Supplement: Local Dependence in Random Graph Models: Characterization, Properties, and Statistical Inference

Michael Schweinberger

Department of Statistics, Rice University, Houston, TX, USA

Mark S. Handcock

Department of Statistics, University of California, Los Angeles, CA, USA

We focus here on auxiliary-variable Markov chain Monte Carlo updates of $\boldsymbol{\theta}_W$ and $\boldsymbol{Z}$ and provide details concerning α, $\boldsymbol{\mu}_W$, $\boldsymbol{\Sigma}_W^{-1}$, $\boldsymbol{\pi}$, and $\boldsymbol{\theta}_B$ in Supplement A. A basic auxiliary-variable Metropolis-Hastings update of $\boldsymbol{\theta}_W$ and $\boldsymbol{z}$ can be described as follows.

(1) Sample $\boldsymbol{\theta}_W^\star$, $\boldsymbol{Z}^\star$, and $\boldsymbol{Y}^\star$:

(1.1) Sample $\boldsymbol{\theta}_W^\star, \boldsymbol{Z}^\star \mid \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W, \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{Y} = \boldsymbol{y} \sim q(\boldsymbol{\theta}_W^\star, \boldsymbol{z}^\star \mid \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W, \boldsymbol{z}, \boldsymbol{y})$.

(1.2) Sample $\boldsymbol{Y}^\star \mid \boldsymbol{\theta}^\star, \boldsymbol{Z}^\star = \boldsymbol{z}^\star \sim \mathbb{P}_{\boldsymbol{\theta}^\star}(\boldsymbol{Y}^\star = \boldsymbol{y}^\star \mid \boldsymbol{Z}^\star = \boldsymbol{z}^\star)$, where $\boldsymbol{\theta}^\star = (\boldsymbol{\theta}_B, \boldsymbol{\theta}_W^\star)$.

(2) Propose to swap the values of $(\boldsymbol{\theta}_W, \boldsymbol{z})$ and $(\boldsymbol{\theta}_W^\star, \boldsymbol{z}^\star)$ and accept the proposal with probability $\min(1, h)$, where

$$\begin{aligned} h &= \frac{\prod_{k=1}^{K_{max}} p(\boldsymbol{\theta}_{W,k}^\star \mid \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1})}{\prod_{k=1}^{K_{max}} p(\boldsymbol{\theta}_{W,k} \mid \boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W^{-1})} \frac{\mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} = \boldsymbol{z}^\star)}{\mathbb{P}_{\boldsymbol{\pi}}(\boldsymbol{Z} = \boldsymbol{z})} \frac{\mathbb{P}_{\boldsymbol{\theta}^\star}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}^\star)}{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z})} \\ &\times \frac{q(\boldsymbol{\theta}_W, \boldsymbol{z} \mid \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W^\star, \boldsymbol{z}^\star, \boldsymbol{y})}{q(\boldsymbol{\theta}_W^\star, \boldsymbol{z}^\star \mid \boldsymbol{\pi}, \boldsymbol{\theta}_B, \boldsymbol{\theta}_W, \boldsymbol{z}, \boldsymbol{y})} \frac{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y}^\star = \boldsymbol{y}^\star \mid \boldsymbol{Z}^\star = \boldsymbol{z})}{\mathbb{P}_{\boldsymbol{\theta}^\star}(\boldsymbol{Y}^\star = \boldsymbol{y}^\star \mid \boldsymbol{Z}^\star = \boldsymbol{z}^\star)}. \end{aligned} \tag{1}$$

Remark: acceptance probability. The acceptance probability (1) of the auxiliary-variable Metropolis-Hastings update depends on the intractable within-neighborhood log normalising constants through the ratios

$$\frac{\mathbb{P}_{\boldsymbol{\theta}^\star}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z}^\star)}{\mathbb{P}_{\boldsymbol{\theta}^\star}(\boldsymbol{Y}^\star = \boldsymbol{y}^\star \mid \boldsymbol{Z}^\star = \boldsymbol{z}^\star)} \frac{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y}^\star = \boldsymbol{y}^\star \mid \boldsymbol{Z}^\star = \boldsymbol{z})}{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z})}. \tag{2}$$

Since the conditional distributions of $\boldsymbol{Y}$ and $\boldsymbol{Y}^\star$ belong to the same exponential family of distributions, all intractable within-neighborhood log normalising constants in (1) cancel. Therefore, the acceptance probability of the auxiliary-variable Metropolis-Hastings algorithm operating on the augmented state space is tractable, whereas the acceptance probability of Metropolis-Hastings algorithms operating on the original state space is intractable.

Remark: sampling $\boldsymbol{\theta}_W^\star$ and $\boldsymbol{Z}^\star$. In Step (1.1), large moves from $\boldsymbol{\theta}_W$ and $\boldsymbol{z}$ may result in low acceptance rates of the auxiliary-variable Metropolis-Hastings algorithm. We therefore consider local moves from $\boldsymbol{\theta}_W$ and $\boldsymbol{z}$ by changing one or more within-neighborhood parameter vectors or one or more memberships. Local moves from $\boldsymbol{\theta}_W$ may be generated from Gaussians centered at the present values, whereas local moves from $\boldsymbol{z}$ may be generated from the full conditional distributions of memberships. It is worth noting that the full conditional distributions of memberships are not, in general, tractable, because the within-neighborhood log normalising constants $\psi_{W,k}(\boldsymbol{\theta}_{W,k})$ of neighborhoods

E-mail: michael.schweinberger@rice.edu

k with $n_k \gg 5$ nodes are intractable. To construct auxiliary distributions which approximate the full conditional distributions of memberships, we approximate the intractable within-neighborhood log normalising constants $\psi_{W,k}(\boldsymbol{\theta}_{W,k})$ by variational methods (Wainwright and Jordan, 2008). Details are provided in Supplement B.

Remark: sampling $\boldsymbol{Y}^\star$. Two remarks are in place.

First, local moves from $\boldsymbol{\theta}_W$ and $\boldsymbol{z}$ require no more than local sampling of $\boldsymbol{Y}^\star$, i.e., sampling subgraphs. Suppose that the Markov chain considers to move from state $(\boldsymbol{\theta}_W, \boldsymbol{z})$ to state $(\boldsymbol{\theta}^\star_W, \boldsymbol{z}^\star)$, where the two states are identical apart from the value of within-block parameter $\boldsymbol{\theta}_{W,k}$. Then the ratio of the probability masses of $\boldsymbol{y}^\star$ in acceptance probability (1) reduces to

$$\begin{aligned} \frac{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y}^\star = \boldsymbol{y}^\star \mid \boldsymbol{Z}^\star = \boldsymbol{z})}{\mathbb{P}_{\boldsymbol{\theta}^\star}(\boldsymbol{Y}^\star = \boldsymbol{y}^\star \mid \boldsymbol{Z}^\star = \boldsymbol{z}^\star)} &= \frac{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y}^\star = \boldsymbol{y}^\star \mid \boldsymbol{Z}^\star = \boldsymbol{z})}{\mathbb{P}_{\boldsymbol{\theta}^\star}(\boldsymbol{Y}^\star = \boldsymbol{y}^\star \mid \boldsymbol{Z}^\star = \boldsymbol{z})} \\ &= \frac{\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y}^\star_{k,k} = \boldsymbol{y}^\star_{k,k} \mid \boldsymbol{Z}^\star = \boldsymbol{z})}{\mathbb{P}_{\boldsymbol{\theta}^\star}(\boldsymbol{Y}^\star_{k,k} = \boldsymbol{y}^\star_{k,k} \mid \boldsymbol{Z}^\star = \boldsymbol{z})}. \end{aligned} \tag{3}$$

In other words, to evaluate the acceptance probability (1), we need to sample a (small) subgraph rather than the whole (large) graph.

Second, direct sampling of subgraphs is infeasible and exact sampling schemes are not known (but see the work in progress by Butts, 2012). However, Liang (2010) and Wang and Atchade (2014) have demonstrated that it is admissible to sample auxiliary variables by suitable reversible Markov chains with the observed data as initial state and the desired distribution as target distribution. We follow the same approach in the incomplete-data problem considered here, i.e., sample subgraphs by suitable reversible Markov chains. The construction of suitable, reversible Markov chains is discussed by, e.g., Snijders (2002); Hunter and Handcock (2006); Handcock et al. (2010).

A. Markov chain Monte Carlo updates

We discuss updating α, $\boldsymbol{\mu}_W$, $\boldsymbol{\Sigma}_W^{-1}$, $\boldsymbol{\pi}$, and $\boldsymbol{\theta}_B$ and note that updating $\boldsymbol{\theta}_W$ and $\boldsymbol{Z}$ is discussed above.

Parameter α. If the hyper-prior of α is Gamma(A, B) with shape parameter A and inverse scale parameter B, then α can be sampled from its full conditional distribution:

$$\alpha \mid A, B, \boldsymbol{\pi} \sim \text{Gamma}(A + K_{max} - 1, B - \log \pi_{K_{max}}). \tag{4}$$

Within-neighborhood mean vector $\boldsymbol{\mu}_W$. If the hyper-priors of the within-neighborhood mean parameters $\mu_{W,i}$ are independent and given by $N(M_i, S_i^2)$, then we can sample $\mu_{W,i}$ from its full conditional distribution:

$$\mu_{W,i} \mid M_i, S_i^2, \boldsymbol{\Sigma}_W^{-1}, \boldsymbol{\theta}_W \sim N\left(\frac{S_i^{-2} M_i + \sigma_i^{-2} \sum_{k=1}^{K_{max}} \theta_{W,k,i}}{S_i^{-2} + K_{max}\sigma_{W,i}^{-2}}, \frac{1}{S_i^{-2} + K_{max}\sigma_{W,i}^{-2}}\right). \tag{5}$$

Within-neighborhood precision matrix $\boldsymbol{\Sigma}_W^{-1}$. If the hyper-prior of the within-neighborhood precision parameters $\sigma_{W,i}^{-2}$ are independent and given by Gamma(A_i, B_i), then we can sample $\sigma_{W,i}^{-2}$ from its full conditional distribution:

$$\sigma_{W,i}^{-2} \mid A_i, B_i, \boldsymbol{\mu}_W, \boldsymbol{\theta}_W \sim \text{Gamma}\left(A_i + \frac{K_{max}}{2}, B_i + \sum_{k=1}^{K_{max}} \frac{(\theta_{W,k,i} - \mu_{W,i})^2}{2}\right). \tag{6}$$

Membership probability vector $\boldsymbol{\pi}$. The membership probability vector $\boldsymbol{\pi}$ can be sampled from its full conditional distribution by sampling

$$\begin{aligned} V_k^\star \mid \boldsymbol{Z}, \alpha &\overset{\text{ind}}{\sim} \text{Beta}\left(1 + n_k, \alpha + \sum_{j=k+1}^{K_{max}} n_j\right),\ k = 1, \ldots, K_{max} - 1 \\ V_{K_{max}} &= 1, \end{aligned} \tag{7}$$

and setting

$$\begin{aligned} \pi_1 &= V_1^\star \\ \pi_k &= V_k^\star \prod_{j=1}^{k-1} (1 - V_j^\star),\ k = 2, \ldots, K_{max}, \end{aligned} \tag{8}$$

where n_k is the number of nodes in neighborhood k.

Between-neighborhood parameter vector $\boldsymbol{\theta}_B$. A Metropolis-Hastings update of $\boldsymbol{\theta}_B$ generates a proposal $\boldsymbol{\theta}_B^\star$ from a proposal distribution $q(\boldsymbol{\theta}_B^\star \mid \boldsymbol{\theta}_B)$ and accepts the proposal with probability $\min(1, h_B)$, where

$$h_B = \frac{p(\boldsymbol{\theta}_B^\star \mid \boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)}{p(\boldsymbol{\theta}_B \mid \boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)} \frac{\prod_{k<l}^{K_{max}} \mathbb{P}_{\boldsymbol{\theta}^\star}(\boldsymbol{Y}_{k,l} = \boldsymbol{y}_{k,l} \mid \boldsymbol{Z} = \boldsymbol{z})}{\prod_{k<l}^{K_{max}} \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y}_{k,l} = \boldsymbol{y}_{k,l} \mid \boldsymbol{Z} = \boldsymbol{z})} \frac{q(\boldsymbol{\theta}_B \mid \boldsymbol{\theta}_B^\star)}{q(\boldsymbol{\theta}_B^\star \mid \boldsymbol{\theta}_B)}. \tag{9}$$

B. Variational approximations of within-neighborhood log normalising constants

To construct proposal distributions which generate acceptable proposals of memberships (see Section 4.2 of the manuscript), we approximate the intractable full conditional distributions of memberships by approximating the intractable within-neighborhood log normalising constants $\psi_{W,k}(\boldsymbol{\theta}_{W,k})$ by variational methods (Wainwright and Jordan, 2008).

The basic idea is to lower bound $\psi_{W,k}(\boldsymbol{\theta}_{W,k})$ and make the lower bound as tight as possible. A lower bound on $\psi_{W,k}(\boldsymbol{\theta}_{W,k})$ can be derived by Jensen's inequality:

$$\begin{aligned} \psi_{W,k}(\boldsymbol{\theta}_{W,k}) &= \log \sum_{\boldsymbol{y}_{k,k} \in \mathcal{Y}_{k,k}} \frac{\exp[\langle \boldsymbol{\theta}_{W,k}, s_{W,k}(\boldsymbol{y}_{k,k}) \rangle]}{A_{\boldsymbol{\xi}}(\boldsymbol{Y}_{k,k} = \boldsymbol{y}_{k,k})} A_{\boldsymbol{\xi}}(\boldsymbol{Y}_{k,k} = \boldsymbol{y}_{k,k}) \\ &\geq \sum_{\boldsymbol{y}_{k,k} \in \mathcal{Y}_{k,k}} \log \left[\frac{\exp[\langle \boldsymbol{\theta}_{W,k}, s_{W,k}(\boldsymbol{y}_{k,k}) \rangle]}{A_{\boldsymbol{\xi}}(\boldsymbol{Y}_{k,k} = \boldsymbol{y}_{k,k})} \right] A_{\boldsymbol{\xi}}(\boldsymbol{Y}_{k,k} = \boldsymbol{y}_{k,k}), \end{aligned} \tag{10}$$

where $A_{\boldsymbol{\xi}}(\boldsymbol{Y}_{k,k} = \boldsymbol{y}_{k,k})$ belongs to a family of auxiliary distributions with support $\mathcal{Y}_{k,k}$ indexed by a vector of parameters $\boldsymbol{\xi}$. It can be shown that the difference between the left-hand side and right-hand side of (10) is equal to the Kullback-Leibler divergence from the auxiliary PMF to the within-neighborhood PMF. Therefore, making the lower bound (10) as tight as possible is equivalent to minimising the Kullback-Leibler divergence from the auxiliary PMF to the within-neighborhood PMF. Minimising the Kullback-Leibler divergence is feasible provided the auxiliary PMF is fully factorised:

$$A_{\boldsymbol{\xi}}(\boldsymbol{Y}_{k,k} = \boldsymbol{y}_{k,k}) = \prod_{i<j} A_{\xi_{i,j}}(Y_{i,j} = y_{i,j}), \tag{11}$$

where the product is taken with respect to all $i < j$ in neighborhood k and $A_{\xi_{i,j}}(Y_{i,j} = y_{i,j})$ is the Bernoulli$(\xi_{i,j})$ distribution with mean-value parameter $\xi_{i,j} = \mathbb{E}(Y_{i,j})$.

A simple example is given by the hierarchical ERGM of Section 5 of the manuscript. The lower bound (10) can be written as

$$\theta_{W,k,1} \sum_{i<j} \xi_{i,j} + \theta_{W,k,2} \sum_{i<j<h} \xi_{i,j}\, \xi_{jh}\, \xi_{ih} - \sum_{i<j} \left[\xi_{i,j} \log \xi_{i,j} + (1 - \xi_{i,j})\, \log(1 - \xi_{i,j}) \right], \tag{12}$$

where the sums are taken with respect to all $i < j$ and $i < j < h$ in neighborhood k. The lower bound (12) can be maximised by initialising the mean-value parameters $\xi_{i,j}$ and updating them as follows:

$$\xi_{i,j} \quad = \quad \text{logit}^{-1}(\lambda_{i,j}),\ i < j, \tag{13}$$

where $\text{logit}^{-1}(\lambda_{i,j})$ is the inverse logit of $\lambda_{i,j}$ and $\lambda_{i,j}$ is given by

$$\lambda_{i,j} \quad = \quad \theta_{W,k,1} + \theta_{W,k,2} \sum_{h \neq i,j} \xi_{ih}\, \xi_{jh},\ i < j. \tag{14}$$

A variational algorithm along these lines converges to a local maximum of the lower bound (12) (Wainwright and Jordan, 2008, pp. 134–138).

C. Non-identifiability of within-neighborhood parameters and membership indicators

A Bayesian Markov chain Monte Carlo approach along the lines of Section 4 of the manuscript suffers from the so-called label-switching problem (Stephens, 2000). The label-switching problem is rooted in the invariance of the likelihood function to switching the labels of neighborhoods, resulting in non-identifiable within-neighborhood parameters $\boldsymbol{\theta}_{W,1}, \ldots, \boldsymbol{\theta}_{W,K_{max}}$ and membership indicators $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$. As a result, in un-processed Markov chain Monte Carlo samples from the posterior, the labels of neighborhoods may have switched multiple times and statistical inference which depends on the labels of neighborhoods cannot be based on un-processed samples. We follow the Bayesian decision-theoretic approach of Stephens (2000) to undo the label-switching, but introduce a stochastic version of the relabeling algorithm of Stephens (2000), which is based on Simulated Annealing (Liu, 2008) and reduces computing time when K_{max} is moderate or large.

A simple loss function due to Stephens (2000), which can be used when the full conditional probability of the membership of a node depends on the memberships of other nodes, is given by

$$L(\boldsymbol{S}; \boldsymbol{Z}) \quad = \quad \min_{\nu} L_0[\boldsymbol{S}; \nu(\boldsymbol{Z})], \tag{15}$$

where

$$L_0(\boldsymbol{S}; \boldsymbol{Z}) \quad = \quad -\log \prod_{i=1}^{n} q_{ic_i}, \tag{16}$$

where $\boldsymbol{S} = (q_{ik})$, q_{ik} is the probability that node i is reported to be member of neighborhood k, c_i is the index of the neighborhood of which node i is a member, and $\nu(\boldsymbol{Z})$ is a permutation of $\boldsymbol{Z}$. We want to report the $\boldsymbol{S}$ such that the posterior expectation of loss (15) is minimal. The posterior expected loss can be approximated by a Markov chain Monte Carlo sample average and, starting with initial permutations $\nu_1(\boldsymbol{Z}_1), \ldots, \nu_N(\boldsymbol{Z}_N)$ of Markov chain Monte Carlo sample points $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$, be minimised by iterating the following steps:

(1) Minimise $\sum_{m=1}^{N} L_0[\boldsymbol{S}; \nu_m(\boldsymbol{Z}_m)]$ with respect to q_{ik} subject to $\sum_{k=1}^{K_{max}} q_{ik} = 1$, $i = 1, \ldots, n$.

(2) Minimise $L_0[\boldsymbol{S}; \nu_m(\boldsymbol{Z}_m)]$ with respect to ν_m, $m = 1, \ldots, N$.

By an application of Lagrange multipliers, the minimisation in Step (1) amounts to computing q_{ik} as the proportion of times node i is member of neighborhood k under the present permutations $\nu_1(\boldsymbol{Z}_1), \ldots, \nu_N(\boldsymbol{Z}_N)$ of the Markov chain Monte Carlo sample points $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$. Step (2) corresponds to N minimisation problems, each requiring $O(K_{max}!)$ operations, where N is the Markov chain Monte Carlo sample size and $K_{max}!$ is the number of permutations of K_{max} neighborhood labels. Unless K_{max} is small, Step (2) is infeasible. We propose Simulated Annealing (Liu, 2008) to implement Step (2): observe that minimising $L_0[\boldsymbol{S}; \nu_m(\boldsymbol{Z}_m)]$ with respect to ν_m is equivalent to maximising, for any $T > 0$,

$$f(\nu_m; T) \quad = \quad \exp\left\{-\frac{L_0[\boldsymbol{S}; \nu_m(\boldsymbol{Z}_m)]}{T}\right\} \tag{17}$$

with respect to ν_m. Simulated Annealing explores $f(\nu_m; T)$ by Markov chains: given a large value $T_1 > 1$, a Markov chain is constructed to explore $f(\nu_m; T_1)$; then T_1 is decreased to T_2, and a Markov chain is constructed to explore $f(\nu_m; T_2)$; etc. If T decreases monotonically and sufficiently slowly to 0, then Simulated Annealing finds the global minimiser of $L_0[\boldsymbol{S}; \nu_m(\boldsymbol{Z}_m)]$ with probability 1 (Liu, 2008, pp. 209–210). Step (2) requires to run N (parallel) Simulated Annealing algorithms to solve N minimisation problems. Upon convergence of the stochastic relabeling algorithm, the minimising classification probabilities and the minimising permutations can be reported and used to relabel the Markov chain Monte Carlo sample of $\boldsymbol{\theta}_{W,1}, \ldots, \boldsymbol{\theta}_{W,K_{max}}$.

References

Butts, C. T. (2012). A perfect sampling method for exponential random graph models. Department of Sociology, University of California, Irvine.

Handcock, M. S., D. R. Hunter, C. T. Butts, S. M. Goodreau, M. Morris, and P. Krivitsky (2010). *R package ergm version 2.2-2: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks.* http://CRAN.R-project.org/package=hergm.

Hunter, D. R. and M. S. Handcock (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics 15*, 565–583.

Liang, F. (2010). A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computing and Simulation 80*, 1007–1022.

Liu, J. S. (2008). *Monte Carlo Strategies in Scientific Computing.* New York: Springer.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure 3*, 1–40.

Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B 62*, 795–809.

Wainwright, M. J. and M. I. Jordan (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning 1*, 1–305.

Wang, J. and Y. F. Atchade (2014). Bayesian inference of exponential random graph models for large social networks. *Communications in Statistics - Simulation and Computation 43*, 359–377.