Bayesian Analysis for Exponential Random Graph Models Using the Double Metropolis-Hastings Sampler

Article · January 2009

CITATION
1

READS
151

2 authors:

Ick Hoon Jin
University of Notre Dame
18 PUBLICATIONS **95** CITATIONS
SEE PROFILE

Faming Liang
University of Florida
129 PUBLICATIONS **2,073** CITATIONS
SEE PROFILE

Bayesian Analysis for Exponential Random Graph Models Using the Double Metropolis-Hastings Sampler

Ick Hoon Jin and Faming Liang*

July 16, 2009

Abstract

Social network analysis has received much attention in the recent literature. In this paper, we consider a fully Bayesian analysis for the exponential random graph models (ERGMs) using the double Metropolis-Hastings sampler, which overcomes the unknown normalizing constant problem encountered in Metropolis-Hastings simulations by augmenting the state space of the Markov chain to include an auxiliary variable. The method is illustrated using a network of friendship relations among high school students from the National Longitudinal Study of Adolescent Health (AddHealth). The results indicates that for parameter estimation, the new method can significantly outperform the Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE) method. Moreover, we apply the new method to the problems of variable selection for the ERGMs and parameter estimation for the ERGMs with missing edges. To the best of our knowledge, this is the first work that addresses the problem of variable selection for social networks under the Bayesian framework.

Keywords: Exponential Random Graph Model; Social Network; Double Metropolis-Hasting Algorithm; Markov Chain Monte Carlo Maximum Likelihood Estimation; Missing Data; Variable Selection.

*To whom correspondence should be addressed. Liang is Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA. Email: fliang@stat.tamu.edu; Tel: 979-845-8885; Fax: 979-845-3144. Jin is graduate student, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA. Email: kentjin@stat.tamu.edu

1 Introduction

Social network analysis has emerged as a key technique in modern sociology. It has also gained significant followings in biology, communication studies, economics, etc. The exponential family of random graphs is among the most widely-used, flexible models for social network analysis. This family includes edge and dyadic independence models, Markov random graphs (Frank and Strauss, 1986), exponential random graphs (also known as p^* models) (Snijders et al., 2006), and many other models. The model that is of particular interest to current researchers is the exponential random graph model (ERGM). The ERGM is a generalization of the Markov random graph model by incorporating some higher order specifications. Not only does the ERGM show improvements in goodness of fit for various datasets, but also it helps to avoid the problem of near degeneracy that often afflicts the fitting of Markov random graphs. Refer to Robins et al. (2007) for a recent review on the ERGMs.

Consider a social network with n actors. Let X_{ij} be a network tie variable; $X_{ij} = 1$ if there is a network tie from i to j and 0 otherwise. Then, we can specify X as a total set of X_{ij}'s in a matrix, the so-called adjacency matrix. Note that X can be directed or non-directed. Let x_{ij} denote the (i, j)-th entry of an observed matrix x. The likelihood function of the ERGM is given by

$$f(x|\theta) = \frac{1}{\kappa(\theta)} \exp\left\{\sum_{a \in A} \theta_a s_a(x)\right\}, \tag{1}$$

where A denotes the set of configuration types, different sets of configuration types representing different models; $s_a(x)$ is an explanatory variable/statistics, θ_a is the coefficient of $s_a(x)$; $\theta = \{\theta_a : a \in A\}$; and $\kappa(\theta)$ is the normalizing constant which makes (1) a proper probability distribution.

Since $\kappa(\theta)$ depends on the parameter θ, estimation of θ is rather difficult. For example, the Metropolis-Hastings (MH) algorithm (Metropolis *et al.*, 1953; Hastings, 1970) cannot be directly applied to simulate from the posterior of θ, because the acceptance probability would involve the unknown ratio $\kappa(\theta)/\kappa(\theta')$, where θ' denotes the proposed parameter value. To overcome this difficulty, two methods are usually used in the literature, namely, the pseudo-likelihood method (Strauss and Ikeda, 1990) and Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE) method (Snijders, 2002; Hunter and Handcock, 2006). The pseudo-likelihood method works on a simplified, analytic form of the likelihood function under an assumption of dyadic independence. Since the dyadic independence assumption is often violated by real networks, the maximum pseudo-likelihood

estimator can perform very badly in practice and its theoretical properties are poorly understood (Handcock, 2003). The MCMCMLE method originates in Geyer and Thompson (1992), and works as follows. Let $\theta^{(0)}$ denote an arbitrary point in the parameter space of (1), and let $z_1, \ldots, z_m$ denote a sample of random networks simulated from $f(z|\theta^{(0)})$, which can be obtained via a MCMC simulation. Then

$$\log f_m(x|\theta) = \sum_{a \in A} \theta_a s_a(x) - \log(\kappa(\theta^{(0)})) - \log\left(\frac{1}{m}\sum_{i=1}^{m} \exp\left\{\sum_{a \in A} \theta_a S_a(z_i) - \sum_{a \in A} \theta_a^{(0)} s_a(z_i)\right\}\right),$$

approaches to $\log f(x|\theta)$ as $m \to \infty$. The estimator $\widehat{\theta} = \arg\max_\theta \log P_m(x|\theta)$ is called the MCMCMLE of θ. It is known that the performance of the method depends on the choice of $\theta^{(0)}$. If $\theta^{(0)}$ lies in the attractive region of the MLE, the method usually produces a good estimation of θ. Otherwise, the method may converge to a local optimal solution. To resolve the difficulty in choosing $\theta^{(0)}$, Geyer and Thompson (1992) recommended an iterative approach, which drew new samples at the current estimate of θ and then re-estimate:

(a) Initialize with a point $\theta^{(0)}$, usually taking to be the maximum pseudo-likelihood estimator.

(b) Simulate m samples from $f(z|\theta^{(t)})$ using MCMC, e.g., the MH algorithm.

(c) Find $\theta^{(t+1)} = \arg\max_\theta \log f_m(x|\theta^{(t)})$.

(d) Stop if a specified number of iterations has reached, or the termination criterion $\max_\theta f_m(x|\theta^{(t)}) - f_m(x|\theta^{(t)}) < \epsilon$ has reached. Otherwise, go back to step (b).

Even with this iterative procedure, as pointed out by Bartz et al. (2008), nonconvergence is still quite common in the ERGMs. The reason is that the starting point, the maximum pseudo-likelihood estimator, is often too far from the MLE.

Recently, Liang (2009) proposed a double Metropolis-Hasting (DMH) sampler for simulating from the distributions with intractable normalizing constants. The DMH sampler overcomes the unknown normalizing constant problem encountered in the MH simulations by augmenting the state space of the Markov chain to include an auxiliary variable. Refer to Section 3 for the details of the algorithm. The DMH sampler has achieved great successes in spatial statistical models, however, its performance to the social networks is unknown. In this paper, we consider a fully Bayesian analysis for the ERGMs using the DMH sampler. Firstly, we explore the use of the DMH sampler for parameter estimation

and compare it with the MCMCMLE method on some simulated and real datasets. Our numerical results indicate that the new method can produce more accurate estimates than can MCMCMLE. Secondly, we consider the parameter estimation for the ERGMs with missing edges. Thirdly, we consider the variable selection for the ERGMs under the framework of the reversible jump MCMC algorithm (Green, 1995). To the best of our knowledge, this is the first work on Bayesian variable selection for social networks.

The remainder of this paper is organized as follows. In section 2, we give a brief description of the ERGMs. In Section 3, we give a brief review of the DMH sampler. In Section 4, we explore the use of the DMH sampler for parameter estimation for the ERGMs. In section 5, we apply the DMH sampler to parameter estimation for the ERGMs with missing edges. In section 6, we apply the DMH sampler to variable selection for the ERGMs. In section 7, we conclude the paper with a brief discussion.

2 Exponential Random Graph Models

To define explicitly the ERGM, the explanatory statistics $s_a(x)$ need to be specified. Since the number of possible specifications is large, we consider only a few key statistics here, including the edge, degree distribution and shared partnership distribution. Refer to Robins et al. (2007) for a recent review on ERGMs. The edge, denoted by $e(x)$, counts the number of edges in the network. The other two statistics are described as follows.

Degree Distribution Let $D_i(x)$ denote the number of nodes in the network x whose degree, the number of edges incident to the node, equals i. For example, $D_{n-1}(x) = n$ when x is the complete graph and $D_0(x) = n$ when x is the empty graph. Note that $D_0(x), \ldots, D_{n-1}(x)$ satisfy the constraint $\sum_{i=0}^{n-1} D_i(x) = n$, and the number of edges in x can be expressed as

$$e(x) = \frac{1}{2} \sum_{i=1}^{n-1} i D_i(x).$$

The degree distribution statistic (Snijders, 2006; Hunter and Handcock, 2006; Hunter, 2007) is defined as

$$u(x|\tau) = e^{\tau} \sum_{i=1}^{n-2} \left\{ 1 - \left(1 - e^{-\tau} \right)^i \right\} D_i(x), \tag{2}$$

where the additional parameter τ specifies the decreasing rate of the weights put on the higher order terms. This statistic is also called the geometrically weighted degree (GWD) statistic. Following Hunter *et al.* (2008), we set $\tau = 0.25$ throughout this paper. Fixing τ to be a constant is sensible, as $u(x|\tau)$ plays a role of explanatory variable for the ERGMs as specified below.

Shared Partnership Following Hunter and Handcock (2006) and Hunter (2007), we define one type of shared partner statistics, the edgewise shared partner statistic, which are denoted by $EP_0(x)$, $\cdots$, $EP_{n-2}(x)$. The $EP_k(x)$ is the number of unordered pairs (i, j) such that $X_{ij} = 1$ and i and j have exactly k common neighbors. The geometrically weighted edgewise shared partnership (GWESP) statistic is defined as

$$v(x|\tau) = e^{\tau} \sum_{i=1}^{n-2} \left\{ 1 - \left(1 - e^{-\tau}\right)^{i} \right\} EP_i(x), \tag{3}$$

where the parameter τ specifies the decreasing rate of the weights put on the higher order terms. Again, we follow Hunter et al. (2008) to set $\tau = 0.25$ throughout this paper.

Based on the above summary statistics, we consider three ERGMs in this paper. They are defined, respectively, as follows.

$$\begin{aligned}
f(X = x|\theta) &= \frac{1}{\kappa(\theta)} \exp\{\theta_1 e(x) + \theta_2 u(x|\tau)\} \quad \text{(Model 1)} \\
f(X = x|\theta) &= \frac{1}{\kappa(\theta)} \exp\{\theta_1 e(x) + \theta_2 v(x|\tau)\} \quad \text{(Model 2)} \\
f(X = x|\theta) &= \frac{1}{\kappa(\theta)} \exp\{\theta_1 e(x) + \theta_2 u(x|\tau) + \theta_3 v(x|\tau)\} \quad \text{(Model 3)}
\end{aligned}$$

3 The Double Metropolis-Hastings Sampler

The DMH sampler was introduced by Liang (2009) for simulating from the distributions with intractable normalizing constants. Let

$$f(x|\theta) = \frac{1}{\kappa(\theta)} \psi(x, \theta),$$

denote the likelihood function of a model, where θ denotes the parameter of the model, and $\kappa(\theta)$ is an intractable normalizing constant. Suppose that a prior distribution, $\pi(\theta)$, has been specified for θ. The DMH sampler can be described as follows.

Let t denote the index iterations, and let θ_t denote the current state of the Markov chain. The DMH sampler iterates between the following steps.

The DMH Sampler

(a) Simulate a new sample θ' from $\pi(\theta)$ using the MH algorithm starting with θ_t.

(b) Generate an auxiliary variable y from $f(y|\theta')$ through m MH updates starting with x. The probability of transition from x to y is

$$P_{\theta'}^{(m)}(y|x) = K_{\theta'}(x \to x_1) \cdots K_{\theta'}(x_{m-1} \to y) \tag{4}$$

Accept y with probability $\min\{1, r(\theta_t, \theta', y|x)\}$, where

$$r(\theta_t, \theta', y|x) = \frac{f(y|\theta_t)P_{\theta'}^{(m)}(x|y)}{f(x|\theta_t)P_{\theta'}^{(m)}(y|x)} = \frac{f(y|\theta_t)f(x|\theta')}{f(x|\theta_t)f(y|\theta')} = \frac{\psi(y,\theta_t)\psi(x,\theta')}{\psi(x,\theta_t)\psi(y,\theta')}. \tag{5}$$

(c) Set $\theta_{t+1} = \theta'$ if the auxiliary variable is accepted in step (b), and set $\theta_{t+1} = \theta_t$ otherwise.

Since two types of MH updates are performed in step (b), one is for generation of the auxiliary variable y and the other for acceptance of θ', the algorithm is called the double MH sampler by Liang (2009). Note that the second equality of (5) follows from the detailed balance condition $f(x|\theta')P_{\theta'}^{(m)}(y|x) = f(y|\theta')P_{\theta'}^{(m)}(x|x)$. Therefore, (5) holds regardless the value of m, the number of MH updates performed for generating y. When y is generated exactly from $f(y|\theta')$ by an exact sampler (Propp and Wilson, 1996), or equivalently, by an infinite number of MH updates, the algorithm is reduced to the exchange auxiliary-variable algorithm (Murray et al., 2006), which is specially designed for simulating samples from distributions with intractable normalizing constants. Hence, the double MH algorithm can also be viewed as an approximation to the exchange auxiliary-variable algorithm. As mentioned in Liang (2009), a remarkable feature of this algorithm is that it removes the need of exact sampling with a delicate use of the detailed balance condition.

In practice, the value of m is not necessarily large. For all examples of this paper, we generated y by a single cycle of Gibbs sampler (Geman and Geman, 1984) starting with x; that is, setting m to $n(n-1)/2$, the total number of possible edges of an undirected graph. Two or more cycles have been tried for each of the ERGMs specified in Section 2, but the results are similar.

Suppose that a sequence of samples $\theta_1, \cdots, \theta_n$ has been simulated from the posterior distribution

$$\pi(\theta|x) \propto \pi(\theta)f(x|\theta)$$

using the DMH sampler. Then, an approximate Bayesian estimator can be calculated by

$$\widehat{\theta}_B = \frac{1}{n}\sum_{i=1}^{n}\theta_i. \tag{6}$$

For an efficient implementation of the DMH sampler for the ERGMs, two issues need to be considered. The first issue is on the choice of the proposal distribution for generating θ' and the proposal distribution for generating the auxiliary variable y. Like other MCMC algorithms, the proposal should be adjusted to have a reasonable acceptance rate, e.g., 0.2 to 0.4 as suggested by Gelman et al. (1996). In the double MH sampler, a proposed sample θ' will be counted as acceptance only when both the MH moves in steps (a) and (b) are accepted. The second issue is on convergence diagnostics for the sampler. Since the double MH sampler falls into the class of MCMC algorithms, the convergence diagnostic tools developed for general MCMC algorithms are still applicable to the DMH sampler.

4 Parameter Estimation for ERGMs

To conduct a Bayesian analysis for the models, we impose on θ the prior distribution $\pi(\theta) = N_d(0, 10^2 I_d)$, where d is the dimension of the parameter vector θ, and I_d is an identity matrix of size $d \times d$. The value of d varies from model to model. Then, the DMH sampler can be used to simulate from the posterior distribution $\pi(\theta|x)$. In step (a), θ_t is updated by a Gaussian random walk proposal $N_d(\theta_t, s^2 I_d)$, where s is called the step size. In step (b), the auxiliary variable y is generated through a single cycle of Gibbs sampling. For each edge y_{ij}, we set $y_{ij} = 1$ with probability

$$P(y_{ij} = 1|y_{-[i,j]}) = \frac{\exp\left\{\sum_{a\in A} \theta_a S_a(y_{ij} = 1, y_{-[i,j]})\right\}}{\exp\left\{\sum_{a\in A} \theta_a S_a(y_{ij} = 0, y_{-[i,j]})\right\} + \exp\left\{\sum_{a\in A} \theta_a S_a(y_{ij} = 1, y_{-[i,j]})\right\}}, \qquad (7)$$

and 0 with the probability $1 - P(y_{ij} = 1|y_{-[i,j]})$, where $y_{-[i,j]}$ denotes all elements of y except for the entry (i, j). Two or more cycles have also been tried for each of the models, but the results are similar. Since, in step (b), there is no a tunable parameter for adjusting the acceptance rate of the auxiliary network y, the acceptance rate of the DMH sampler is controlled by the value of s. Given a sequence of samples simulated from the posterior distribution by the DMH sampler, an approximate Bayesian estimate of θ can be calculated as in (6).

4.1 A Simulated Example

We first considered a simulated example. The dataset consists of 50 networks, and each network consists of 50 nodes. Each network was simulated using the Gibbs sampler from model 2 with

Statistics	True Value	DMH	MCMCMLE
Edges	-4.5	-4.5866(0.056)	-4.7041(0.060)
GWESP	1.5	1.6886(0.043)	1.9878(0.069)

Table 1: Estimation results for the simulated example. The number in the parentheses denotes the standard error of the corresponding estimate.

the parameters $\theta_1 = -4.5$ and $\theta_2 = 1.5$. The DMH sampler and the MCMCMLE method were both applied to the 50 networks. The DMH sampler was run once for each network with the step size $s = 0.2$. Each run consisted of 10500 iterations, where the first 500 iterations were discarded for the burn-in process, and the remaining iterations were used for estimation. The results were summarized in Table 1. The MCMCMLEs were obtained using a package `statnet`, which is available at `http://cran.r-project.org/web/packages/statnet/index.html`. The comparison indicates that the DMH sampler significantly outperforms the MCMCMLE method for this example. The estimates produced by the DMH sampler are much closer to the true parameter values than those produced by the MCMCMLE method.

4.2 AddHealth Data

In this section, we considered a real network data collected during the first wave (1994-1995) of National Longitudinal Study of Adolescent Health(AddHealth). The data were collected through a stratified sampling survey in the US schools containing grades 7 through 12. To collect the friendship, the school administrator made a roster of all students in each school and asked students to nominate five close male and female friends. Students were allowed to nominate their friends who were not in their school or stop to nominate if they did not have five close male or female students. A detailed description of the dataset can be found in Resnick et al. (1997), Udry and Bearman (1998), or at http://www.cpc.unc.edu/projects/addhealth.

The full dataset contains 86 schools and 90,118 students. In this paper, we analyzed a single school, school 10, which has 205 students. Also, we considered only the undirected network for the case of mutual friendship which means both students A and B check a close friendship each other, although the true data is a directed network.

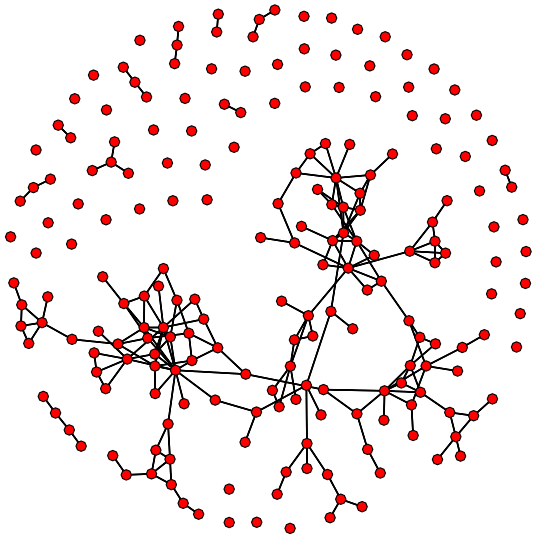

Figure 1: AddHealth Data at School 10

The DMH sampler was first applied to this dataset. It was run 5 times. Each run started with (0,0) and was iterated for 10,500 iterations, where the first 500 iterations were discarded for the burn-in process and the sample collected from the remaining iterations were used for estimation. The results were summarized in Table 2. For comparison, we also included the MCMCMLEs in the table, where the parameter estimates of models 1 and 2 were from Hunter et al. (2008) and those of model 3 were calculated using the package `statnet`. For models 2 and 3, the two methods produced similar estimates, but the estimates produced by the DMH sampler have smaller standard errors. For model 1, the estimates produced by the two methods are quite different. As shown later, the MCMCMLE method fails for model 1; the estimates produced by it may be very far from the true value.

As mentioned before, the DMH sampler falls into the class of MCMC algorithms. The diagnostic tools developed for the MCMC algorithms are also applicable to it. Figure 2 shows the Gelman-Rubin $\hat{R}$-statistic (Gelman and Rubin, 1992) against iterations for model 2. The simulations are usually considered as converged, when $\hat{R}$ falls below 1.1. Figure 2 indicates that the DMH sampler

Methods	Statistics	Model 1	Model 2	Model 3
DMH	edges	$-3.895(0.003)$	$-5.545(0.004)$	$-5.450(0.015)$
	GWD	$-1.563(0.006)$		$-0.131(0.011)$
	GWESP		$1.847(0.004)$	$1.797(0.008)$
MCMCMLE	edges	$-1.423(0.50)$	$-5.280(0.10)$	$-5.266(0.070)$
	GWD	$-1.305(0.20)$		$-0.252(0.173)$
	GWESP		$1.544(0.10)$	$1.635(0.022)$

Table 2: Parameter estimation for the AddHealth data. MCMCMLE: the results of model 1 and model 2 are from Hunter et. al. (2008), and the results of model 3 are from the software ERGM.

converged very fast for this example, usually with a few hundred of iterations. This is similar for the other two models.

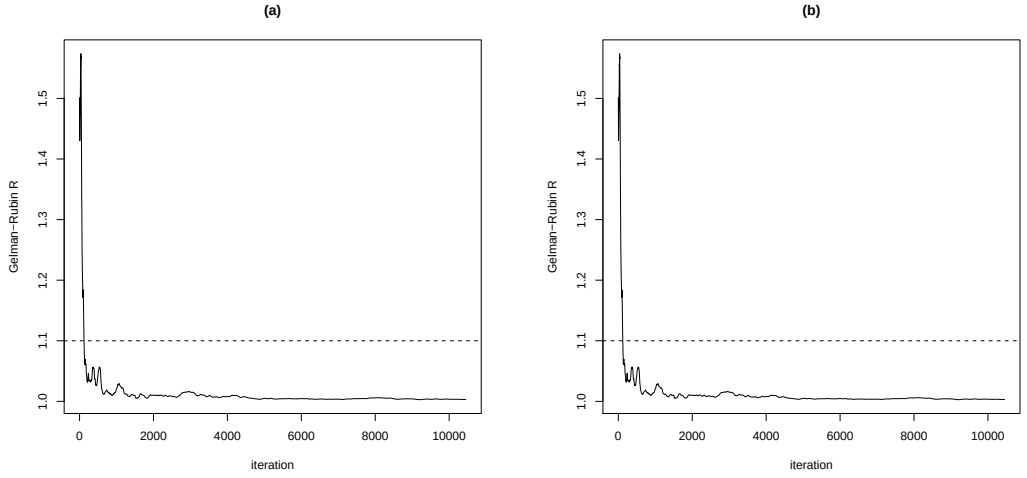

Figure 2: Convergence diagnostic of the DMH sampler for ADDHealth school 10 with model 2. (a) Gelman-Rubin diagnostic based on the sample of θ_1 in independent 5 runs; (b) Gelman-Rubin diagnostic based on the sample of θ_2 in independent 5 runs.

To assess accuracy of the MCMCMLE and DMH estimates, we proposed the following procedure based on the principle of parametric bootstrap methods (Efron and Tibshirani, 1993). Note that the statistics $\{S_a(x) : a \in A\}$ are sufficient for θ. If an estimate $\widehat{\theta}$ is accurate, then $S_a(x)$'s can be reversely estimated by simulated networks from the distribution $f(x|\widehat{\theta})$. The proposed procedure

calculated the root mean squared errors of the estimates of $S_a(x)$'s:

(a) Given the estimate $\hat{\theta}$, simulate m networks, $x_1, \ldots, x_m$, independently using the Gibbs sampler.

(b) Calculate the statistics $S_a(x)$, $a \in A$ for each of the simulated networks.

(c) Calculate RMSE by following equation.

$$RMSE(S_a) = \sqrt{\sum_{i=1}^{m} \left[S_a(x_i) - S_a(x)\right]^2 / m}, \quad a \in A, \tag{8}$$

where $S_a(x)$ is the corresponding statistic calculated from the network x.

For each of the estimates shown in Table 2, the RMSEs were calculated using the above procedure. The results were summarized in Table 3. The results indicate that the DMH sampler produced much more accurate estimates than the MCMCMLE method for all the three models. For model 1, the MCMCMLE method even failed; the corresponding estimates have large standard errors (as shown in Table 2) and very large RMSEs (as shown in Table 3).

Methods	**Coefficients**	**Model 1**	**Model 2**	**Model 3**
Double MH	Edges	22.187	19.204	20.449
	GWD	10.475		9.668
	GWESP		22.094	22.820
MCMCMLE	Edges	4577.2	20.756	22.372
	GWD	90.011		10.045
	GWESP		40.333	30.308

Table 3: Root mean square errors of the MCMCMLE and DMH estimates for the ADDHealth School 10 data.

5 Bayesian Analysis for the ERGMs with Missing Edges

The DMH sampler can be easily applied to the networks with missing edges, because the missing edges can be simply imputed using a MCMC procedure. Let X_{obs} denote the collection of the

observed edges of a network, and let X_{mis} denote the collection of missing edges. Then, the ERGM with missing edges can be written as

$$f(x_{obs}, x_{mis}|\theta) = \frac{1}{\kappa(\theta)} \exp\left\{\sum_{a\in A} \theta_a S_a(x_{obs}, x_{mis})\right\}. \tag{9}$$

Let t denote the index of iterations, and let θ_t denote the current estimate of θ. For the networks with missing edges, the DMH sampler can be run as follows:

(a) Impute the missing edges at the current estimate θ_t using the Gibbs sampler. Denote the imputed edges by $x_{mis}^{(t)}$.

(b) Simulate a new sample θ' from $\pi(\theta)$ using the MH algorithm starting with θ_t.

(c) Generate an auxiliary variable $y \sim P_{\theta'}^{(m)}(y|x_{obs}, x_{mis}^{(t)})$, and accept it with probability $\min\{1, r(\theta_t, \theta', y, x_{mis}^{(t)}|x_{obs})\}$, where the ratio is

$$r(\theta_t, \theta', y, x_{mis}^{(t)}|x_{obs}) = \frac{f(y|\theta_t) f(x_{obs}, x_{mis}^{(t)}|\theta')}{f(x_{obs}, x_{mis}^{(t)}|\theta_t) f(y|\theta')}. \tag{10}$$

(d) Set $\theta_{t+1} = \theta'$ if the auxiliary network is accepted in step (c), and set $\theta_{t+1} = \theta_t$ otherwise.

To illustrate this procedure, we randomly selected 10, 25, and 50 nodes of the AddHealth school 10 network and deleted all edges connected with these nodes, and then applied the above procedure to re-estimate the network parameters. The results were shown in Table 4. For comparison, we also included in Table 4 the estimates for the network without missing edges, which have been reported in Table 2. As expected, the estimates shift away gradually from the estimates without missing edges when the number of missing edges increases.

6 Bayesian Variable Selection for ERGMs

6.1 The algorithm

The DMH sampler can be easily applied to the problem of variable selection for the ERGMs under the framework of reversible jump MCMC (Green, 1995). Let A denote the set of explanatory variables included in the full model, and let $|A|$ denotes the total number of variables in A, where each variable corresponds to a statistic as described in the next subsection. Let M_k denote the current model which

Missing Edges	Coefficients	Model 1	Model 2	Model 3
0	edges	−3.895(0.003)	−5.545(0.004)	−5.450(0.015)
	GWD	−1.563(0.006)		−0.131(0.011)
	GWESP		1.847(0.004)	1.797(0.008)
10	edges	−3.879(0.024)	−5.678(0.027)	−5.336(0.021)
	GWD	−1.816(0.064)		−0.165(0.020)
	GWESP		1.871(0.024)	1.749(0.009)
25	edges	−3.803(0.032)	−5.713(0.034)	−5.158(0.026)
	GWD	−1.816(0.052)		−0.226(0.013)
	GWESP		1.901(0.027)	1.680(0.009)
50	edges	−3.747(0.076)	−5.908(0.050)	−4.960(0.034)
	GWD	−1.910(0.120)		−0.285(0.017)
	GWESP		1.941(0.108)	1.594(0.010)

Table 4: Coefficients and Standard Error of Estimation with Double MH Sampler for Network Data with Missing Variables

includes k variables, and let θ_k denote the coefficients associated with the model. For simplicity, we here imposed a uniform prior on the model space; that is, each model is subject to the same prior probability. Given the current model M_k, with probability $q_{k,b} = 1/3$ ($k \neq 1, |A|$), $q_{1,b} = 2/3$, and $q_{|A|,b} = 0$, we propose to add ("birth") one variable to the current model; with probability $q_{k,d} = 1/3$ ($k \neq 1, |A|$), $q_{1,d} = 0$, and $q_{|A|,d} = 2/3$, we propose to delete ("death") one variable from the current model; and with probability $1 - q_{k,b} - q_{k,d}$ to update the coefficients of the current model.

If we decide to add one variable to the current model, we pick one of the $|A| - k$ variables in $A \setminus M_k$ for addition with equal probability $1/(|A| - k)$. Here, with a slight abuse of notation, we use M_k to denote the set of variables included the model M_k. Say variable S_m is chosen for addition, then we change $k' = k + 1$, set $M_{k'} = M_k + \{S_m\}$, draw a coefficient β_m from $N(0, s^2)$ for the new variable, and denote the new coefficient vector by $\theta' = (\theta_k, \beta_m)$. The proposal is accepted with probability $\min\{1, A(\text{birth})\}$, where

$$A(\text{birth}) = \frac{\pi(\theta')\psi(x, \theta')\psi(y, \theta_k)}{\pi(\theta_k)\psi(x, \theta_k)\psi(y, \theta')} \frac{q_{k',d}}{q_{k,b}} \frac{|A| - k}{k'} \frac{s}{\phi(\beta_m/s)},$$

where $\phi(\cdot)$ denotes the density of the standard normal distribution, and y is an auxiliary network simulated using the Gibbs sampler with parameters θ'. If we decide to delete one variable from the current model, we pick one of the k variables in M_k for deletion with equal probability $1/k$. Say the variable S_m is chosen for deletion, then we change $k' = k - 1$, set $M_{k'} = M_k \setminus \{S_m\}$, and set $\theta' = \theta_k \setminus \{\beta_m\}$, where β_m denotes the coefficient of S_m in the current model. The proposal is accepted with probability $\min\{1, A(\text{death})\}$, where

$$A(\text{death}) = \frac{\pi(\theta')\psi(x,\theta')\psi(y,\theta_k)}{\pi(\theta_k)\psi(x,\theta_k)\psi(y,\theta')} \frac{q_{k',b}}{q_{k,d}} \frac{k}{|A| - k'} \frac{\phi(\beta_m/s)}{s}.$$

If we decide to update the coefficient of the current model, the move is reduced to the DMH sampler as used in Section 4; that is, to accept the proposed move with probability $\min\{1, A(\text{coef. updating})\}$, where

$$A(\text{coef. updating}) = \frac{\pi(\theta')\psi(x,\theta')\psi(y,\theta_k)}{\pi(\theta_k)\psi(x,\theta_k)\psi(y,\theta')} \frac{T(\theta_k|\theta')}{T(\theta'|\theta_k)},$$

where $T(\cdot|\cdot)$ denotes the proposal distribution of θ.

6.2 The nodal covariates

In this section, we extend the model 3 to include some nodal covariates for selection. Following Hunter et al. (2008), we expressed each nodal covariate as a dyadic independence statistic of the form

$$\sum\sum_{i<j} y_{ij} h(X_i, X_j), \tag{11}$$

for a suitably chosen function $h(X_i, X_j)$. Two types of nodal covariates are considered here, which are defined as follows.

Nodal factor effects Given a particular level of a particular factor (i.e., categorical variable), the nodal factor effect counts the number of nodes with that level for each edge in the network. That is,

$$h(X_i, X_j) = \begin{cases} 2 & \text{if both nodes } i \text{ and } j \text{ have the specified factor level,} \\ 1 & \text{if exactly one of } i,\ j \text{ has the specified factor level,} \\ 0 & \text{if neither } i \text{ nor } j \text{ has the specified factor level.} \end{cases} \tag{12}$$

For example, consider the grade factor, which has levels 7-12. The six levels of grade factor requires five separate statistics for the nodal factor effect; one level must be excluded because the sum of all

six equals twice the number of edges in the network, thus creating a linear dependency among the statistics.

Homophily factor effect A homophily statistic for a particular factor gives each edge in the network a score 0 or 1, depending on whether the two end nodes of the edge have matching values of the factor. Two types of homophily factor effects are considered in this paper, the uniform homophily factor and the differential homophily factor. For the uniform homophily factor, we have a single statistic defined by

$$h(X_i, X_j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ have the same level of factor level,} \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

For the differential homophily, we have a set of statistics, one for each level of the factor, with each being defined by

$$h(X_i, X_j) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ have the specified factor level,} \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

For the AddHealth data, the full model includes three network statistics, edges, GWD and GWESP, and 20 nodal covariates related to the features "Grade", " Race", and "Sex". "Grade" has six levels, "Race" has five levels, and "Sex" has two levels. To remove linear dependency, we exclude one level of covariates for each nodal factor. For sex (a two-level factor), we include only the nodal factor and exclude the differential homophily factor. As explained by Hunter *et al.* (2008), in an undirected network, there are only three types of edges – male and male, male and female, and female and female – so that two statistics are enough to show all characteristics of the sexes of both endpoints of an edge. A differential homophily effect (two statistics) plus a nodal factor effect (one statistic) would entail redundant information. In addition, the differential homophily factor for "Black" in "Race" was removed since the corresponding statistic has a value of 0.

For this example, the algorithm was run 5 times independently. Each run consisted of 10500 iterations, where the first 500 iterations were discarded for the burn-in process, and the samples produced in the remaining iterations were used for inference. Table 5 shows the estimates of the marginal inclusion probability of each variable, which is calculated by $\sum_{k=501}^{10500} I(S_r a \in M_k)/10000$ for each $a \in A$, and $I(\cdot)$ is the indicator function. On average, 15 variables are selected at each

Variable	Prob.	Variable	Prob.	Variable	Prob.	Variable	Prob.
Edges	1.000	GWD	0.340	GWESP	1.000	DH(G=7)	1.000
DH(G=8)	1.000	DH(G=9)	0.651	DH(G=10)	0.980	DH(G=11)	1.000
DH(G=12)	0.809	DH(R=W)	0.683	DH(R=H)	0.252	DH(R=N)	1.000
NF(G=8)	0.400	NF(G=9)	0.600	NF(G=10)	0.400	NF(G=11)	0.400
NF(G=12)	1.000	NF(R=W)	0.331	NF(R=H)	0.010	NF(R=N)	0.423
NF(R=O)	0.997	NF(S=M)	0.176	UH(Sex)	0.985		

Table 5: Marginal inclusion probability for the AddHealth School 10 data. The first column gives the name of covariates, and the second column reports the estimates of the marginal inclusion probability of each covariate.

iteration by the DMH sampler. The results are reasonable. For example, Table 5 indicates that the statistic GWD is not necessarily included in the model. This is consistent with Table 3, where model 2 has smaller RMSEs than model 3 and thus better prediction ability.

7 Conclusion

In this paper, we considered a fully Bayesian analysis for the exponential random graph models using the double Metropolis-Hastings sampler. The method was illustrated using a network of friendship relations among high school students from the National Longitudinal Study of Adolescent Health (AddHealth). The results indicates that for parameter estimation, the new method can significantly outperform the MCMCMLE method. Moreover, the new method was applied to the problems of variable selection for the ERGMs and parameter estimation for the ERGHMs with missing edges. To the best of our knowledge, this is the first work that addresses the variable selection problems for social networks under the Bayesian framework.

In this paper, the DMH sampler was only used for undirected networks. Application of DMH to directed networks is straightforward. We note that for the directed networks, the explanatory statistics can be defined as in Robins et al. (2009).

Acknowledgment

Liang's research was supported in part by the grant (DMS-0607755) made by the National Science Foundation and the award (KUS-C1-016-04) made by King Abdullah University of Science and Technology (KAUST).

References

Bartz, K., Blitzstein, J., Liu, J. Monte Carlo Maximum Likelihood for Exponential Random Graph Models: From Snowballs to Umbrella Densities. Technical Report, Department of Statistics, Harvard University.

Efron, B., Tibshirani, R.J.. 1993. An introduction to the bootstrap. *Chapman & Hall.*

Frank, I., Strauss, D. 1986. Markov Graphs. *J. Amer. Statist. Assoc.*, 81, 832-842.

Geman, S., Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." IEEE Trans. Pattern Analysis and Machine Intelligence 6: 721-741.

Geyer, C., Thompson, E. 1992. Constrained Monte Carlo Maximum Likelihood for Dependent Data. *Journal of the Royal Statistical Society. Series B* 54, 657-699.

Gelman, A., Rubin, D. B. 1992. Inference from Iterative Simulation Using Multiple Sequences (with discussion). *Statistical Science*, 7, 457-472.

Goodreau, S. M. 2007. Advances in Exponential Random Graph Models Applied to a Large Social Network. *Social Networks*, 29, 231-248.

Green, P., J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711-732.

Handcock, et. al. 2003. statnet: Software tools for the Statistical Modeling of Network Data. http://statnetproject.org.

Handcock, M. 2003. Statistical Models for Social Networks: Degeneracy and Inference. In: Breiger, R., Carley, K., Pattison, P. (Eds.) Dynamic Social Network Modeling and Analysis. National Academies Press, Washington, DC, pp. 229-240.

Hastings, W. 1970. Monte Carlo Sampling Methods using Markov Chains and Their Applications. *Biometrika* 57, 97-109.

Hunter, D. 2007. Curved Exponential Family Models for Social Network. *Social Networks*, 29, 216-230.

Hunter, D., Handcock, M. 2006. Inference in Curved Exponential Family Models for Network. *Journal of Computational and Graphical Statistics*, 15, 565-583.

Hunter, D., Handcock, M., Butts, C., Goodreau, S., Morris, M. 2008. ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *Journal of Statistical Software*, 24(3).

Liang, F. 2009. A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computing and Simulation*, in press.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. 1953. Equation of State Calculations by Fast Computing Machines. *The Journal of chemical physics*, 21, 1087-1092.

Murray, I. A., Ghahramani, Z., MacKay, D. J. C. 2006. MCMC for doubly-intractable distributions. *Proc. 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI).*

Propp, J. G., Wilson, D. B. 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9, 223-252.

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

Robins, G., Pattison, P., Kalish, Y., Lusher, D. 2007. An Introduction to Exponential Random Graph (p^*) Models for Social Networks. *Social Networks*, 29, 173-191.

Robins, G., Pattison, P., Wang, P. 2009. Closure, connectivity and degree distributions: Exponential random graph (p^*) models for directed social networks. *Social Networks*, 31, 105-117.

Robins, G., Snijers, T., Wang, P., Handcock, M., Pattison, P. 2007. Recent Development in Exponential Random Graph Models for Social Networks. *Social Networks*, 29, 192-215.

Snijders, T.A.B. 2002. Markov Chain Monte Carlo Estimation of Exponential Random Graph Models. *Journal of Social Structure*, 3, 2.

Snijders, T.A.B., Pattison, P. E., Robins, G. L., Handcock, M. S. 2006. New Specifications for Exponential Random Graph Models. *Sociological Methodology*, 99-153.

Strauss, D., Ikeda, M., 1990. Pseudo-likelihood Estimation for Social Network. *J. Amer. Statist. Assoc.*, 82, 204-212.

Udry, J. R., Bearman, P. S. 1998. New Methods for New Research on Adolescent Sexual Behavior, in New Perspectives on Adolescent Risk Behavior, R. Jessor, ed. New York: Cambridge University Press, pp.241-269.