STOCHASTIC BLOCK MODELS FOR LATENT NETWORKS AND MULTIPLE TYPES OF DATA TO INFORM EDGES

RICHARD MCELREATH AND DANIEL J. REDHEAD

August 5, 2019

1. Overview

We develop a generative model structure for social network inference in which:

(1) The true network is assumed to be unobserved.
(2) Each node belongs to a possibly unobserved clique (or block).
(3) Probability of a directed edge from node i to node j can be modeled flexibly using any combination of variables, whether at the node or clique (block) level.
(4) Multiple kinds of data can be used simultaneously to inform the network, each having its own parameters to express the association between the data and the underlying true network.

We develop a fully Bayesian estimation solution, using Hamiltonian Monte Carlo, that can be readily modified.

Our initial scientific objective is to investigate the reliability of different methods for eliciting social network ties in human communities. Survey methods are easier than observing the behavioral consequences of ties, such as gifts and instances of helping. But given that we are often interested in predicting helping behavior, to what extent can survey methods provide reliable information?

2. Model

2.1. **Basic structure.** To make the description simpler, first let's consider a model with no individual or block covariates. Assume that a community of N individuals is divided into K blocks. Each individual belongs to only one block. The true (unobserved) network comprises the presence or absence of directed ties between pairs of individuals. The probability of a tie y_{ij} from an individual i in block b_i to an individual j in block b_j is given by the entry

Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

E-mail address: richard_mcelreath@eva.mpg.de.

in the square matrix $B[b_i, b_j]$. For example, suppose there are $K = 3$ blocks. If individuals in the same block are more likely to form ties, the matrix B might be:

$$\begin{bmatrix} 0.5 & 0.1 & 0.1 \\ 0.1 & 0.5 & 0.1 \\ 0.1 & 0.1 & 0.5 \end{bmatrix}$$

However this structure is arbitrary and can be made a function of parameters and data.

The true ties y_{ij} generate observable variables x_{ijvt}, where v is an index for the specific type of observable variable and t is the time point. These x variables can have any arbitrary distribution and relation to the ties y_{ij}. We consider as an example two types.

(1) Survey data for which each i nominates a set of alters j who have either provided aid to i or been given aid by i. Such data may be unreliable, and the reliability may vary by the direction, $i \rightarrow j$ versus $i \leftarrow j$. Note that i's report of j's aid may disagree with j's report of the same relationship.

(2) Behavioral data on directed exchanges $i \rightarrow j$, such as gifts or shares of a resource. These data may be more reliable, but resource constraints may also make it impossible to share with all ties.

For binary data, the model assumes:

$$\begin{aligned} x_{ijvt} &\sim \text{Bernoulli}(p_{ijv}) \\ \text{logit}(p_{ijv}) &= \alpha_v + \beta_v y_{ij} \end{aligned}$$

where α_v is a baseline log-odds of a reported tie or gift, in the absence of a true tie, and β_v is the marginal gain in log-odds when there is a true tie. This allows each variable v to have a unique relationship—or lack of relationship—to the underlying network. These parameters α_v and β_v can also be constructed as functions of (time-varying) covariates specific to the individuals, dyads, or blocks.

All together, the generative model can be expressed:

$$\begin{aligned} x_{ijvt} &\sim \text{Bernoulli}(p_{ijv}) \\ \text{logit}(p_{ijv}) &= \alpha_v + \beta_v y_{ij} \\ y_{ij} &\sim \text{Bernoulli}(B[b_i, b_j]) \\ b_i &\sim \text{Categorical}(\Phi) \\ \Phi &\sim \text{Dirichlet}(\theta) \end{aligned}$$

The elements of the matrix B also require priors. In a typical case, the diagonal elements, which indicate ties within a block, will have higher prior mean than the off-diagonal elements. For example:

$$B_{kk} \sim \text{Beta}(6, 10)$$
$$B_{k\bar{k}} \sim \text{Beta}(1, 10)$$

where kk indicates a diagonal element and $k\bar{k}$ indicates an off-diagonal element.

2.2. **Individual effects.** Individual nodes may have unique tendencies to form ties or receive ties across blocks. Nodes may also have unique α and β effects. We allow these effects by specifying the ties as:

$$y_{ij} \sim \text{Bernoulli}(q_{ij})$$
$$\text{logit}(q_{ij}) = \text{logit}^{-1}\big(B[b_i, b_j]\big) + g_i + r_j$$

where g_i is a parameter that measures i's tendency to form ties and r_j measures j's tendency to receive directed ties. These effects are in addition to the block effect. Individual effects on the observable x variables have similar influence:

$$x_{ijvt} \sim \text{Bernoulli}(p_{ijv})$$
$$\text{logit}(p_{ijv}) = \alpha_v + G_{iv} + R_{jv} + \beta_v y_{ij}$$

where G_{iv} is the tendency of i to perform the behavior in general, R_{jv} is j's tendency to receive the behavior, both independent of the network. How can these effects be interpreted? Suppose for example the behavior is reporting help, and there is a node who reports helping everyone. In that case, G_{iv} would be large. It would similarly be possible to specify individual effects on β_v.

We model the individual effects as standard partially pooled parameters. For example, if there are three x variables, then there are 8 parameters unique to each node. This defines an 8-by-8 covariance matrix. We set priors independently on the scale parameters and correlation matrix.

3. Computation

We implement the statistical model in Stan (mc-stan.org), a library for Hamiltonian Monte Carlo simulation. Stan does not allow discrete parameters, but we recover posterior distributions for the discrete y_{ij} tie and b_i block parameters nevertheless. In this brief section, we explain how.

To compute probabilities of observed variables, we marginalize over the unknown discrete variables y_{ij} and b_i and b_j. Consider the simplest case in which both b_i and b_j are observed. Then the probability of x_{ij} is given by:

$$\begin{aligned}\Pr(x_{ij}|\Theta) = {} & \Pr(y_{ij} = 1|\Theta)\Pr(x_{ij}|\Theta, y_{ij} = 1) \\ & + \Pr(y_{ij} = 0|\Theta)\Pr(x_{ij}|\Theta, y_{ij} = 0)\end{aligned}$$

where Θ is a vector of relevant parameters. Then after sampling, we can recover the posterior distribution of y_{ij} is:

$$\Pr(y_{ij} = 1 | x_{ij}, \Theta) = \frac{\Pr(y_{ij} = 1 | \Theta) \Pr(x_{ij} | y_{ij} = 1, \Theta)}{\Pr(x_{ij} | \Theta)}$$

from Bayes rule. This can be calculated using MCMC samples of the parameters Θ.

The same approach works for the block assignments b_i and b_j. In that case, there are more terms to sum over in the mixture. For example if there are 3 blocks and neither b_i nor b_j are observed, then there will be 6 terms in the mixture probability $\Pr(x_{ij}|\Theta)$:

$$\begin{aligned} \Pr(x_{ij}|\Theta) = & \sum_T \sum_m \sum_n \Pr(b_i = m) \Pr(b_j = n) \\ & \times \Pr(y_{ij} = T | b_i = m, b_j = n) \Pr(x_{ij} | y_{ij} = 1, b_i = m, b_j = n) \end{aligned}$$

omitting the Θ notation in the above for brevity.

Again we recover posterior distributions for b_i by inverting the probabilities post sampling. Note however that computing the posterior probability of each node's b_i requires considering all j alters at once.

$$\Pr(b_i = m | x_{ij}) = \frac{\Pr(b_i = m) \sum_j \Pr(x_{ij} | b_i = m)}{\sum_n \Pr(b_i = n) \sum_j \Pr(x_{ij} | b_i = n)}$$

where $\Pr(x_{ij}|b_i)$ marginalizes over y_{ij} and b_j.

Our Stan code marks the sections corresponding to each of these calculations.

4. Validating the model

Model validation proceeds by first simulating data from the model.