A generalized model-based approach to estimating networked population attributes with snowball sampling

Kyle Vincent* and Steve Thompson†

Abstract

We develop a novel Bayesian inference procedure for estimating the size and distribution of networked populations when sampling is based on a complete one-wave snowball design. Our data augmentation procedure is based on an elaborate and generalized graph model that facilitates modeling the erratic clustering effects commonly seen in empirical populations like those comprised of injection drug-users. We evaluate our new strategy over a series of simulated populations and an empirical population of individuals at risk for HIV/AIDS. The estimators from both studies demonstrate that efficient estimates of the size and distribution of the population can be achieved with the new strategy.

Keywords: Bayesian inference; Ignoring unit labels; Markov chain Monte Carlo; Missing data; Multiple imputation; Population size estimator

*Currency Department, Bank of Canada, 234 Wellington Street, Ottawa, Ontario, CANADA, K1A 0G9, email: kvincent@bankofcanada.ca
†Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia, CANADA, V5A 1S6, email: thompson@sfu.ca
1 Introduction

In this article we introduce a novel model-based approach for estimating the size and distribution of networked populations when data is obtained through a complete one-wave snowball sampling design (Frank and Snijders 1994). With the use of a generalized graph model and by deliberately ignoring unit labels (see Scott and Smith (1973) and Cassel et al. (1977) for a discussion on the topic) our inference procedure permits for estimation of the population size and the covariate information that influences the topology of the network graph.

There has been a recent surge in the literature related to the use of link-tracing designs for estimating attributes of networked populations. For example, Thompson and Frank (2000) developed a model-based approach for estimating proportions as well as the pattern of links between members of a networked population when the population size is known and data collection is based on a snowball sampling design. Thompson and Chow (2003) extended on this work by taking a Bayesian approach to overcome some of the difficulties, like obtaining interval estimates, that are commonly seen in the frequentist approach. Handcock and Gile (2010) have developed a theoretical framework for basing inference of population unknowns on exponential random graph models when using an adaptive sampling design.

Approaches using elaborate graph models for inference of networked populations are gaining popularity. For example, Hoff et al. (2002) and Krivitsky et al. (2009) develop a latent space approach to analyzing social networks of known size. In these approaches they posit an underlying graph model from which the probability of links between members of the population depends on their location in an unobserved social
space. Their analytical procedure results in one which gives the audience a visual representation of the population as a networked graph.

Several model-based approaches for estimating population sizes with the use of an adaptive sampling design have been developed for when a subset of the target population is accessible from a sampling frame. Felix-Medina and Thompson (2004) developed an approach that assumes that links from the partial sampling frame are made in a homogenous pattern that facilitates a mark-recapture likelihood style of inference for estimating the number of members located outside of the partial sampling frame. Felix-Medina and Monjardin (2006) extend this work by proposing a Bayesian-assisted approach to overcome some of the difficulties with the maximum likelihood estimators, and Felix-Medina and Monjardin (2009) further extend this work by allowing for a method based on an initial sequential sample that gives control over the final sample sizes.

Empirically, the snowball sampling design has been a popular/natural choice for studying hidden populations, and numerous studies have been based on the use of the sampling design. For example, Frank and Snijders (1994) applied their inference techniques based to a snowball sample of heroin drug-users. Boys et al. (2001) used a snowball sampling method to obtain a sample of young poly-drug users to study reasons for their choice of using psychoactive drugs. More recently, Petersen and Valdez (2011) explored the use of a snowball sampling design to study the behavior of gang-affiliated adolescents.

Kwanisai (2004) developed a Bayesian data augmentation procedure based on the use of a snowball sampling design and a simple graph model, termed the stochastic

3
**block model**, for making inference on hard-to-reach networked populations when the population size is known. In this article, a new, generalized, and more elaborate graph model, termed the *stochastic cluster model*, is proposed to facilitate making inference for both the model parameters and the population size. In the flavor of the latent space approach that [Hoff et al. (2002)](https://doi.org/10.1002/0471868355) and [Krivitsky et al. (2009)](https://doi.org/10.1002/sim.3291) took for modeling the link structure of a network graph, our approach incorporates the covariate information that influences the link structure of the graph via a distance measure. An extended data augmentation routine is developed for making inference on the unknown population size and model parameters corresponding with both the stochastic block model and the stochastic cluster model when the classic one-wave snowball sampling design is employed. With the stochastic cluster model-based approach, our strategy allows the analyst to take into account the covariate information that drives the underlying structure of the graph which in turn facilitates in obtaining more precise estimates of the population size and characteristics.

The article is organized as follows. In Section 2, we introduce the stochastic cluster model, complete one-wave snowball sampling design, and the observed likelihoods based on observing and ignoring unit labels. In Section 3, we provide an overview of the extended data augmentation routine. Section 4 presents the results from a simulation study based on samples obtained from population graphs that are generated from the stochastic cluster model. Section 5 presents the results from a simulation study based on samples obtained from an empirical population of hard-to-reach drug users that are at risk for HIV/AIDS. Section 6 provides a general discussion of the methods and results presented in this article as well as some direction for future
2 Graph Model Setup and Sampling Design, Resulting Likelihoods

The stochastic cluster model is summarized as follows. The stochastic cluster model is a conditional model with three layers. The first layer accounts for more prominent/influential characteristics that can serve to partition the population into homogenous groups/stratum. Conditional on the information provided in the first layer, the second layer accounts for extraneous covariate information. Conditional on the first two layers, the third layer accounts for the link activity between pairs of individuals in the population.

In this section, we outline a specific variation of the stochastic cluster model, namely one that utilizes well-known distributions like the multinomial and the multivariate normal distributions, and the Euclidean distance function to facilitate understanding of the essence of the stochastic cluster model. We refer to this as a variation of the model as it is not necessarily restricted to using such well-known distributions. We also present the complete one-wave snowball sampling design and the observed likelihoods based on observing and ignoring unit labels. Technical Appendix A contains the corresponding mathematical details based on the stochastic block model.
2.1 The stochastic cluster model

Define a population $U$ to be of size $N$ where the unit labels can be indexed as $1, 2, ..., N$. It shall be understood that there is no relationship between the unit labels and the population size.

We will first posit that all units $i = 1, 2, ..., N$ are assigned to a group $C_i \in \{1, 2, ..., G\}$ according to a multinomial distribution based on the vector of parameters $\lambda = (\lambda_1, \lambda_2, ..., \lambda_G)$ where $G$ is the number of groups and $\lambda_j$ is the probability of a unit being assigned to group $j$. In a hard-to-reach population like those comprised of injection drug-users, the group memberships could be based on information related to the type or combination of drugs the members use.

Second, we will posit that, conditional on the population vector of group memberships $C$, each unit $i = 1, 2, ..., N$ independently realizes a $K$-dimensional vector of covariate information $Z_i$ via

$$Z_i|C = Z_i|C_i \sim MVN(\mu_{C_i}, \sum_{C_i}),$$

(1)

where $Z_i = (Z_{1,i}, Z_{2,i}, ..., Z_{K,i})$, $\mu_{C_i} = (\mu_{1,C_i}, \mu_{2,C_i}, ..., \mu_{K,C_i})$ is the centre of group $C_i$’s covariate information and $\sum_{C_i}$ is the $K \times K$ variance-covariance matrix of group $C_i$’s covariate information. In our study we will take the variance-covariance matrix to be $\sum_{C_i} = \sigma^2_{C_i} I_d$ where $\sigma^2_{C_i}$ is the dispersion parameter associated with group $C_i$ and $I_d$ is the identity matrix of size $K \times K$. We will also take $K = 2$, primarily for illustrative purposes. In an injection drug-using population, typical covariate
information could come in the form of gender, age, and other characteristics.

Define \( Y \) to be the symmetric adjacency matrix of the population that records the presence and absence of links between all pairs of units. We will posit that conditional on group memberships \( C \) and the population vector of covariate information \( Z \), links between pairs of units occur independently and in a logistic fashion that depends on the group memberships and covariate information of the pairs of individuals as follows. We introduce the parameters \((\alpha, \beta)\) where if \( C_i = C_j \) then \( \beta_{C_i,C_j} = \beta_0 \) and \( \alpha_{C_i,C_j} = \alpha_0 \), and if \( C_i \neq C_j \) then \( \beta_{C_i,C_j} = \beta_1 \) and \( \alpha_{C_i,C_j} = \alpha_1 \). As the covariate information in this study comes in the form of continuous values, we will make use of the Euclidean norm that measures the distance, in terms of covariate (location) values, between units \( i \) and \( j \), that is \( ||Z_i - Z_j|| \). Define

\[
\chi(C, Z, C^*, Z^*) = \frac{\exp(\beta_{C,C^*} + \alpha_{C,C^*}||Z - Z^*||)}{1 + \exp(\beta_{C,C^*} + \alpha_{C,C^*}||Z - Z^*||)}. \quad (2)
\]

Now, for any \( i, j = 1, 2, ..., N \), if \( i \neq j \) then

\[
P(Y_{ij} = 1 | C, Z) = P(Y_{ij} = 1 | C_i, C_j, Z_i, Z_j) = \chi(C_i, Z_i, C_j, Z_j), \quad (3)
\]

and if \( i = j \) then

\[
P(Y_{ii} = 1) = 0. \quad (4)
\]
Links between individuals in an injection drug-using population could come in the form of nominations based on the sharing of drug using paraphernalia and/or coming into sexual contact.

Under the aforementioned variation of the stochastic cluster model, the likelihood function for the population parameters based on a full graph realization is

\[
L(\lambda, \mu, \sigma^2, \beta, \alpha | C, Z, Y) \propto \prod_{k=1}^{G} \lambda_k^{N_k} \cdot \prod_{i=1}^{N} \mathcal{B}V N(Z_{C_i}; \mu_{C_i}, \sigma^2_{C_i} I_d) \\
\times \prod_{i=1}^{N} \left[ \prod_{Y_{ij} = 1: j > i} \chi(C_i, Z_i, C_j, Z_j) \right] \cdot \prod_{i=1}^{N} \left[ \prod_{Y_{ij} \neq 1: j > i} (1 - \chi(C_i, Z_i, C_j, Z_j)) \right].
\]

(5)

where \( N_k \) is the size of group \( k \).

### 2.2 The complete one-wave snowball sampling design

The complete one-wave snowball sampling design consists of selecting an initial sample \( S_0 \) of size \( n_0 \) through a random sampling design where all units of the population have equal probability of being selected. From the initial sample, all links are traced out to all neighbors. Those neighbors outside the initial sample comprise the first wave, denoted as \( S_1 \), which is of a random size \( n_1 \). The data observed from the sample \( S = S_0 \cup S_1 \) is \( d_0 = \{ S, C_{S}, Z_{S}, Y_{S_0,S}, Y_{S_0,S} \} \) where \( \bar{S} \) is the set of members not selected for the sample and \( Y_{S_0,S} \equiv 0 \). \( C_{S} \) is the vector of the observed group memberships of the sampled members, \( Z_{S} \) is the vector of the observed covariate information of the sampled members, \( Y_{S_0,S} \) is the recorded observations of the presence
or absence of links between the initial sample and the final sample, and \( Y_{S_0, S} \equiv 0 \) is understood to be the absence of links between the initial sample and the unobserved members (for which there is an unknown number of them).

### 2.3 The observed likelihoods based on observing and omitting unit labels

Thompson and Frank (2000) presented the likelihood for the stochastic block model parameters when a snowball sampling design is used. With data collection based on the complete one-wave snowball sampling design the observed likelihood under the stochastic cluster model is expressed as

\[
L_0(N, \Lambda, \mu, \sigma^2, \alpha, \beta \mid d_0) = p(S \mid N, Y_{S_0, U}) \sum_{C_S, Z_S, Y_{S_1 \cup S, S_1 \cup S}} f(C, Z, Y \mid N, \Lambda, \mu, \sigma^2, \alpha, \beta)
\]

\[
= p(S_0 \mid N) \sum_{C_S, Z_S, Y_{S_1 \cup S, S_1 \cup S}} f(C, Z, Y \mid N, \Lambda, \mu, \sigma^2, \alpha, \beta)
\]

\[
\propto \frac{1}{N_{n_0}} \prod_{i \in S_0} \left[ \lambda_{C_i} \cdot \text{BVN}(Z_{C_i} \mid \mu_{C_i}, \sigma_{C_i}^2 I_d) \right]
\]

\[
\times \prod_{j \in S_1} \left[ (\chi(C_i, Z_i, C_j, Z_j))^{Y_{ij}} \cdot (1 - \chi(C_i, Z_i, C_j, Z_j))^{(1 - Y_{ij})} \right]
\]

\[
\times \prod_{j \in S_1} \left[ \lambda_{C_j} \cdot \text{BVN}(Z_j \mid \mu_{C_j}, \sigma_{C_j}^2 I_d) \prod_{i \in S_0} \left[ (\chi(C_i, Z_i, C_j, Z_j))^{Y_{ij}} \cdot (1 - \chi(C_i, Z_i, C_j, Z_j))^{(1 - Y_{ij})} \right] \right]
\]

\[
\times \left[ \sum_{k=1}^{G} \left( \lambda_k \int_{-\infty}^{\infty} \prod_{i \in S_0} (1 - \chi(C_i, Z_i, k, Z)) \text{BVN}(Z \mid \mu_k, \sigma_k^2 I_d) dZ \right)^{N - n_0 - n_1} \right],
\]
where the last component of the likelihood is equivalent to

\[
\prod_{j \in \bar{S}} \left[ \sum_{k=1}^{G} \left( \lambda_k \int_{-\infty}^{\infty} \prod_{i \in S_0} (1 - \chi(C_i, Z_i, k, Z)) \text{BVN}(Z; \mu_k, \sigma_k^2 I_d) dZ \right) \right].
\]

(7)

The observed likelihood based on \(d_0\) will not provide meaningful inference for the population size; consider holding all parameters constant, then the likelihood is a monotonically decreasing function of \(N\) when \(N \geq n_0 + n_1\). Instead, one can ignore unit labels to derive a more suitable likelihood for the population size and model parameters as follows (see Royall (1968) for a discussion on how ignoring unit labels can lead to meaningful inference in some sampling contexts). We shall let \(d_I = \{ C_S, Z_S, Y_{S_0,S}, Y_{S_0,S} \}\) be the corresponding observed data where unit labels are ignored and the adjacency matrix is known only up to permutations of the original observations. As the population can be partitioned into three sets corresponding with \(S_0, S_1,\) and \(\bar{S}\), each of size \(n_0, n_1,\) and \(N - n_0 - n_1,\) respectively, in \(\binom{N}{n_0, n_1, N - n_0 - n_1}\) ways the resulting observed likelihood is

\[
L_0(N, \lambda, \mu, \sigma^2, \alpha, \beta | d_I) \equiv \binom{N}{n_0, n_1, N - n_0 - n_1} L_0(N, \lambda, \mu, \sigma^2, \alpha, \beta | d_0)
\]

\[
\propto \binom{N - n_0}{n_1} \prod_{i=1}^{n_0} \left[ \lambda_{C_i} \cdot \text{BVN}(Z_{C_i}; \mu_{C_i}, \sigma_{C_i}^2 I_d) \right]
\]

10
\[ \times \prod_{i,j=1,2,\ldots,n_0} \left[ (\chi(C_i, Z_i, C_j, Z_j))^{Y_{ij}} \cdot (1 - \chi(C_i, Z_i, C_j, Z_j))^{(1-Y_{ij})} \right] \]

\[ \times \prod_{j=n_0+1}^{n} \left[ \lambda_{C_j} \cdot BVN(Z_j; \mu_{C_j}, \sigma^2_{C_j}) \prod_{i \in S_0} (\chi(C_i, Z_i, C_j, Z_j))^{Y_{ij}} \cdot (1 - \chi(C_i, Z_i, C_j, Z_j))^{(1-Y_{ij})} \right] \]

\[ \times \left[ \sum_{k=1}^{G} \left( \lambda_k \int_{-\infty}^{\infty} \prod_{i \in \bar{S}_0} (1 - \chi(C_i, Z_i, k, Z)) BVN(Z; \mu_k, \sigma^2_k I_d) dZ \right)^{N-n_0-n_1} \right] \]

(8)

where for convenience we make use of the labels \((1, 2, \ldots, n)\) for presenting the corresponding observed likelihood that ignores specific unit labels (that is, only the structure of the observed subset of the graph is retained in the observed likelihood). Notice that, in contrast to the observed likelihood that depends on unit labels, the observed likelihood based on ignoring unit labels is now one that is more suitable for estimating \(N\). The reason is that by definition of the stochastic cluster model, links between \(S_0\) and \(\bar{S}_0\) occur independently between all pairs of units given the corresponding group memberships and covariate information. These Bernoulli-type outcomes can be regarded as independent and, by symmetry, identically distributed. Hence, embedded within Expression (8) is a Binomial type of experiment where a “success” occurs if a unit outside of \(S_0\) is linked to at least one unit in \(S_0\) and is a “failure” otherwise.
3 Data Augmentation

Estimation is carried out using a computational Bayes approach. The Markov Chain Monte Carlo (MCMC) method includes a data augmentation step in which the full set of population node and link values are imputed from the predictive posterior distribution (see Tanner and Wong (1987) for further details on Bayesian data augmentation). In this section we outline the sequential steps taken in our extended data augmentation procedure that incorporates the unknown population size into the routine with the stochastic cluster model. For the reader’s convenience, Technical Appendix A outlines the corresponding extended data augmentation procedure that can be used with the stochastic block model. Mathematical proofs and derivations of the probability mass/density functions of the missing values and posterior distributions of the model parameters that correspond with the stochastic cluster model are provided in Technical Appendix B.

3.1 Distributions of missing data

1) For the purposes of formulating a suitable likelihood for the population size $N$, we consider the reduced data $d_I = \{C_S, Z_S, Y_{S_0,S}, Y_{S_0,S}\}$. It can be shown that with a prior distribution on $N$ to be $\pi(N) \propto 1$ when $N \geq n_0 + n_1$ and $\pi(N) = 0$ when $N < n_0 + n_1$, the posterior distribution of $N$ given the data ignoring the unit labels is
\[
P(N|d_I) \propto \binom{N - n_0}{n_1} (1 - p)^{N - n_0 - n_1} I[N \geq n_0 + n_1]
\] (9)

where I is the indicator function that takes on a value of one if the condition within the brackets holds and zero otherwise, and

\[
1 - p = \sum_{k=1}^{G} \left( \lambda_k \int_{-\infty}^{\infty} \prod_{i=1}^{n_0} (1 - \chi(C_i, Z_i, k, Z)) \text{BVN}(Z; \mu_k, \sigma^2_k I_d) dZ \right).
\] (10)

2) Let \(d_0 = \{S, C_S, Z_S, Y_{s_0, u}\}\) where \(U\) is a hypothetical population of known size. Now, for any \(i \in \bar{S}\), and for any group \(k = 1, 2, ... G\), it can be shown that

\[
P(C_i = k|d_0) = \frac{\lambda_k \cdot \int_{-\infty}^{\infty} \prod_{j=1}^{n_0} (1 - \chi(C_j, Z_j, k, Z)) \text{BVN}(Z; \mu_k, \sigma^2_k I_d) dZ}{\sum_{\ell=1}^{G} \left[ \lambda_\ell \cdot \int_{-\infty}^{\infty} \prod_{j=1}^{n_0} (1 - \chi(C_j, Z_j, \ell, Z)) \text{BVN}(Z; \mu_\ell, \sigma^2_\ell I_d) dZ \right]}.
\] (11)

3) Let \(d_1 = \{S, C, Z_S, Y_{s_0, u}\}\) where \(C\) represents the hypothetical full graph realization of group memberships. For any \(i \in \bar{S}\) and for any \(z^* \in \mathbb{R}^2\), it can be
shown that the density of $Z_i$ at this point is evaluated as

$$
P(Z_i = z^*|d_1) = \frac{\prod_{j=1}^{n_0} (1 - \chi(C_j, Z_j, z^*)) \text{BVN}(z^*; \mu_{C_i}, \sigma^2_{C_i} I_d)}{\int_{-\infty}^{\infty} \prod_{j=1}^{n_0} (1 - \chi(C_j, Z_j, C_i, Z)) \text{BVN}(Z; \mu_{C_i}, \sigma^2_{C_i} I_d) \, dZ}.
$$

(12)

4) Let $d_2 = \{S, C, Z, Y_{S_0,0}\}$. For any $(i, j) \in (\bar{S}_0, \bar{S}_0)$, $i \neq j$ it can be shown that

$$
P(Y_{ij} = 1|d_2) = \chi(C_i, Z_i, C_j, Z_j).
$$

(13)

3.2 Posterior distributions of model parameters

Based on the use of independent and conjugate/non-informative prior distributions for each of the model parameters, the resulting posterior distributions are outlined below. The product of the posterior distributions comprises the (hypothetical) full graph likelihood based on a (hypothetical) full graph realization $d = \{C, Z, Y\}$ and prior distributions.

1) $\lambda$:

$$
\pi(\lambda|d) \sim \text{Dirichlet}(N_1 + 1, \ldots, N_G + 1).
$$

(14)

2) $\sigma^2$:

Let $J_k$ be the units in group $k$. For $k = 1, 2, \ldots, G$, 


\( \pi(\sigma_k^2|d) \sim \Gamma^{-1}(A, B) \), where

\[
A = N_k + 3, \quad \text{and} \\
B = 1 + \frac{1}{2} \left[ \sum_{i \in J_k} Z_{1,i}^2 - \frac{\left( \sum_{i \in J_k} Z_{1,i} \right)^2}{N_k + 1} + \sum_{i \in J_k} Z_{2,i}^2 - \frac{\left( \sum_{i \in J_k} Z_{2,i} \right)^2}{N_k + 1} \right].
\] (15)

3) \( \mu \):

For \( k = 1, 2, \ldots, G \), and \( j = 1, 2, \)

\[
\mu_{j,k}|d \sim N \left( \frac{\sum_{i \in J_k} Z_{j,i}}{N_k + 1}, \frac{N_k}{N_k + 1} \sigma_k^2 \right).
\] (16)

4) \( \alpha \) and \( \beta \):

\[
\pi(\beta_0, \alpha_0, \beta_1, \alpha_1|d) = \\
\frac{e^{\beta_0 + \alpha_0}}{1 + e^{\beta_0 + \alpha_0}} \cdot \left( \frac{1}{1 + e^{\beta_0 + \alpha_0}} \right) \prod_{k=1}^{G} \prod_{i<j} \left[ (\chi(C_i, Z_i, C_j, Z_j))^{Y_{ij}} (1 - \chi(C_i, Z_i, C_j, Z_j))^{(1-Y_{ij})} \right] \\
\times \frac{e^{\beta_1 + \alpha_1}}{1 + e^{\beta_1 + \alpha_1}} \left( \frac{1}{1 + e^{\beta_1 + \alpha_1}} \right) \prod_{k,\ell=1}^{G} \prod_{i<j} \left[ (\chi(C_i, Z_i, C_j, Z_j))^{Y_{ij}} (1 - \chi(C_i, Z_i, C_j, Z_j))^{(1-Y_{ij})} \right].
\] (17)
4 Simulation Study

To explore the model-based features of the new inference strategy (see Thompson and Seber (1996) for a discussion on model-based inference) we evaluated the new strategy as follows. 1000 realizations of a network graph under the stochastic cluster model were obtained where the parameter values were set to be $N = 300, \Lambda = (0.5, 0.3, 0.2), \mu_1 = (-6, -9), \mu_2 = (6, -9), \mu_3 = (0, 5), \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 16, \beta_0 = -1.5, \beta_1 = -2.5$, and $\alpha_0 = -0.5, \alpha_1 = -0.5$. An illustration of a typical realization of a population graph with these model parameters can be found in Figure 1. One sample from each graph was obtained where the initial sample sizes were held constant at 60 (the average final sample size was approximately 148 with a standard deviation of 8). Figure 1 presents two plots that show a typical complete one-wave snowball sample. The data presented in the illustrations reflect the sample data (namely the group status, covariate information, and links) that is required to be observed for the inferential procedures outlined in this article.

From each sample, approximate Bayes estimates of the population size and model parameters were obtained based on Markov chains of length 1000. Non-informative seeds that complement the prior distributions were used. Figure 2 provides four augmented graphs based on the use of the stochastic cluster model with iterations 1, 5, 50, and 1000 that are based on the data collected in the sample presented in Figure 1. Notice how quickly the augmented graphs begin to resemble the true population graph.

Table 1 presents the Bayes estimates of the population size and model parameters. It was found that the Bayes estimates were approximately unbiased for the true
population size and model parameters based on the stochastic block model. Therefore, for network populations that exhibit a structure of a stochastic cluster model type, these results suggest that the strategy based on this reduced model approach may be a robust choice for inferring on the population size and model parameters. The Bayes estimates of the population size and group model parameters based on the stochastic cluster model very nearly coincided with the true values and with a significantly smaller deviation than the corresponding Bayes estimates obtained with the stochastic block model. The Bayes estimates of the parameters governing the behaviour of the covariate information and links between units also came out close to the true values.

5 Empirical Study

Figure 3 presents an empirical population at risk for HIV/AIDS in the Colorado Springs area (Darrow et al., 1999; Klovdahl et al., 1994; Rothenberg et al., 1995). This figure represents the information available for this population, namely the group status of individuals based on drug-using habits and the presence of links between members based on drug-sharing relationships. Therefore only the stochastic block model is utilized in this analysis.

1000 samples each with an initial sample size of 100 were obtained (the average final sample sizes were approximately 227 with a standard deviation of 10). Figure 3 presents two plots that show a typical complete one-wave snowball sample. The data presented in the illustrations reflect the sample data that is required to be
observed for the inferential procedures outlined in this article and that are based on the stochastic block model. Bayes estimates of the population size and model parameters were obtained in a similar fashion to those for the simulation study.

Table 2 presents the Bayes estimates of the population size and model parameters. It was found that the Bayes estimates of the population size came out with a relatively small amount of bias. The estimates for the model parameters based on the group memberships were approximately unbiased.

6 Implications

In this article we have presented our Bayesian inference strategy in full detail and have explored the procedure via a simulation study based on network graphs generated from a complicated network model and an empirical study based on a population at risk for HIV/AIDS. We have demonstrated that our inference strategy results in efficient estimates of population attributes like the population size and those related to the topology of the network.

The methods presented in this article can be extended to the case where the covariate information of the members that comprise the first wave is not observed, that is, when only nominations are required. Kwanisai (2004) outlined a strategy for making inference for the two-group stochastic block model parameters for such a case. Therefore, extending this method to work over the stochastic cluster model when the population size is unknown should be considered. Furthermore, applying a similar strategy to that presented in this article for analyses that involve latent
space approaches where the data is based on a sample of the population should be attempted.

New work based on sampling over additional waves is deserving of attention. Given the novel graph model and inference strategy presented in this article this can be achieved for populations of an unknown size as binomial style likelihoods can be developed by deliberately ignoring unit labels. Imputation procedures based on other variations of the stochastic cluster model should also be explored.

7 Acknowledgements

This work was supported through a Natural Sciences and Engineering Research Council Postgraduate Scholarship PGS D and Discovery Grant. The authors wish to thank Carl Schwarz, Laura Cowen, Richard Lockhart, Charmaine Dean, Chris Henry, and Kim Huynh, for their helpful comments, and John Potterat and Steve Muth for making the Colorado Springs data available.

All views expressed in this manuscript are solely those of the authors and should not be attributed to the Bank of Canada.
References

Boys, A., Marsden, K., and Strang, J. (2001). Understanding reasons for drug use amongst young people: a functional perspective. *Health Education Research* **16**, 457–469.

Cassel, C.-M., Särndal, C.-E., and Wretman, J. H. (1977). *Foundations of inference in survey sampling*. Wiley New York.

Darrow, W. W., Potterat, J. J., Rothenberg, R. B., Woodhouse, D. E., Muth, S. Q., and Klovdahl, A. S. (1999). Using knowledge of social networks to prevent human immunodeficiency virus infections: The colorado springs study. *Sociological Focus* **32**, 143–158.

Felix-Medina, M. and Monjardin, P. (2006). Combining linktracing sampling and cluster sampling to estimate the size of hidden populations: A bayesian assisted approach. *Catalogue no. 12-001-XIE* pages 187–195.

Felix-Medina, M. H. and Monjardin, P. E. (2009). Link-tracing sampling with an initial sequential sample of sites: Estimating the size of a hidden human population. *Statistical Methodology* **6**, 490 – 502.

Felix-Medina, M. H. and Thompson, S. K. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics* **20**, 19–38.

Frank, O. and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics* **10**, 53–67.
Handcock, M. S. and Gile, K. J. (2010). Modeling social networks from sampled data. *Annals of applied statistics* **4**, 5–25.

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.

Klovdahl, A., Potterat, J., Woodhouse, D., Muth, J., Muth, S., and Darrow, W. (1994). Social networks and infectious disease: The colorado springs study. *Social Science & Medicine* **38**, 79 – 88.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks* **31**, 204 – 213.

Kwanisai, M. (2004). *Estimation in link-tracing designs with subsampling*. PhD thesis, The Pennsylvania State University.

Petersen, R. and Valdez, A. (2011). Using snowball-based methods in hidden populations to generate a randomized community sample of gang-affiliated adolescents. *Journal of Peace Research* **48**, 423–435.

Rothenberg, R. B., Potterat, J. J., Woodhouse, D. E., Darrow, W. W., Muth, S. Q., and Klovdahl, A. S. (1995). Choosing a centrality measure: Epidemiologic correlates in the colorado springs study of social networks. *Social Networks* **17**, 273 – 297. Social networks and infectious disease: HIV/AIDS.
Royall, R. (1968). An old approach to finite population sampling theory. *Journal of the American Stat 63*, 1269–1279.

Scott, A. and Smith, T. M. F. (1973). Survey design, symmetry and posterior distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* 35, pp. 57–60.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82, 528–540.

Thompson, S. and Chow, M. (2003). Estimation with link tracing sampling designs a bayesian approach. *Survey Methodology* 29, 197–205.

Thompson, S. K. and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology* 26, 87–98.

Thompson, S. K. and Seber, G. A. F. (1996). *Adaptive Sampling*. Wiley Series in Probability Statistics, New York.
Figure 1: Top: A simulated network population under the stochastic cluster model. The x and y axes refer to the first and second dimension of covariate information, respectively. Bottom: A complete one-wave snowball sample where the initial sample size is 60 and the final sample size is 152. The figure on the left displays the initial sample and links within the initial sample. The figure on the right displays the full sample and links between the initial sample and the first wave.
Figure 2: The augmented missing data for the population graph for augmentation iterations 1, 5, 50, and 1000 under the stochastic cluster model based on the sample presented in Figure I.
Figure 3: Top: The empirical study population of 595 individuals at risk for HIV/AIDS in the Colorado Springs area (Darrow et al., 1999; Klovdahl et al., 1994; Rothenberg et al., 1995) where the dark nodes represent individuals who are injection drug-users and the links between pairs of nodes represent a drug using relationship. Bottom: A complete one-wave snowball sample where the initial sample size is 100 and the final sample size is 225. The figure on the left displays the initial sample and links within the initial sample. The figure on the right displays the full sample and links between the initial sample and the first wave.
Table 1: Simulated study results: Relative bias and standard deviation of parameters based on the stochastic block and cluster model.

| Parameter | Relative Bias | Standard Deviation |
|-----------|---------------|--------------------|
| $N$       | 0.027         | 57.95              |
| $\lambda_1$ | -0.001     | 0.055              |
| $\lambda_2$ | 0.032       | 0.049              |
| $\lambda_3$ | -0.047     | 0.047              |

Stochastic Cluster Model

| Parameter | Relative Bias | Standard Deviation |
|-----------|---------------|--------------------|
| $N$       | 0.000         | 35.71              |
| $\lambda_1$ | 0.016       | 0.045              |
| $\lambda_2$ | 0.001       | 0.042              |
| $\lambda_3$ | -0.043     | 0.041              |

Table 2: Empirical study results: Relative bias and standard deviation of parameters based on the stochastic block model.

| Parameter | Relative Bias | Standard Deviation |
|-----------|---------------|--------------------|
| $N$       | 0.105         | 309.2              |
| $\lambda_1$ | -0.013      | 0.064              |
| $\lambda_2$ | 0.009       | 0.064              |
A  Technical Appendix A

Technical Appendix A provides an outline of the stochastic block model with derivations of the probability mass/density functions of the missing data and posterior distributions for the corresponding model parameters.

A.1  The stochastic block model

We first posit that all units $i = 1, 2, ..., N$ are assigned to a group $C_i \in \{1, 2, ..., G\}$ according to a multinomial distribution based on the vector of parameters $\lambda = (\lambda_1, \lambda_2, ..., \lambda_G)$ where $G$ is the number of groups and $\lambda_j$ is the probability of a unit being assigned to group $j$.

Define $Y$ to be the symmetric adjacency matrix of the population where for all $i, j = 1, 2, ..., N, Y_{ij} = 1$ if a link is present between units $i$ and $j$ and 0 otherwise. Now posit that conditional on the population vector of group memberships $\underline{C} = (C_1, C_2, ..., C_N)$, links occur independently between all pairs of units in the population where for any two units $i, j = 1, 2, ..., N$, if $i \neq j$ then

$$P(Y_{ij} = 1 | \underline{C}) = P(Y_{ij} = 1 | C_i, C_j) = \beta_{C_i, C_j},$$

(18)

and if $i = j$ then

$$P(Y_{ii} = 1) = 0.$$ 

(19)
It shall be understood that for all \( k, \ell = 1, 2, ..., G \), \( \beta_{k, \ell} = \beta_{\ell, k} \).

Under the stochastic block model, the likelihood function for the population parameters based on an entire graph realization is

\[
L(\lambda, \beta | C, Y) \propto \prod_{k=1}^{G} \lambda_{k}^{N_{k}} \prod_{i,j=1: i < j}^{N} \beta_{C_{i}, C_{j}}^{Y_{ij}} (1 - \beta_{C_{i}, C_{j}})^{1 - Y_{ij}}
\]

\[
= \prod_{k=1}^{G} \lambda_{k}^{N_{k}} \prod_{k=1}^{G} \beta_{k,k}^{M_{k,k}} \prod_{k=1}^{G} (1 - \beta_{k,k})^{(N_{k})^{-\frac{1}{2}}} \prod_{k=1}^{G} \beta_{k,\ell}^{M_{k,\ell}} \prod_{k,\ell=1: k < \ell}^{G} (1 - \beta_{k,\ell})^{N_{k}N_{\ell} - M_{k,\ell}}
\]

where \( N_{k} \) is the size of group \( k \), \( M_{k,\ell} \) is the number of links between groups \( k \) and \( \ell \), where \( k, \ell \in \{1, 2, ..., G\} \).

The observed data is \( d_{0} = \{ S, C_{S}, Y_{S_{0}, S}, Y_{S_{0}, \bar{S}} \} \). The observed likelihood based on the data that ignores unit labels, that is \( d_{I} = \{ C_{S}, Y_{S_{0}, U} \} \), can be shown to be

\[
L_{0}(\lambda, \beta | d_{I}) = \left( \frac{N - n_{0}}{n_{1}} \right)^{n_{0}} \prod_{i=1}^{n_{1}} \left[ \lambda_{C_{i}} \right] \times \left[ \prod_{i,j=1,2,...,n_{0}: i < j}^{n} \beta_{C_{i}, C_{j}}^{Y_{ij}} (1 - \beta_{C_{i}, C_{j}})^{1 - Y_{ij}} \right] \times \left[ \prod_{j=n_{0}+1}^{n} \left( \lambda_{C_{j}} \prod_{i=1}^{n_{0}} \beta_{C_{i}, C_{j}}^{Y_{ij}} (1 - \beta_{C_{i}, C_{j}})^{1 - Y_{ij}} \right) \right] \times \left[ \sum_{j=1}^{G} \lambda_{j} \prod_{i=1}^{n_{0}} (1 - \beta_{C_{i}, j}) \right]^{N - n_{0} - n_{1}}.
\]
A.2 The extended data augmentation procedure based on the stochastic block model

Kwanisai (2004) developed proofs of the results presented in the following subsections for the two-group stochastic block model. We have omitted the proofs since they are a straightforward extension of those based on the two-group stochastic block model. In this section it shall be understood that all probability mass and density functions are conditional on the (most recently sampled) model parameters.

A.2.1 The probability mass function of the population size

First consider $d_I = \{S, C_S, Y_{S_0,s}, Y_{S_0,\bar{s}}\}$. Let $\pi(N) \propto 1$ when $N \geq n_0 + n_1$ and $\pi(N) = 0$ when $N < n_0 + n_1$. By Expression (21) the resulting posterior distribution of $N$ is

$$P(N|d_I) = \frac{P(d_I|N)P(N)}{P(d_I)}$$

$$= \frac{(N-n_0)(1-p)^{N-n_0-n_1}I[N \geq n_0 + n_1]}{\sum_{N' \geq n_0+n_1} \binom{N'-n_0}{n_1} (1-p)^{N'-n_0-n_1}}$$

$$\propto \binom{N-n_0}{n_1} (1-p)^{N-n_0-n_1} I[N \geq n_0 + n_1] \tag{22}$$

where
\[1 - p = \sum_{k=1}^{G} \left( \lambda_k \prod_{i=1}^{n_0} (1 - \beta_{C_i,k}) \right).\]  

(23)

### A.2.2 Augmentation of the missing values

Let \(d_0 = \{S, C_S, Y_{S_0}, U\}\), where \(U\) is a hypothetical population of size equal to the most recent \(N\) that was sampled from its probability distribution. For any \(i \in \bar{S}\) and \(k = 1, 2, \ldots, G\),

\[
P(C_i = k|d_0) = \frac{\lambda_k \prod_{j=1}^{n_0} (1 - \beta_{C_j,k})}{\sum_{\ell=1}^{G} \left( \lambda_\ell \prod_{j=1}^{n_0} (1 - \beta_{C_j,\ell}) \right)}.
\]

(24)

Values of the missing group memberships \(C_S\) are then assigned independently according to the distribution outlined in Expression (24).

After generating \(C_S\), the graph data is updated from \(d_0\) to \(d_1\) where \(d_1 = \{S, C, Y_{S_0}, U\}\) and \(C\) represents the full hypothetical graph realization of group memberships. Now, for any \(i, j \in S_0\) where \(i \neq j\),

\[
P(Y_{ij} = 1|d_1) = \beta_{C_i,C_j},
\]

(25)

and hence links between each pair of units \(i, j \in \bar{S}_0\) for \(i \neq j\) are assigned according to the probability distribution found in Expression (25) to generate a hypothetical full
graph realization of $Y$. This gives a hypothetical full graph realization $d = \{C, Y\}$.

A.2.3 The posterior distributions of the stochastic block model parameters based on the augmented data

The factorization theorem asserts that, with the use of independent prior distributions on the model parameters, the posterior distributions of the parameters are all independent under the (hypothetical) full graph realization $d = \{C, Y\}$. In our study, we shall place independent conjugate Dirichlet and Beta priors on $\lambda$ and $\beta$, respectively. That is,

\[
\pi(\lambda) \propto \prod_{k=1}^{G} \lambda_k^{\alpha_k - 1}\]

and

\[
\pi(\beta) \propto \prod_{k, \ell=1: k \leq \ell}^{G} \left( \beta_{k, \ell}^{\gamma_1 - 1}(1 - \beta_{k, \ell})^{\gamma_2 - 1} \right). \tag{27}
\]

We shall take the prior distributions to be noninformative by setting $\alpha_k = 1$ for $k = 1, 2, \ldots, G$ and $\gamma_j = 1$ for $j = 1, 2$. The resulting posterior distribution of $\lambda$ is then

\[
\pi(\lambda|d) \sim \text{Dirichlet}(N_1 + 1, \ldots, N_G + 1). \tag{28}
\]
The resulting posterior distribution of $\beta_{k,\ell}$ for $k, \ell = 1, 2, ..., G, k \neq \ell$, is

$$\pi(\beta_{k,\ell}|d) \sim \text{Beta}(M_{k,\ell} + 1, N_k N_\ell - M_{k,\ell} + 1),$$  \hspace{1cm} (29)$$

and for $k = 1, 2, ..., G$,

$$\pi(\beta_{k,k}|d) \sim \text{Beta}(M_{k,k} + 1, \left(\frac{N_k}{2}\right) - M_{k,k} + 1).$$  \hspace{1cm} (30)$$
B Technical Appendix B

Technical Appendix B provides the derivations of the probability mass/density functions of the missing data and the posterior distributions for the stochastic cluster model parameters.

B.1 Probability Mass/Density Functions of the Population Size and Missing Values

In this section it shall be understood that all probability mass and density functions are conditional on the (most recently sampled) model parameters.

B.1.1 The probability mass function of the population size

Recall that when deliberately ignoring unit labels the data observed is \( d_I = \{C_S, Z_S, Y_{S_0,S}, Y_{S_0,S}\} \). For a choice of prior distribution on the population size \( N \) we shall take \( \pi(N) \propto 1 \) when \( N \geq n_0 + n_1 \) and \( \pi(N) = 0 \) when \( N < n_0 + n_1 \). Now, it can be shown that

\[
P(N|d_I) = \frac{P(d_I|N)P(N)}{P(d_I)} = \frac{\binom{N-n_0}{n_1} (1-p)^{N-n_0-n_1} I[N \geq n_0 + n_1]}{\sum_{N' \geq n_0 + n_1} \binom{N'-n_0}{n_1} (1-p)^{N'-n_0-n_1}} \
\propto \binom{N-n_0}{n_1} (1-p)^{N-n_0-n_1} I[N \geq n_0 + n_1]
\]

(31)
where

\[ 1 - p = \sum_{k=1}^{G} \left( \lambda_k \int_{-\infty}^{\infty} \prod_{i=1}^{n_0} \left( 1 - \frac{\exp(\beta_{C_i,k} + \alpha_{C_i,k} \mid Z_i - Z \mid)}{1 + \exp(\beta_{C_i,k} + \alpha_{C_i,k} \mid Z_i - Z \mid)} \right) \text{BVN}(Z; \mu_k, \sigma_k^2 I_d) \mathrm{d}Z \right) \]

(32)

B.1.2 The probability mass function of group memberships

After sampling an \( N \) from the binomial style posterior distribution, we shall make use of the labels solely for augmentation purposes and continue the data augmentation process with the observed graph data

\[ d_0 = \{ S, C_S, Z_S, Y_{S_0,U} \} \]

(33)

where \( U \) is a hypothetical population of size \( N \).

The probability of obtaining a realization of \( C_S \) given \( d_0 \) is obtained as

\[
P(C_S|d_0) = \frac{P(S, C_S, Z_S, Y_{S_0,U}) \cdot P(C_S)}{P(d_0)} \]

\[
= \frac{P(S|C_S, Z_S, Y_{S_0,U}, C_S) \cdot P(C_S, Z_S, Y_{S_0,U}|C_S) \cdot P(C_S)}{P(d_0)} \]

\[
= \frac{P(S|C_S, Z_S, Y_{S_0,U}) \cdot P(Y_{S_0,U}|C_S, Z_S, C_S) \cdot P(C_S, Z_S, |C_S) \cdot P(C_S)}{P(d_0)} \]
\[
= \frac{P(S|C_S, Z_S, Y_{S_{0}}, U)}{P(d_0)} \cdot P(Y_{S_{0}}, U|C, Z_S) \cdot P(Z_S|C_S, C) \cdot P(C_S|C) \cdot P(C) \\
= \frac{P(S|C_S, Z_S, Y_{S_{0}}, U)}{P(d_0)} \cdot P(Y_{S_{0}}, U|C, Z_S) \cdot P(Z_S|C) \cdot P(C).
\] (34)

Note that \(C_S\) is dropped from the first term since the adaptive sampling design only depends on the information collected in the sample (as outlined in Thompson and Seber (1996)). We now have,

\[
P(C_S|d_0) = \frac{P(S|C_S, Z_S, Y_{S_{0}}, U)}{P(d_0)} \cdot P(Y_{S_{0}}, U|C, Z_S) \cdot P(Z_S|C_S) \cdot P(C) \\
= \frac{P(S|C_S, Z_S, Y_{S_{0}}, U)}{P(d_0)} \cdot P(Y_{S_{0}}, U|C, Z_S) \cdot P(Z_S|C) \cdot \prod_{i=1}^{N} P(C_i) \\
= \frac{P(S|C_S, Z_S, Y_{S_{0}}, U)}{P(d_0)} \cdot P(Y_{S_{0}}, U|C, Z_S) \cdot \prod_{i=1}^{n} \text{BVN}(Z_i; \mu_{C_i}, \sigma_{C_i}^2 I_d) \cdot \prod_{i=1}^{N} P(C_i),
\] (35)

where

\[
P(Y_{S_{0}}, U|C, Z_S) = \\
\prod_{i=1}^{n_0} \prod_{j=1: i < j}^{n} P(Y_{ij}|C_i, C_j, Z_i, Z_j) \cdot \prod_{k=n+1}^{N} \int_{Z_k} \prod_{i=1}^{n_0} P(Y_{ik}|C_i, C_k, Z_i, Z_k) \text{BVN}(Z_k; \mu_k, \sigma_k^2 I_d) \, dZ_k.
\] (36)

Therefore, by the factorization theorem we have that
\[ P(C_s|d_0) = \prod_{i \in \bar{S}} P(C_i|d_0) \]
\[ = \prod_{i \in \bar{S}} P(C_i|S, C_s, Z_s, Y_{s_0, i}) \]
\[ = \prod_{i \in \bar{S}} P(C_i|S, C_{s_0}, Z_{s_0}, Y_{s_0, i}) . \] (37)

Again, we note that the last equality comes from the use of the adaptive sampling design [Thompson and Seber 1996].

Now, take any \( i \in \bar{S} \). Then for any group \( k = 1, 2, \ldots G \),

\[ P(C_i = k|S, C_{s_0}, Z_{s_0}, Y_{s_0, i}) \]
\[ = \frac{P(C_i = k, S, C_{s_0}, Z_{s_0}, Y_{s_0, i})}{\sum_{\ell=1}^{G} P(C_i = \ell, S, C_{s_0}, Z_{s_0}, Y_{s_0, i})} \]
\[ = \frac{\lambda_k \cdot \int \prod_{j=1}^{n_0} \left( 1 - \frac{\exp(\beta_{C_j, k} + \alpha_{C_j, k}||Z_j - \mu||)}{1 + \exp(\beta_{C_j, k} + \alpha_{C_j, k}||Z_j - \mu||)} \right) \text{BVN}(Z; \mu_k, \sigma_k^2 I_d) \, dZ}{\sum_{\ell=1}^{G} \left[ \lambda_{\ell} \cdot \int \prod_{j=1}^{n_0} \left( 1 - \frac{\exp(\beta_{C_j, \ell} + \alpha_{C_j, \ell}||Z_j - \mu||)}{1 + \exp(\beta_{C_j, \ell} + \alpha_{C_j, \ell}||Z_j - \mu||)} \right) \text{BVN}(Z; \mu_\ell, \sigma_\ell^2 I_d) \, dZ \right]} . \] (38)
Values of the missing group memberships $C_{s}$ are then assigned independently according to the distribution outlined in Expression (38).

### B.1.3 The probability density function of the covariate information

For a specific realization $Z_{s}$, the density given $d_{1} = \{S, C, Z_{s}, Y_{s_0, u}\}$ is evaluated as

\[
P(Z_{s} | d_{1}) = \frac{P(Z_{s} | S, C, Z_{s}, Y_{s_0, u})}{P(d_{1})} = \frac{P(S, C, Z_{s}, Y_{s_0, u} | Z_{s}) \cdot P(Z_{s})}{P(d_{1})} = \frac{P(S | C_{s}, Z_{s}, Y_{s_0, u}) \cdot P(Y_{s_0, u} | C_{s}, Z_{s}) \cdot P(C_{s} | Z_{s}) \cdot P(Z_{s})}{P(d_{1})} \cdot \prod_{i=1}^{N} \text{BVN}(Z_{i}; \mu_{C_{i}}, \sigma_{C_{i}}^{2} I_{d}) \cdot \prod_{i=1}^{N} P(C_{i}) \cdot \prod_{j=1}^{n_{0}} \prod_{i=1}^{N} P(Y_{ij} | C_{i}, C_{j}, Z_{i}, Z_{j}) \cdot \prod_{i=1}^{N} \text{BVN}(Z_{i}; \mu_{C_{i}}, \sigma_{C_{i}}^{2} I_{d}) \cdot \prod_{i=1}^{N} P(C_{i}).
\]

(39)

Note that $Z_{s}$ is dropped in the first term by use of the adaptive sampling design \cite{Thompson and Seber 1996}. Once again, by the factorization theorem, we have
that

\[
P(Z_{\bar{S}} | d_1) = \prod_{i \in \bar{S}} P(Z_i | d_1)
\]

\[
= \prod_{i \in \bar{S}} P(Z_i | S, C_i, Z_S, Y_{S_0}, U)
\]  \hspace{1cm} \text{(40)}

Now, take any \(i \in \bar{S}\) and \(z^* \in \mathbb{R}^2\). The density at this point is evaluated as

\[
P(Z_i = z^* | d_1) = P(Z_i = z^* | S, C_i, Z_S, Y_{S_0}, i)
\]

\[
= \frac{P(Z_i = z^* | S, C_i, Z_S, Y_{S_0}, i)}{\int_{-\infty}^{\infty} P(Z_i = Z, S, C_i, Z_S, Y_{S_0}, Z_i = z^*) dZ}
\]

\[
= \frac{\prod_{j=1}^{n_0} P(Y_{ij} | C_j, C_i, Z_j, Z_i = z^*) \text{BVN}(z^*; \mu_{C_i}, \sigma_{C_i}^2 I_d)}{\int_{-\infty}^{\infty} \prod_{j=1}^{n_0} P(Y_{ij} | C_j, C_i, Z_j, Z_i = Z) \text{BVN}(Z; \mu_{C_i}, \sigma_{C_i}^2 I_d) dZ}
\]

\[
= \prod_{j=1}^{n_0} \left(1 - \frac{\exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i} \|Z_j - z^*\|)}{1 + \exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i} \|Z_j - z^*\|)} \right) \text{BVN}(z^*; \mu_{C_i}, \sigma_{C_i}^2 I_d)
\]

\[
= \prod_{j=1}^{n_0} \left(1 - \frac{\exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i} \|Z_j - Z\|)}{1 + \exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i} \|Z_j - Z\|)} \right) \text{BVN}(Z; \mu_{C_i}, \sigma_{C_i}^2 I_d) dZ
\]  \hspace{1cm} \text{(41)}

Values of the covariate information \(Z_{\bar{S}}\) are then assigned independently according to the distribution outlined in Expression (41).
B.1.4 The probability density function of the links

Let \( d_2 = \{S, C, Z, Y_{S_0,U}\} \), where \( Z \) is the hypothetical full graph realization of covariate information. Recall that by definition of the model, we have that conditional on \( C, Z \), for any \((i, j), (i^*, j^*) \in (\bar{S}_0, \bar{S}_0)\), \( Y_{ij} \) is independent of \( Y_{i^*,j^*} \). Hence, for any \((i, j) \in (\bar{S}_0, \bar{S}_0), i \neq j\), we have that

\[
P(Y_{ij} = 1|C_i, C_j, Z_i, Z_j) = \left( \frac{\exp(\beta_{C_i,C_j} + \alpha_{C_i,C_j}||Z_i - Z_j||)}{1 + \exp(\beta_{C_i,C_j} + \alpha_{C_i,C_j}||Z_i - Z_j||)} \right).
\] (42)

Hence links between each pair of units \(i, j \in \bar{S}_0\) for \( i \neq j \) are assigned according to the probability distribution found in Expression (42) to generate a hypothetical full graph realization of \( Y \). This gives a hypothetical full graph realization \( d = \{C, Z, Y\} \).

B.2 The Posterior Distributions of the Model Parameters

All derivations in this section are based on a (hypothetical) full graph realization \( d = \{C, Z, Y\} \). With the use of independent prior distributions for the model parameters the factorization theorem asserts that the set of parameters \( \lambda, (\mu, \sigma^2), (\beta_0, \alpha_0), \) and \( (\beta_1, \alpha_1) \) are all independent. Furthermore, for all \( k, \ell = 1, 2, ..., G, k \neq \ell \), \((\mu_k, \sigma_k^2)\) is independent of \((\mu_\ell, \sigma_\ell^2)\).
B.2.1 The posterior distribution of $\lambda$

For $\lambda$ we shall use a conjugate Dirichlet prior where

$$\pi(\lambda) \propto \prod_{k=1}^{G} \lambda_k^{\alpha_k-1}.$$  \hfill (43)

We shall use a noninformative prior by setting $\alpha_k = 1$ for $k = 1, 2, ..., G$. The resulting posterior distribution is

$$\pi(\lambda | d) \sim \text{Dirichlet}(N_1 + 1, \ldots, N_G + 1).$$  \hfill (44)

B.2.2 The joint posterior distribution of $(\sigma^2_k, \mu_k)$

Take any $k = 1, 2, \ldots, G$ and let $(Z_{1,i}, Z_{2,i})$ represent the position in $\mathbb{R}^2$ of the $i^{th}$ unit in group $k$ (it shall be understood that, for notational convenience, the units in group $k$ are temporarily indexed to be the first $N_k$ units of the population where $N_k$ is the size of group $k$). We shall determine the posterior distributions of $\pi(\sigma^2_k | Z_k)$ and $\pi(\mu_k | \sigma^2_k, Z_k)$. Recall that we have that

$$Z_{1,1}, Z_{1,2}, \ldots, Z_{1,N_k} \overset{iid}{\sim} N(\mu_{1,k}, \sigma^2_k),$$

$$Z_{2,1}, Z_{2,2}, \ldots, Z_{2,N_k} \overset{iid}{\sim} N(\mu_{2,k}, \sigma^2_k),$$  \hfill (45)

all of which arise independently given the group memberships.

In order to evaluate the posterior distributions we will need
\[
\begin{align*}
f(Z_k|\mu_{1,k}, \mu_{2,k}, \sigma^2_k) \\
&= \prod_{i=1}^{N_k} \left[ \frac{1}{\sqrt{2\pi\sigma^2_k}} \exp \left\{ -\frac{1}{2\sigma^2_k} (Z_{1,i} - \mu_{1,k})^2 \right\} \right] \cdot \prod_{i=1}^{N_k} \left[ \frac{1}{\sqrt{2\pi\sigma^2_k}} \exp \left\{ -\frac{1}{2\sigma^2_k} (Z_{2,i} - \mu_{2,k})^2 \right\} \right] \\
&= \frac{1}{(2\pi\sigma^2_k)^{N_k}} \cdot \exp \left\{ -\frac{1}{2\sigma^2_k} \cdot \sum_{i=1}^{N_k} (Z_{1,i} - \mu_{1,k})^2 \right\} \cdot \exp \left\{ -\frac{1}{2\sigma^2_k} \cdot \sum_{i=1}^{N_k} (Z_{2,i} - \mu_{2,k})^2 \right\} .
\end{align*}
\]

For a choice of conjugate prior on \( \sigma^2_k \), we will let \( \pi(\sigma^2_k) \sim \Gamma^{-1}(\alpha, \beta) \) so that

\[
\pi(\sigma^2_k) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2_k)^{-\alpha-1} \exp \left\{ -\frac{\beta}{\sigma^2_k} \right\} .
\]

For \( \mu_{j,k}, j = 1, 2 \), we shall take the independent conjugate priors of \( \pi(\mu_{j,k}|\sigma^2_k) \sim N(\gamma_j, \frac{\sigma^2_k}{\nu_j}) \). We then have,

\[
\pi(\mu_{1,k}, \mu_{2,k}|\sigma^2_k) = \pi(\mu_{1,k}|\sigma^2_k)\pi(\mu_{2,k}|\sigma^2_k) \\
= \frac{1}{\sqrt{2\pi\frac{\sigma^2_k}{\nu_1}}} \exp \left\{ -\frac{1}{2} \frac{(\mu_{1,k} - \gamma_1)^2}{\frac{\sigma^2_k}{\nu_1}} \right\} \cdot \frac{1}{\sqrt{2\pi\frac{\sigma^2_k}{\nu_2}}} \exp \left\{ -\frac{1}{2} \frac{(\mu_{2,k} - \gamma_2)^2}{\frac{\sigma^2_k}{\nu_2}} \right\} .
\]

Therefore we have the following posterior distribution,
\[
\pi(\sigma_k^2, \mu_{1,k}, \mu_{2,k}|Z_k) \propto \pi(\sigma_k^2) \cdot \pi(\mu_{1,k}, \mu_{2,k}|\sigma_k^2) \cdot f(Z_k|\mu_{1,k}, \mu_{2,k}, \sigma_k^2)
\]
\[
= \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma_k^2)^{-\alpha-1} \exp \left\{ -\frac{\beta}{\sigma_k^2} \right\} 
\times \frac{1}{\sqrt{2\pi} \frac{\sigma_k}{\sqrt{v_1}}} \exp \left\{ -\frac{1}{2} \frac{(\mu_{1,k} - \gamma_1)^2}{\sigma_k^2} \right\} \cdot \frac{1}{\sqrt{2\pi} \frac{\sigma_k}{\sqrt{v_2}}} \exp \left\{ -\frac{1}{2} \frac{(\mu_{2,k} - \gamma_2)^2}{\sigma_k^2} \right\}
\times \frac{1}{(2\pi \sigma_k^2)^{N_k}} \exp \left\{ -\frac{1}{2\sigma_k^2} \sum_{i=1}^{N_k} (Z_{1,i} - \mu_{1,k})^2 + \frac{N_k}{2\sigma_k^2} \sum_{i=1}^{N_k} (Z_{2,i} - \mu_{2,k})^2 \right\}
\propto (\sigma_k^2)^{-\alpha-1} \cdot \exp \left\{ -\frac{\beta}{\sigma_k^2} \right\}
\times \frac{1}{\sqrt{2\pi} \frac{\sigma_k}{\sqrt{v_1}}} \exp \left\{ -\frac{1}{2} \frac{(\mu_{1,k} - \gamma_1)^2}{\sigma_k^2} \right\} \cdot \frac{1}{\sqrt{2\pi} \frac{\sigma_k}{\sqrt{v_2}}} \exp \left\{ -\frac{1}{2} \frac{(\mu_{2,k} - \gamma_2)^2}{\sigma_k^2} \right\}
\times \frac{1}{(\sigma_k^2)^{N_k}} \exp \left\{ -\frac{1}{2\sigma_k^2} \sum_{i=1}^{N_k} (Z_{1,i} - \mu_{1,k})^2 \right\} \cdot \exp \left\{ -\frac{1}{2\sigma_k^2} \sum_{i=1}^{N_k} (Z_{2,i} - \mu_{2,k})^2 \right\}.
\]

(49)

Notice that, by the factorization theorem, \(\mu_{1,k}|\sigma_k^2, Z_k\) is independent of \(\mu_{2,k}|\sigma_k^2, Z_k\).

**B.2.3 The posterior distribution of \(\sigma_k^2\)**

We wish to determine \(\pi(\sigma_k^2|Z_k) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \pi(\sigma_k^2, \mu_{1,k}, \mu_{2,k}|Z_k) \, d\mu_{1,k} \, d\mu_{2,k}\). Integrating over Expression (49), we have

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \pi(\sigma_k^2, \mu_{1,k}, \mu_{2,k}|Z_k) \, d\mu_{1,k} \, d\mu_{2,k}
\]
By completing the square over $\mu_{1,k}$ and $\mu_{2,k}$, it can be shown that

$$
\pi(\sigma_k^2|Z_k) \propto (\sigma_k^2)^{-\alpha - 1} \cdot \frac{1}{\sigma_k^2} \cdot \frac{1}{\sigma_k} \cdot \frac{1}{(\sigma_k^2)^{N_k}} \cdot \frac{1}{(\sigma_k^2)^{N_k/2}}
$$

$$
\times \exp \left\{ - \frac{1}{2\sigma_k^2} \left[ \nu_1 \gamma_1^2 + \sum_{i=1}^{N_k} Z_{1,i}^2 - \left( \frac{\nu_1 \gamma_1 + \sum_{i=1}^{N_k} Z_{1,i}}{\nu_1 + N_k} \right)^2 \right] \right\} \cdot \frac{1}{\sigma_k}
$$

$$
\times \exp \left\{ - \frac{1}{2\sigma_k^2} \left[ \nu_2 \gamma_2^2 + \sum_{i=1}^{N_k} Z_{2,i}^2 - \left( \frac{\nu_2 \gamma_2 + \sum_{i=1}^{N_k} Z_{2,i}}{\nu_2 + N_k} \right)^2 \right] \right\} \cdot \frac{1}{\sigma_k}
$$

$$
= (\sigma_k^2)^{-\alpha - 1 - N_k - 1} \cdot \exp \left\{ - \frac{1}{\sigma_k^2} \left[ \beta + \frac{1}{2} \left[ \nu_1 \gamma_1^2 + \sum_{i=1}^{N_k} Z_{1,i}^2 - \left( \frac{\nu_1 \gamma_1 + \sum_{i=1}^{N_k} Z_{1,i}}{\nu_1 + N_k} \right)^2 \right] \right. \right.
$$

$$
+ \frac{\nu_2 \gamma_2^2 + \sum_{i=1}^{N_k} Z_{2,i}^2 - \left( \frac{\nu_2 \gamma_2 + \sum_{i=1}^{N_k} Z_{2,i}}{\nu_2 + N_k} \right)^2}{\nu_2 + N_k} \right]\right\}.
$$

Therefore, we have $\pi(\sigma_k^2|Z_k) \sim \Gamma^{-1}(A, B)$ where

$$
A = \alpha + N_k + 2, \quad \text{and}
$$

$$
B = \beta + \frac{1}{2} \left[ \nu_1 \gamma_1^2 + \sum_{i=1}^{N_k} Z_{1,i}^2 - \left( \frac{\nu_1 \gamma_1 + \sum_{i=1}^{N_k} Z_{1,i}}{\nu_1 + N_k} \right)^2 \right]
$$

and

$$
B = \beta + \frac{1}{2} \left[ \nu_2 \gamma_2^2 + \sum_{i=1}^{N_k} Z_{2,i}^2 - \left( \frac{\nu_2 \gamma_2 + \sum_{i=1}^{N_k} Z_{2,i}}{\nu_2 + N_k} \right)^2 \right]
$$
\[ + \nu_2 \gamma_2^2 + \sum_{i=1}^{N_k} Z_{2,i}^2 \]

In this study we shall take \( \alpha = \beta = \nu_1 = \nu_2 = 1, \) and \( \gamma_1 = \gamma_2 = 0, \) resulting in a non-informative prior.

### B.2.4 The posterior distribution of \( \mu_k | \sigma_k^2 \)

To find the posterior distribution of \( \mu_{1,k}, \) we condition on \( \sigma_k^2 \) and hence

\[
\pi(\mu_{1,k} | Z_k, \sigma_k^2) \propto \exp \left\{ -\frac{1}{2} \frac{(\mu_{1,k} - \gamma_1)^2}{\sigma_k^2} \right\} \exp \left\{ -\frac{1}{2} \frac{\sum_{i=1}^{N_k} (Z_{1,i} - \mu_{1,k})^2}{\sigma_k^2} \right\}
\]

\[
= \exp \left\{ -\frac{1}{2} \frac{(\mu_{1,k} - 2\gamma_1 \mu_{1,k} + \gamma_1^2)}{\nu_1} \right\} \exp \left\{ -\frac{1}{2} \frac{\sum_{i=1}^{N_k} (Z_{1,i} - \mu_{1,k})^2}{\sigma_k^2} \right\}
\]

\[
= \exp \left\{ -\frac{1}{2} \frac{\nu_1 \mu_{1,k}^2 - 2\gamma_1 \nu_1 \mu_{1,k} + \gamma_1^2}{\nu_1} \right\} \exp \left\{ -\frac{1}{2} \frac{\sum_{i=1}^{N_k} (Z_{1,i} - \mu_{1,k})^2}{\sigma_k^2} \right\}
\]

\[
= \exp \left\{ -\frac{1}{2} \frac{(\nu_1 + N_k) \mu_{1,k}^2 + \mu_{1,k} \cdot (2\gamma_1 \nu_1 - \sum_{i=1}^{N_k} 2Z_{1,i})}{\nu_1 + N_k} \right\}
\]

\[
= \exp \left\{ -\frac{1}{2} \frac{(\nu_1 + N_k) \left( \mu_{1,k} - \gamma_1 \nu_1 + \sum_{i=1}^{N_k} Z_{1,i} \right)^2}{\nu_1 + N_k} \right\}
\]

\[
= \exp \left\{ -\frac{1}{2} \frac{\sigma_k^2}{\nu_1 + N_k} \left( \mu_{1,k} - \frac{\gamma_1 \nu_1 + \sum_{i=1}^{N_k} Z_{1,i}}{\nu_1 + N_k} \right)^2 \right\}.
\]
\[ \mu_{1,k} | \sigma_k^2, Z_k \sim N \left( \frac{\gamma_1 \nu_1 + \sum_{i=1}^{N_k} Z_{1,i}}{\nu_1 + N_k}, \frac{\sigma_k^2}{\nu_1 + N_k} \right), \tag{55} \]

and similarly,

\[ \mu_{2,k} | \sigma_k^2, Z_k \sim N \left( \frac{\gamma_2 \nu_2 + \sum_{i=1}^{N_k} Z_{2,i}}{\nu_2 + N_k}, \frac{\sigma_k^2}{\nu_2 + N_k} \right). \tag{56} \]

In this study we shall take \( \nu_1 = \nu_2 = 1 \), and \( \gamma_1 = \gamma_2 = 0 \), resulting in a non-informative prior.

**B.2.5 The probability density function of \( \alpha \) and \( \beta \)**

We shall place independent prior distributions on \((\beta_0, \alpha_0)\) and \((\beta_1, \alpha_1)\) that contribute as one success and one failure in observing links for both within and between groups, assuming these units are located one unit distance from each other. Recall that for any \( \gamma \in \mathbb{R} \), \( 1 - \frac{e^{-\gamma}}{1 + e^{-\gamma}} = \frac{1}{1 + e^{\gamma}} \), and hence the resulting posterior distribution can be shown to be

\[
\pi(\beta_0, \alpha_0, \beta_1, \alpha_1 | d) = \frac{e^{\beta_0 + \alpha_0}}{1 + e^{\beta_0 + \alpha_0}} \cdot \left( \frac{1}{1 + e^{\beta_0 + \alpha_0}} \right) \\
\times \prod_{k=1}^{G} \prod_{i,j \in J_k: i < j} \left[ \left( \frac{e^{\beta_0 + \alpha_0||Z_i - Z_j||}}{1 + e^{\beta_0 + \alpha_0||Z_i - Z_j||}} \right)^{Y_{ij}} \left( \frac{1}{1 + e^{\beta_0 + \alpha_0||Z_i - Z_j||}} \right)^{(1-Y_{ij})} \right] 
\]
\[
\times \frac{e^{\beta_1 + \alpha_1}}{1 + e^{\beta_1 + \alpha_1}} \left( \frac{1}{1 + e^{\beta_1 + \alpha_1}} \right) \\
\times \prod_{k, \ell = 1: i \in J_k, \ k < \ell \ j \in G_\ell} \prod \left[ \left( \frac{\exp(\beta_1 + \alpha_1 \||Z_i - Z_j\||)}{1 + \exp(\beta_1 + \alpha_1 \||Z_i - Z_j\||)} \right)^{Y_{ij}} \left( \frac{1}{1 + \exp(\beta_1 + \alpha_1 \||Z_i - Z_j\||)} \right)^{(1 - Y_{ij})} \right]
\]

(57)

Note that this prior is used in order to ensure that the resulting posterior distribution integrates to a value of 1, since, in the event that we observe and augment all successes or failures of links between nodes within and/or between groups, we will have a posterior distribution with infinite area.