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

RICHARD MCELREATH AND DANIEL J. REDHEAD

August 4, 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

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 in the square

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

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

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, $i \to j$ 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 $i \to j$, 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:

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

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.

3. Validating the model

Model validation proceeds by first simulating data from the model. This requires plugging in values for all of the variables, except for y_{ij} and υ_{ijk}. The y_{ij} values are simulated first as Bernoulli random numbers from the definition of p_{ij}. Then the observable υ_{ijk} values are simulated from the definition of π_{ijk}.

We programmed the statistical model in Stan and drew samples from the posterior distribution of the model and simulated data. This allows us to validate both theoretical usefulness of the approach as well as the validity of our code. Because Stan does not allow sampling for discrete parameters, we used the definition of the posterior distribution $P(y_{ij} = 1|\upsilon_{ij})$ to compute these in Stan's generated quantities block.