “Learning Hidden Curved Exponential Random Graph Models to Infer Face-to-Face Interaction Networks from Situated Speech Data,” *Proceedings of AAAI*.
- Zanghi, H., Picard, F., Miele, V., and Ambroise, C. (2010), “Strategies for online inference of model-based clustering in large and growing networks,” *Annals of Applied Statistics*, 4, 687–714.
- Zijlstra, B. J. H., van Duijn, M. A. J., and Snijders, T. A. B. (2009), “MCMC estimation for the p2 network regression model with crossed random effects,” *British Journal of Mathematical and Statistical Psychology*, 62, 143–166.