STRATEGIES FOR ESTIMATING THE SIZE AND DISTRIBUTION OF HARD-TO-REACH POPULATIONS WITH ADAPTIVE SAMPLING

by

Kyle Shane Vincent
B.Sc. (Honours) (Mathematics), University of Winnipeg, 2006
M.Sc. (Statistics), Simon Fraser University, 2008

a Thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in the Department of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY
Spring 2012

APPROVAL

Name: Kyle Shane Vincent

Degree: Doctor of Philosophy

Title of Thesis: Strategies for Estimating the Size and Distribution of Hard-to-Reach Populations with Adaptive Sampling

Examining Committee:

Dr. Tim Swartz, Professor
Chair

Dr. Steve Thompson, Professor
Senior Supervisor

Dr. Carl Schwarz, Professor
Supervisor

Dr. Richard Lockhart, Professor
Supervisor

Dr. Charmaine Dean, Professor
SFU Examiner

Dr. Laura Cowen, Associate Professor
External Examiner
University of Victoria

Date Approved: April 20, 2012

Partial Copyright Licence

Abstract

This thesis develops new methods for estimating the size and distribution of hard-to-reach populations when employing an adaptive sampling design. Hard-to-reach populations, like those comprised of injection drug-users, are usually not covered by a sampling frame. Hence, the sampler may desire to exploit the social links between its members to adaptively sample individuals for the study. We have developed three novel procedures based on various adaptive sampling designs for estimating the population unknowns.

The first project introduces a complex graph model that accounts for the erratic clustering behaviour commonly seen in hard-to-reach populations through observed covariate information. Our novel approach bases inference for the population size and model parameters on a Bayesian data augmentation routine.

The second project explores a new design-based approach that is based on a multi-sample study. Preliminary estimates of population unknowns are based on the initial random selections made for each sample. The adaptively selected members of the sample are included in the inference procedure through Rao-Blackwellization of the preliminary estimator based on sample reorderings which are consistent with a sufficient statistic.

The third project extends the design-based approach to inference that was introduced by Frank and Snijders (1994) where inference is based on the links originating from members selected for a Bernoulli sample. We propose new estimators of the population size that are based on one wave selected after the initial sample is obtained. We also introduce a Rao-Blackwellization procedure that is similar to that found in the second project for obtaining improved estimates.

The fourth project offers new methods for approximating the Rao-Blackwellized estimates obtained with a design-based approach to inference. We introduce a method termed *improved importance sampling*, based on a single-stage cluster sampling procedure, to obtain improved estimates over the preliminary importance sampling estimates.

For our thesis study population we use a networked population that was simulated from the complex graph model. We conduct a series of simulation studies based on several different adaptive sampling designs to evaluate the performance of the estimators from each of the projects.

Keywords: Adaptive sampling, Bayesian inference, Capture-recapture, Markov chain Monte Carlo, Network sampling, Rao-Blackwellization

Acknowledgments

First and foremost I would like to thank my supervisor, Professor Steve Thompson, for all of the guidance and support that he has provided me during my studies at Simon Fraser University. The many illuminating conversations we had about our research and experiences have been invaluable to my academic career. Thank you.

I would like to thank my examining committee for all of their feedback on my thesis and for taking the time out of their busy schedules to meet with me in person on many occasions. There were many invaluable conversations I had with each of you, the kind that motivates a Ph.D student to continue striving to be the best academic that they can be.

To Mike, Soheil, Will, and Ryan, a guy could not have possibly asked for better friends than each of you have been. The amount of support and encouragement that you guys had for me when times got tough was nothing short of amazing.

To Carolyn and Saman, you guys were at SFU when I first arrived. I am glad that I found a couple of excellent student mentors in each of you. You both provided a lot of inspiration for me and helped bring a sense of enthusiasm to my work. It was a pleasure working along side of you in IRMACS.

Finally, I cannot forget to thank Professor Michael Stephens for educating me on the responsibilities of being a Ph.D student. Thank you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Adaptive sampling methods are typically used to recruit individuals from hard-to-reach populations or to increase sampling effort where observed units are revealed to be members that are of high interest to the researcher. For the purpose of estimating the size of a target population there is an abundance of literature based on the use of capture-recapture methods. Most of these methods rely on the use of conventional sampling designs like random or stratified sampling because of the tractable likelihood functions they induce for the population size. This thesis presents methods for estimating the size and distribution of populations when an initial sample(s) is selected from the target population completely at random and members are then adaptively recruited, via following social links from the current sample, to be in the final sample. The adaptive sampling and inference procedures outlined in this thesis can assist in the recruitment from, and estimation of the size and characteristics of, hard-to-reach populations like those comprised of injection drug-users, men who have sex with men, illegal sex workers and the homeless, to name a few.

1.1 Background

For study purposes, a typical social network can be conceived of as a graph where the nodes correspond with the members of the population and the links (commonly

referred to as arks, nominations, or mentions) correspond with the presence of some predefined relationship between all pairs of the individuals. For example, in a population of injection drug-users a link may occur between two individuals if there is a mutual nomination between them. These nominations may come in the form of the sharing of drug-using paraphernalia and/or coming into sexual contact. As the individuals of such hard-to-reach populations may not be covered by a sampling frame and/or may be difficult to locate or identify, a researcher may benefit from the use of an adaptive sampling strategy to recruit individuals for study purposes. With a typical adaptive sampling design, a subset of the graph/population is initially recruited, ideally or conceivably through a probability sampling design, and members of the population are then adaptively recruited by tracing some of the social network links out of the nodes in the current sample.

There are two common inferential methods for estimating unknown population quantities in sampling. In a model-based approach the population graph is assumed to have arisen from an underlying model. With a model-based approach estimates of a population quantity, y_0 say, are said to be *model unbiased* if the expectation of the estimator $\hat{y}_0$ given any sample S equals the expectation of y_0. To clarify, suppose that the realization of a population quantity is y_0 where y_0 is a function of the realization of the responses of interest $\mathbf{y} = (y_1, y_2, ..., y_N)$ according to some joint distribution F where N is the population size. We then refer to $\hat{y}_0$ as being model-unbiased if for any sample S, $E[\hat{y}_0|S] = E[y_0]$. Model-based approaches may be preferred for obtaining estimates of unknown population quantities like the population mean or average node degree (that is, the average number of neighbours a member possesses) as the model assumptions may facilitate obtaining estimates of the population unknowns.

With a design-based approach to inference in sampling there is an absence of an assumed underlying model from which the population graph has arisen. Instead, all responses are regarded as fixed and only known for those which are observed in the sample. As probability only enters the inferential procedure through the sampling design, an estimator for a population quantity is said to be *design-unbiased* if $E[\hat{y}_0|\mathbf{y}] = y_0$. Notice that a design-unbiased estimator holds the attractive feature of being unbiased for the population quantity that is taken on at the time of the survey.

However, with a design-based approach it is likely to be a more complicated task to obtain estimators with desirable features like unbiasedness as the absence of an underlying model will not permit exploiting accompanying mathematical assumptions which may facilitate obtaining such estimators. For more information, Thompson (2002) provides a discussion of the use of model-based and design-based approaches for inference in sampling.

There is a growing body of literature on both model-based and design-based approaches to making inference for population unknowns with the use of adaptive sampling strategies when the population size is known. Thompson and Frank (2000) described an approach to likelihood-based inference for adaptive sampling (also known as link-tracing) designs. This approach was used in further development in Thompson and Chow (2003), and Handcock and Gile (2010) developed a theoretical framework for basing inference of population unknowns on exponential random graph models when using an adaptive sampling design. Thompson (2006b) developed a design-based method for estimating population proportions when using a targeted random walk sampling design that is based on the use of Markov chain theory. Thompson (2006a) generalized the design-based method for estimating population proportions based on an adaptive sampling strategy termed adaptive web sampling that allows for fixed sample sizes as well as the flexibility to allocate as much random or adaptive effort as desired at each step in the sample selection procedure. For additional information, Fienberg (2010a,b) provides a summary and discussion of some of the work on the modeling and analysis of networked populations, as well as a general introduction to papers with applications towards sampling and analyzing hidden populations. Spreen (1992) also provides some review of link-tracing designs and their applicability to sampling hard-to-reach populations.

Heckathorn (1997, 2002) developed a procedure termed respondent-driven sampling which bases estimates of population proportions on Markov chain theory. Abdul-Quader et al. (2006) describes empirical findings based on a respondent-driven sampling design to collect data on an HIV-related population in the New York City area. Recent work by Gile and Handcock (2011) proposes a modified estimator of

population means when employing respondent-driven sampling that utilizes a model-assisted approach to help overcome the bias introduced with the selection of an initial sample that is not a probability sample. Since their inference procedure requires knowing the population size, they show that this is a robust estimator when the inference procedure substitutes a relatively large or small estimated value for the true population size.

There are many methods for estimating population sizes through a capture-recapture style of study (see Schwarz and Seber (1999) and Chao et al. (2001) for a summary of the existing methods), and some of these classic methods have been implemented for estimating the size of hidden drug-using populations (see Frischer et al. (1993), Mastro et al. (1994) and Hook and Regal (1995)). 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 combines model-based and design-based inference. It assumes that links from the partial sampling frame are made in a homogenous pattern that facilitates a capture-recapture likelihood style of inference. Felix-Medina and Monjardin (2006) extend this work by proposing a Bayesian-assisted approach to overcome some of the bias that the maximum likelihood estimators possess, and Felix-Medina and Monjardin (2009) further extend the aforementioned work by allowing for a method based on an initial sequential sample that gives control over the final sample sizes. Frank and Snijders (1994) were able to develop a design-based approach that formulates consistent moment-based estimates of the population size based on the links originating from the members of a Bernoulli sample.

The model-based methods outlined in the existing literature for estimating the size of a population with adaptive sampling strategies make relaxed/generic mathematical assumptions for which the nomination probabilities are based upon (for example, see Frank and Snijders (1994) for an approach based on assuming a Bernoulli graph model and Thompson and Frank (2000) for an approach based on assuming a simple graph model known as the stochastic block model). Also, none of the existing model-based and design-based methods permit inference based on members who

are sampled beyond those linked to the initial sample. Furthermore, most of the existing methods have not harnessed or adopted some of the well-known capture-recapture methods, such as those used in the well-known eight closed population models (see Schwarz and Seber (1999) for a description), for estimating population sizes. These issues serve as the primary motivation for the work found in this thesis and are addressed throughout chapters 2, 3, and 4. Chapters 2, 3, and 4 are each written in a stand-alone fashion, and chapter 5 is written as an extension of chapter 4. All simulation experiments that were performed in this thesis were done in the R programming language. A copy of the code can be provided upon request.

1.2 Organization of the Thesis

Chapter 2 introduces a complex graph model, termed the *stochastic cluster model*, that is developed to account for the erratic clustering behaviour commonly seen in hard-to-reach populations. Kwanisai (2004) explored the use of the two-group stochastic block model for modeling a networked population, and the stochastic cluster model extends on this model by permitting additional covariate information to be incorporated into the model. The additional covariate information is utilized to help capture and explain the clustering behaviour in the population via a logistic regression model that governs the presence of links between members. We base our analysis on a Bayesian data augmentation routine that is applied to the data observed from a complete one-wave snowball sampling design. We compare the performance of the Bayes estimates of the population size and model parameters corresponding with the stochastic block model (which is now generalized for inference on as many groups as is desired) and the stochastic cluster model based on our strategy that is applied to a networked population that was simulated from the stochastic cluster model (for which we deem the *thesis study population*). Estimates of the population size and model parameters perform reasonably well in both cases with estimates based on the stochastic cluster model outperforming those based on the stochastic block model.

Chapter 3 explores a design-based approach for estimating the size and average node degree of a population based on a multi-sample study. Preliminary estimates

of the population unknowns are based on the initial random selections for each sample and the estimates of the population size are of a classic/conventional capture-recapture type. The adaptively selected members of the sample are included in the analysis through a Rao-Blackwellization procedure based on sample reorderings which are consistent with a sufficient statistic. As the number of sample reorderings may become prohibitively large for computational tabulation of their contributions to the Rao-Blackwellized estimator, a Markov chain resampling process is utilized to make computation of the improved estimates feasible. A simulation study demonstrates that gains in efficiency over the existing classic capture-recapture estimators are certain.

Chapter 4 extends the design-based approach introduced by Frank and Snijders (1994) for estimating population sizes with moment-based estimators based on the observations made from the selection of a Bernoulli sample. We propose new estimators of the population size that are based on one wave that is selected after the initial wave. We also show that Rao-Blackwellized estimators can be obtained in a manner similar to those which were obtained in the multi-sample study that was explored in chapter 3. A simulation study shows that the new estimators perform well and that gains in efficiency with the improved estimators are certain.

Chapter 5 presents a new method for approximating the Rao-Blackwellized estimates that are obtained with a design-based approach to inference in sampling. We view the sample space, that is all of the sample reorderings from a sample, as a sampling frame and use an importance sampling method to obtain preliminary approximations of the Rao-Blackwellized estimates. We introduce a method based on a single-stage cluster sampling design, which we term *improved importance sampling*, that entails defining neighbourhoods over the sample reorderings. The approach in this strategy rests on being able to observe the necessary responses from the units in the neighbourhoods of those members sampled under the importance sampler with relative ease once the corresponding responses from one of the units in the neighbourhood is observed. We can then improve on the preliminary approximations using the additional observations to obtain improved approximations. This method

is tried on the Rao-Blackwellized estimates based on one of the one-sample studies that is explored in the third project. The simulation study demonstrates that the improved importance sampling method will outperform the existing importance sampling method.

Chapter 6 is reserved for a discussion and general conclusions based on the thesis work. We also provide some direction for future research.

Chapter 2

The Stochastic Cluster Model-Based Approach

2.1 Introduction

Kwanisai (2004) developed a Bayesian data augmentation routine to make inference for model parameters based on a sample selected via a one-wave snowball subsampling design, that is obtained from a population with known size, when working under the basic two-group stochastic block model. In this chapter, a new and more elaborate graph model, termed the *stochastic cluster model*, is proposed to make inference for model parameters. It incorporates into the model the clustering behaviour commonly seen in networked populations through the use of observed covariate information. An extended data augmentation routine is developed for making inference on the unknown population size and the model parameters corresponding with both the stochastic block model and the stochastic cluster model.

In Section 2.2, we introduce the stochastic block model and the stochastic cluster model as well as the complete one-wave snowball sampling design used in our study. In Section 2.3, we introduce the extended data augmentation routine which entails developing the necessary posterior distributions for the unknown population size and model parameters as well as the probability distributions for the missing data values

for both graph models. Section 2.4 presents the results from a simulation study based on samples obtained from a population that is generated from the stochastic cluster model. Section 2.5 provides a general discussion of the results presented in this chapter.

2.2 The Graph Models and Sampling Design

This section introduces the stochastic block model and the stochastic cluster model. The likelihood functions based on each model for a full graph realization are also presented in this section. We conclude with an introduction to the complete one-wave snowball sampling design.

2.2.1 The stochastic block model

We define a population U to consist of the set of units/individuals $U = \{1, 2, ..., N\}$ where N is the population size. All units $i = 1, 2, ..., N$ are assigned to a group $C_i \ \epsilon \ \{1, 2, ..., G\}$ according to a multinomial distribution based on the vector of parameters $\underline{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_G)$ where G is the number of groups and λ_k is the probability of a unit being assigned to group k.

We shall 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. 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 $i \neq j$ we assume

$$P(Y_{ij} = 1|\underline{C}) = P(Y_{ij} = 1|C_i, C_j) = \beta_{C_i, C_j}, \tag{2.1}$$

and if $i = j$ then

$$P(Y_{ii} = 1) = 0. \tag{2.2}$$

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

$$\begin{aligned} L(\underline{\lambda}, \underline{\beta}|\underline{C}, Y) &\propto \prod_{k=1}^{G} \lambda_k^{N_k} \prod_{\substack{i,j=1: \\ i<j}}^{N} \beta_{C_i,C_j}^{Y_{ij}} (1 - \beta_{C_i,C_j})^{1-Y_{ij}} \\ &= \prod_{k=1}^{G} \lambda_k^{N_k} \prod_{\substack{k,\ell=1: \\ k<\ell}}^{G} \beta_{k,\ell}^{M_{k,\ell}} \prod_{\substack{k,\ell=1: \\ k<\ell}}^{G} (1 - \beta_{k,\ell})^{N_k N_\ell - M_{k,\ell}} \prod_{k=1}^{G} \beta_{k,k}^{M_{k,k}} \prod_{k=1}^{G} (1 - \beta_{k,k})^{\binom{N_k}{2} - M_{k,k}} \end{aligned} \tag{2.3}$$

where N_k is the size of group k, and $M_{k,\ell}$ is the number of links from group k to group ℓ, $k, \ell = 1, 2, ..., G$. $\underline{C}$ and Y are the full graph realizations of the group memberships and adjacency matrix, respectively. The first component of the likelihood describes the group memberships, the second and third components of the likelihood describe the links between groups, and the fourth and fifth components describe the links within the groups.

2.2.2 The stochastic cluster model

Again we define a population U to consist of the units/individuals $U = \{1, 2, ..., N\}$ where N is the population size. 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 $\underline{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_G)$ where G is the number of groups and λ_k is the probability of a unit being assigned to group k.

Conditional on the population vector of group memberships $\underline{C}$, each unit $i = 1, 2, ..., N$ independently realizes a K-dimensional vector of covariate information Z_i via

$$Z_i|\underline{C} = Z_i|C_i \sim MVN(\mu_{C_i}, \sum_{C_i}), \tag{2.4}$$

where $Z_i = (Z_{1,i}, Z_{2,i}, ..., Z_{K,i})$, $\mu_{C_i} = (\mu_{1,C_i}, \mu_{2,C_i}, ..., \mu_{K,i})$ 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 took the variance-covariance matrix to be $\sum_{C_i} = \sigma^2_{C_i} \mathrm{I}_d$ 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 also took $K = 2$, primarily for illustrative purposes. The covariate information is then based on the realization

$$Z_i|\underline{C} = Z_i|C_i \sim MVN(\mu_{C_i}, \sigma^2_{C_i}\mathrm{I}_d). \tag{2.5}$$

Similar to the approach that Hoff et al. (2002) took in modeling links among members of the population, we will posit that conditional on $\underline{C}$ and the population vector of covariate information $\underline{Z}$, links between units occur independently and in a logistic fashion depending on parameters $(\underline{\alpha}, \underline{\beta})$ as follows. First, if $C_i = C_j$ then we shall let $\beta_{C_i,C_j} = \beta_0$ and $\alpha_{C_i,C_j} = \alpha_0$, and if $C_i \neq C_j$ we shall let $\beta_{C_i,C_j} = \beta_1$ and $\alpha_{C_i,C_j} = \alpha_1$. Also, $\|Z_i - Z_j\|$ is taken to be the Euclidean norm that measures the distance, in terms of covariate (location) values, between units i and j. Now for any $i, j = 1, 2, ..., N$, if $i \neq j$ we assume

$$P(Y_{ij} = 1|\underline{C}, \underline{Z}) = P(Y_{ij} = 1 \mid C_i, C_j, Z_i, Z_j) = \frac{\exp(\beta_{C_i,C_j} + \alpha_{C_i,C_j}\|Z_i - Z_j\|)}{1 + \exp(\beta_{C_i,C_j} + \alpha_{C_i,C_j}\|Z_i - Z_j\|)}, \tag{2.6}$$

and if $i = j$ then

$$P(Y_{ii} = 1) = 0. \tag{2.7}$$

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

$$\begin{aligned}
&L(\underline{\lambda}, \underline{\mu}, \underline{\sigma}^2, \underline{\beta}, \underline{\alpha} | \underline{C}, \underline{Z}, Y) \propto \\
&\prod_{k=1}^{G} \lambda_k^{N_k} \cdot \prod_{i=1}^{N} \left[\frac{1}{2\pi\sigma_{C_i}^2} \cdot \exp\left\{ -\frac{1}{2} \left(\frac{(Z_{1,i} - \mu_{1,C_i})^2}{\sigma_{C_i}^2} + \frac{(Z_{2,i} - \mu_{2,C_i})^2}{\sigma_{C_i}^2} \right) \right\} \right] \\
&\cdot \prod_{i=1}^{N} \left[\prod_{\substack{Y_{ij}=1: \\ j>i}} \left(\frac{\exp(\beta_{C_i,C_j} + \alpha_{C_i,C_j} ||Z_i - Z_j||)}{1 + \exp(\beta_{C_i,C_j} + \alpha_{C_i,C_j} ||Z_i - Z_j||)} \right) \right] \\
&\cdot \prod_{i=1}^{N} \left[\prod_{\substack{Y_{ij}\neq 1: \\ j>i}} \left(1 - \frac{\exp(\beta_{C_i,C_j} + \alpha_{C_i,C_j} ||Z_i - Z_j||)}{1 + \exp(\beta_{C_i,C_j} + \alpha_{C_i,C_j} ||Z_i - Z_j||)} \right) \right].
\end{aligned} \tag{2.8}$$

The first component of the likelihood describes the group memberships, the second component of the likelihood describes the covariate information of the units, and the third and fourth components describe the links between the units.

2.2.3 The complete one-wave snowball sampling design

The complete one-wave snowball sampling design consists of selecting an initial sample, which we shall denote S_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 the first wave, denoted as S_1. Working under the stochastic block model, the data from the sample $S = S_0 \cup S_1$ we observe is $d_0' = \{S, \underline{C}_S, Y_{S_0,S}, Y_{S_0,\bar{S}}\}$ where $Y_{S_0,\bar{S}} \equiv 0$. Here $\underline{C}_S$ is the vector of the observed group memberships of the sampled members, $Y_{S_0,S}$ is the recorded observations of

the presence or absence of links from the initial sample to the final sample, and $Y_{S_0,\bar{S}} \equiv 0$ is understood to be the absence of links from the initial sample to the unknown number of unobserved members. Similarly, working under the stochastic cluster model the data we observe is $d_0' = \{S, \underline{C}_S, \underline{Z}_S, Y_{S_0,S}, Y_{S_0,\bar{S}}\}$ where $\underline{Z}_S$ is the vector of the observed covariate information of the sampled members.

We shall reorder the units in the population so that the first $|S_0| = n_0$ units are those selected for the initial sample and the next $|S_1| = n_1$ units are those linked to at least one unit in the initial sample. Also, we shall let $|S| = n_0 + n_1 = n$.

2.3 Data Augmentation

This section presents an outline of the general data augmentation procedure that was developed by Kwanisai (2004). We then outline the extended data augmentation procedure that is used in this chapter for incorporating the unknown population size with both the stochastic block model and the stochastic cluster model.

2.3.1 The data augmentation routine

Suppose we base a population model on a set of parameters $\underline{\theta}$. Through some sampling design let $\underline{X}_{obs}$ be the subset of the population data that we observe from the sample (for example, $\underline{X}_{obs} = d_0'$). Also, let $\underline{X}_{mis}$ be the missing subset of the population values that are not observed. If we intend to make inference upon $\underline{\theta}$ then, in a Bayesian context, we would like to obtain $\pi(\underline{\theta}|\underline{X}_{obs})$, the posterior distribution of $\underline{\theta}|\underline{X}_{obs}$. This may be challenging if the design is unconventional and instead we can consider $\pi(\underline{\theta}|\underline{X}_{obs}, \underline{X}_{mis})$ and $\mathrm{P}(\underline{X}_{mis}|\underline{X}_{obs}, \underline{\theta})$, the posterior distribution of $\underline{\theta}$ based on a (hypothetical) full graph realization and the probability distribution of the missing values given the observed data and the (estimated) model parameter values, respectively. Using a Gibbs sampling approach, we can iteratively "sample/update" the $\underline{\theta}$ and $\underline{X}_{mis}$ values as outlined below. By the results presented in Tanner and Wong (1987), the sampled $\underline{\theta}$'s will converge to $\pi(\underline{\theta}|\underline{X}_{obs})$ and the sampled $\underline{X}_{mis}$'s will converge to $\mathrm{P}(\underline{X}_{mis}|\underline{X}_{obs})$. The procedure, working over a known population size, is

summarized as follows:

Step 1: Assign initial values $\underline{\theta}^{(0)}$ to the parameters.

For $t = 1, 2, ..., T$, T sufficiently large, do the following:

Step 2: Generate a value of $\underline{X}_{mis}$ from the distribution $\mathrm{P}(\underline{X}_{mis}|\underline{X}_{obs}, \underline{\theta}^{(t-1)})$, call this $\underline{X}_{mis}^{(t)}$.

Step 3: Sample $\underline{\theta}$ from the posterior distribution $\pi(\underline{\theta}|\underline{X}_{obs}, \underline{X}_{mis}^{(t)})$, call this $\underline{\theta}^{(t)}$.

Revert to step 2 until $t = T$.

Step 4: Make inference on the model parameters and unknown population values with $\underline{\theta}^{(0)}, \underline{\theta}^{(1)}, ..., \underline{\theta}^{(T)}$ and $\underline{X}_{mis}^{(1)}, \underline{X}_{mis}^{(2)}, ..., \underline{X}_{mis}^{(T)}$, respectively.

2.3.2 The extended data augmentation procedure based on the stochastic block model

This subsection outlines the extended data augmentation procedure used with 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 now extended these results to cover the general case but have omitted the proofs since they are a straightforward extension of those based on the two-group stochastic block model. We shall retain the notation of the true population size N, the model parameters $\underline{\theta}$, and the missing subset of data $\underline{X}_{mis}$ for each of the iteratively sampled and augmented values.

Augmenting the missing values

With the stochastic block model, we can append an additional step to the aforementioned data augmentation routine to sample over a suitable posterior distribution of the population size as follows. Recall that the sampling design used in our study is the complete one-wave snowball sampling design and for the purposes of formulating

a likelihood for the population size N, consider the data $d_0^* = \{S_0, \underline{C}_{S_0}\}$. We shall consider a binomial experiment in the following manner. If a unit in $\bar{S}_0$ (that is, those units outside of S_0) is linked to at least one individual in S_0 then this shall be deemed as a success (which occurs with probability p). For any unit located in $\bar{S}_0$ we can determine the probability of this unit not being to any unit in the initial sample, conditional on d_0^*, as

$$1 - p = \sum_{k=1}^{G} [\lambda_k \prod_{i=1}^{n_0} (1 - \beta_{C_i,k})] \tag{2.9}$$

since, by definition of the model and given d_0^*, all group memberships of units in $\bar{S}_0$ arise independently and according to the multinomial distribution. As well, links from S_0 to $\bar{S}_0$ occur independently between all pairs of units, given the group memberships. Note that expression (2.9) averages over the probability that a unit in $\bar{S}_0$ is not linked to any units in S_0, where averages are taken over the weighted probability of being in each group (since we only condition on d_0^*).

Now, since all of the Bernoulli outcomes can be regarded as independent and, by symmetry, identically distributed, a likelihood function for N conditional on $d_0' = \{d_0^*, S_1, Y_{S_0,S_0 \cup S_1}, Y_{S_0,\bar{S}} \equiv 0\}$ is

$$L(N|d_0') = \binom{N - n_0}{n_1} p^{n_1} (1 - p)^{N - n_0 - n_1}. \tag{2.10}$$

In this study we shall take the prior distribution of N to be $\pi(N) \propto 1$ when $N \geq n_0 + n_1$ and $\pi(N) \propto 0$ when $N < n_0 + n_1$, and hence the resulting posterior distribution of N is

$$\pi(N|d_0') \propto \binom{N - n_0}{n_1} (1 - p)^{N - n_0 - n_1} \mathrm{I}[N \geq n_0 + n_1] \tag{2.11}$$

where I is the indicator function that takes on a value of one if the condition within the

brackets holds and zero otherwise. The new data augmentation routine now appends an additional Step 1.5 that samples a N from the updated posterior distribution presented in expression (2.11).

Upon sampling a N from the resulting posterior distribution, we shall consider the updated information to be $d_0 = \{S, \underline{C}_S, Y_{S_0,U}\}$, where U is a hypothetical population of size equal to the most recent N that was sampled from its probability distribution. As shown by Kwanisai (2004), for any $i \epsilon \bar{S}$ and $k = 1, 2, ..., G$,

$$P(C_i = k|d_0) = \frac{\lambda_k \prod_{j=1}^{n_0}(1 - \beta_{C_j,k})}{\sum_{\ell=1}^{G} \lambda_l \prod_{j=1}^{n_0}(1 - \beta_{C_j,\ell})}. \tag{2.12}$$

Values of the missing group memberships $\underline{C}_{\bar{S}}$ are then assigned according to the distribution outlined in expression (2.12).

After generating $\underline{C}_{\bar{S}}$, the graph data is updated from d_0 to d_1 where $d_1 = \{S, \underline{C}, Y_{S_0,U}\}$ and $\underline{C}$ represents the full hypothetical graph realization of group memberships. Again, as shown by Kwanisai (2004), for any $i, j \epsilon \bar{S}_0$ where $i \neq j$,

$$P(Y_{ij} = 1|d_1) = \beta_{C_i,C_j}, \tag{2.13}$$

and hence links between each pair of units $i, j \epsilon \bar{S}_0$ for $i \neq j$ are assigned according to the probability distribution found in expression (2.13) to generate a hypothetical full graph realization of Y.

The posterior distributions of the stochastic block model parameters based on a hypothetical realization of the full population graph

As shown by Kwanisai (2004), 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

$d = \{\underline{C}, Y\}$ (see subsection 2.2.1 for the likelihood based on a full graph realization). In our study, we shall place independent conjugate Dirichlet and Beta priors on $\underline{\lambda}$ and $\underline{\beta}$, respectively. That is,

$$\pi(\underline{\lambda}) \propto \prod_{k=1}^{G} \lambda_k^{\alpha_k - 1} \tag{2.14}$$

and

$$\pi(\underline{\beta}) \propto \prod_{\substack{k \leq \ell: \\ k,\ell=1}}^{G} \beta_{k,\ell}^{\gamma_1 - 1}(1 - \beta_{k,\ell})^{\gamma_2 - 1}. \tag{2.15}$$

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

$$\pi(\underline{\lambda}|d) \sim \text{Dirichlet}(N_1 + 1, ..., N_G + 1). \tag{2.16}$$

The resulting posterior distribution of $\beta_{k,\ell}$ for $k, \ell = 1, 2, ..., G, k \neq \ell$, is

$$\pi(\beta_{k,\ell}|d) \sim \text{Beta}(M_{k,\ell} + 1, N_k N_\ell - M_{k,\ell} + 1), \tag{2.17}$$

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

$$\pi(\beta_{k,k}|d) \sim \text{Beta}(M_{k,k} + 1, \binom{N_k}{2} - M_{k,k} + 1). \tag{2.18}$$

2.3.3 The extended data augmentation procedure based on the stochastic cluster model

This subsection outlines the data augmentation procedure that corresponds with the stochastic cluster model. Mathematical proofs and derivations of the probability mass/density functions of the missing values and posterior distributions of the model parameters are shown in Appendix A of the thesis.

Augmenting the missing values

We shall follow the same approach that was used to develop a posterior distribution for the population size N under the stochastic block model. Under the stochastic cluster model with $d_0^* = \{S_0, \underline{C}_{S_0}, \underline{Z}_{S_0}\}$ we have that for any unit in $\bar{S}_0$ the probability of observing no links to the initial sample, conditional on d_0^*, is

$$1 - p = \sum_{k=1}^{G} \lambda_k \int_{-\infty}^{\infty} \prod_{i=1}^{n_0} \left(1 - \frac{\exp(\beta_{C_i,k} + \alpha_{C_i,k}||Z_i - Z||)}{1 + \exp(\beta_{C_i,k} + \alpha_{C_i,k}||Z_i - Z||)}\right) \mathrm{BVN}(Z; \mu_k, \sigma_k^2 \mathrm{I}_d) \mathrm{d}Z \tag{2.19}$$

where Z is a specific realization of covariate information. This holds as, by definition of the model and given d_0^*, all group memberships, and then covariate information, of units in $\bar{S}_0$ arise independently and according to the multinomial distribution, and then the corresponding bivariate normal distribution. As well, links from S_0 to $\bar{S}_0$ occur independently between all pairs of units given the corresponding group memberships and covariate information. We will take the prior distribution of N to be $\pi(N) \propto 1$ when $N \geq n_0 + n_1$ and $\pi(N) \propto 0$ when $N < n_0 + n_1$. Now, with $d_0' = \{d_0^*, S_1, Y_{S_0, S_0 \cup S_1}, Y_{S_0, \bar{S}} \equiv 0\}$ we have that

$$\pi(N|d_0') \propto \binom{N - n_0}{n_1} (1 - p)^{N - n_0 - n_1} \mathrm{I}[N \geq n_0 + n_1] \tag{2.20}$$

is a binomial type of posterior distribution for N.

After sampling a value of N from the (updated) posterior distribution found in expression (2.20), we can consider the updated information to be $d_0 = \{S, \underline{C}_S, \underline{Z}_S, Y_{S_0,U}\}$ where U is a hypothetical population of size equal to the sampled N from the posterior distribution. Now, for any $i \in \bar{\mathrm{S}}$, and for any group $k = 1, 2, ...G$, it can be shown (by subsection A.1.1 of Appendix A) that

$$\mathrm{P}(C_i = k|\mathrm{S}, \underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,i}) = \frac{\lambda_k \cdot \int_{-\infty}^{\infty} \prod_{j=1}^{n_0} \left(1 - \frac{\exp(\beta_{C_j,k} + \alpha_{C_j,k}||Z_j - Z||)}{1+\exp(\beta_{C_j,k} + \alpha_{C_j,k}||Z_j - Z||)}\right) \mathrm{BVN}(Z; \mu_k, \sigma_k^2 \mathrm{I}_d) \ \mathrm{d}Z}{\sum_{\ell=1}^{G} \left[\lambda_\ell \cdot \int_{-\infty}^{\infty} \prod_{j=1}^{n_0} \left(1 - \frac{\exp(\beta_{C_j,\ell} + \alpha_{C_j,\ell}||Z_j - Z||)}{1+\exp(\beta_{C_j,\ell} + \alpha_{C_j,\ell}||Z_j - Z||)}\right) \mathrm{BVN}(Z; \mu_\ell, \sigma_\ell^2 \mathrm{I}_d) \ \mathrm{d}Z\right]}. \tag{2.21}$$

Values of the missing group memberships $\underline{C}_{\bar{S}}$ are then assigned according to the distribution outlined in expression (2.21).

After generating $\underline{C}_{\bar{\mathrm{S}}}$, the graph data is updated from d_0 to d_1, where $d_1 = \{\mathrm{S}, \underline{C}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U}\}$ and $\underline{C}$ represents the hypothetical full graph realization of group memberships. For any $i \in \bar{\mathrm{S}}$ and for any $z^* \in \mathbb{R}^2$, it can be shown (by subsection A.1.2 of Appendix A) that the density of Z_i at this point is evaluated as

$$\mathrm{P}(Z_i = z^*|d_1) = \frac{\prod_{j=1}^{n_0} \left(1 - \frac{\exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i}||Z_j - z^*||}{1+\exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i}||Z_j - z^*||}\right) \mathrm{BVN}(z^*; \mu_{C_i}, \sigma_{C_i}^2 \mathrm{I}_d)}{\int_{-\infty}^{\infty} \prod_{j=1}^{n_0} \left(1 - \frac{\exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i}||Z_j - Z||}{1+\exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i}||Z_j - Z||}\right) \mathrm{BVN}(Z; \mu_{C_i}, \sigma_{C_i}^2 \mathrm{I}_d) \ \mathrm{d}Z}. \tag{2.22}$$

Values of the missing covariate information $\underline{Z}_{\bar{S}}$ are then assigned according to the distribution outlined above.

After generating $\underline{Z}_{\bar{\mathrm{S}}}$, the graph data is updated to $d_2 = \{S, \underline{C}, \underline{Z}, Y_{S_0,U}\}$, where $\underline{Z}$ is the hypothetical full graph realization of covariate information. It now remains to assign links between all pairs of nodes (i, j) in $(\bar{S}_0, \bar{S}_0)$. Recall that by definition

of the model, we have that conditional on $\underline{C}, \underline{Z}$, for any $(i,j), (i^*,j^*) \, \epsilon \, (\bar{S}_0, \bar{S}_0)$, Y_{ij} is independent of $Y_{i^*j^*}$. Hence, for any $(i,j) \, \epsilon \, (\bar{S}_0, \bar{S}_0)$, $i \neq j$, we have that

$$\mathrm{P}(Y_{ij} = 1 | C_i, C_j, Z_i, Z_j) = \left(\frac{\exp(\beta_{C_i,C_j} + \alpha_{C_i,C_j} ||Z_i - Z_j||)}{1 + \exp(\beta_{C_i,C_j} + \alpha_{C_i,C_j} ||Z_i - Z_j||)} \right) \ , \tag{2.23}$$

and links between each pair of units $(i,j) \, \epsilon \, (\bar{S}_0, \bar{S}_0)$ for $i \neq j$ are assigned according to the probability distribution found in expression (2.23) to generate a hypothetical full graph realization of Y.

The posterior distributions of the stochastic cluster model parameters based on a hypothetical realization of the full population graph

With the use of independent prior distributions for the model parameters the factorization theorem asserts that the set of parameters $\underline{\lambda}, (\underline{\mu}, \underline{\sigma}^2), (\beta_0, \alpha_0)$, and (β_1, α_1) are all independent (see subsection 2.2.2 for the likelihood based on a full graph realization). Furthermore, for all $k, \ell = 1, 2, ..., G$, $k \neq \ell$, (μ_k, σ_k^2) is independent of $(\mu_\ell, \sigma_\ell^2)$. We now outline the prior distributions and resulting posterior distributions for each set of parameters.

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

$$\pi(\underline{\lambda}) \propto \prod_{k=1}^{G} \lambda_k^{\alpha_k - 1} \ . \tag{2.24}$$

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

$$\pi(\underline{\lambda} \, | d) \sim \ \mathrm{Dirichlet}(N_1 + 1, \ldots, N_G + 1) \ . \tag{2.25}$$

For the choice of a conjugate prior distribution for each σ_k^2, $k = 1, 2, ..., G$, we

shall let $\pi(\sigma_k^2) \sim \Gamma^{-1}(\alpha, \beta)$ which gives

$$\pi(\sigma_k^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma_k^2)^{-\alpha-1} \exp\left\{\frac{-\beta}{\sigma_k^2}\right\} . \tag{2.26}$$

For each group's center of covariate information $\mu_k = (\mu_{1,k}, \mu_{2,k})$ and dispersion parameter σ_k^2, $k = 1, 2, ..., G$, we shall take the conditional conjugate prior distribution of $\pi(\mu_{j,k}|\sigma_k^2) \sim \mathrm{N}(\gamma_j, \frac{\sigma_k^2}{\nu_j})$, where $j = 1, 2$. Since, $\mu_{1,k}$ and $\mu_{2,k}$ are independent (by the factorization theorem), it can be shown (by subsection A.2.1 of Appendix A) that

$$\begin{aligned}
&\pi(\sigma_k^2, \mu_{1,k}, \mu_{2,k}|\underline{Z}_k) \propto \\
&(\sigma_k^2)^{-\alpha-1} \cdot \exp\left\{-\frac{\beta}{\sigma_k^2}\right\} \\
&\cdot \frac{1}{\frac{\sigma_k}{\sqrt{\nu_1}}} \exp\left\{-\frac{1}{2}\frac{(\mu_{1,k}-\gamma_1)^2}{\frac{\sigma_k^2}{\nu_1}}\right\} \cdot \frac{1}{\frac{\sigma_k}{\sqrt{\nu_2}}} \exp\left\{-\frac{1}{2}\frac{(\mu_{2,k}-\gamma_2)^2}{\frac{\sigma_k^2}{\nu_2}}\right\} \\
&\cdot \frac{1}{(\sigma_k^2)^{N_k}} \exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i=1}^{N_k}(Z_{1,i}-\mu_{1,k})^2\right\} \cdot \exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i=1}^{N_k}(Z_{2,i}-\mu_{2,k})^2\right\} ,
\end{aligned} \tag{2.27}$$

where 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.

We shall first work over the posterior distribution of σ_k^2. It can be shown (by subsection A.2.2 of Appendix A) that

$$\begin{aligned}
&\pi(\sigma_k^2|\underline{Z}_k) \propto (\sigma_k^2)^{-\alpha-1-1-N_k-1} \\
&\cdot \exp\left\{ -\frac{1}{\sigma_k^2} \cdot \left[\beta + \frac{1}{2}\left[\nu_1\gamma_1^2 + \sum_{i=1}^{N_k} Z_{1,i}^2 - \frac{(\nu_1\gamma_1 + \sum_{i=1}^{N_k} Z_{1,i})^2}{\nu_1 + N_k}\right.\right.\right. \\
&\qquad \left.\left.\left. + \nu_2\gamma_2^2 + \sum_{i=1}^{N_k} Z_{2,i}^2 - \frac{(\nu_2\gamma_2 + \sum_{i=1}^{N_k} Z_{2,i})^2}{\nu_2 + N_k}\right]\right]\right\} .
\end{aligned} \tag{2.28}$$

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

$$\begin{aligned} A &= \alpha + N_k, \text{ and} \\ B &= \beta + \frac{1}{2}\Bigg[\nu_1\gamma_1^2 + \sum_{i=1}^{N_k} Z_{1,i}^2 - \frac{(\nu_1\gamma_1 + \sum_{i=1}^{N_k} Z_{1,i})^2}{\nu_1 + N_k} \\ &+ \nu_2\gamma_2^2 + \sum_{i=1}^{N_k} Z_{2,i}^2 - \frac{(\nu_2\gamma_2 + \sum_{i=1}^{N_k} Z_{2,i})^2}{\nu_2 + N_k}\Bigg]. \end{aligned} \tag{2.29}$$

To obtain the posterior distribution of $\mu_{1,k}$, we shall condition on the σ_k^2 sampled from the the distribution in (2.29). It can then be shown (by subsection A.2.3 of Appendix A) that

$$\mu_{1,k}|\sigma_k^2, \underline{Z}_k \sim \mathrm{N}\left(\frac{\gamma_1\nu_1 + \sum_{i=1}^{N_k} Z_{1,i}}{\nu_1 + N_k}, \frac{\sigma_k^2}{\nu_1 + N_k}\right), \tag{2.30}$$

and similarly

$$\mu_{2,k}|\sigma_k^2, \underline{Z}_k \sim \mathrm{N}\left(\frac{\gamma_2\nu_2 + \sum_{i=1}^{N_k} Z_{2,i}}{\nu_2 + N_k}, \frac{\sigma_k^2}{\nu_2 + N_k}\right). \tag{2.31}$$

In this study we shall set $\alpha = \beta = 1$, $\gamma_j = 0$, and $\nu_j = 1$ for $j = 1, 2$, which result in noninformative prior distributions.

Finally, we shall place independent prior distributions on (β_0, α_0) and (β_1, α_1) that contribute as one success and one failure in observing links within and between groups, respectively, 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

$$\pi(\beta_0, \alpha_0, \beta_1, \alpha_1|d) = \frac{e^{\beta_0+\alpha_0}}{1+e^{\beta_0+\alpha_0}} \cdot \left(\frac{1}{1+e^{\beta_0+\alpha_0}}\right)$$
$$\cdot \prod_{k=1}^{G} \prod_{\substack{i,j\in G_k: \\ i<j}} \left[\left(\frac{\exp(\beta_0+\alpha_0||Z_i-Z_j||)}{1+\exp(\beta_0+\alpha_0||Z_i-Z_j||)}\right)^{Y_{ij}} \left(\frac{1}{1+\exp(\beta_0+\alpha_0||Z_i-Z_j||)}\right)^{1-Y_{ij}}\right]$$
$$\cdot \frac{e^{\beta_1+\alpha_1}}{1+e^{\beta_1+\alpha_1}} \left(\frac{1}{1+e^{\beta_1+\alpha_1}}\right)$$
$$\cdot \prod_{\substack{k,\ell=1: \\ k<\ell}}^{G} \prod_{\substack{i\in G_k, \\ j\in G_\ell}} \left[\left(\frac{\exp(\beta_1+\alpha_1||Z_i-Z_j||)}{1+\exp(\beta_1+\alpha_1||Z_i-Z_j||)}\right)^{Y_{ij}} \left(\frac{1}{1+\exp(\beta_1+\alpha_1||Z_i-Z_j||)}\right)^{1-Y_{ij}}\right] . \tag{2.32}$$

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 a group(s), we will have a posterior distribution with infinite area.

2.4 Simulation Study

We will use the thesis study population to evaluate the new inference procedures outlined in this chapter. The population was generated according to the stochastic cluster model with a choice of three groups whose centers of covariate information form an approximate equilateral triangle in $\mathbb{R}^2$. We set the parameter values to be $N = 300, \underline{\lambda} = (0.5, 0.3, 0.2), \mu_1 = (-6, -9), \mu_2 = (6, -9), \mu_3 = (0, 5), \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 16, \beta_0 = -1.5, \beta_1 = -2.5$, and $\alpha_0 = -0.5, \alpha_1 = -0.5$. An illustration of the population can be found in Figure 2.1.

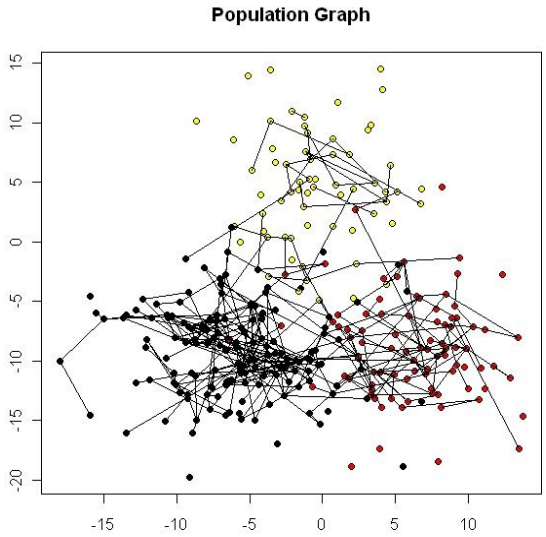

Figure 2.1: The simulated thesis study population. The values on the x-axis refer to the first dimension of covariate information and the values on the y-axis refer to the second dimension of covariate information. The axes are provided to compare the dispersion between groups as well as the relative occurrence of links between members within and between groups.

Figure 2.2 presents two plots that show a typical complete one-wave snowball sample. The initial sample size was 60 and the final sample size was 152. The data presented in the illustrations reflect the sample data (namely the group colour, covariate information, and links) that is required to be observed for the inferential procedures outlined in this chapter.

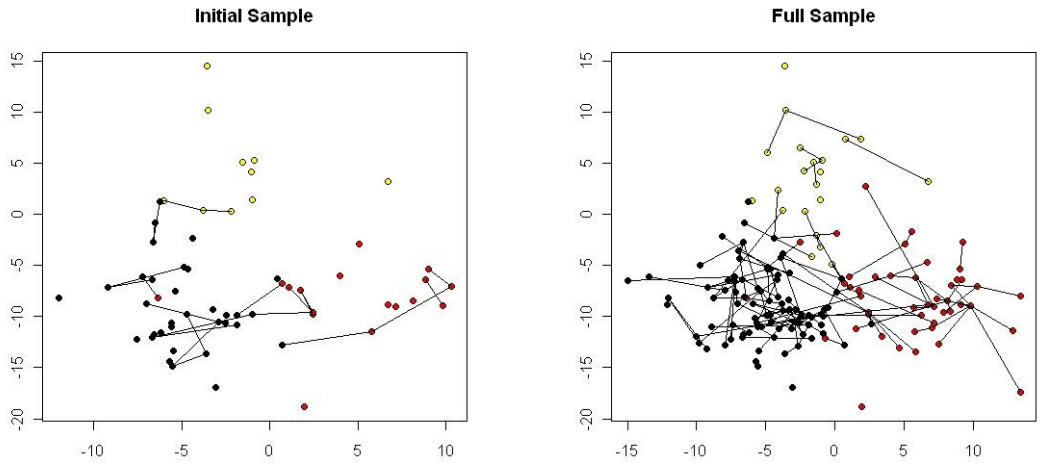

Figure 2.2: A complete one-wave snowball sample where the initial sample size is 60 and the final sample size is 152. The figure on the left displays the initial sample and links within the initial sample. The figure on the right displays the full sample and links from the initial sample to the first wave.

We conducted a simulation study based on each of the stochastic block model and the stochastic cluster model as follows. In each study we selected 500 samples each with an initial sample size of 60. Figure 2.3 shows a histogram of the final sample sizes of the 500 samples. The solid triangle on the x-axis indicates the average of the final sample sizes.

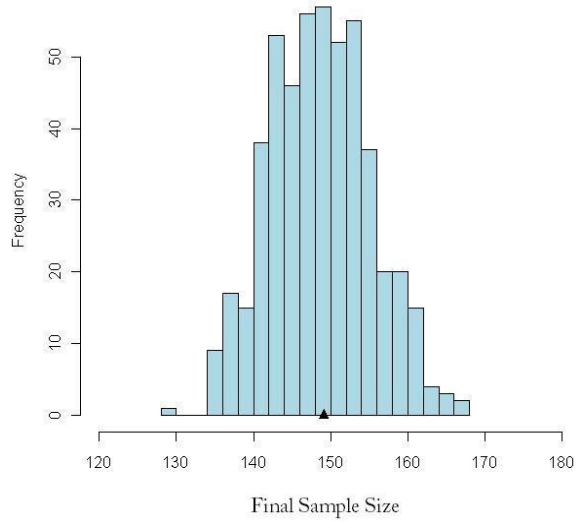

Figure 2.3: Histogram of the final sample sizes based on 500 simulations. In each case the initial sample size was 60. The dark triangle on the x-axis represents the average final sample size.

Bayes estimates of the population size and model parameters based on both the stochastic block model and the stochastic cluster model are presented in the following subsections. For each simulation, Markov chains of length 1000 were run with a 10% burn-in to obtain approximate Bayes estimates. Although the estimates are developed under a Bayesian framework, we employ a frequentist style of evaluation to measure the efficiency of the estimates under the two models.

2.4.1 Simulation study based on the use of the stochastic block model

This subsection contains a presentation and discussion of the Bayes estimates of the population size and model parameters based on the use of the stochastic block model. We commence with the Markov chain Monte Carlo (MCMC) trace plots of

the population size and model parameters based on the sample shown in Figure 2.2. This is appropriate as the behaviour of the MCMC trace plots based on the data from several other samples were found to be similar. The MCMC trace plots of the population size and model parameters can be found in Figures 2.4, 2.5, 2.6, and 2.7. As our study is based on a model-based approach to inference and since we are exploring our sampling and inference strategy applied to only one simulated network graph, we will use the maximum likelihood estimates (MLEs) of the population parameters based on the full graph realization as the target values for our Bayes estimates. Note that the MLEs of the model parameters will likely change between simulated networked graphs and hence our choice of using the MLEs based on this network graph is justified.

The solid lines in the figures indicate the maximum likelihood estimate(s) (MLE) based on a full graph realization. Note that the full graph realization permits observing N to be the true value of 300. Also, note that in the following figures shown the trace plots are for only one sample and hence the trace plots may not average to the corresponding MLEs.

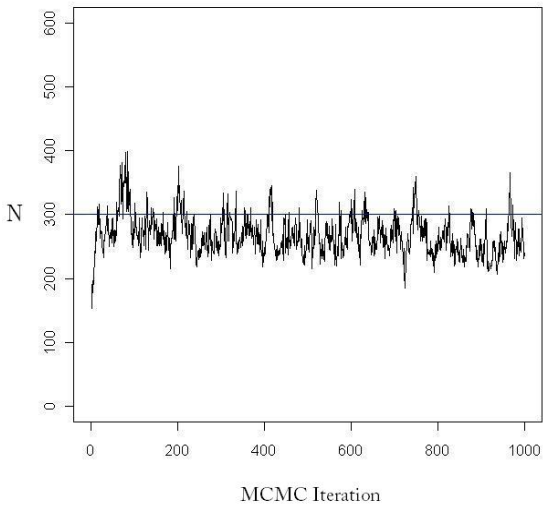

Figure 2.4: The sample dependent MCMC trace plot of the N values under the stochastic block model based on the sample presented in Figure 2.2. The solid line represents the MLE of the population size based on a full graph realization.

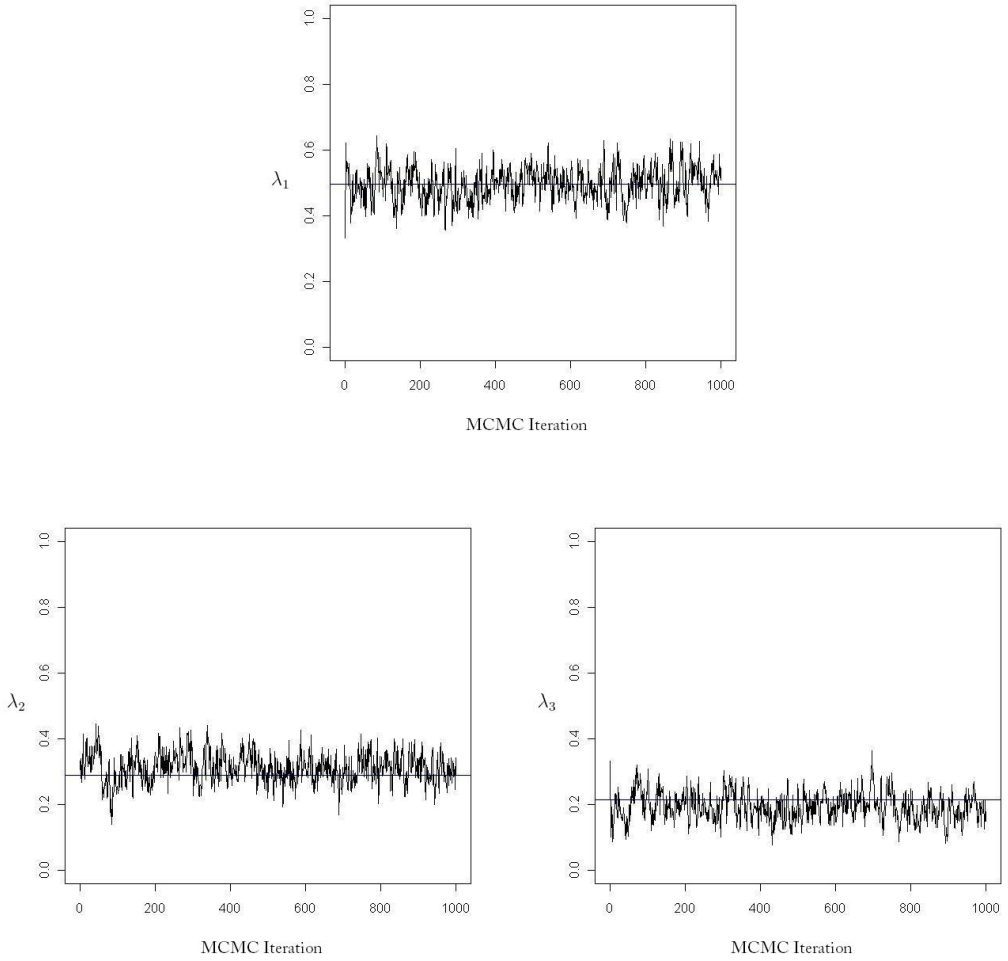

Figure 2.5: The sample dependent MCMC trace plots of λ_1, λ_2, and λ_3 under the stochastic block model based on the sample presented in Figure 2.2. The solid lines represent the MLEs of the population parameters based on a full graph realization.

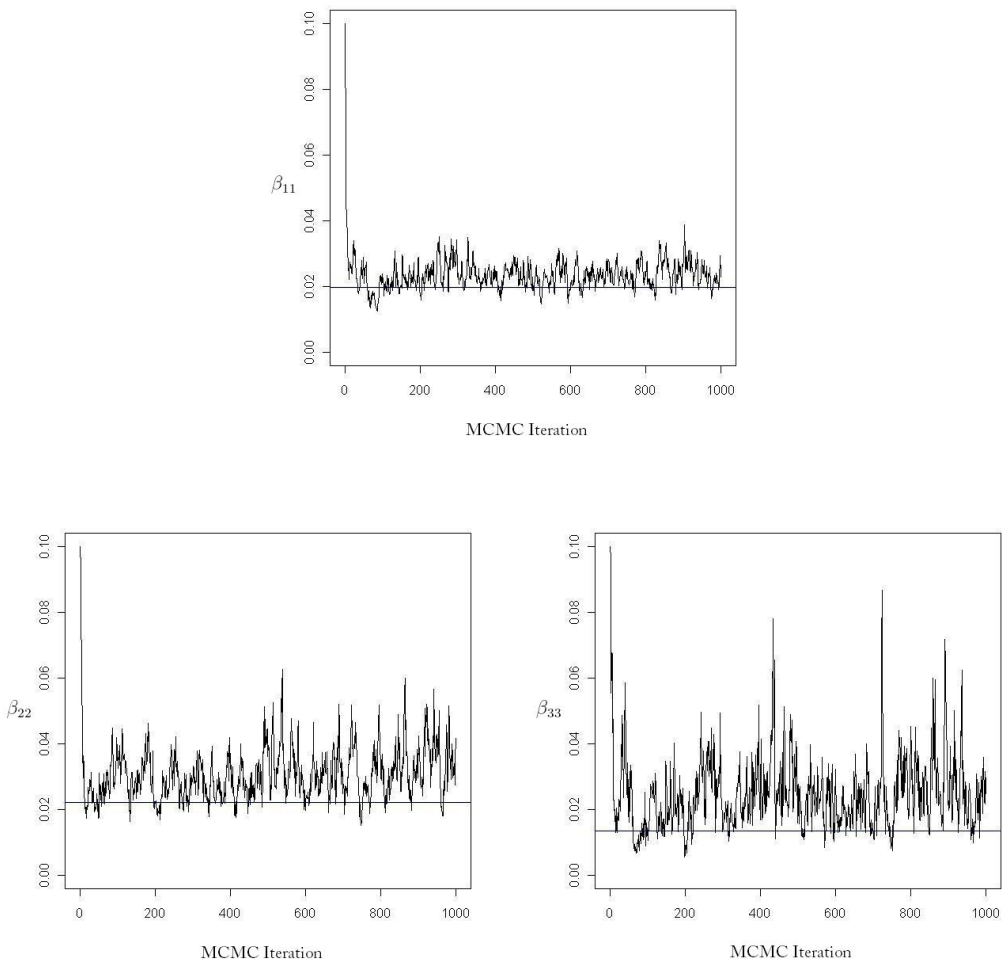

Figure 2.6: The sample dependent MCMC trace plots of β_{11}, β_{22}, and β_{33} under the stochastic block model based on the sample presented in Figure 2.2. The solid lines represent the MLEs of the population parameters based on a full graph realization.

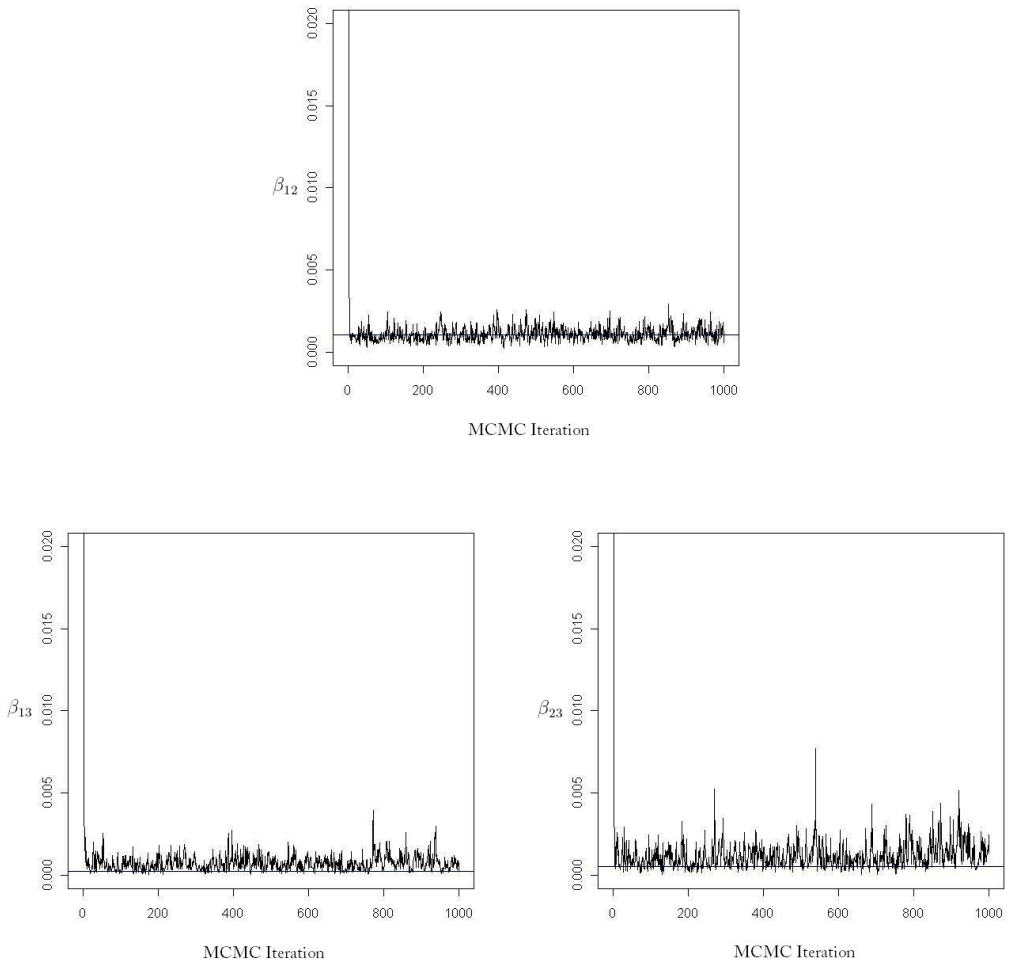

Figure 2.7: The sample dependent MCMC trace plots of β_{12}, β_{13}, and β_{23} under the stochastic block model based on the sample presented in Figure 2.2. The solid lines represent the MLEs of the population parameters based on a full graph realization.

The MCMC trace plots of the population size and model parameters, given the data found in the sample presented in Figure 2.2, appear to have sufficiently explored the posterior space indicating that a chain of length 1000 is sufficient. Notice how quickly the MCMC chains break away from the (noninformative) seeds for each of the model parameters where the seeds are chosen to reflect the noninformative prior

distributions (that is, we set $\underline{\lambda}^{(0)} = (\underline{1/3})$, and $\underline{\beta}^{(0)} = (\underline{0.10})$ to generate a relatively dense graph on the first augmentation iteration).

Histograms of the Bayes estimates of the population size and model parameters can be found in Figures 2.8, 2.9, 2.10, and 2.11. The solid triangle on the x-axis represents the MLE of the population parameter based on a full graph realization, and the transparent triangle represents the average of the Bayes estimates of the population size and model parameters.

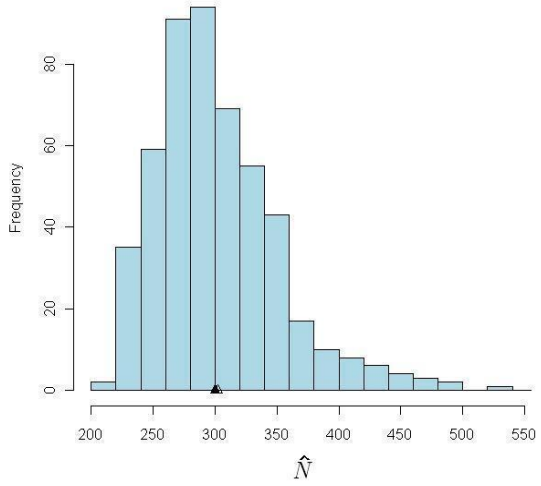

Figure 2.8: Histogram of the Bayes estimates of the N values under the stochastic block model based on 500 simulations. The solid triangle represents the MLE of the population size based on a full graph realization (which is the true population size of 300), and the transparent triangle represents the average of the Bayes estimates of the population size.

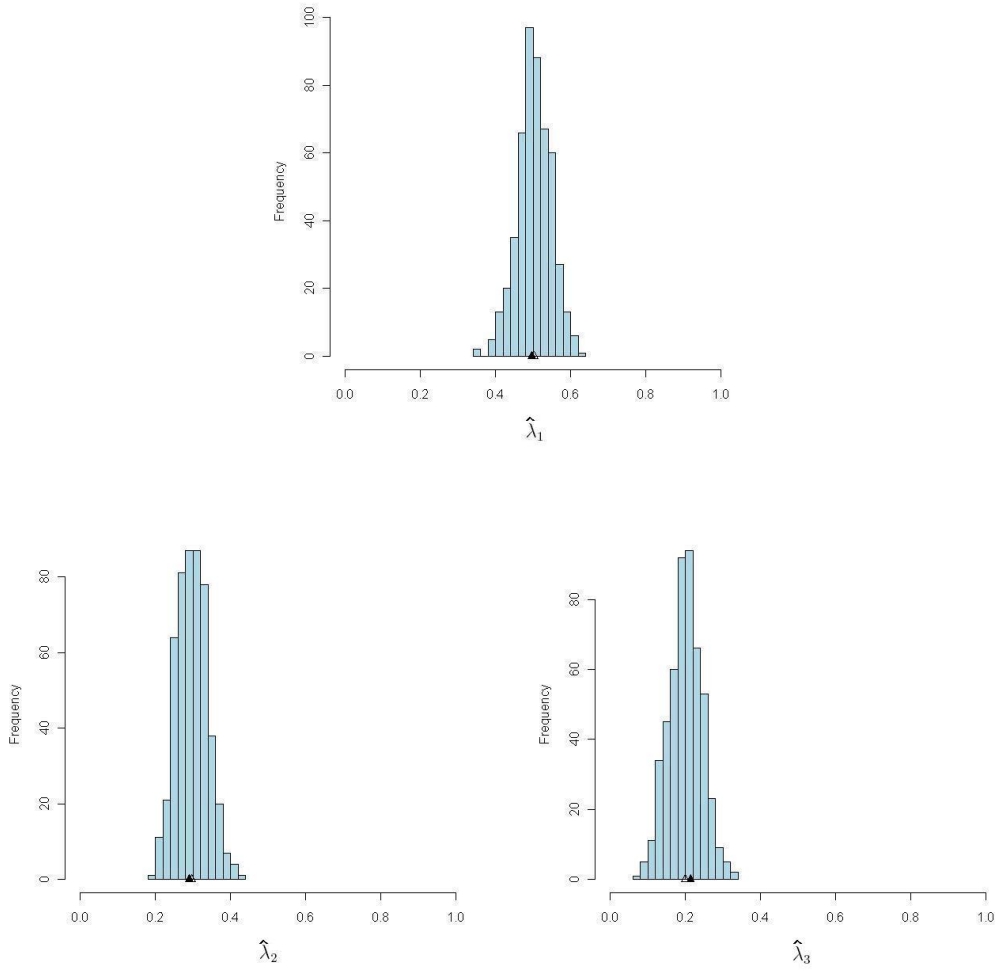

Figure 2.9: Histograms of the Bayes estimates of λ_1, λ_2, and λ_3 under the stochastic block model based on 500 samples. The solid triangles represent the MLEs of the model parameters based on a full graph realization, and the transparent triangles represent the average of the Bayes estimates of the model parameters.

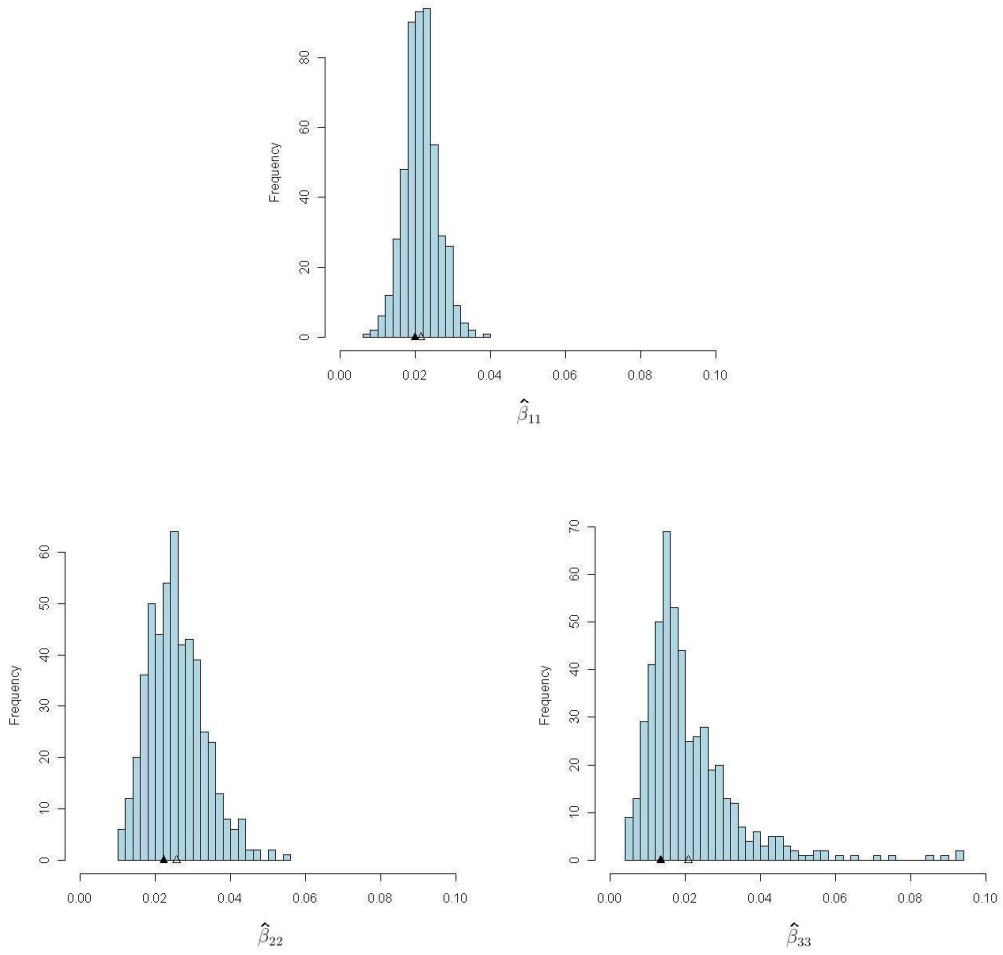

Figure 2.10: Histograms of the Bayes estimates of β_{11}, β_{22}, and β_{33} under the stochastic block model based on 500 samples. The solid triangles represent the MLEs of the model parameters based on a full graph realization, and the transparent triangles represent the average of the Bayes estimates of the model parameters.

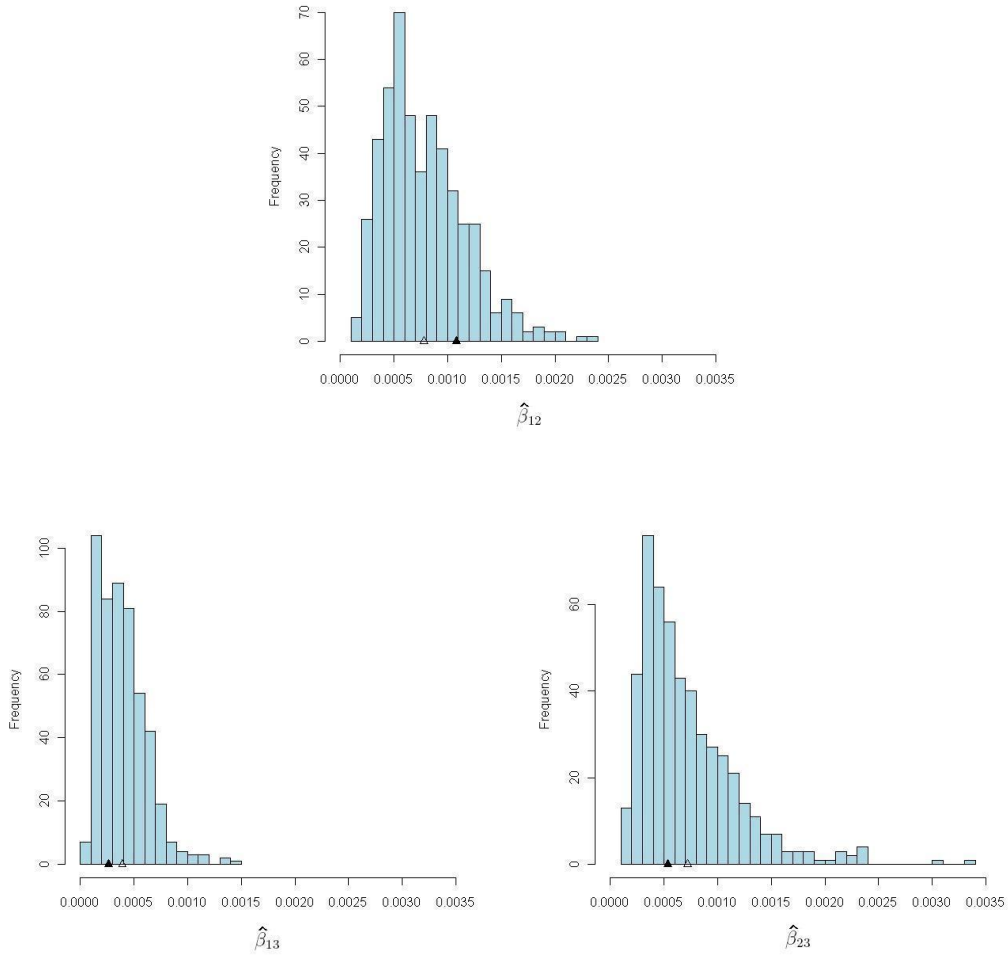

Figure 2.11: Histograms of the Bayes estimates of β_{12}, β_{13}, and β_{23} under the stochastic block model based on 500 samples. The solid triangles represent the MLEs of the model parameters based on a full graph realization, and the transparent triangles represent the average of the Bayes estimates of the model parameters.

It appears that the Bayes estimates come out with little to no bias for the true population size and MLEs of the model parameters. For populations that exhibit a similar behaviour to realizations based on the stochastic cluster model, these results

suggest that using the stochastic block model in conjunction with a complete one-wave snowball sampling design may be a robust choice for inferring on the population size and parameters. One reason for this is to notice that the stochastic block model can be regarded as a special case of the stochastic cluster model. There is an exception in that we allow for links within and between groups under the stochastic block model to be governed by their own parameter $\beta_{k,l}, k, l = 1, 2, ..., G$. Note that for the stochastic cluster model the corresponding parameters are (β_0, α_0) and (β_1, α_1), however notice that the parameters $\underline{\mu}$ and $\underline{\sigma}^2$ indirectly contribute to the modeling of the links. That is, we indirectly account for the locations of the covariate information in the modeling of the presence of links.

2.4.2 Simulation study based on the use of the stochastic cluster model

This subsection contains a presentation and discussion of the Bayes estimates of the population size and parameters based on the use of the stochastic cluster model. Again, we commence with the MCMC trace plots for the population size and model parameters based on the sample presented in Figure 2.2 as the behaviour of the MCMC trace plots based on the data from several other samples were found to be similar. The MCMC trace plots of the population size and model parameters can be found in Figures 2.12, 2.13, 2.14, 2.15, and 2.16. We shall also reiterate here that our study is based on a model-based approach to inference and since we are exploring our sampling and inference strategy applied to only one simulated network graph, we will use the maximum likelihood estimates (MLEs) of the population parameters based on the full graph realization as the target values for our Bayes estimates. As the MLEs of the model parameters will likely change between simulated networked graphs our choice of using the MLEs based on this network graph is justified.

The solid lines in the figures indicate the maximum likelihood estimate(s) (MLE) of the population parameters based on a full graph realization. Note that the full graph realization permits observing N to be the true value of 300 (since we have observed all units and their information relevant to the model). Also, note that in

the following figures the trace plots are for only one sample and hence the trace plots may not average to the corresponding MLEs.

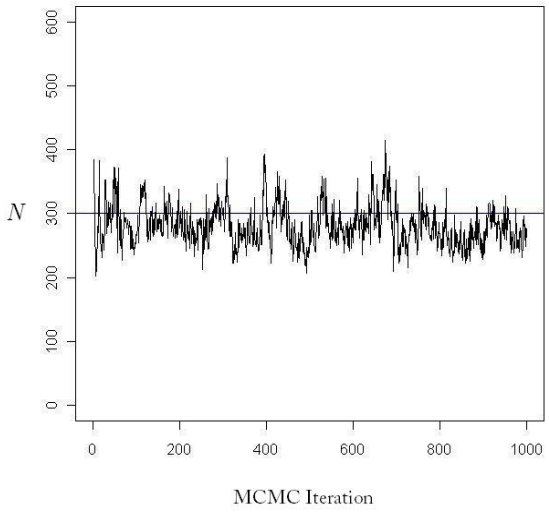

Figure 2.12: The sample dependent MCMC trace plot of the N values under the stochastic cluster model based on the sample presented in Figure 2.2. The solid line represents the MLE of the population size based on a full graph realization.

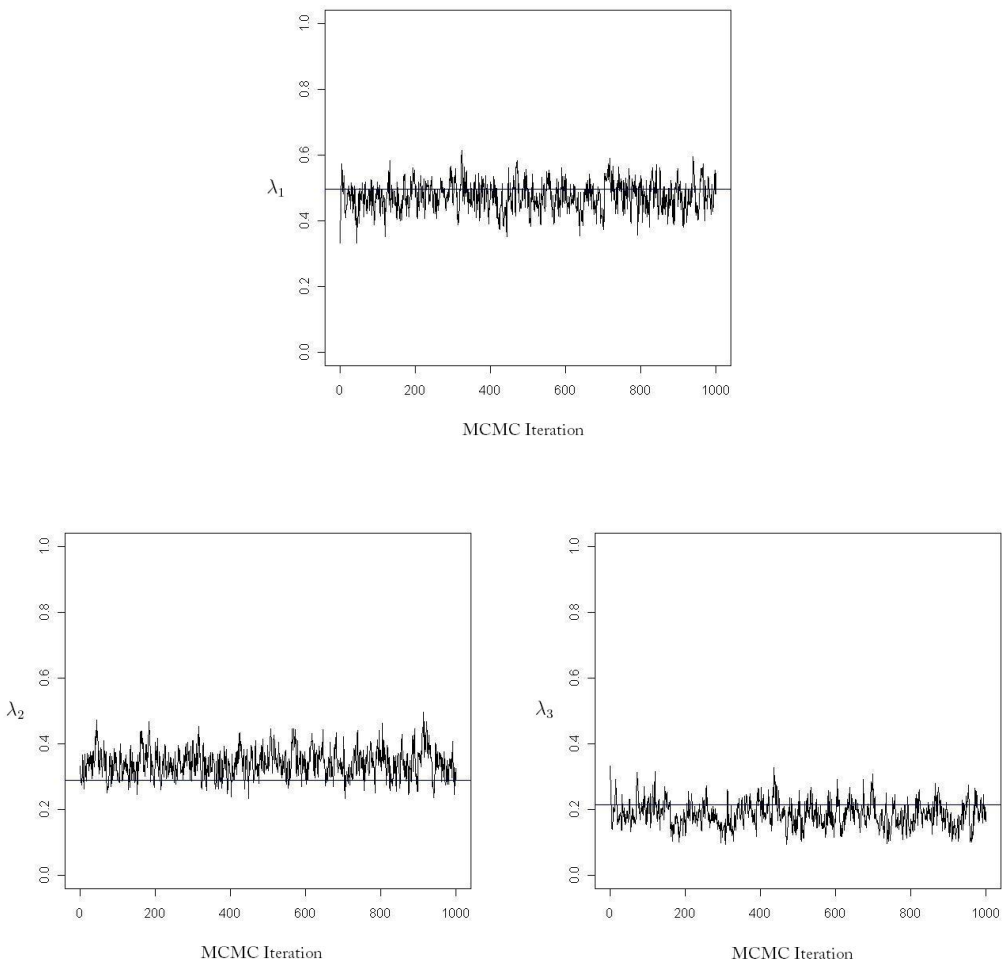

Figure 2.13: The sample dependent MCMC trace plots of λ_1, λ_2, and λ_3 under the stochastic cluster model based on the sample presented in Figure 2.2. The solid lines represent the MLEs of the model parameters based on a full graph realization.

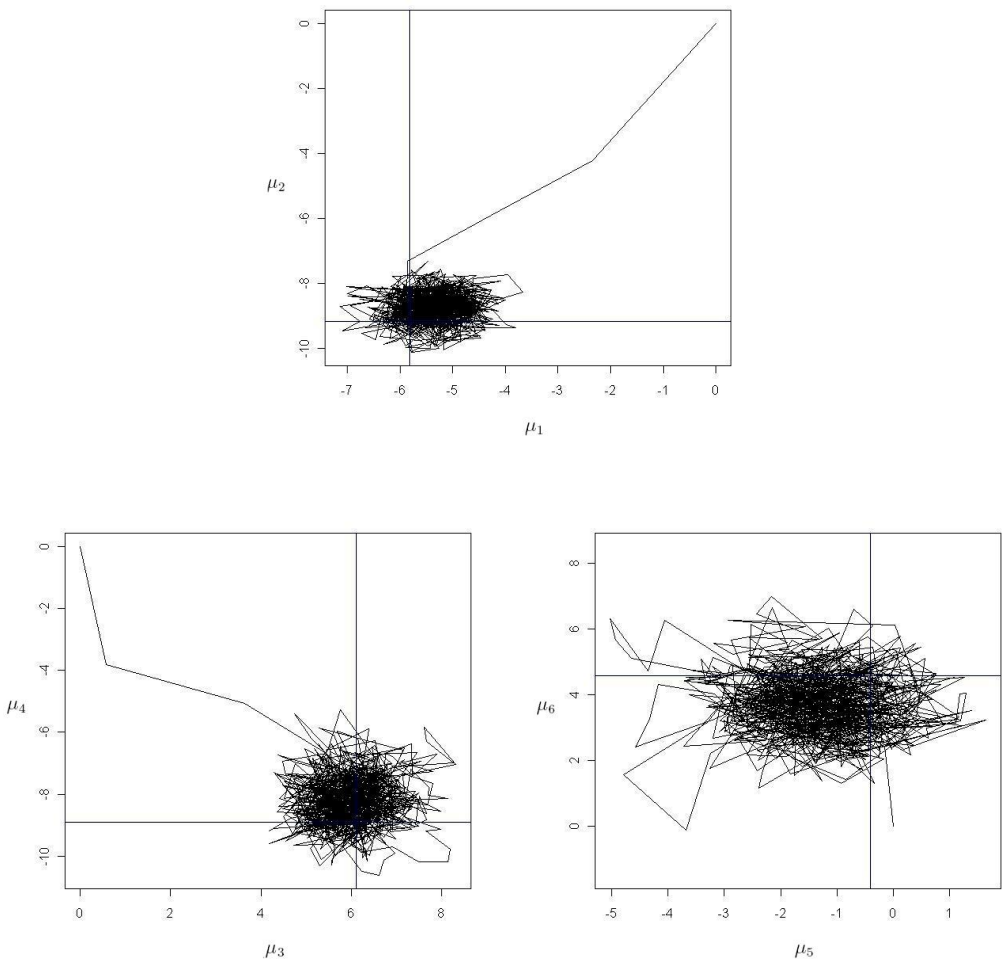

Figure 2.14: The sample dependent MCMC trace plots of $(\mu_1, \mu_2), (\mu_3, \mu_4)$, and (μ_5, μ_6) under the stochastic cluster model based on the sample presented in Figure 2.2. The solid lines represent the MLEs of the model parameters based on a full graph realization.

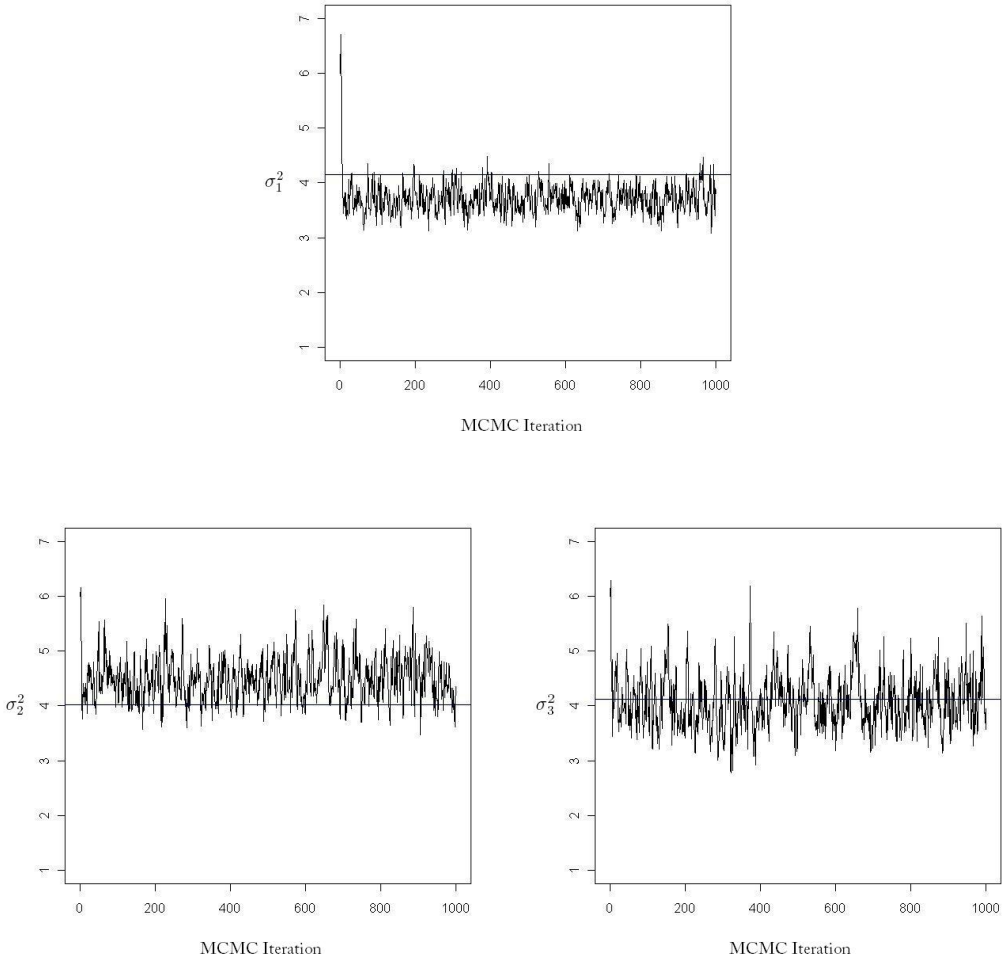

Figure 2.15: The sample dependent MCMC trace plots of σ_1^2, σ_2^2, and σ_3^2 under the stochastic cluster model based on the sample presented in Figure 2.2. The solid lines represent the MLEs of the model parameters based on a full graph realization.

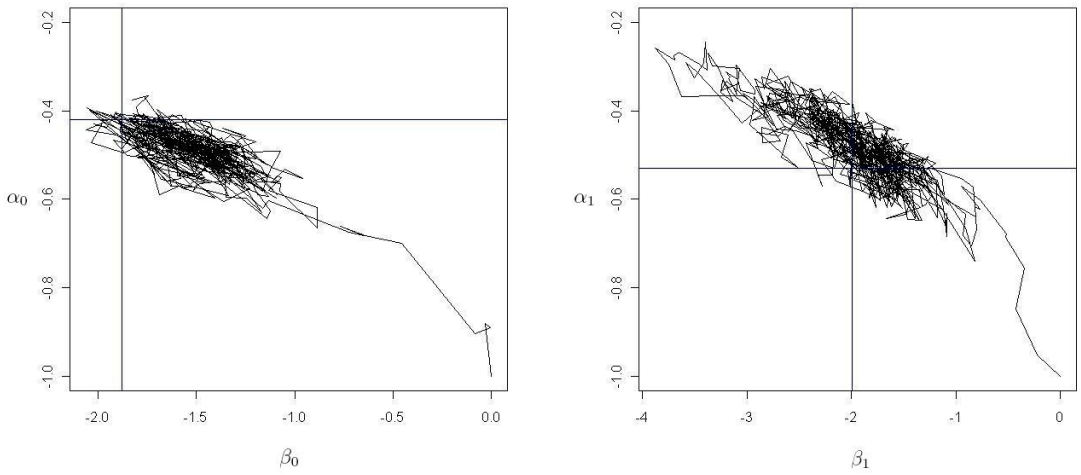

Figure 2.16: The sample dependent MCMC trace plots of (β_0, α_0), and (β_1, α_1) under the stochastic cluster model based on the sample presented in Figure 2.2. The solid lines represent the MLEs of the model parameters based on a full graph realization.

The MCMC trace plots for the population size and model parameters for the stochastic cluster model exhibit a behaviour similar to the MCMC trace plots for the population size and model parameters corresponding with the stochastic block model, namely fast convergence and insensitivity to the initial seeds which are reflective of the noninformative prior distributions (we set $\underline{\lambda}^{(0)} = (\underline{1/3}), \underline{\mu}^{(0)} = (\underline{0})$ and $\underline{\sigma}^{2\ (0)} = (\underline{36})$). Again, it appears that the posterior space has been adequately explored and that a chain of length 1000 is sufficient.

The MCMC trace plots based on (β_0, α_0) and (β_1, α_1) appear to exhibit a diagonal trend. This is to be anticipated as the logistic probability function that we base links between units on is linear in the parameters and hence this is a result of the sampling correlation between the estimates. Notice that the trace plots show that the posterior space has been explored well and that a chain of length 1000 is sufficient. Also notice that the chain has exhibited insensitivity to the noninformative seeds of $(\alpha_0^{(0)}, \beta_0^{(0)}) = (\alpha_1^{(0)}, \beta_1^{(0)}) = (0, -1)$, which give rise to a relatively dense population on the first augmentation iteration. A summary of the convergence of the MCMC chains

can be found in Figure 2.17. This figure presents the four hypothetical realizations of the population graph at augmentation iterations 1, 5, 50, and 1000. Notice how quickly the augmented graphs resemble the shape of the true population graph.

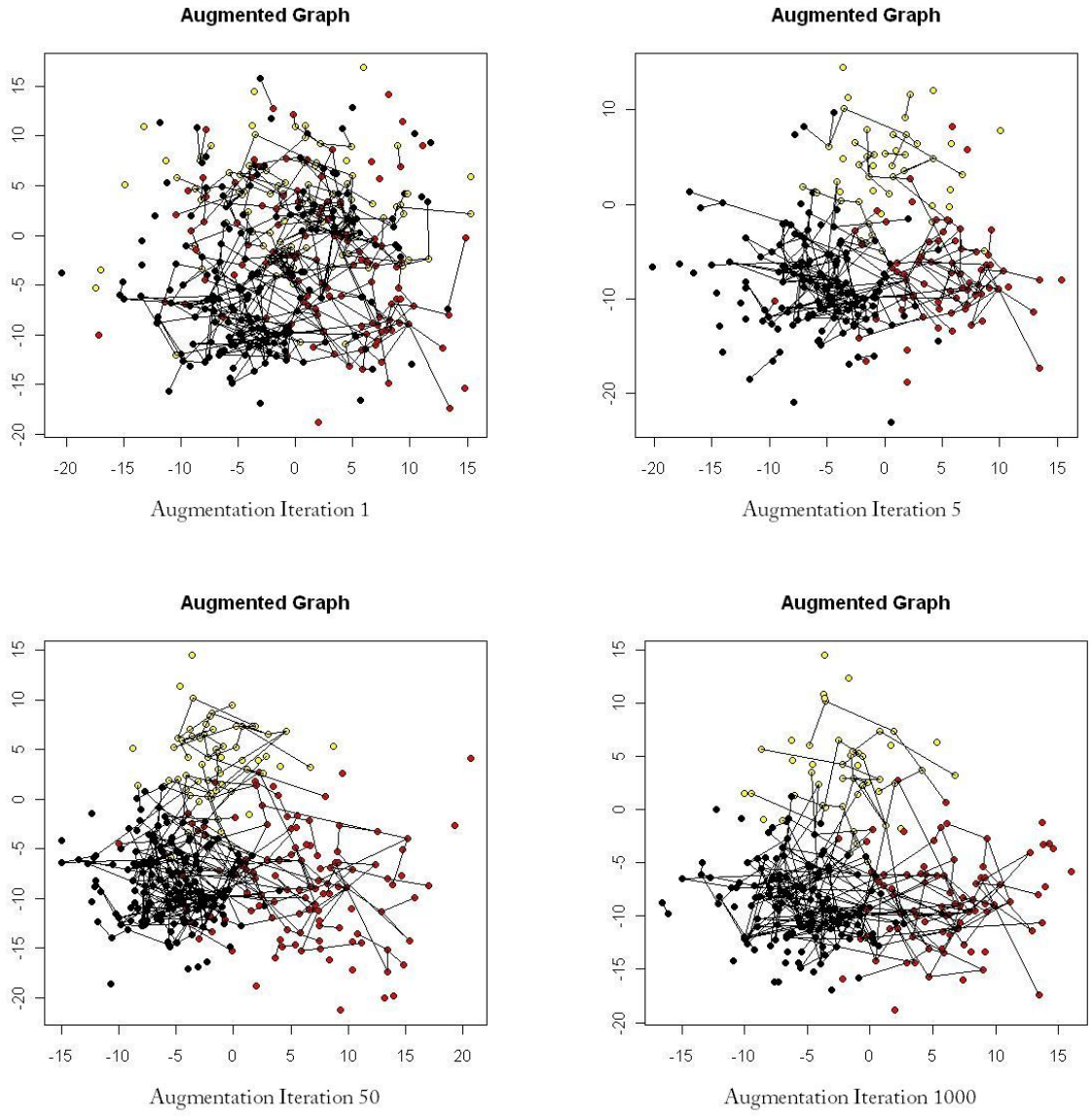

Figure 2.17: The augmented missing data for the population graph for augmentation iterations 1, 5, 50, and 1000 under the stochastic cluster model based on the sample presented in Figure 2.2.

Histograms/scatterplots of the Bayes estimates of the population size and model parameters can be found in Figures 2.18, 2.19, 2.20, 2.21, and 2.22. The solid triangle for the histograms in Figures 2.18, 2.19 and 2.21 and dark lines for the scatterplots in Figures 2.20 and 2.22 represent the corresponding MLEs of the population parameters based on a full graph realization. The transparent triangle for the histograms and shaded lines for the scatterplots represent the average of the Bayes estimates for the population size and model parameters.

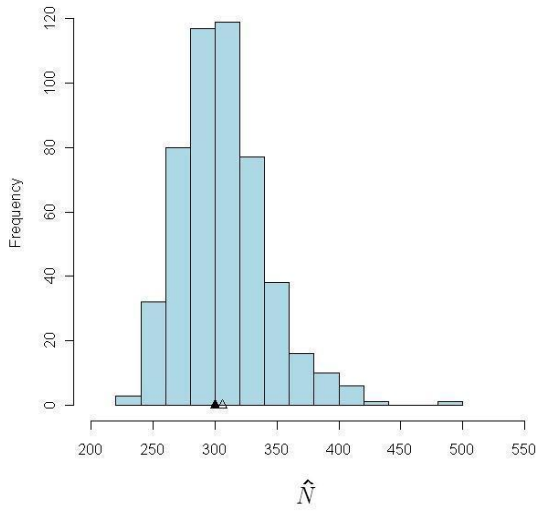

Figure 2.18: Histogram of the Bayes estimates of the N values under the stochastic cluster model based on 500 samples. The solid triangle represents the MLE of the population size based on a full graph realization (which is the true population size of 300), and the transparent triangle represents the average of the Bayes estimate of the population size.

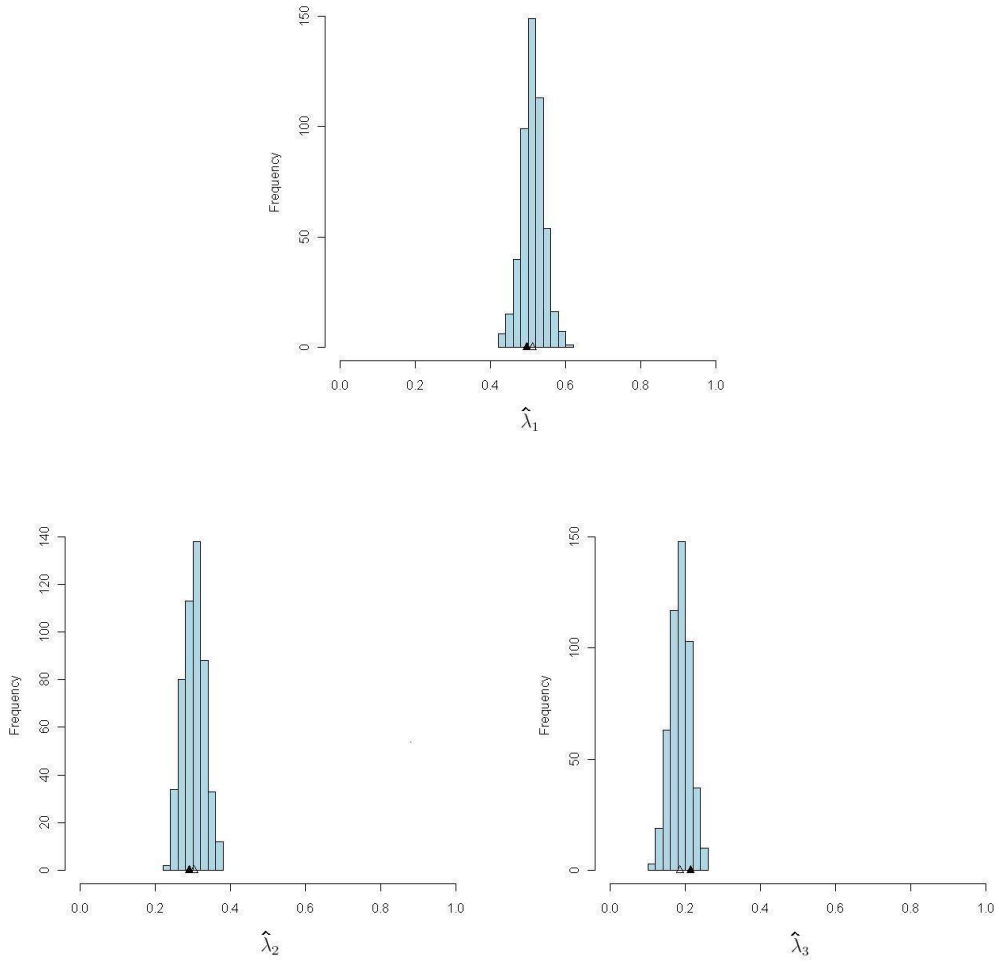

Figure 2.19: Histograms of the Bayes estimates of λ_1, λ_2, and λ_3 under the stochastic cluster model based on 500 samples. The solid triangles represent the MLEs of the model parameters based on a full graph realization, and the transparent triangles represent the average of the Bayes estimates of the model parameters.

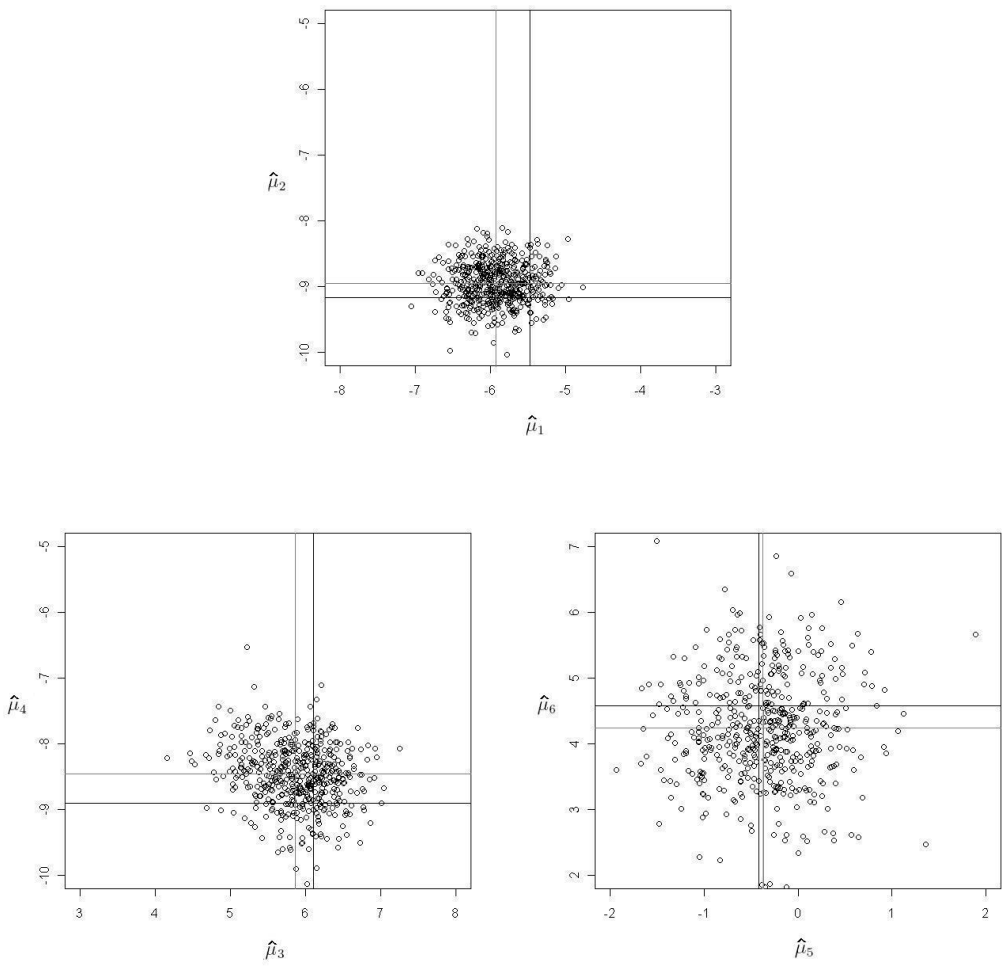

Figure 2.20: Scatterplot of the Bayes estimates of $(\mu_1, \mu_2), (\mu_3, \mu_4)$, and (μ_5, μ_6) under the stochastic cluster model based on 500 samples. The solid lines represent the MLEs of the model parameters based on a full graph realization, and the shaded lines represent the average of the Bayes estimates of the model parameters.

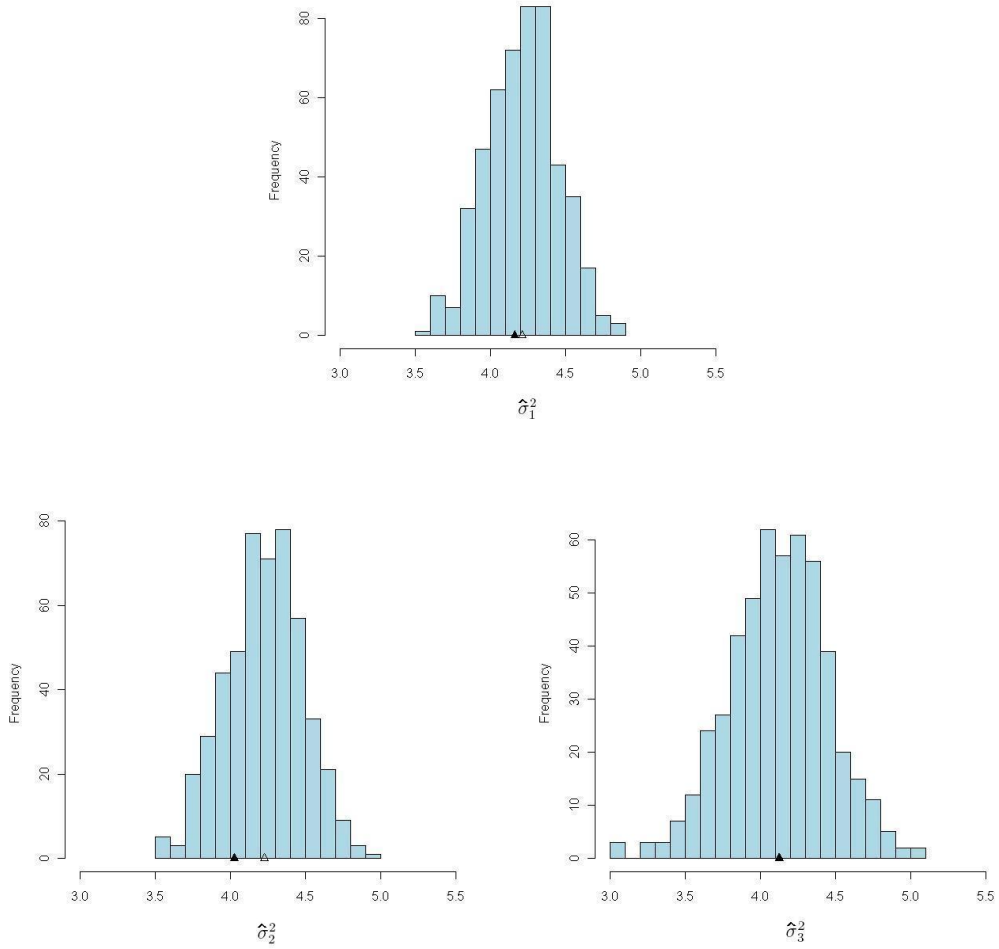

Figure 2.21: Histograms of the Bayes estimates of σ_1^2, σ_2^2, and σ_3^2 under the stochastic cluster model based on 500 samples. The solid triangles represent the MLEs of the model parameters based on a full graph realization, and the transparent triangles represent the average of the Bayes estimates of the model parameters.

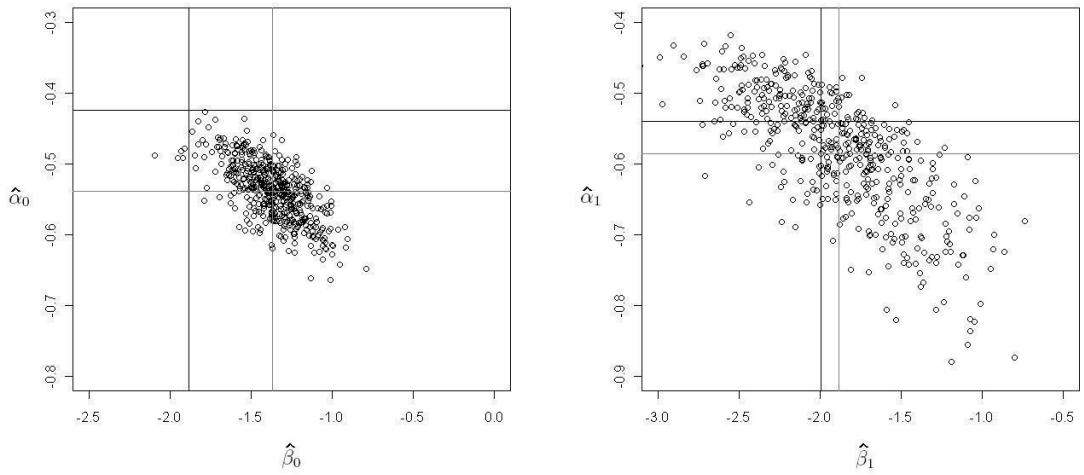

Figure 2.22: Scatterplot of the Bayes estimates of (β_0, α_0), and (β_1, α_1) under the stochastic cluster model based on 500 samples. The solid lines represent the MLEs of the model parameters based on a full graph realization, and the shaded lines represent the average of the Bayes estimates of the model parameters.

As expected, the Bayes estimates of the population size and $\underline{\lambda}$ appear to come out approximately unbiased 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 also come out approximately unbiased. However, the estimates for the parameters which govern the links within groups (that is (α_0, β_0)) appear to come out with some bias. The diagonal trend illustrates the sampling correlation between these parameters, and therefore the behaviour of the estimates, is not to be of concern as a less than expected influence of one parameter is likely to be compensated by its counterpart (notice that as the estimates of α_0 increase the corresponding estimates of β_0 decrease, and vice-versa).

2.5 Discussion

The results presented in this chapter show that the Bayes estimates of the population size and model parameters for both the stochastic block model and the stochastic cluster model have performed well and are approximately unbiased for the MLEs based on the full graph realization of the simulated population. A reasonable amount of variability is found for Bayes estimates of the population size and model parameters corresponding to the stochastic block model, whereas the Bayes estimates for the population size and $\underline{\lambda}$ parameters corresponding to the stochastic cluster model came out superior. Hence, incorporating the covariate information into the model has proven to be very beneficial as it appears to provide better estimates of the population size and parameters.

The methods presented in this chapter can be extended to a more general case where the covariate information of the members that comprise the first wave is not observed. Kwanisai (2004) outlined a strategy for making inference for the two-group stochastic block model parameters for such a case. Extending this method to work over the stochastic cluster model when the population size is unknown should be considered.

As an adaptive sampling procedure may result in a reduction in the effort required to recruit additional units from the target population once an initial sample is obtained, new work based on sampling over additional waves is deserving of attention. For example, extending these methods to work over an adaptive web sampling design (Thompson, 2006a) may prove to be highly useful when studying empirical populations. The immediate challenge is presented where we may not be able to exploit the conceptually straightforward use of the binomial distribution, as outlined in expressions (2.11) and (2.20), that accompanies the complete one-wave snowball sampling design in the inference procedure.

In contrast, further recruitment from the initial sample may prove to be a cumbersome task for the sampler in some empirical settings and hence extending the methods presented in this chapter to work over a snowball subsampling design, where a subset of the nominated members from the initial sample cannot be identified, should be

considered.

Chapter 3

The Multi-Sample Design-Based Approach

3.1 Introduction

In this chapter we introduce a new design-based method for estimating the size of networked hard-to-reach populations based on independent samples selected through a link-tracing design. Our method permits adaptively selected members of the target population to be included in the inference procedure through a Rao-Blackwellization method based on a sufficient statistic given the observed data. Moreover, our method possesses an additional advantage over the existing inferential methods based on estimating population sizes with link-tracing designs; our method permits the possibility of obtaining members not within arms reach of the initial sample (that is, those not immediately linked to the initial sample).

In Section 3.2, we introduce the notation that is used in this chapter. In Section 3.3, we outline the link-tracing sampling designs that are explored, namely one which is analogous to the original adaptive web sampling design that was introduced by Thompson (2006a) as well as a nearest neighbours adaptive web sampling design that has the potential for more practical use for sampling from a hidden population. Section 3.4 develops estimators of the population size and average node degree

of the population as well as the estimates of the variances of these estimators. As tabulating the preliminary estimates from all reorderings of the final samples is computationally cumbersome for the samples selected in this study, in Section 3.5 we outline a Markov chain resampling procedure to obtain approximations of the Rao-Blackwellized estimates. In Section 3.6, we perform a two-sample simulation study for the two sampling designs on a simulated networked population, and then draw conclusions and provide a general discussion of the novel methods developed in this chapter in Section 3.7.

3.2 Sampling Setup

We define a population U to consist of the set of units/individuals $U = \{1, 2, ..., N\}$ where N is the population size. Each pair of units (i, j), $i, j = 1, 2, ..., N$, is associated with a weight w_{ij}. We set $w_{ij} = 1$ if there is a link (or predetermined relationship) from unit i to unit j, and zero otherwise. We define $w_{ii} = 0$ for all $i = 1, 2, ..., N$.

We shall refer the reader to Section B.1 of Appendix B for an illustration to help clarify the notation and designs that are outlined in the following sections. Now, an adaptive web sampling design that is selected without replacement consists of two stages. Suppose a study is based on K samples. For each sample $k = 1, 2, ..., K$, the sample selection procedure commences with the selection of n_{0k} members completely at random and then $n_k - n_{0k}$ members are added adaptively without replacement. The adaptively selected members are added as follows. For each step t_k, $t_k = 1, 2, ..., n_k - n_{0k}$, any member i not yet chosen is selected with probability $q_{t_k,i} = d\frac{w_{a_{t_k},i}}{w_{a_{t_k},+}} + (1-d)\frac{1}{N-(n_{0k}+t_k-1)}$ where $w_{a_{t_k},i}$ is the number of links from the current active set $a_{t_k} \subseteq s_{t_k}$ (where s_{t_k} the current sample at time t_k) out to unit i at step t_k and $w_{a_{t_k},+}$ is the number of links out of the current active set to members not yet selected at step t_k. Hence, with probability $0 \leq d \leq 1$ a unit is added via tracing a link from the active set and with probability $1 - d$ a unit is added completely at random (a random jump is taken), given that $w_{a_{t_k},+} > 0$. In the event that $w_{a_{t_k},+} = 0$, a member is selected completely at random (that is, a random jump is taken) amongst those not yet selected with probability $\frac{1}{N-(n_{0k}+t_k-1)}$.

The observed data is $d_0 = \{(i, w_{ij}, w_i^+, t_{i,k}), \underline{J}_k : i, j \in s_k,\ k = 1, 2, ..., K\}$ where s_k refers to sample k for $k = 1, 2, ..., K$; w_i^+ is the out-degree of individual i (that is, the number of members acknowledged by individual i); $t_{i,k}$ is the time (or step) in the sampling sequence that unit i is selected for sample k; $\underline{J}_k$ is an indicator vector of length $L = \max_{j=1,2,...,K}\{n_j\}$ that records the sequence of jumps after the initial sample is selected for sample $k, k = 1, 2, ..., K$. It shall be understood that for all $k = 1, 2, ..., K$, $J_{1,k}, ..., J_{n_{0k},k} = 0$ and if $n_k < \max_{j=1,2,...,K}\{n_j\}$ then $J_{n_k+1,k}, ..., J_{L,k} = 0$. As demonstrated in web appendix B, a sufficient statistic for the population size and unobserved adjacency data can be shown to be $d_r = \{(i, w_{ij}, w_i+), \underline{\mathcal{J}} : i, j \in s_k,\ k = 1, 2, ..., K\}$ where $\underline{\mathcal{J}} = (\sum_{k=1}^{K} J_{1,k}, \sum_{k=1}^{K} J_{2,k}, ..., \sum_{k=1}^{K} J_{L,k}) = (\mathcal{J}_1, \mathcal{J}_2, ..., \mathcal{J}_L)$. In summary, the sufficient statistic removes the time element that is assigned to each unit that is selected for each sample and reduces the records of when random jumps are taken to a sum of the number of random jumps that are taken at each step in the sample selection procedure. We shall note here that Thompson (2006) showed that, when the population size is known, the observed data for a set of adaptive web samples selected independently is $d_0 = \{(i, w_{ij}, w_i^+, t_{i,k}) : i, j \in s_k,\ k = 1, 2, ..., K\}$ and hence a sufficient statistic is $d_r = \{(i, w_{ij}, w_i+) : i, j \in s_k,\ k = 1, 2, ..., K\}$. Notice that in our study, since the population size is unknown we require additional information in the observed data to utilize a sufficient statistic for the purposes of formulating Rao-Blackwellized estimates.

In our study we consider the special case where each sample is selected based on a design that does not allow for random jumps after the initial samples are selected (that is, $d = 1$). As our study does not permit for random jumps, the observed data is reduced to $d_0 = \{(i, w_{ij}, w_i^+, t_{i,k}) : i, j \in s_k,\ k = 1, 2, ..., K\}$ and hence a sufficient statistic based on the full data set can be shown to be $d_r = \{(i, w_{ij}, w_i^+) : i, j \in s_k,\ k = 1, 2, ..., K\}$ (notice that this corresponds with the reduced case from the preceding sufficiency result where $d = 1$ and $\underline{\mathcal{J}} \equiv 0$).

3.3 The Sampling Designs

In this chapter, we explore the use of two different adaptive web sampling designs, the first being the original design outlined by Thompson (2006a) and the second being a nearest neighbours adaptive web sampling design. In this section, we will outline the selection process we use in our study for the two adaptive web sampling designs for the special case of when $d = 1$.

The first sampling design selects independently K adaptive web samples as follows. The sampling procedure commences with the selection of an initial sample s_0 of size n_0 of members from the population completely at random. A predetermined maximum number of individuals, $n-n_0$ say, are further selected sequentially to bring the sample size up to $n' \leq n$ by tracing links, when available, out of the current active set as follows (notice that we place $d = 1$ in our study design). For any step t, $t = 1, 2, ..., n - n_0$, any member i that has not yet been selected is selected for inclusion with probability $q_{t,i} = \frac{w_{a_t,i}}{w_{a_t,+}}$, where $w_{a_t,i}$ is the number of links from the current active set a_t to unit i, and $w_{a_t,+}$ is the number of links out of the current active set to members not yet selected. Hence, if the number of links to trace out of the current active set is exhausted at any intermediate step in the sampling process then sampling stops at the most recent step $t - 1$ so that the final sample size is $n' = n_0 + t - 1$. In the original adaptive web sampling design, the active set consists of all members that have been selected for the current sample so that recruitment at any intermediate step is based on all links stemming out of the current sample. That is, for any step t, $t = 1, 2, ..., n - n_0$, any member i not yet chosen is selected with probability $q_{t,i} = \frac{w_{s_t,i}}{w_{s_t,+}}$ where s_t represents the current sample. Notice that with the original adaptive web sampling design it is possible to select members that are not directly linked to the initial sample.

The nearest neighbours adaptive web sampling design restricts the active set to consist only of those members who are selected for the initial sample so that only the units that are linked to the initial sample have a positive probability of being selected for the final sample. To clarify, for any step t, $t = 1, 2, ..., n - n_0$, any member i not yet chosen is selected with probability $q_{t,i} = \frac{w_{s_{0,t},i}}{w_{s_{0,t},+}}$ where $w_{s_{0,t},i}$ is the number of

links from the initial sample out to unit i at step t and $w_{s_{0,t},+}$ is the number of links out of the initial sample to members not yet selected at step t.

For each sample $k = 1, 2, ..., K$ we shall let s_{0k} represent the initial random sample corresponding with sample k where $|s_{0k}| = n_{0k}$. We shall also let s_k represent the final sample k in the order it was selected, where $|s_k| = n'_k = n_{01}, n_{01} + 1, ..., n_k$. For inferential purposes we shall define $s_{(0_1,0_2,...,0_K)}$ to be the full ordered sample of the samples in the respective original order they were selected. The probability of selecting the sample $s_{(0_1,0_2,...,0_K)}$ can then be expressed as

$$p(s_{(0_1,0_2,...,0_K)}) = \prod_{k=1}^{K} \left(\frac{1}{\binom{N}{n_{0k}}} \prod_{t_k=0}^{n'_k - n_{0k}} q_{t_k}^{s_k} \right).$$

The first terms in the expression correspond with the random selection of the initial samples and $q_{t_k}^{s_k}$ is the probability of adaptively selecting the unit that was selected at step t_k for sample k. It shall be understood that for $t_k = 0$, $q_{t_k}^{s_k} = 1$ for $k = 1, 2, ..., K$.

3.4 Estimation

3.4.1 Population size estimators

Suppose that $\hat{N}_0$ is a preliminary estimate of the population size based on the K initial random samples. For example, in a two sample study a preliminary estimate of the population size based on the initial random samples is the Lincoln-Petersen estimator (Petersen, 1896), or its more practical counterpart, the bias-adjusted Lincoln-Petersen (LP) estimator proposed by Chapman (1951). This estimator is of the form

$$\hat{N}_0 = \frac{(n_{01} + 1)(n_{02} + 1)}{m + 1} - 1, \tag{3.1}$$

where m denotes the number of individuals that are selected for both initial samples s_{01} and s_{02}. An improved estimator based on the sufficient statistic d_r is

$$E[\hat{N}_0|d_r] = \hat{N}_{RB} = \sum_{r_1=1}^{n_1'!} \sum_{r_2=1}^{n_2'!} \cdots \sum_{r_K=1}^{n_K'!} \hat{N}_0^{(r_1,r_2,...,r_K)} p(s_{(r_1,r_2,...,r_K)}|d_r) \tag{3.2}$$

where $\hat{N}_0^{(r_1,r_2,...,r_K)}$ is the preliminary population size estimate based on the hypothetical initial samples corresponding with reorderings $(r_1, r_2, ..., r_K)$ of samples $1, 2, ..., K$, respectively, and $p(s_{(r_1,r_2,...,r_K)}|d_r)$ is the conditional probability of obtaining the sample reorderings $r_1, r_2, ..., r_K$ given the data observed for d_r. Notice that for any specific sample reordering $s_{(x_1,x_2,...,x_K)}$ (that is, $(x_1, x_2, ..., x_K)$ are specific indices of the corresponding reorderings of the samples $(s_1, s_2, ..., s_K)$, respectively), the conditional probability of obtaining these sample reorderings can be expressed as

$$\begin{aligned}
&p(s_{(x_1,x_2,...,x_K)}|d_r) = p(s_{(x_1,x_2,...,x_K)}) / \sum_{r_1=1}^{n_1'!} \sum_{r_2=1}^{n_2'!} \cdots \sum_{r_K=1}^{n_K'!} p(s_{(r_1,r_2,...,r_K)}) \\
&= \frac{1}{\binom{N}{n_{01}}} \prod_{t_1=0}^{n_1'-n_{01}} q_{t_1}^{s_{x_1}} \times \frac{1}{\binom{N}{n_{02}}} \prod_{t_2=0}^{n_2'-n_{02}} q_{t_2}^{s_{x_2}} \times \cdots \times \frac{1}{\binom{N}{n_{0K}}} \prod_{t_K=0}^{n_K'-n_{0K}} q_{t_K}^{s_{x_K}} / \\
&\sum_{r_1=1}^{n_1'!} \sum_{r_2=1}^{n_2'!} \cdots \sum_{r_K=1}^{n_K'!} \left(\frac{1}{\binom{N}{n_{01}}} \prod_{t_1=0}^{n_1'-n_{01}} q_{t_1}^{s_{r_1}} \times \frac{1}{\binom{N}{n_{02}}} \prod_{t_2=0}^{n_2'-n_{02}} q_{t_2}^{s_{r_2}} \times \cdots \times \frac{1}{\binom{N}{n_{0K}}} \prod_{t_K=0}^{n_K'-n_{0K}} q_{t_K}^{s_{r_K}} \right) \\
&= \prod_{t_1=0}^{n_1'-n_{01}} q_{t_1}^{s_{x_1}} \times \prod_{t_2=0}^{n_2'-n_{02}} q_{t_2}^{s_{x_2}} \times \cdots \times \prod_{t_K=0}^{n_K'-n_{0K}} q_{t_K}^{s_{x_K}} / \\
&\sum_{r_1=1}^{n_1'!} \sum_{r_2=1}^{n_2'!} \cdots \sum_{r_K=1}^{n_K'!} \left(\prod_{t_1=0}^{n_1'-n_{01}} q_{t_1}^{s_{r_1}} \times \prod_{t_2=0}^{n_2'-n_{02}} q_{t_2}^{s_{r_2}} \times \cdots \times \prod_{t_K=0}^{n_K'-n_{0K}} q_{t_K}^{s_{r_K}} \right).
\end{aligned} \tag{3.3}$$

The essence of using the sufficient statistic is highlighted by bringing to the reader's attention that all terms involving the unknown population size N are factored out of the expression and can be canceled to make computation of the Rao-Blackwellized estimates possible. Notice that we have proved that the statistic d_r is sufficient for N since the ratio of the probability of selecting any two data points from the same

partition that the statistic induces (that is, the data points are simply reorderings of the respective samples) does not depend on the unknown population size N.

3.4.2 Alternative population size estimator for the two-sample study

An estimator of the population size can be obtained by taking the analogue of the Lincoln-Petersen estimator based on the initial random selection for the first sample and the full sample for the second sample (Seber, 1982). This estimator can be expressed as

$$\hat{N}_{0,1} = \frac{(n_{01}+1)(n_2'+1)}{m_1+1} - 1, \tag{3.4}$$

where n_{01} is the initial sample size of the first sample, n_2' is the size of the second sample, and m_1 is the number of individuals selected for both initial sample 1 and final sample 2. The Rao-Blackwellized version of this estimator is expressed as

$$E[\hat{N}_{0,1}|d_r] = \hat{N}_{RB,1} = \sum_{r_1=1}^{n_1'!} \hat{N}_{0,1}^{(r_1,r_2)} p(s_{(r_1,r_2)}|d_r), \tag{3.5}$$

where $\hat{N}_{0,1}^{(r_1,r_2)}$ is the estimate of N obtained with the hypothetical initial sample of reordering r_1 of sample 1 and all of (reordered) sample 2 (notice that this estimate does not actually depend on the order in which sample 2 was selected).

Another estimator is formulated by taking the average of the two corresponding estimators, that being

$$\hat{N}_{0,1,2} = \frac{\hat{N}_{0,1} + \hat{N}_{0,2}}{2}. \tag{3.6}$$

The corresponding Rao-Blackwellized version of this estimator is expressed as

$$E[\hat{N}_{0,1,2}|d_r] = \hat{N}_{RB,1,2} = \frac{\sum_{r_1=1}^{n_1'!} \hat{N}_{0,1}^{(r_1,r_2)} p(s_{(r_1,r_2)}|d_r) + \sum_{r_2=1}^{n_2'!} \hat{N}_{0,2}^{(r_1,r_2)} p(s_{(r_1,r_2)}|d_r)}{2}. \quad (3.7)$$

We shall note here that in the event that a fixed list of n_2' individuals from the target population has previously been obtained, whether it be through a probability sampling design or not, one can still utilize $\hat{N}_1$ and $\hat{N}_{RB,1}$ to estimate the population size. The reason for this is that the Lincoln-Petersen estimator only requires one of the two samples to be obtained completely at random (Seber, 1982).

3.4.3 Average node degree estimators

Estimates of the distribution of individual responses like the out-degree of the population members can be useful to the researcher when inferring on hard-to-reach populations. We can obtain estimates of such characteristics like the average out-degree of the population members as follows. For notational convenience, we shall let $M = \bigcup_{k=1}^{K} s_{0k}$. We can then estimate the average out-degree of the population,

$$w_\mu = \frac{\sum_{i=1}^{N} w_i^+}{N}, \quad (3.8)$$

with the estimator based on the unique members selected for the initial samples, namely

$$\hat{w_0} = \frac{\sum_{i \epsilon M} w_i^+}{|M|}. \quad (3.9)$$

Conditional on $|M|$ this estimator can be viewed as being based on a random sample of $|M|$ individuals selected without replacement. Therefore, conditional on $|M|$,

$\hat{w}_0$ can be shown to be an unbiased estimator for w_μ. The Rao-Blackwellized version of the preliminary estimator of the average out-degree is made possible through the same procedure as obtaining the Rao-Blackwellized version of a preliminary estimator of the population size. The corresponding formula used for obtaining the Rao-Blackwellized version of $\hat{w}_0$ is

$$E[\hat{w}_0|d_r] = \hat{w}_{RB} = \sum_{r_1=1}^{n_1'!} \sum_{r_2=1}^{n_2'!} \cdots \sum_{r_K=1}^{n_K'!} \hat{w}_0^{(r_1,r_2,\ldots,r_K)} p(s_{(r_1,r_2,\ldots,r_K)}|d_r). \tag{3.10}$$

We shall note here that estimates of the average node degree which are based on the two-sample study and that are analogous to $\hat{N}_{0,1}$ will introduce some bias into the estimator and therefore are not explored in this chapter.

3.4.4 Variance estimators

Schwarz and Seber (1999) outlined several methods for obtaining estimates of the variance of the population size estimates based on a K sample capture-recapture study. In our two-sample study we shall take, as an estimator of the variance of the preliminary estimators $\hat{N}_0$ and $\hat{N}_{0,1}$, the estimator that was proposed by Seber (1970). These estimators are of the form

$$\hat{\text{Var}}(\hat{N}_0) = \frac{(n_{01}+1)(n_{02}+1)(n_{01}-m)(n_{02}-m)}{(m+1)^2(m+2)} \tag{3.11}$$

and

$$\hat{\text{Var}}(\hat{N}_{0,1}) = \frac{(n_{01}+1)(n_2'+1)(n_{01}-m_1)(n_2'-m_1)}{(m_1+1)^2(m_1+2)}. \tag{3.12}$$

As $\text{Var}(\hat{N}_{0,1,2}) = \text{Var}(\frac{\hat{N}_{0,1}+\hat{N}_{0,2}}{2}) = \frac{1}{4}\text{Var}(\hat{N}_{0,1}) + \frac{1}{4}\text{Var}(\hat{N}_{0,2}) + \frac{1}{2}\text{Cov}(\hat{N}_{0,1}, \hat{N}_{0,2})$, we will take a conservative estimate of this value to be

$$\hat{\text{Var}}(\hat{N}_{0,1,2}) = \frac{1}{4}\hat{\text{Var}}(\hat{N}_{0,1}) + \frac{1}{4}\hat{\text{Var}}(\hat{N}_{0,2}) + \frac{1}{2}\hat{\text{Var}}(\hat{N}_{0,1}). \tag{3.13}$$

An estimate of the variance of $\hat{w}_0$ is the conditionally unbiased estimate

$$\hat{\text{Var}}(\hat{w}_0|M) = \left(\frac{N - |M|}{N}\right)\frac{s^2}{|M|}, \tag{3.14}$$

where $\frac{N-|M|}{N}$ corresponds with the finite population correction factor and $s^2 = \frac{1}{|M|-1}\sum_{i\epsilon M}(w_i^+ - \hat{w}_0)^2$. As the population size is not known in advance we shall substitute N with $\hat{N}_0$ in the finite population correction factor.

To estimate the variance of the Rao-Blackwellized estimators, Thompson (2006a) proposed the following unbiased estimator. For any estimator $\hat{\theta}_{RB} = E[\hat{\theta}_0|d_r]$ for some population unknown θ, where $\hat{\theta}_0$ is the preliminary estimate, the conditional decomposition of variances gives

$$\text{Var}(\hat{\theta}_{RB}) = \text{Var}(\hat{\theta}_0) - E[\text{Var}(\hat{\theta}_0|d_r)]. \tag{3.15}$$

An unbiased estimator of $\text{Var}(\hat{\theta}_{RB})$ is

$$\hat{\text{Var}}(\hat{\theta}_{RB}) = E[\hat{\text{Var}}(\hat{\theta}_0)|d_r] - \text{Var}(\hat{\theta}_0|d_r). \tag{3.16}$$

This estimator is the difference of the expectation of the estimated variance of the preliminary estimator over all reorderings of the data and the variance of the preliminary estimator over all the reorderings of the data. As this estimator can result in negative estimates of the variance, a conservative approach is take the estimate of $\text{Var}(\hat{\theta}_{RB})$ to be $E[\hat{\text{Var}}(\hat{\theta}_0)|d_r]$ when such a scenario arises.

3.5 Markov Chain Resampling Estimators

Due to the large number of sample permutations that are obtained with the sample sizes used in this study, a Markov chain resampling procedure similar to the one found in Thompson (2006a) is implemented to obtain approximations of the Rao-Blackwellized estimates. As the sampling strategy presented in this paper selects multiple independent adaptive web samples, the Markov chain resampling strategy needs to be modified. We outline the modified Markov chain accept/reject (Hastings, 1970) resampling procedure below.

Suppose θ is an unknown population quantity we wish to estimate with the improved estimator $\hat{\theta}_{RB} = E[\hat{\theta}_0|d_r]$ where d_r is a sufficient statistic.

Step 0: Let $\hat{\theta}_0^{(0)}$ be the estimated value of θ and $\hat{\text{Var}}(\hat{\theta}_0^{(0)})$ be the estimated value of $\text{Var}(\hat{\theta}_0)$ that is obtained from selecting K adaptive samples in the original order they were selected. Also, let $t^{(0)} = s_{(0_1,0_2,...,0_K)}$ be the ordered original samples in the order they were selected.

For step $l = 1, 2, ..., R$, where R is sufficiently large:
Draw a candidate sample reordering, $t^{(l)}$ say, from a candidate distribution (that is, $t^{(l)}$ is an ordered set of reorderings of each sample). Suppose the most recently accepted candidate reordering is $t^{(y)}$ for some ordered set of reorderings of the samples where $y = 0, 1, 2, ..., l-1$. Let $p(t^{(l)})$ be the probability of obtaining $t^{(l)}$ under the true population and $q(t^{(l)})$ be the probability of obtaining reordering $t^{(l)}$ under the candidate distribution. Generate a uniform random number between 0 and 1, and if this value is less than

$$\min\left\{\frac{p(t^{(l)})}{p(t^{(y)})}\frac{q(t^{(y)})}{q(t^{(l)})}, 1\right\}, \tag{3.17}$$

let $\hat{\theta}_0^{(l)}$ and $\hat{\text{Var}}(\hat{\theta}_0^{(l)})$ be the estimates of θ and $\text{Var}(\hat{\theta}_0)$, respectively, obtained with the ordered set of sample reorderings $t^{(l)}$. Otherwise, take $\hat{\theta}_0^{(l)} = \hat{\theta}_0^{(l-1)}$ and $\hat{\text{Var}}(\hat{\theta}_0^{(l)}) =$

$\hat{\text{Var}}(\hat{\theta}_0^{(l-1)})$. Recall that $p(t^{(l)})$ needs only be known for the (hypothetical) adaptive recruitment probabilities found in the corresponding ordered set of sample reorderings as all terms involving the unknown population size N can be factored out of the ratio of the true probabilities of obtaining sample reorderings and cancelled from the expression.

Final step: Take estimates of $\hat{\theta}_{RB}$ to be

$$\tilde{\theta}_{RB} = \frac{1}{R+1} \sum_{l=0}^{R} \hat{\theta}_0^{(l)}, \tag{3.18}$$

and similarly take the estimate of $\hat{\text{Var}}(\hat{\theta}_{RB})$ to be

$$\begin{aligned} \tilde{\text{Var}}(\hat{\theta}_{RB}) &= \tilde{E}[\hat{\text{Var}}(\hat{\theta}_0)|d_r] - \tilde{\text{Var}}(\hat{\theta}_0|d_r) \\ &= \frac{1}{R+1} \sum_{l=0}^{R} \hat{\text{Var}}(\theta_0^{(l)}) - \frac{1}{R+1} \sum_{l=0}^{R} (\hat{\theta}_0^{(l)} - \tilde{\theta}_{RB})^2. \end{aligned} \tag{3.19}$$

With the adaptive web sampling designs restricted to only recruiting members that are linked to the current active set, and not allowing for random jumps (that is $d = 1$), a large number of the sample reorderings will likely have zero probability of being selected in the full population setting. One primary reason for this is that the sample reorderings that consist of at least one member added after the hypothetical current sample, with whom do not share a link to any previously selected members that are in the active set, result in a sample that is not sequentially obtainable under an adaptive web sampling design that does not permit for random jumps.

We used a candidate distribution which works over each sample individually and first places all their sampled units that are not nominated by any other sampled units into their hypothetical initial sample with probability one (notice that these members must be in the corresponding original initial sample). The candidate distribution then selects the remaining members for each individual sample based on the original

adaptive web sampling design, with a small probability of jumps allowed, applied to the reduced population that consists only of those sampled members.

3.6 Simulation Study

We will use the thesis study population to evaluate the new inference procedures outlined in this chapter. The population was generated according to the stochastic cluster model that was outlined in chapter 2. An illustration of the simulated population can be found in Figure 3.1.

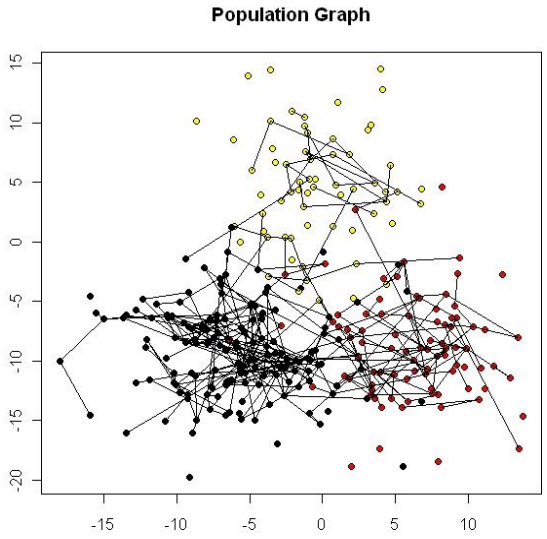

Figure 3.1: The simulated thesis study population.

Figure 3.2 shows two samples where the graph nodes are enlarged for ease of visualization. These members were selected under the original adaptive web sampling design and nearest neighbours adaptive web sampling design where 40 members are selected for the initial samples with (up to) 10 members added adaptively to each sample. The first sample is represented by light coloured nodes and the second

sample is represented by dark coloured nodes. Nodes that are selected for both samples are highlighted as shaded nodes. Notice the disproportionate increase in the overlap between the adaptively recruited members from the samples for both designs illustrating the additional information that may be harnessed for inferential purposes.

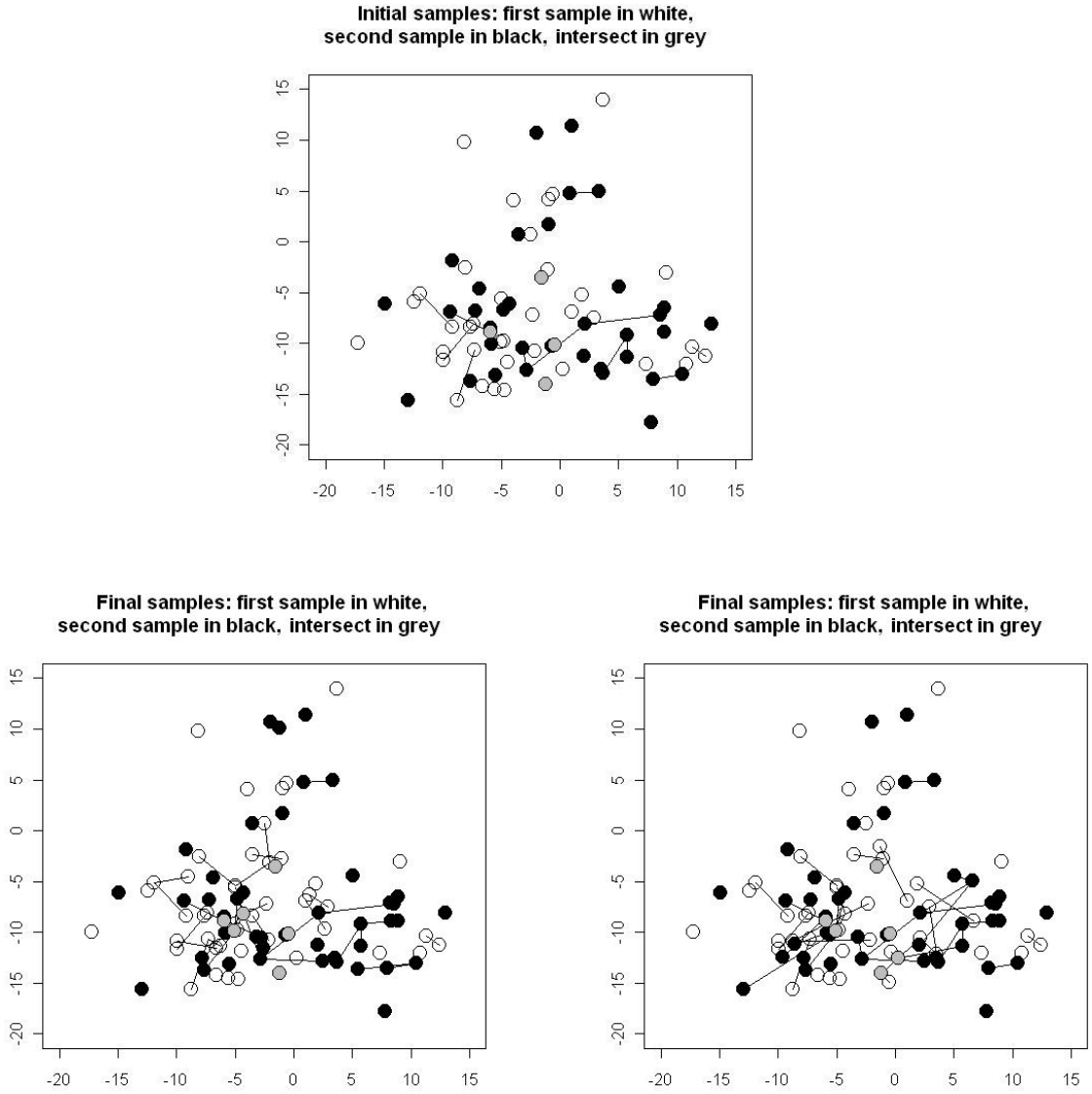

Figure 3.2: Two initial samples selected at random on the top. Two original adaptive web samples on the bottom left and two nearest neighbours adaptive web samples on the bottom right where both samples start with the two initial samples selected at random on the top. The size of each of the initial samples was 40 and (up to) 10 members are added adaptively. The first sample is represented by the light colored nodes, the second sample is represented by the dark colored nodes, and the intersection between the two samples is represented by the shaded nodes.

A simulation study was conducted as follows. A total of 1000 pairs of samples with 5000 resamples from each pair of samples for the Markov chain resampling procedure were obtained with each sampling design. Initial samples of size 40 with (up to) 10 members recruited adaptively were selected for each sample. Histograms of the estimates of the population size and average node degree are shown in Figures 3.3 and 3.4, respectively. The true population size of 300 and average node degree of 2.8 are indicated by the solid triangles on the x-axis of the corresponding graphs. All estimates came out approximately unbiased. Table 3.1 provides the standardized mean squared error (MSE) scores for each of the estimates. The scores are standardized by the MSE score obtained with the preliminary estimates.

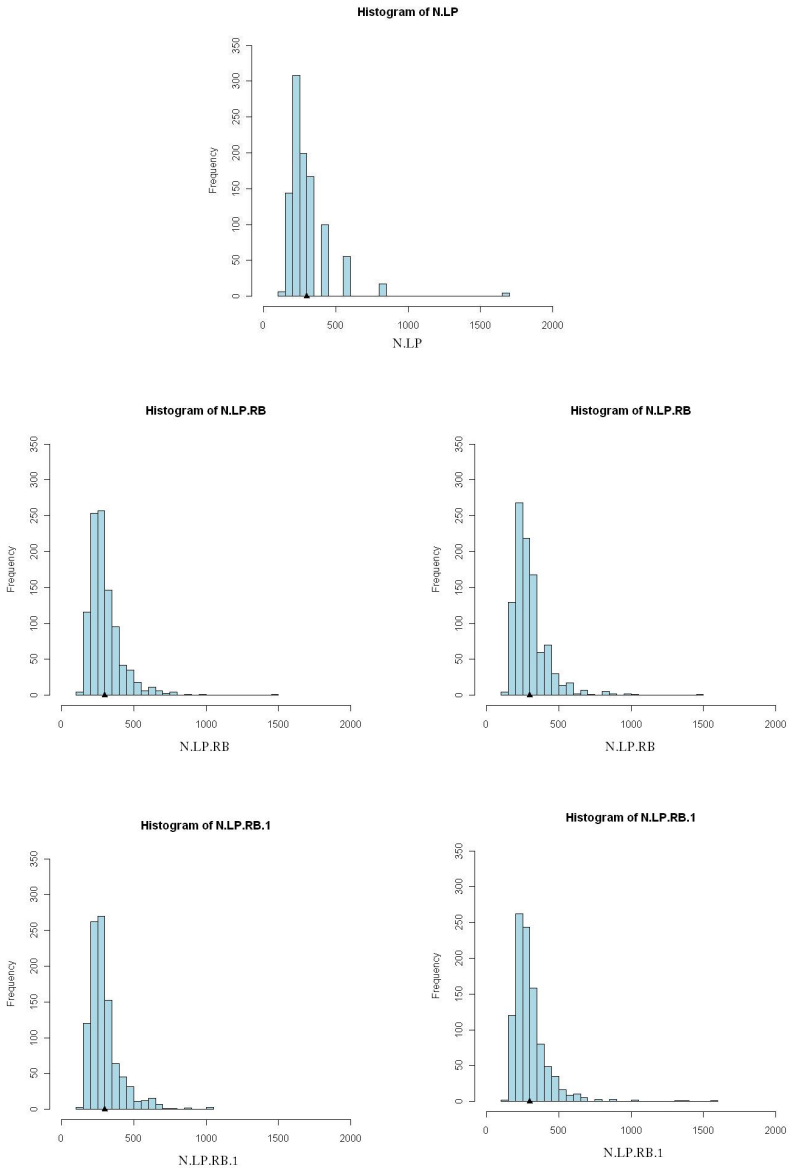

Figure 3.3: Histograms of the population size estimates with $\hat{N}_0$ on top and $\hat{N}_{RB}$ based on the original adaptive web sampling design and the nearest neighbours adaptive web sampling design, respectively, in the middle. Histograms of $\hat{N}_{RB,1}$ based on the original web adaptive sampling design and the nearest neighbours adaptive web sampling design, respectively, on the bottom. The dark triangles on the x-axis indicate the true population size of 300. All estimates came out approximately unbiased.

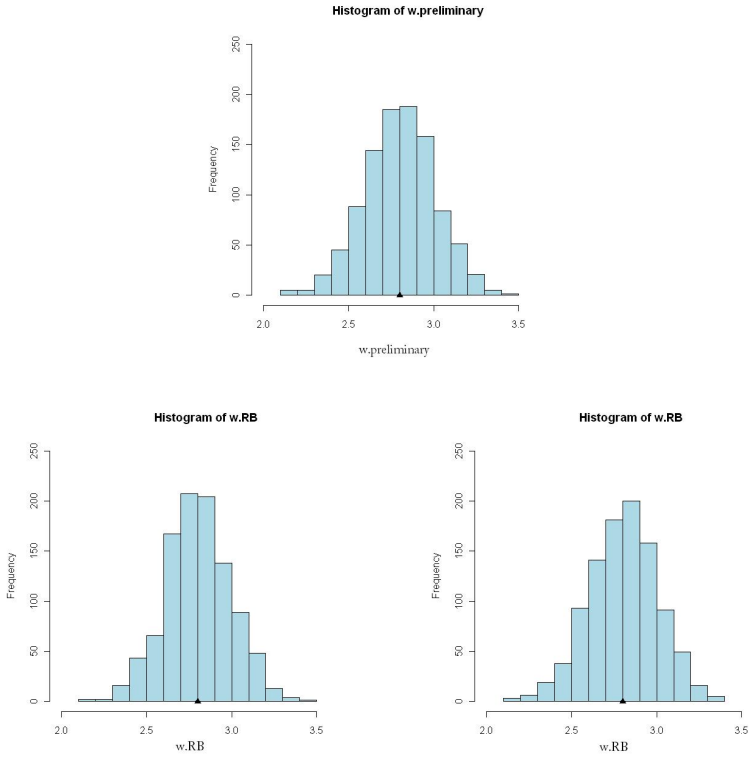

Figure 3.4: Histograms of the average node degree estimates with $\hat{w}_0$ on top and $\hat{w}_{RB}$ on the bottom based on the original adaptive web sampling design and the nearest neighbours adaptive web sampling design, respectively. The dark triangles on the x-axis indicate the true population average node degree of 2.8. All estimates came out approximately unbiased.

Table 3.1: Standardized MSE scores for the estimates of the population size and average node degree where OR refers to the original adaptive web sampling design and NN refers to the nearest neighbours adaptive web sampling design.

	Estimator				
	$\hat{N}_0$	$\hat{N}_{RB}, \hat{w}_{RB}$ OR	$\hat{N}_{RB}, \hat{w}_{RB}$, NN	$\hat{N}_{RB,1}$, OR	$\hat{N}_{RB,1}, NN$
Parameter					
N	1	0.566	0.637	0.518	0.594
w_μ	1	0.857	0.947		

As illustrated in the histograms and table of scores, gains in precision are made over the preliminary estimates for the population size when using the improved estimates. It also appears that the improved estimates for the average out-degree of the population made significant gains in precision over their preliminary estimator counterparts. $\hat{N}_{RB,1}$ exhibited the best performance, offering some improvement over $\hat{N}_{RB}$. Another estimator of the population size can be obtained by taking the average of $\hat{N}_{RB,1}$ and $\hat{N}_{RB,2}$ (where $\hat{N}_{RB,2}$ is the Rao-Blackwellized estimator based on the preliminary estimator $\frac{(n_1'+1)(n_{02}+1)}{m_2+1}$). In this study, these estimates came out highly correlated and offered minimal improvement over $\hat{N}_{RB,1}$.

Tables 3.2 and 3.3 give the coverage rates of the population size using nominal 95% confidence intervals based on the Central Limit Theorem (CLT) and the log transformation strategy outlined in Chao (1987), respectively. As can be seen, the coverage rates for the population size based on the central limit theorem are smaller than 95%, possibly due to the skewed shape of the distribution of the estimates which in turn may be influenced by the small sample sizes used in the study. However, it is apparent that the log transformation strategy has helped to improve the coverage rates of the population size.

Table 3.2: Coverage rates of the population size using nominal 95% confidence intervals based on the CLT where OR refers to the original adaptive web sampling design and NN refers to the nearest neighbours adaptive web sampling design.

	Estimator				
	$\hat{N}_0$	$\hat{N}_{RB}$ OR	$\hat{N}_{RB}$, NN	$\hat{N}_{RB,1}$, OR	$\hat{N}_{RB,1}, NN$
Parameter					
N	0.850	0.854	0.851	0.884	0.887

Table 3.3: Coverage rates of the population size using confidence intervals based on a log transformation where OR refers to the original adaptive web sampling design and NN refers to the nearest neighbours adaptive web sampling design.

	Estimator				
	$\hat{N}_{0,1}$ OR	$\hat{N}_{0,1}$, NN	$\hat{N}_{RB,1}$, OR	$\hat{N}_{RB,1}$, NN	
Parameter					
N	0.902	0.903	0.901	0.913	0.911

Table 3.4 gives the nominal 95% coverage rates for the average out-degree based on the Central Limit Theorem. The coverage rates came out close to 95%, indicating that substituting the estimate of the population size into the corresponding variance expression found in expression (3.14) is a suitable choice.

Table 3.4: Coverage rates of the average node degree using the nominal 95% confidence intervals based on the CLT where OR refers to the original adaptive web sampling design and NN refers to the nearest neighbours adaptive web sampling design.

	Estimator		
	$\hat{w}_0$	$\hat{w}_{RB}$ OR	$\hat{w}_{RB}$, NN
Parameter			
w_μ	0.938	0.931	0.922

We shall note here that in our study we did not encounter any negative estimates of the variance of the Rao-Blackwellized estimates. Therefore, we did not have to resort to using the conservative approach that was suggested in subsection 3.4.4.

3.7 Discussion

In this chapter we have outlined a new inferential method that uses link-tracing strategies and a sufficient statistic to estimate the size of hard-to-reach populations. The new method possesses the ability to adaptively recruit hard-to-reach members for the study, through an adaptive sampling probability mechanism that can be tailored to meet the sampler's needs, without introducing additional bias into the improved estimates while allowing for control over sample sizes. As the theoretical results and simulation studies showed, the new methods outlined in this chapter will give rise to more precise estimators relative to those based on the Lincoln-Petersen estimator which is based on the initial samples.

One additional advantage the new methods presented in this chapter possess over some of the existing capture-recapture methods is outlined as follows. In some empirical settings when sampling from a large population with relatively small sample sizes, the selection of two random samples may give rise to little or no overlap in the samples, hence rendering an undesirable estimate of the population size when using a capture-recapture type of estimator. With the methods outlined in this chapter,

overlap between the adaptive recruitment stages of the samples is more certain and hence the use of the new inferential procedure should result in a much more reliable estimate of the population size.

Extending the methods outlined in this chapter to be compatible with the eight closed population models commonly used in capture-recapture studies, namely the $M_0, M_t, M_b, M_h, M_{tb}, M_{th}, M_{bh}, M_{tbh}$ models (see Schwarz and Seber (1999) for a description of these models), is certainly deserving of future attention. The methods presented in this chapter may serve as a foundation for the theory required to achieve these goals.

Chapter 4

The Single-Sample Design-Based Approach

4.1 Introduction

Frank and Snijders (1994) developed a design-based approach for estimating the size of a networked population based on the links (commonly referred to as arks or nominations) originating from the members selected from a Bernoulli sample. In this chapter we extend this method by allowing sampling to continue beyond the Bernoulli sample (also known as the initial wave) and base new estimates for the population size on the corresponding observations made from a succeeding wave. Estimates of the variance of the new estimators are based on a jackknife approach similar to that developed by Frank and Snijders (1994). A Rao-Blackwellization method for improving the preliminary estimators is also developed in this chapter. A simulation study is conducted on the thesis study population to evaluate the extended sampling and inferential methods outlined in this chapter.

In Section 4.2, we introduce the notation that is used in this chapter and we outline the link-tracing sampling designs that are explored as well as the statistics that are observed. Section 4.3 is reserved for developing moment-based estimators of the population size as well as the estimates of the variance of the population size

estimators. As tabulating the preliminary estimates from all reorderings of the final samples is computationally cumbersome due to the sample sizes used in this study, we outline a Markov chain resampling method in Section 4.4 to approximate the improved estimators. Section 4.5 provides a simulation study based on the thesis study population, and Section 4.6 provides a discussion of the novel methods that are introduced in this chapter.

4.2 Sampling Setup and Design

We define a population U to consist of the units/individuals $U = \{1, 2, ..., N\}$ where N is the population size. Following the approach developed by Frank and Snijders (1994), for all $i, j = 1, 2, ..., N$ we will let $w_{ij} = 1$ if there is a link from member i to member j and $w_{ij} = 0$ otherwise, and we will let $w_{ii} = 1$ for all $i = 1, 2, ..., N$. The adjacency neighbourhood of unit i, A_i, is defined as the set of all individuals with links to unit i. We can express A_i as $A_i = \{j : w_{ji} = 1\}$ and we will let $|A_i| = a_i$.

A diagram summarizing the statistics introduced in this section is provided below to aid understanding of the sampling procedure that is based on selecting one additional wave after the initial wave is selected. Estimates of the population size are based on two types of statistics which we will refer to as class A and class B statistics.

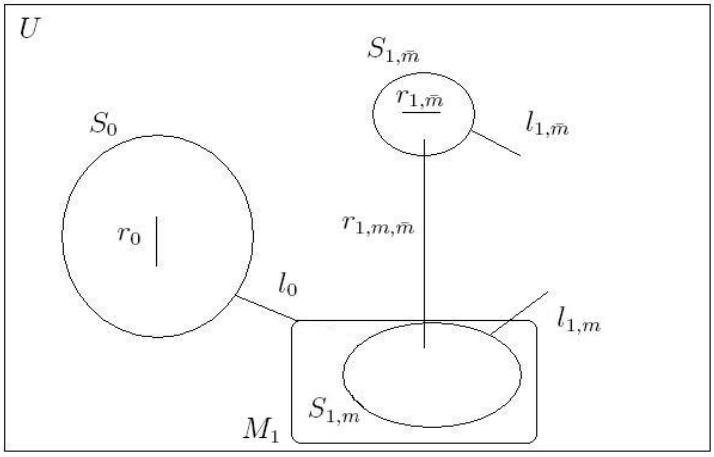

Figure 4.1: The class A statistics. The circles represent those members who are selected for the sample, the square represents those members nominated from the initial sample and who are outside the initial sample. The lines indicate nominations made from/between those members selected for the sample.

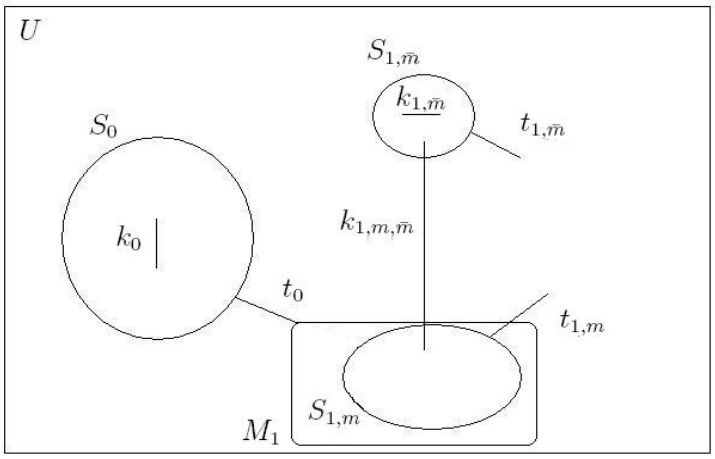

Figure 4.2: The class B statistics. The circles represent those members who are selected for the sample, the square represents those members nominated from the initial sample and who are outside the initial sample. The lines indicate nominations made from/between those members selected for the sample.

4.2.1 Statistics based on the initial wave

The sampling procedure starts with the selection of an initial sample/wave S_0 of size n_0 completely at random (that is, all units have the same probability of being selected for the initial sample) from the population.

The class A statistics based on the initial wave

We will let $X_i^{(0)} = 1$ if unit i is selected for the initial sample and 0 otherwise, and $L = \sum_{i \neq j} w_{ij}$ will be the total number of non-loop arcs (that is, non-self nominations) in the population. We will let r_0 be the number of non-loop arcs within S_0, that is, the number of nominations within S_0. We will also let l_0 be the number of links leading out of S_0. We can conveniently express the class A statistics r_0 and l_0 as

$$r_0 = \sum_{i \neq j} X_i^{(0)} X_j^{(0)} w_{ij} \tag{4.1}$$

and

$$l_0 = \sum_{i,j} X_i^{(0)} (1 - X_j^{(0)}) w_{ij}. \tag{4.2}$$

The class B statistics based on the initial wave

We will define k_0 to be the number of members selected for the initial sample that are mentioned by at least one other individual selected for the initial sample, and t_0 to be the number of members not selected for the initial sample that are mentioned by at least one member selected for the initial sample. The class B statistics can be expressed as

$$k_0 = \sum_j X_j^{(0)} \max_{i \epsilon A_j \setminus \{j\}} \{X_i^{(0)}\} \tag{4.3}$$

and

$$t_0 = \sum_j (\max_{i \epsilon A_j} \{X_i^{(0)}\} - X_j^{(0)}). \tag{4.4}$$

Notice that in a two sample capture-recapture study, the first sample corresponds with the set of members in the initial wave, the second sample corresponds with all members mentioned from the initial wave (not including self-nominations), and the recaptured members in the second sample correspond with the set of members mentioned by at least one other member in the initial wave and whom are also in the initial wave.

4.2.2 Statistics based on wave one

When sampling is permitted to continue beyond the selection of the initial wave, we can formulate parallel statistics based on those members selected for the succeeding wave with the following design. After selecting the initial wave, we will let M_1 be the set of individuals that are mentioned/nominated by those individuals in S_0 and that are not selected for S_0 (note that $|M_1| = t_0$). Now, each of the M_1 members are selected for wave one independently and with probability β_1 (that is, we select a Bernoulli sample from M_1) and those whom are selected comprise the set $S_{1,m}$. Also, an additional $\bar{m}_1$ members are selected completely at random for wave one from $U \setminus (S_0 \cup M_1)$ and these members comprise the set $S_{1,\bar{m}}$. Finally, we will define wave one to be $S_1 = S_{1,m} \cup S_{1,\bar{m}}$.

The class A statistics based on wave one

We will let $X_i^{(1,m)} = 1$ if unit i in M_1 is selected for $S_{1,m}$ and 0 otherwise. Similarly, we will let $X_i^{(1,\bar{m})} = 1$ if unit i in $U \setminus (S_0 \cup M_1)$ is selected for $S_{1,\bar{m}}$ and 0 otherwise. We will let $r_{1,\bar{m}}$ be the number of non-loop arks within $S_{1,\bar{m}}$ and $r_{1,m,\bar{m}}$ be the number of links from $S_{1,m}$ to $S_{1,\bar{m}}$. We will also let $l_{1,\bar{m}}$ be the number of links from $S_{1,\bar{m}}$ to $U \setminus (S_0 \cup M_1 \cup S_{1,\bar{m}})$ and $l_{1,m}$ be the number of links from $S_{1,m}$ to $U \setminus (S_0 \cup M_1 \cup S_{1,\bar{m}})$. We can express the aforementioned class A statistics as

$$r_{1,\bar{m}} = \sum_{\substack{i \neq j: \\ i,j \epsilon \overline{S_0 \cup M_1}}} X_i^{(1,\bar{m})} X_j^{(1,\bar{m})} w_{ij}, \tag{4.5}$$

$$r_{1,m,\bar{m}} = \sum_{\substack{i \epsilon M_1, \\ j \epsilon \overline{S_0 \cup M_1}}} X_i^{(1,m)} X_j^{(1,\bar{m})} w_{ij}, \tag{4.6}$$

$$l_{1,\bar{m}} = \sum_{i,j \epsilon \overline{S_0 \cup M_1}} X_i^{(1,\bar{m})}(1 - X_j^{(1,\bar{m})})w_{ij}, \tag{4.7}$$

and

$$l_{1,m} = \sum_{\substack{i \epsilon M_1, \\ j \epsilon \overline{S_0 \cup M_1}}} X_i^{(1,m)}(1 - X_j^{(1,\bar{m})})w_{ij}. \tag{4.8}$$

The class B statistics based on wave one

We will now define $k_{1,\bar{m}}$ to be the number of individuals in $S_{1,\bar{m}}$ mentioned by at least one other individual in $S_{1,\bar{m}}$ and $k_{1,m,\bar{m}}$ to be the number of individuals in $S_{1,\bar{m}}$ that are mentioned by at least one individual in $S_{1,m}$. We will also let $t_{1,\bar{m}}$ be the number of individuals in $U \setminus (S_0 \cup M_1 \cup S_{1,\bar{m}})$ that are mentioned by at least one individual in $S_{1,\bar{m}}$ and $t_{1,m}$ to be the number of individuals in $U \setminus (S_0 \cup M_1 \cup S_{1,\bar{m}})$ that are mentioned by at least one individual in $S_{1,m}$. We will now let $A_j^{(1,\bar{m})} = \{i : i \, \epsilon \, \overline{S_0 \cup M_1}, w_{ij} = 1\}$ and $A_j^{(1,m)} = \{i : i \, \epsilon \, M_1, w_{ij} = 1\}$. The aforementioned class B statistics can be expressed as

$$k_{1,\bar{m}} = \sum_{j \epsilon \overline{S_0 \cup M_1}} X_j^{(1,\bar{m})} \max_{i \epsilon A_j^{(1,\bar{m})} \setminus \{j\}} \{X_i^{(1,\bar{m})}\}, \tag{4.9}$$

$$k_{1,m,\bar{m}} = \sum_{\substack{j \epsilon \overline{S_0 \cup M_1}, \\ i \epsilon M_1}} X_j^{(1,\bar{m})} \max_{i \epsilon A_j^{(1,m)}} \{X_i^{(1,m)}\}, \tag{4.10}$$

$$t_{1,\bar{m}} = \sum_{j \epsilon \overline{S_0 \cup M_1}} \left(\max_{i \epsilon A_j^{(1,\bar{m})}} \{X_i^{(1,\bar{m})}\} - X_j^{(1,\bar{m})} \right), \tag{4.11}$$

and

$$t_{1,m} = \sum_{\substack{j \epsilon \overline{S_0 \cup M_1}, \\ i \epsilon M_1}} (1 - X_j^{(1,\bar{m})}) \max_{i \epsilon A_j^{(1,m)}} \{X_i^{(1,m)}\}. \tag{4.12}$$

We note here that additional observational effort is required by the sampler to identify nominations outside of the sample for the purpose of obtaining the class B statistics. To clarify, notice that for the class B statistics we require that individuals nominated outside of the sample be identifiable.

The probability of selecting a sample s with this sampling design can be expressed as

$$p(s) = \frac{1}{\binom{N}{n_0}} \beta_1^{m_1} (1 - \beta_1)^{|M_1| - m_1} \frac{1}{\binom{N - n_0 - |M_1|}{\bar{m}_1}}. \tag{4.13}$$

The first term corresponds with the random selection of the initial wave and the second and third terms correspond with the Bernoulli sample selected from those members mentioned from the initial wave. The fourth term corresponds with the random selection of those members not selected for the initial wave and that are not linked to any individuals in the initial wave.

Notice that in a two sample capture-recapture study, the first sample corresponds with the set of members in the completely random component of the first wave, the second sample corresponds with all members mentioned from the first wave (not including self-nominations), and the recaptured members in the second sample correspond with the set of members mentioned by at least one other member in the first wave and whom are also in the random component of the first wave.

4.3 Estimation

This section outlines the preliminary and Rao-Blackwellized estimators for the population size based on the initial wave and wave one as well as the jackknife procedure that Frank and Snijders (1994) used to estimate the variance of the estimates based on the initial wave. Note that the estimators defined in this section are moment-based estimators of the population size. We will make the definition of the moment expectation E_M to be $E_M[g_1(X_1)g_2(X_2)] = g_1(E[X_1])g_2(E[X_2])$ for functions g_1 and g_2 of the random variables X_1 and X_2, respectively.

We shall note here that Frank and Snijders (1994) developed an inference procedure based on an initial wave selected via a Bernoulli sampling design. In our study the design selects a predetermined number of individuals for the initial wave.

4.3.1 Population size estimators based on the initial wave

Frank and Snijders (1994) showed that a moment-based consistent estimator of the population size N based on the selection of a Bernoulli sample and the class A statistics is

$$\hat{N}_{A,0} = n_0\left(\frac{r_0 + l_0}{r_0}\right) \tag{4.14}$$

since

$$\begin{aligned} E_M[r_0] &= \sum_{i \neq j} E[X_i^{(0)}]E[X_j^{(0)}]w_{ij} \\ &= \left(\frac{n_0}{N}\right)^2 \sum_{i \neq j} w_{ij} \\ &= \left(\frac{n_0}{N}\right)^2 L, \end{aligned} \tag{4.15}$$

and

$$\begin{aligned} E_M[l_0] &= \sum_{i \neq j} E[X_i^{(0)}] E[1 - X_j^{(0)}] w_{ij} \\ &= \left(\frac{n_0}{N}\right)\left(\frac{N - n_0}{N}\right) \sum_{i \neq j} w_{ij} \\ &= \left(\frac{n_0}{N}\right)\left(\frac{N - n_0}{N}\right) L. \end{aligned} \tag{4.16}$$

Frank and Snijders (1994) also showed that a second moment-based consistent estimator of the population size N based on the class B statistics is

$$\hat{N}_{B,0} = n_0 \left(\frac{k_0 + t_0}{k_0}\right) \tag{4.17}$$

since

$$\begin{aligned} E_M[k_0] &= \sum_j E[X_j^{(0)}] E[\max_{i \epsilon A_j \setminus \{j\}} \{X_i^{(0)}\}] \\ &= \sum_j E[X_j^{(0)}] E[1 - \min_{i \epsilon A_j \setminus \{j\}} \{1 - X_i^{(0)}\}] \\ &= n_0 - \frac{n_0}{N} \sum_j \left(1 - \frac{n_0}{N}\right)^{(a_j - 1)}, \end{aligned} \tag{4.18}$$

and

$$\begin{aligned} E_M[t_0] &= \sum_j (E[\max_{i \epsilon A_j}\{X_i^{(0)}\}] - E[X_j^{(0)}]) \\ &= \sum_j (E[1 - \min_{i \epsilon A_j}\{1 - X_i^{(0)}\}] - E[X_j^{(0)}]) \\ &= N - n_0 - \sum_j \left(1 - \frac{n_0}{N}\right)^{a_j}. \end{aligned} \tag{4.19}$$

4.3.2 Population size estimators based on wave one

To obtain a moment-based estimator of the population size N based on wave one, we first condition on the sizes of the sets $S_{1,m}$ and $S_{1,\bar{m}}$ to be $|S_{1,m}| = m_1$ and $|S_{1,\bar{m}}| = \bar{m}_1$. For notational convenience, we will also let $N_1 = N - n_0 - |M_1|$, and L_{U_1} and L_{U_1,U_2} will denote the number of non-loop arks within U_1 and the number of links from U_1 to U_2 for $U_1, U_2 \subseteq U = \{1, 2, ..., N\}$, respectively. We will also let $|A_j^{(1,m)}| = a_j^{(1,m)}$ and $|A_j^{(1,\bar{m})}| = a_j^{(1,\bar{m})}$.

A conditional moment-based estimator of the population size N based on the class A statistics can now be shown to be

$$\hat{N}_{A,1} = n_0 + |M_1| + \bar{m}_1 \left(\frac{r_{1,\bar{m}} + r_{1,m,\bar{m}} + l_{1,\bar{m}} + l_{1,m}}{r_{1,\bar{m}} + r_{1,m,\bar{m}}} \right) \tag{4.20}$$

since

$$
\begin{aligned}
E_M[r_{1,\bar{m}}] &= \sum_{\substack{i\neq j,\\ i,j\epsilon\overline{S_0\cup M_1}}} E[X_i^{(1,\bar{m})}]E[X_j^{(1,\bar{m})}]w_{ij} \\
&= \left(\frac{\bar{m}_1}{N_1}\right)^2 \sum_{\substack{i\neq j,\\ i,j\epsilon\overline{S_0\cup M_1}}} w_{ij} \\
&= \left(\frac{\bar{m}_1}{N_1}\right)^2 L_{\overline{S_0\cup M_1}}, \qquad (4.21)
\end{aligned}
$$

$$
\begin{aligned}
E_M[r_{1,m,\bar{m}}] &= \sum_{\substack{i\epsilon M_1,\\ j\epsilon\overline{S_0\cup M_1}}} E[X_i^{(1,m)}]E[X_j^{(1,\bar{m})}]w_{ij} \\
&= \left(\frac{m_1}{|M_1|}\right)\left(\frac{\bar{m}_1}{N_1}\right) L_{M_1,\overline{S_0\cup M_1}}, \qquad (4.22)
\end{aligned}
$$

$$
\begin{aligned}
E_M[l_{1,\bar{m}}] &= \sum_{\substack{i\neq j,\\ i,j\epsilon\overline{S_0\cup M_1}}} E[X_i^{(1,\bar{m})}]E[(1-X_j^{(1,\bar{m})})]w_{ij} \\
&= \left(\frac{\bar{m}_1}{N_1}\right)\left(1-\frac{\bar{m}_1}{N_1}\right) L_{\overline{S_0\cup M_1}}, \qquad (4.23)
\end{aligned}
$$

and

$$
\begin{aligned}
E_M[l_{1,m}] &= \sum_{\substack{i\epsilon M_1,\\ j\epsilon\overline{S_0\cup M_1}}} E[X_i^{(1,m)}]E[(1-X_j^{(1,\bar{m})})]w_{ij} \\
&= \left(\frac{m_1}{|M_1|}\right)\left(1-\frac{\bar{m}_1}{N_1}\right) L_{M_1,\overline{S_0\cup M_1}}. \qquad (4.24)
\end{aligned}
$$

Another conditional moment-based estimator of the population size N based on the

class B statistics can also be shown to be

$$\hat{N}_{B,1} = n_0 + |M_1| + \bar{m}_1 \left(\frac{k_{1,\bar{m}} + k_{1,m,\bar{m}} + t_{1,\bar{m}} + t_{1,m,\bar{m}}}{k_{1,\bar{m}} + k_{1,m,\bar{m}}} \right) \tag{4.25}$$

since

$$\begin{aligned} E_M[k_{1,\bar{m}}] &= \sum_{j \epsilon \overline{S_0 \cup M_1}} E[X_j^{(1,\bar{m})}] E[\max_{i \epsilon A_j^{(1,\bar{m})} \setminus \{j\}} \{X_i^{(1,\bar{m})}\}] \\ &= \sum_{j \epsilon \overline{S_0 \cup M_1}} E[X_j^{(1,\bar{m})}] E[1 - \min_{i \epsilon A_j^{(1,\bar{m})} \setminus \{j\}} \{1 - X_i^{(1,\bar{m})}\}] \\ &= \bar{m}_1 - \frac{\bar{m}_1}{N_1} \sum_{j \epsilon \overline{S_0 \cup M_1}} \left(1 - \frac{\bar{m}_1}{N_1}\right)^{a_j^{(1,\bar{m})} - 1}, \end{aligned} \tag{4.26}$$

$$\begin{aligned} E_M[k_{1,m,\bar{m}}] &= \sum_{\substack{j \epsilon \overline{S_0 \cup M_1}, \\ i \epsilon M_1}} E[X_j^{(1,\bar{m})}] E[\max_{i \epsilon A_j^{(1,m)}} \{X_i^{(1,m)}\}] \\ &= \sum_{\substack{j \epsilon \overline{S_0 \cup M_1}, \\ i \epsilon M_1}} E[X_j^{(1,\bar{m})}] E[1 - \min_{i \epsilon A_j^{(1,m)}} \{1 - X_i^{(1,m)}\}] \\ &= |M_1| \frac{m_1}{N_1} \left(N_1 - \sum_{j \epsilon \overline{S_0 \cup M_1}} \left(1 - \frac{m_1}{|M_1|}\right)^{a_j(1,m)} \right), \end{aligned} \tag{4.27}$$

$$
\begin{aligned}
E_M[t_{1,\bar{m}}] &= \sum_{j \epsilon \overline{S_0 \cup M_1}} (E[\max_{i \epsilon A_j^{(1,\bar{m})}} \{X_i^{(1,\bar{m})}\}] - E[X_j^{(1,\bar{m})}]) \\
&= \sum_{j \epsilon \overline{S_0 \cup M_1}} (E[1 - \min_{i \epsilon A_j^{(1,\bar{m})}} \{1 - X_i^{(1,\bar{m})}\}] - E[X_j^{(1,\bar{m})}]) \\
&= N_1 - \bar{m}_1 - \sum_{j \epsilon \overline{S_0 \cup M_1}} \left(1 - \frac{\bar{m}_1}{N_1}\right)^{a_j^{(1,\bar{m})}}, \qquad (4.28)
\end{aligned}
$$

and

$$
\begin{aligned}
E_M[t_{1,m,\bar{m}}] &= \sum_{\substack{j \epsilon \overline{S_0 \cup M_1}, \\ i \epsilon M_1}} E[(1 - X_j^{(1,\bar{m})})] E[\max_{i \epsilon A_j^{(1,m)}} \{X_i^{(1,m)}\}] \\
&= \sum_{\substack{j \epsilon \overline{S_0 \cup M_1}, \\ i \epsilon M_1}} E[(1 - X_j^{(1,\bar{m})})] E[1 - \min_{i \epsilon A_j^{(1,m)}} \{1 - X_i^{(1,m)}\}] \\
&= |M_1| \left(N_1 - \bar{m}_1 - \left(1 - \frac{\bar{m}_1}{N_1}\right) \sum_{j \epsilon \overline{S_0 \cup M_1}} \left(1 - \frac{m_1}{|M_1|}\right)^{a_j(1,m)} \right). \qquad (4.29)
\end{aligned}
$$

Note that the population size estimators based on wave one require knowing the size of the set M_1. This will likely entail the use of the class B statistics at the initial wave (that is, members outside of S_0 must be identifiable).

For all estimators presented in this chapter, we will add a value of 1 to each statistic and subtract a value of 1 from the final estimator in the same manner as Chapman (1951) had proposed for the Lincoln-Petersen estimator. This bias-adjusted estimator will ensure that, in the event that a set of statistics based on the internal nominations (namely those statistics of type r and k) of a wave all take a value of zero, a value of zero will not show up in the denominator of any of the estimators and hence will not result in an unstable estimate of the population size.

4.3.3 The Rao-Blackwellized estimators

Two different Rao-Blackwellized estimators based on the initial wave are explored. In the first case sampling stops after the initial sample S_0 and the subset $S_{1,m}$ of M_1 are selected (that is, $S_{1,\bar{m}}$ is not selected). We shall refer to this sample as the *restricted sample* so that only estimators based on the initial wave are determined. In this case the data observed is $d_0 = \{(i, A_i, t_i) : i \epsilon S\}$ where t_i is the time or step that unit i is selected for the sample. We shall assign values of $t_i = 0$ if unit i is selected for S_0 and $t_i = 1$ if unit i is selected for $S_{1,m}$. A sufficient statistic for the population size N is then $d_r = \{(i, A_i) : i \epsilon S\}$.

Rao-Blackwellized versions of the preliminary estimators $\hat{N}_{A,0}$ and $\hat{N}_{B,0}$ based on the restricted sample can be obtained as follows. Suppose that $\hat{N}_0$ represents either of these two preliminary estimators. Then based on a final sample of size $n = |S_0| + |S_{1,m}| = n_0 + m_1$ the Rao-Blackwellized estimator is

$$
\begin{aligned}
&E[\hat{N}_0|d_r] = \hat{N}_{RB} = \sum_{k=1}^{n!} \hat{N}_0^{(k)} p(s^{(k)}|d_r) \\
&= \sum_{k=1}^{n!} \hat{N}_0^{(k)} p(s^{(k)}) / \sum_{k=1}^{n!} p(s^{(k)}) \\
&= \sum_{k=1}^{n!} \left(\hat{N}_0^{(k)} \frac{1}{\binom{N}{n_0}} \beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} I[S_{1,m}^{(k)} \subseteq M_1^{(k)}, |M_1^{(k)}| \geq m_1] \right) / \\
&\sum_{k=1}^{n!} \left(\frac{1}{\binom{N}{n_0}} \beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} I[S_{1,m}^{(k)} \subseteq M_1^{(k)}, |M_1^{(k)}| \geq m_1] \right) \\
&= \sum_{k=1}^{n!} \left(\hat{N}_0^{(k)} \beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} I[S_{1,m}^{(k)} \subseteq M_1^{(k)}, |M_1^{(k)}| \geq m_1] \right) / \\
&\sum_{k=1}^{n!} \left(\beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} I[S_{1,m}^{(k)} \subseteq M_1^{(k)}, |M_1^{(k)}| \geq m_1] \right)
\end{aligned}
\tag{4.30}
$$

where k represents the corresponding sample reordering, $\hat{N}_0^{(k)}$ is the corresponding

estimate of the population size, $M_1^{(k)}$ is the corresponding set of members that are mentioned from the hypothetical initial wave $S_0^{(k)}$, and $S_{1,m}^{(k)}$ is the corresponding set of members that are hypothetically adaptively recruited from $M_1^{(k)}$. Note that only the size of $M_1^{(k)}$ is permitted to vary over the reorderings. The number of members selected at random for the initial wave (that is, n_0) as well as the number of members who are adaptively recruited for the sample reordering (that is, m_1) must remain fixed in order for the sample reordering to be consistent with the sufficient statistic (that is the reduced data must coincide with d_r).

In the second case, if sampling continues so that we also obtain $S_{1,\bar{m}}$ (that is, we are considering the *full sample*), the data assumed to be observed is $d_0 = \{(i, A_i, t_i) : i \epsilon S\}$. We shall assign values of $t_i = 0$ if unit i is selected for S_0, $t_i = 1$ if unit i is selected for $S_{1,m}$, and $t_i = 2$ if unit i is selected $S_{1,\bar{m}}$. A sufficient statistic for the population size N is then $d_r = \{(i, A_i), |M_1| : i \epsilon S\}$ (notice that M_1 is a function of $\{A_i : i \epsilon S\}$). The Rao-Blackwellized version of the preliminary estimators of the population size based on the full sample can be obtained as follows. Suppose that $\hat{N}_0$ represents any of the preliminary estimators $\hat{N}_{A,0}, \hat{N}_{B,0}, \hat{N}_{A,1}, \hat{N}_{B,1}$ based on a final sample of size $n = |S_0| + |S_{1,m}| + |S_{1,\bar{m}}| = n_0 + m_1 + \bar{m}_1$. Then the Rao-Blackwellized estimator is

$$
\begin{aligned}
&E[\hat{N}_0|d_r] = \hat{N}_{RB} = \sum_{k=1}^{n!} \hat{N}_0^{(k)} p(s^{(k)}|d_r) \\
&= \sum_{k=1}^{n!} \hat{N}_0^{(k)} p(s^{(k)}) / \sum_{k=1}^{n!} p(s^{(k)}) \\
&= \sum_{k=1}^{n!} \left(\hat{N}_0^{(k)} \frac{1}{\binom{N}{n_0}} \beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} \frac{1}{\binom{N-n_0-|M_1^{(k)}|}{\bar{m}_1}} \times \right. \\
&\left. I[S_{1,m}^{(k)} \subseteq M_1^{(k)}, S_{1,\bar{m}}^{(k)} \cap M_1^{(k)} = \phi, |M_1^{(k)}| = |M_1|] \right) / \\
&\sum_{k=1}^{n!} \left(\frac{1}{\binom{N}{n_0}} \beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} \frac{1}{\binom{N-n_0-|M_1^{(k)}|}{\bar{m}_1}} \times \right. \\
&\left. I[S_{1,m}^{(k)} \subseteq M_1^{(k)}, S_{1,\bar{m}}^{(k)} \cap M_1^{(k)} = \phi, |M_1^{(k)}| = |M_1|] \right) \\
&= \sum_{k=1}^{n!} \hat{N}_0^{(k)} \left(\beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} I[S_{1,m}^{(k)} \subseteq M_1, S_{1,\bar{m}}^{(k)} \cap M_1^{(k)} = \phi, |M_1^{(k)}| = |M_1|] \right) / \\
&\sum_{k=1}^{n!} \left(\beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} I[S_{1,m}^{(k)} \subseteq M_1^{(k)}, S_{1,\bar{m}}^{(k)} \cap M_1^{(k)} = \phi, |M_1^{(k)}| = |M_1|] \right) \quad (4.31)
\end{aligned}
$$

where k represents the corresponding sample reordering, $\hat{N}_0^{(k)}$ is the corresponding estimate of the population size, $M_1^{(k)}$ is the corresponding set of members that are mentioned from the hypothetical initial wave $S_0^{(k)}$, $S_{1,m}^{(k)}$ is the corresponding members that are hypothetically adaptively recruited from $M_1^{(k)}$, $S_{1,\bar{m}}^{(k)}$ is the corresponding set of members that are recruited at random for wave one, and ϕ is the empty set. Notice that it is required that $|M_1^{(k)}| = |M_1|$ for reordering k to be consistent with the reduced data, which in turn will guarantee that all terms involving the unknown population size N can be factored out and canceled from the expression. Hence, the size of $M_1^{(k)}$ is not permitted to vary over the reorderings. Also, the number of members selected at random for the initial wave (that is, n_0), the number of members who are adaptively recruited (that is, m_1), and the number of members selected at

random from $U \setminus (S_0^{(k)} \cup M_1^{(k)})$ (that is, $m\bar{n}_1$) must remain fixed in order for a sample reordering to be consistent with the sufficient statistic.

When using the full sample a large number of the reorderings may have zero probability of being selected. The reason for this is found in the stringent requirement that for reorderings to be consistent with the original data they must have exactly $|M_1|$ members nominated from the hypothetical initial sample. Instead, we have adopted an alternative sampling design that selects members for $S_{1,\bar{m}}$ from $U \setminus (S_0 \cup S_{1,m})$ (as opposed to $U \setminus (S_0 \cup M_1)$). The probability of selecting a sample s with this sampling design can be expressed as

$$p(s) = \frac{1}{\binom{N}{n_0}} \beta_1^{m_1} (1-\beta_1)^{|M_1|-m_1} \frac{1}{\binom{N-n_0-m_1}{\bar{m}_1}}. \tag{4.32}$$

A sufficient statistic for the population size N is then $d_r = \{(i, A_i) : i \in s\}$. With this approach it is more likely that reorderings will be consistent with the original data as $|M_1^{(k)}|$ is now permitted to vary over values greater than or equal to m_1 (compare this approach and sufficient statistic with the corresponding approach and sufficient statistic used in the restricted sample case). However, it is not as straightforward to develop estimators of the population size based on the first wave as some members from M_1 (which is a random variable) may now be selected for $S_{1,\bar{m}}$. Hence, accounting for this additional component may prove a cumbersome task when constructing estimators of the population size. Nevertheless, in our simulation study we have kept with the estimators based on the statistics defined in Section 4.2 while implementing the alternative sampling strategy.

With the alternative approach, suppose $\hat{N}_0$ is an estimator of the population size. Then the Rao-Blackwellized estimator is

$$
\begin{aligned}
&E[\hat{N}_0|d_r] = \hat{N}_{RB} = \sum_{k=1}^{n!} \hat{N}_0^{(k)} p(s^{(k)}|d_r) \\
&= \sum_{k=1}^{n!} \hat{N}_0^{(k)} p(s^{(k)}) / \sum_{k=1}^{n!} p(s^{(k)}) \\
&= \sum_{k=1}^{n!} \left(\hat{N}_0^{(k)} \frac{1}{\binom{N}{n_0}} \beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} \frac{1}{\binom{N-n_0-m_1}{\bar{m}_1}} I[S_{1,m}^{(k)} \subseteq M_1^{(k)}, |M_1^{(k)}| \geq m_1] \right) / \\
&\sum_{k=1}^{n!} \left(\frac{1}{\binom{N}{n_0}} \beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} \frac{1}{\binom{N-n_0-m_1}{\bar{m}_1}} I[S_{1,m}^{(k)} \subseteq M_1^{(k)}, |M_1^{(k)}| \geq m_1] \right) \\
&= \sum_{k=1}^{n!} \left(\hat{N}_0^{(k)} \beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} I[S_{1,m}^{(k)} \subseteq M_1^{(k)}, |M_1^{(k)}| \geq m_1] \right) / \\
&\sum_{k=1}^{n!} \left(\beta_1^{m_1} (1-\beta_1)^{|M_1^{(k)}|-m_1} I[S_{1,m}^{(k)} \subseteq M_1^{(k)}, |M_1^{(k)}| \geq m_1] \right). \qquad (4.33)
\end{aligned}
$$

The essence of using the sufficient statistic is highlighted by bringing to the reader's attention that in each of the three cases all terms involving the unknown population size N factor out of the expression and cancel therefore making computation of the Rao-Blackwellized estimates possible. Notice that we have proved that the aforementioned statistics are sufficient for N since the ratio of the probability of selecting any two data points from the same partition that the respective statistic induces (that is, the data points are simply reorderings of each other) does not depend on the unknown population size N.

We mention here that, in order to obtain the Rao-Blackwellized version of the estimators outlined in this chapter, all nominations within the final sample $S = S_0 \cup S_{1,m} \cup S_{1,\bar{m}}$, as well as those nominations outside of S, must be identifiable. The justification for this can be reasoned by observing that, for any two members in S that are reordered to hypothetically be selected for the initial sample, we require the ability to observe if either member nominates the other. Furthermore, nominations outside of the sample must be known due to the need to identify the number of

individuals who comprise the set of $M_1^{(k)}$ for all reorderings $k = 1, 2, ..., n!$.

4.3.4 Variance estimators

Frank and Snijders (1994) proposed using a variant of the jackknife method for estimating the variances of the estimators based on the initial wave that they developed. This same method is used for the new estimators based on wave one that is developed in this chapter. We will outline the procedure in this subsection.

Consider the current wave upon which an estimator $\hat{N}$ for the population size N is based on. We calculate $\hat{N}$ for each individual i deleted from the corresponding wave. If we denote this estimate as $\hat{N}_{(i)}$, then the estimate of the variance of the estimator is taken to be

$$\hat{\text{Var}}_J(\hat{N}) = \frac{n-2}{2n} \sum_{i=1}^{n} (\hat{N}_{(i)} - \hat{N}_{(.)})^2, \tag{4.34}$$

where $\hat{N}_{(.)} = \sum_{i=1}^{n} \hat{N}_{(i)}$ and n is the size of the corresponding wave.

With respect to the variance estimates of the Rao-Blackwellized estimators, Thompson (2006a) proposed the following unbiased estimator. For any estimator $\hat{N}_{RB} = E[\hat{N}_0|d_r]$, the conditional decomposition of variances gives

$$\text{Var}(\hat{N}_{RB}) = \text{Var}(\hat{N}_0) - E[\text{Var}(\hat{N}_0|d_r)]. \tag{4.35}$$

An unbiased estimator of $\text{Var}(\hat{N}_{RB})$ is

$$\hat{\text{Var}}(\hat{N}_{RB}) = E[\hat{\text{Var}}(\hat{N}_0)|d_r] - \text{Var}(\hat{N}_0|d_r). \tag{4.36}$$

This estimator is the difference of the expectation of the estimated variance of the preliminary estimator over all reorderings of the data and the variance of the preliminary estimator over all the reorderings of the data. As this estimator can result in

negative estimates of the variance, a conservative approach would take the estimate of $\text{Var}(\hat{N}_{RB})$ to be $E[\hat{\text{Var}}(\hat{N}_0)|d_r]$ when such a scenario arises.

4.4 Markov Chain Resampling Estimators

Due to the large number of sample permutations that are obtained with the sample sizes used in this study, a Markov chain resampling procedure similar to the one found in Thompson (2006a) is implemented to obtain estimates of the improved estimates. As the sampling strategy presented in this chapter selects a sample through a snowball sampling type of design, the Markov chain resampling strategy needs to be modified. We outline the modified Markov chain accept/reject Hastings (1970) resampling procedure below.

Suppose θ is a population unknown we wish to estimate with the improved estimator $\hat{\theta}_{RB} = E[\hat{\theta}_0|d_r]$ where d_r is a sufficient statistic.

Step 0: Let $\hat{\theta}_0^{(0)}$ be the estimated value of θ and $\hat{\text{Var}}(\hat{\theta}_0^{(0)})$ be the estimated value of $\text{Var}(\hat{\theta}_0)$ that is obtained from selecting the sample in the original order it was selected. Also, let $t^{(0)} = s$ be the original sample in the order it was selected.

For step $l = 1, 2, ..., R$, where R is sufficiently large:

Draw a candidate sample reordering, $t^{(l)}$ say, from a candidate distribution. Suppose the most recently accepted candidate reordering is $t^{(y)}$ for some reordering of the sample where $y = 0, 1, 2, ..., l-1$. Let $p(t^{(l)})$ be the probability of obtaining $t^{(l)}$ under the true population and $q(t^{(l)})$ be the probability of obtaining reordering $t^{(l)}$ under the candidate distribution. Generate a uniform random number between 0 and 1, and if this value is less than

$$\min\left\{\frac{p(t^{(l)})}{p(t^{(y)})}\frac{q(t^{(y)})}{q(t^{(l)})}, 1\right\}, \tag{4.37}$$

let $\hat{\theta}_0^{(l)}$ and $\hat{\mathrm{Var}}(\hat{\theta}_0^{(l)})$ be the estimates of θ and $\mathrm{Var}(\hat{\theta}_0)$, respectively, obtained with sample reordering $t^{(l)}$. Otherwise, take $\hat{\theta}_0^{(l)} = \hat{\theta}_0^{(l-1)}$ and $\hat{\mathrm{Var}}(\hat{\theta}_0^{(l)}) = \hat{\mathrm{Var}}(\hat{\theta}_0^{(l-1)})$. Recall that $p(t^{(l)})$ needs only to be known for the (hypothetical) adaptive recruitment probabilities found in the sample reorderings as all terms involving the unknown population size N can be factored out of the ratio of the true probabilities of obtaining sample reorderings and cancelled from the expression.

Final step: Take estimates of $\hat{\theta}_{RB}$ to be

$$\tilde{\theta}_{RB} = \frac{1}{R+1} \sum_{l=0}^{R} \hat{\theta}_0^{(l)}, \tag{4.38}$$

and similarly take the estimate of $\hat{\mathrm{Var}}(\hat{\theta}_{RB})$ to be

$$\begin{aligned} \tilde{\mathrm{Var}}(\hat{\theta}_{RB}) &= \tilde{E}[\hat{\mathrm{Var}}(\hat{\theta}_0)|d_r] - \tilde{\mathrm{Var}}(\hat{\theta}_0|d_r) \\ &= \frac{1}{R+1} \sum_{l=0}^{R} \hat{\mathrm{Var}}(\theta_0^{(l)}) - \frac{1}{R+1} \sum_{l=0}^{R} (\hat{\theta}_0^{(l)} - \tilde{\theta}_{RB})^2. \end{aligned} \tag{4.39}$$

Several candidate distributions were explored for the purpose of obtaining the Markov chain resampling estimators. These ranged from selecting a permutation of the sample completely at random to strategies that first placed varying amounts of homogenous weight to the elements from each wave (and therefore heterogenous weights between waves) to be selected for the candidate reordering's initial wave. These methods all resulted in chains that promoted very little mixing and hence we explored the use of a more adaptive technique for selecting candidate reorderings. We decided on a method that interchanges one unit from the initial wave and one unit from wave one based on the most recently accepted sample permutation. This style of candidate distribution mimics those adaptive proposal distributions discussed in Atchadé and Rosenthal (2005) where a posterior distribution is sampled from with the use of a random walk type of sampler.

4.5 Simulation Study

We will use the thesis study population to evaluate the new inference procedures outlined in this chapter. The population was generated according to the stochastic cluster model that was outlined in chapter 2. An illustration of the simulated population can be found in Figure 4.3.

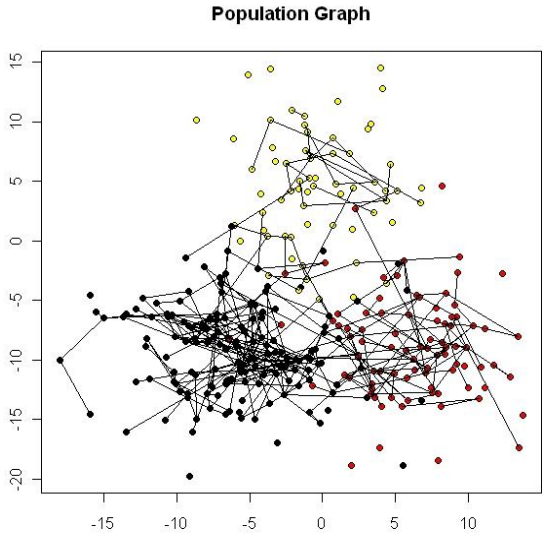

Figure 4.3: The simulated thesis study population.

Figures 4.4 and 4.5 show a sample that is selected under the adaptive sampling design that is outlined in this chapter. Figure 4.4 shows the 50 members that are selected for the initial sample with the nominations originating from the initial wave that are required to be observed for the estimators of the population size based on the initial wave.

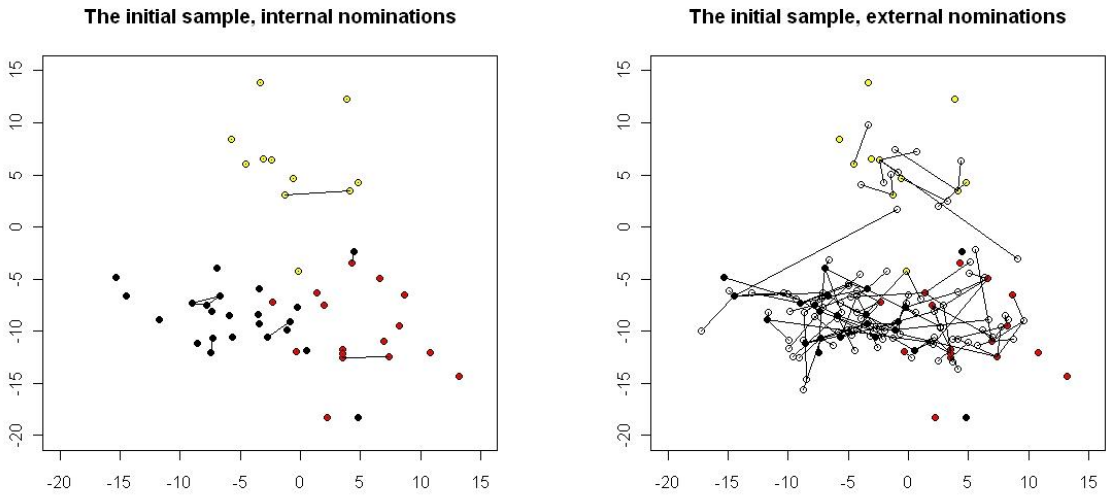

Figure 4.4: A sample selected from the thesis study population showing the initial wave with the internal and external nominations. The size of the initial wave is 50. The illustration on the left shows the nominations made within the initial wave and the second illustration highlights, as white nodes, those members not selected for the initial wave and that are linked to at least one member in the initial wave.

Figure 4.5 depicts those 10 members that are selected at random for wave one as well as those members that were adaptively recruited for wave one (where the probability for these members being recruited for wave one was 50%) and the nominations originating from wave one that are required to be observed for the estimators of the population size based on wave one.

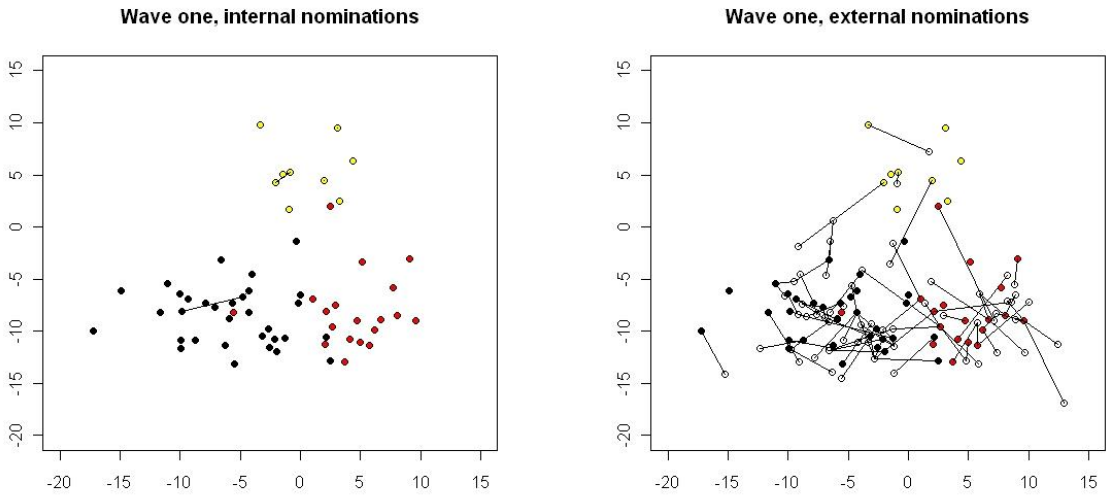

Figure 4.5: A sample selected from the thesis study population showing wave one with the internal and external nominations. Members mentioned from the initial wave are traced with probability 50% and the number of randomly selected members for wave one is 10. The illustration on the left shows the internal nominations required for the wave one statistics and the second illustration highlights, as white nodes, those members outside of wave one (and the initial wave) and that are linked to at least one member in wave one.

We conducted a simulation study as follows. A total of 500 samples with 5000 resamples from each sample for the Markov chain resampling procedure were obtained. We used the alternative sampling strategy outlined in subsection 4.3.3. Initial samples S_0 of size 50 were obtained, and those members in M_1 were recruited for $S_{1,m}$ with a probability of $\beta_1 = 50\%$. An additional 10 members were selected at random for $S_{1,\bar{m}}$ (recall that these may now include members from $M_1 \setminus S_{1,m}$).

Histograms of the estimates of the population size are presented in the following subsections. The true population size of 300 is indicated by the solid triangles on the x-axis of the histograms and the approximate expectation of the estimators is indicated by the transparent triangle. Tables are also provided to display the bias

and mean squared error (MSE) scores for each estimator as well as the average semi-lengths of the nominal 95% confidence intervals based on the Central Limit Theorem of the estimates.

4.5.1 Simulation study of the estimators of the population size based on the restricted sample

Figure 4.6 presents the preliminary and improved estimates based on the initial wave when sampling stops after recruiting a Bernoulli subset of the members nominated from the initial wave to outside of the initial wave. Recall that for these improved estimators, the sampling design requires that a subset of the nominations outside of the initial wave be recruited, that is, requires observing $S_{1,m}$.

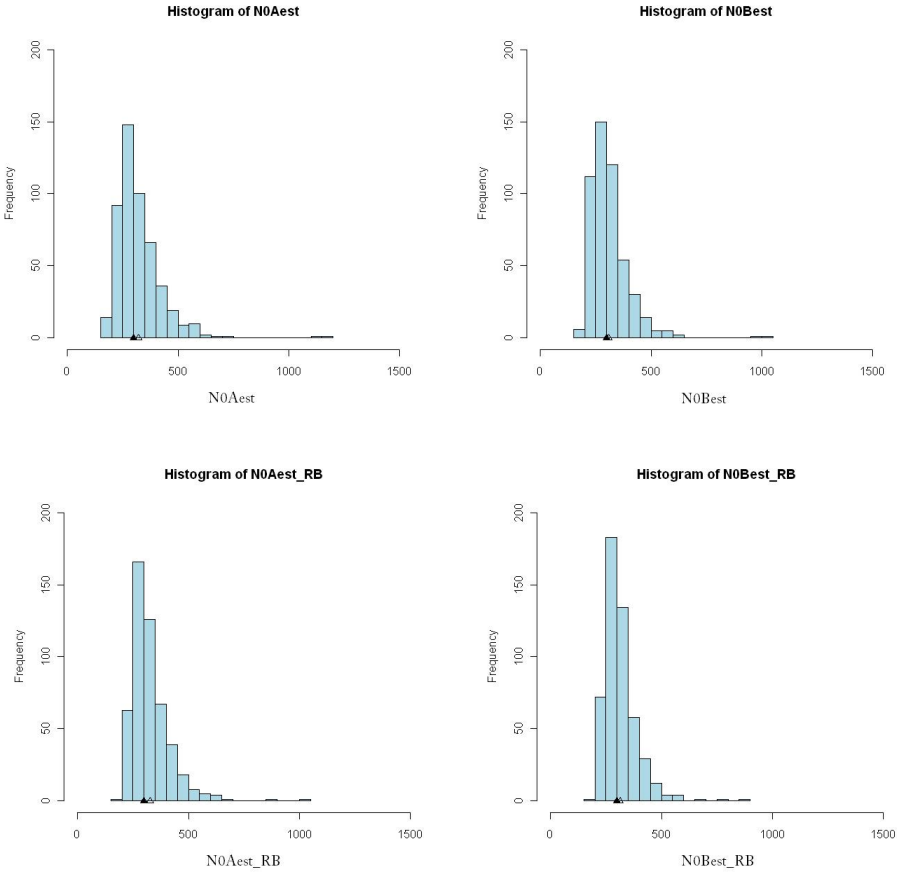

Figure 4.6: Histograms of the population size estimates $\hat{N}_{A,0}, \hat{N}_{B,0}, \hat{N}_{A,RB,0}$, and $\hat{N}_{B,RB,0}$ based on the initial wave from the restricted sample. The dark triangle indicates the true population size of 300 and the transparent triangle indicates the approximate expectation of the distribution of the estimates.

Table 4.1 displays the bias and MSE scores with the coverage rates and average semi-lengths of the nominal 95% confidence intervals based on the Central Limit Theorem for the estimators based on the initial wave. The simulation study did not report any negative estimates for the variance of the Rao-Blackwellized estimates based on the initial wave when only considering the reduced sample.

Table 4.1: Bias and MSE scores of the population size ($N = 300$) with the coverage rates and average semi-lengths in parentheses of the nominal 95% confidence intervals based on the CLT for the estimators based on the initial wave from the restricted sample.

Estimator	Bias	MSE	Coverage rates (semi-length)
$\hat{N}_{A,0}$	21.0	11295	0.950 (232)
$\hat{N}_{A,RB,0}$		8372	0.986 (222)
$\hat{N}_{B,0}$	9.2	7586	0.932 (187)
$\hat{N}_{B,RB,0}$		5663	0.982 (183)

As shown in the histograms and table of scores, Rao-Blackwellization of the preliminary estimators has resulted in significantly improved estimates of the population size, while the reported coverage rates appear to be slightly higher for the improved estimates.

4.5.2 Simulation study of the estimators of the population size based on the full sample

Figure 4.7 presents the histograms of the preliminary and improved estimates based on the initial wave from the full sample when implementing the alternative sampling strategy (as outlined in subsection 4.3.3). Notice that the behaviour of the preliminary estimators based on the initial wave do not change from those based on the restricted sample. However, the improved estimators may change as there are an additional $\bar{m}_1 = 10$ members in the final sample thereby increasing the number of sample reorderings and corresponding population size estimates and sample reordering probabilities that contribute to the improved estimates.

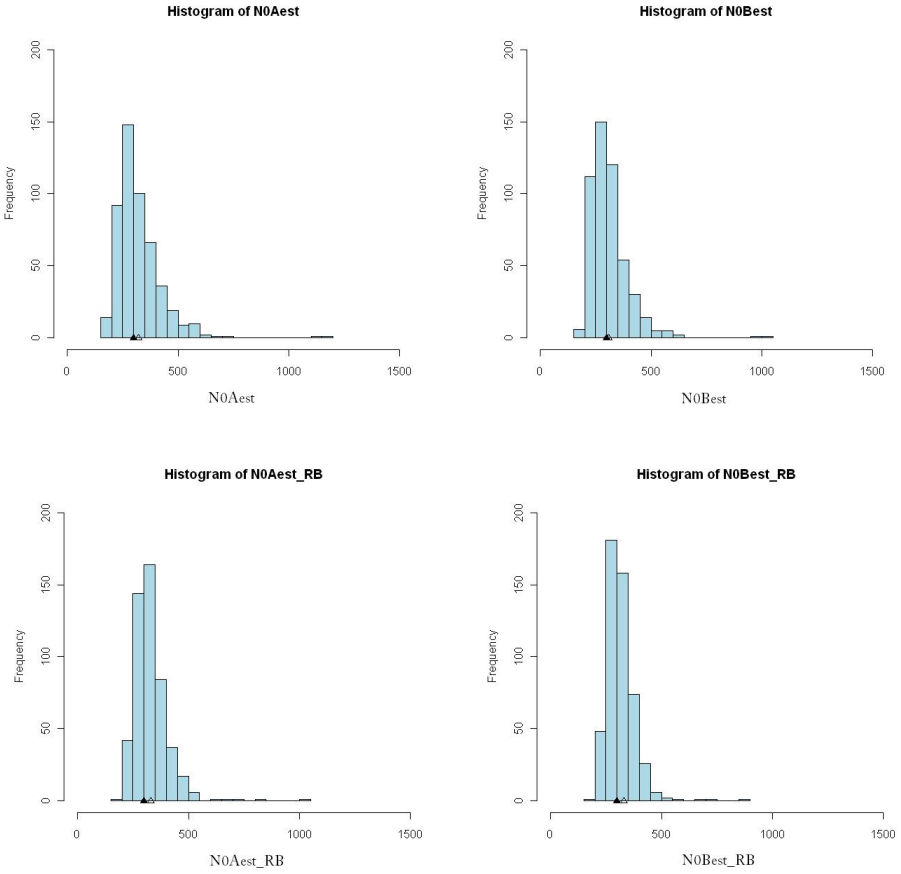

Figure 4.7: Histograms of the population size estimates $\hat{N}_{A,0}, \hat{N}_{B,0}, \hat{N}_{A,RB,0}$, and $\hat{N}_{B,RB,0}$ based on the initial wave from the full sample. The dark triangle indicates the true population size of 300 and the transparent triangle indicates the approximate expectation of the distribution of the estimates.

Table 4.2 displays the bias and MSE scores with the coverage rates and average semi-lengths of the nominal 95% confidence intervals based on the Central Limit Theorem for the estimates based on the initial wave from the full sample. The simulation study did not report any negative estimates for the variance of the Rao-Blackwellized estimates based on the initial wave when considering the full sample.

Table 4.2: Bias and MSE scores of the population size ($N = 300$) with the coverage rates and average semi-lengths in parentheses of the nominal 95% confidence intervals based on the CLT for the estimators based on the initial wave from the full sample.

Estimator	Bias	MSE	Coverage rates (semi-lengths)
$\hat{N}_{A,0}$	21.0	11295	0.950 (232)
$\hat{N}_{A,RB,0}$		7125	0.991 (218)
$\hat{N}_{B,0}$	9.2	7586	0.932 (187)
$\hat{N}_{B,RB,0}$		4508	0.991 (178)

Once again, it appears that Rao-Blackwellization of the preliminary estimators has significantly improved the estimates of the population size with efficiency gains greater than those corresponding estimates obtained with the restricted sample. As in the restricted sample case, the reported coverage rates appear to be slightly higher for the improved estimators.

Figure 4.8 presents the histograms of the preliminary and improved estimates based on wave one from the full sample.

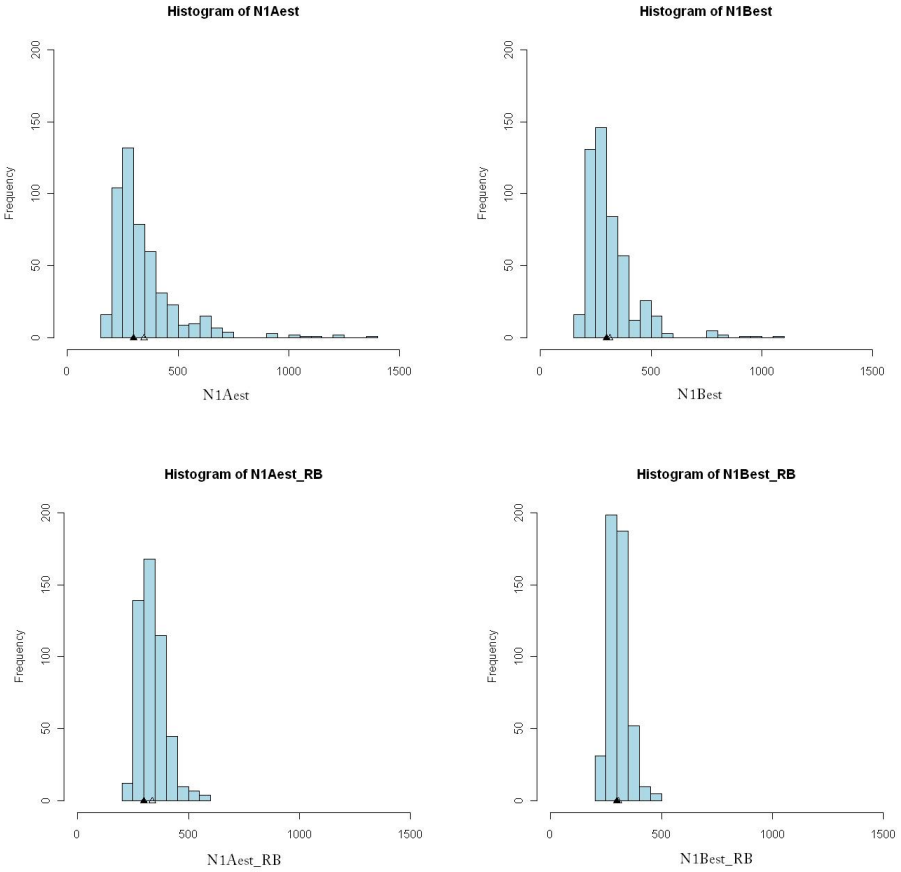

Figure 4.8: Histograms of the population size estimates $\hat{N}_{A,1}, \hat{N}_{B,1}, \hat{N}_{A,RB,1}$, and $\hat{N}_{B,RB,1}$ based on wave one from the full sample. The dark triangle indicates the true population size of 300 and the transparent triangle indicates the approximate expectation of the distribution of the estimates.

Table 4.3 displays the bias and MSE scores with the coverage rates and average semi-lengths of the confidence intervals for the estimators based on wave one. A large number of negative estimates for the variance of the Rao-Blackwellized estimates based on wave one were encountered. However, the conservative approach that was suggested in subsection 4.3.4 was used and appears to have compensated enough to

give reasonable estimates for the variances of the estimates and their corresponding coverage rates.

Table 4.3: Bias and MSE scores of the population size ($N = 300$) with the coverage rates and average semi-lengths in parentheses of the nominal 95% confidence intervals for the estimators based on wave one from the full sample.

Estimator	Bias	MSE	Coverage rates
$\hat{N}_{A,1}$	24.0	14262	0.900 (275)
$\hat{N}_{A,1,RB}$		4801	0.955 (190)
$\hat{N}_{B,1}$	3.0	8279	0.877 (199)
$\hat{N}_{B,1,RB}$		2045	0.932 (134)

As expected the Rao-Blackwellized estimators have significantly improved on the preliminary estimators, even with a small number of recruits selected completely at random for wave one (recall that we set $|S_{1,\bar{m}}| = 10$). It appears that with these sample sizes a reasonable amount of bias for both cases is present. It also appears that the alternative sampling design has worked well with the estimators based on the original sampling design.

4.6 Discussion

The new methods developed in this chapter allow for formulating estimates based on a succeeding wave that is obtained after the initial sample is selected. We have also developed a method to Rao-Blackwellize the preliminary estimators, and as demonstrated in the simulation study, improved estimates of the population size are guaranteed. Extending this method over more waves deserves attention and the methods presented in this chapter could serve as a foundation for determining estimators based on further sampling effort that is similar to that which is found in the sampling designs outlined in this chapter.

The coverage rates of the Rao-Blackwellized estimators appear to have come out higher than their preliminary estimator counterparts. However, the reported estimates of the variance of the Rao-Blackwellized estimators were significantly smaller than those based on the preliminary estimators, indicating that this may not be too much of a concern to the analyst. With respect to the preliminary estimators, it appears that using the method outlined in Frank and Snijders (1994) for obtaining estimates of the variance of the estimators based on wave one will give reasonable approximations of the variance of the estimators. There were a large number of negative estimates of the variance of the Rao-Blackwellized estimates based on wave one from the full sample, and we resorted to using the conservative approach outlined in subsection 4.3.4. Future work on obtaining practical estimates of the variance of Rao-Blackwellized estimators in such scenarios is required, and one strategy that should be attempted is with the use of a bootstrap-based strategy.

Frank and Snijders (1994) showed that the population size estimators based on the initial wave are consistent estimators. Future work on determining if the population size estimators based on wave one also possess such desirable features is deserving of attention.

For practical purposes, it may only be possible to utilize the class A statistics. Notice that in order to determine the class A statistics, only the subset of the data $d_{A,0} \subseteq d_0$ needs to be observed where

$$d_{A,0} = \{(i, w_{ij}, w_i^+, t_i) : i, j \, \epsilon \, S\}, \tag{4.40}$$

and w_i^+ is the degree of unit i (that is, the number of nominations unit i makes). However, for the purposes of Rao-Blackwellization of the estimators $\hat{N}_{A,0}$ and $\hat{N}_{A,1}$, the class B statistics must be observed, as described in subsection 4.3.3.

Future work on determining an estimate based on a weighting of the estimates of the population size from the initial wave and wave one is deserving of attention. We shall note that the estimates came out approximately uncorrelated (primarily due to the random selections made at each wave) and hence determining an optimal choice

of weights, perhaps based on the number of random recruits and adaptive recruits for each wave, should be explored.

Chapter 5

Improved Importance Sampling

5.1 Introduction

In this chapter we introduce a method for obtaining improved versions of the approximations of the Rao-Blackwellized estimates of a population unknown when employing an importance sampling strategy. Recall that with a design-based approach to inference in sampling, we can obtain Rao-Blackwellized estimates of the preliminary estimates of a population unknown by tabulating the preliminary estimates of the population unknown from and weighting against the probability of each sample reordering (see expression (5.1) below). However, when selecting a large sample size this will result in a large number of sample reorderings. This may be computationally cumbersome and will likely require a method like importance sampling or Markov chain Monte Carlo (MCMC) for approximating the Rao-Blackwellized estimates.

We have developed a strategy, termed *improved importance sampling*, based on a single cluster sampling type of sampling design to increase the efficiency of the approximations of the Rao-Blackwellized estimates of a population unknown when using importance sampling. The method entails defining neighbourhoods of the sample space (that is, all of the sample reorderings) at the analyst's discretion and observing the responses associated with all of the units in the neighbourhoods of sampled units to increase the efficiency of the importance sampling estimators. This

method may prove to be highly useful if each of the relevant responses of the neighbours of those units selected under the importance sampler can be evaluated with a reduced amount of computational effort once the corresponding information of one of the units from the neighbourhood is evaluated.

In Section 5.2, we introduce the estimation procedure that is used in this chapter. In Section 5.3 we provide a simulation study based on the study that was presented in subsection 4.5.1. In Section 5.4, we provide a discussion of the results presented in this chapter.

5.2 Estimation

Recall that with a sample s that is selected with an adaptive sampling design, a Rao-Blackwellized estimate of a population unknown, θ say, based on a preliminary estimator $\hat{\theta}_0$ is

$$\begin{aligned}\hat{\theta}_{RB} &= \sum_{k=1}^{n!} \hat{\theta}_0^{(k)} p(s^{(k)}|d_r) \\ &= \sum_{k=1}^{n!} \hat{\theta}_0^{(k)} p(s^{(k)}) / \sum_{k=1}^{n!} p(s^{(k)})\end{aligned} \tag{5.1}$$

where d_r is a sufficient statistic, n is the sample size, $\hat{\theta}_0^{(k)}$ is the preliminary estimate of θ obtained with sample reordering k, and $p(s^{(k)})$ is the probability of obtaining sample reordering k for $k = 1, 2, ..., n!$. One can obtain an estimate of the population unknown using an importance sampling approach (Gelman et al., 2004), as outlined below.

Suppose that for some function g we wish to estimate $E[g(x)] = \mu_{g(x)} = \sum_{k=1}^{n!} g(x_k)p_k$ where x_k is a response of interest of unit k, p_k is the probability of observing unit k under the target distribution, and $n!$ is the number of sample reorderings (that is, the size of the target population). Suppose we take m draws of the sample space with replacement where the probability of selecting unit k on any draw is q_k. Then,

an unbiased estimate of $\mu_{g(x)}$ (Gelman et al., 2004) is

$$\hat{\mu}_{g(x)} = \frac{1}{m}\sum_{k=1}^{m}\frac{g(x_k)p_k}{q_k} \tag{5.2}$$

(where it is understood that the elements of the population are reordered so that the first m units of the population coincide with those units in the sample, with the possibility of replacement).

Now, suppose we define neighbourhoods of the sample space (that is, the $n!$ elements/reorderings) to be $\mathcal{N} = \{\mathcal{N}_k : k = 1, 2, ..., T\}$ where there are a total of T neighbourhoods. For all $k = 1, 2, ..., T$ we will define

$$y_k = \sum_{i \epsilon \mathcal{N}_k} g(x_i)p_i, \tag{5.3}$$

and

$$q'_k = \sum_{i \epsilon \mathcal{N}_k} q_i. \tag{5.4}$$

Suppose we take m draws of the original sample space and for each observation $k = 1, 2, ..., m$ we completely observe $\mathcal{N}_k$. That is, for each $i \, \epsilon \, \mathcal{N}_k$ we observe $g(x_i), p_i$, and q_i. We shall let $\underline{\mathcal{N}}_s = (\mathcal{N}_1, \mathcal{N}_2, ..., \mathcal{N}_m)$ (that is the ordered set of neighbourhoods that correspond with the sample s in the order the sample was selected in). Another estimate of $\mu_{g(x)}$ is of the Hansen-Hurwitz type of estimator (Thompson, 2002) and is

$$\hat{\mu}_y = \frac{1}{m}\sum_{k=1}^{m}\frac{y_k}{q'_k}. \tag{5.5}$$

This estimator is also unbiased since

$$\begin{aligned}
\hat{\mu}_y &= \frac{1}{m}\sum_{k=1}^{m}\frac{y_k}{q'_k} \\
&= \frac{1}{m}\sum_{k=1}^{m}\frac{\sum_{i\epsilon\mathcal{N}_k} g(x_i)p_i}{\sum_{i\epsilon\mathcal{N}_k} q_i} \\
&= \frac{1}{m}\sum_{k=1}^{m}\frac{\sum_{i\epsilon\mathcal{N}_k} \frac{g(x_i)p_i}{q_i}q_i}{\sum_{i\epsilon\mathcal{N}_k} q_i} \\
&= \frac{1}{m}\sum_{k=1}^{m} E\left[\frac{g(x_k)p_k}{q_k}\middle|\mathcal{N}_k\right] \\
&= E\left[\frac{1}{m}\sum_{k=1}^{m}\frac{g(x_k)p_k}{q_k}\middle|\underline{\mathcal{N}}_s\right] \\
&= E[\hat{\mu}_{g(x)}|\underline{\mathcal{N}}_s] \qquad (5.6)
\end{aligned}$$

(where it is understood that the expectation is taken with respect to the importance sampling distribution q). Therefore, as $E[\hat{\mu}_y] = E[E[\hat{\mu}_{g(x)}|\underline{\mathcal{N}}_s]] = E[\hat{\mu}_{g(x)}] = \mu_{g(x)}$ this gives unbiasedness. Furthermore, as $\underline{\mathcal{N}}_s$ is a sufficient statistic, $\hat{\mu}_y$ will result in an improved estimator over $\hat{\mu}_{g(x)}$.

Recall that the Rao-Blackwellized version of the improved estimator $\hat{N}_0$ of the population size N can be expressed as

$$\begin{aligned}
\hat{N}_{RB} &= \sum_{k=1}^{n!}\hat{N}_0^{(k)}p(s^{(k)}|d_r) \\
&= \sum_{k=1}^{n!}\hat{N}_0^{(k)}p(s^{(k)})/\sum_{k=1}^{n!}p(s^{(k)}). \qquad (5.7)
\end{aligned}$$

Now, for all sample reorderings $k = 1, 2, ..., n!$ we can replace $g(x_k)$ with $\hat{N}_0^{(k)}$,

p_k with $p(s^{(k)})$, and q_k with $q(s^{(k)})$ to obtain a preliminary unbiased estimate of $\sum_{k=1}^{n!} \hat{N}_0^{(k)} p(s^{(k)})$. We can also replace $g(x_k)$ with 1, p_k with $p(s^{(k)})$, and q_k with $q(s^{(k)})$ to obtain a preliminary unbiased estimate of $\sum_{k=1}^{n!} p(s^{(k)})$. A preliminary consistent estimate of $\hat{N}_{RB}$ can then be found by taking the ratio of these preliminary estimates. The final preliminary importance sampling estimate of the Rao-Blackwellized estimator is then

$$\tilde{N}_{RB}^{0} = \sum_{k=1}^{m} \frac{\hat{N}_0^{(k)} p(s^{(k)})}{q(s^{(k)})} / \sum_{k=1}^{m} \frac{p(s^{(k)})}{q(s^{(k)})}. \tag{5.8}$$

Similarly, for all $k = 1, 2, ..., T$ we can replace y_k with $\sum_{i \epsilon \mathcal{N}_k} \hat{N}_0^{(i)} p(s^{(i)})$ and q'_k with $\sum_{i \epsilon \mathcal{N}_k} q(s^{(i)})$ to obtain an improved unbiased estimate of $\sum_{k=1}^{n!} \hat{N}_0^{(k)} p(s^{(k)})$. Also, we can replace y_k with $\sum_{i \epsilon \mathcal{N}_k} p(s^{(i)})$ and q'_k with $\sum_{i \epsilon \mathcal{N}_k} q(s^{(i)})$ to obtain an improved unbiased estimate of $\sum_{k=1}^{n!} p(s^{(k)})$. An improved consistent estimate of $\hat{N}_{RB}$ can then be found by taking the ratio of these improved estimates. The final improved importance sampling estimate of the Rao-Blackwellized estimator is then

$$\tilde{N}_{RB} = \sum_{k=1}^{m} \frac{\sum_{i \epsilon \mathcal{N}_k} \hat{N}_0^{(i)} p(s^{(i)})}{\sum_{i \epsilon \mathcal{N}_k} q(s^{(i)})} / \sum_{k=1}^{m} \frac{\sum_{i \epsilon \mathcal{N}_k} p(s^{(i)})}{\sum_{i \epsilon \mathcal{N}_k} q(s^{(i)})}. \tag{5.9}$$

5.3 Simulation Study

We explore the improved importance sampling method outlined in this chapter for making inference for the population size based on the Rao-Blackwellized estimators introduced in subsection 4.5.1. We look at two different cases of how the neighbourhoods of the sample space are defined. In the first case we take any pair of sample

reorderings to be in the same neighbourhood if they share at least $n_0 - 1$ units in the initial wave. To clarify, for any sample reordering, this reordering is in the same neighbourhood as those for which we interchange one unit from the initial wave and one unit from wave one. In the second case we shall take any pair of sample reorderings to be in the same neighbourhood if they share at least $n_0 - 2$ units in the initial wave. To clarify, for any sample reordering this reordering is in the same neighbourhood as those for which we interchange either one unit from the initial wave and one unit from wave one, or, two units from the initial wave and two units from wave one.

With the above definitions of the neighbourhoods, once a sample reordering is selected under the importance sampler and its true probability of being selected and corresponding estimate of the population size is evaluated, the relevant information of any neighbour of this sample reordering can be evaluated readily. Recall that the members that are recruited for wave one are selected independently and with probability β_1, and hence all that is required is the corresponding observations obtained from replacing all internal and external nominations of the units interchanged from the initial wave with the units selected for wave one.

To evaluate the performance of the improved importance sampling method, we conducted a simulation study as follows. We selected 500 samples and $m = 500$ importance sampling draws of reorderings from each sample to make inference for the Rao-Blackwellized estimators. We explored using two different importance samplers where the first sampler gives a relative weight of 10 to 1 for each unit in the original sample's initial wave to be selected for the sampled reordering's initial wave relative to each unit in wave one. The second sampler is identical to the first except that a relative weight is chosen to be 25 to 1 for each unit in the original sample's initial wave to be selected for the sampled reordering's initial wave relative to each unit in wave one.

With the two definitions of neighbourhoods, each neighbourhood will be comprised of $\binom{n_0}{1}\binom{m_1}{1}$ sample reorderings in the first case and $\binom{n_0}{1}\binom{m_1}{1} + \binom{n_0}{2}\binom{m_1}{2}$ sample reorderings in the second case where n_0 is the size of the initial wave and m_1 is the number of members added adaptively to the sample. Instead of evaluating the necessary observations from all reorderings within a sampled neighbourhood, we

took a random sample r_k of size 100 in the first case and 250 in the second case to estimate the corresponding observations required from each neighbourhood k that was sampled. With this approach, the final improved consistent estimator can be shown to be

$$\tilde{N}^r_{RB} = \sum_{k=1}^{m} \frac{\sum\limits_{\substack{i \epsilon \mathcal{N}_k: \\ i \epsilon r_k}} \hat{N}_0^{(i)} p(s^{(i)})}{\sum\limits_{\substack{i \epsilon \mathcal{N}_k: \\ i \epsilon r_k}} q(s^{(i)})} \Big/ \sum_{k=1}^{m} \frac{\sum\limits_{\substack{i \epsilon \mathcal{N}_k: \\ i \epsilon r_k}} p(s^{(i)})}{\sum\limits_{\substack{i \epsilon \mathcal{N}_k: \\ i \epsilon r_k}} q(s^{(i)})}. \tag{5.10}$$

We used the following measure to evaluate the performance of the improved importance sampling estimators. We considered the reported values of

$$\sum_{k=1}^{500} (\tilde{N}^0_{RB,k} - \hat{N}^{(k)}_{RB})^2 \tag{5.11}$$

and

$$\sum_{k=1}^{500} (\tilde{N}^r_{RB,k} - \hat{N}^{(k)}_{RB})^2 \tag{5.12}$$

where $\tilde{N}^0_{RB,k}$ and $\tilde{N}^r_{RB,k}$ are the corresponding preliminary and (estimated) improved importance sampling approximations of the Rao-Blackwellized estimator $\hat{N}^{(k)}_{RB}$, respectively, that corresponds with sample k for $k = 1, 2, ..., 500$ (that is, those 500 samples selected for the simulation study in subsection 4.5.1). Our rationale is as follows: each of the values $(\tilde{N}^0_{RB,k} - \hat{N}^{(k)}_{RB})^2$ and $(\tilde{N}^r_{RB,k} - \hat{N}^{(k)}_{RB})^2$ is an (approximately) unbiased estimate for the variance of $\tilde{N}^0_{RB,k}$ and $\tilde{N}^r_{RB,k}$, respectively (since these are both unbiased estimators for $\hat{N}^{(k)}_{RB}$ when m is large enough), and hence this will provide us with a global measure of the performance of the preliminary and improved importance sampling inference methods.

Tables 5.1 and 5.2 give the output of the values corresponding with expressions (5.11) and (5.12) where the values are standardized by

$$\sum_{k=1}^{500}(\hat{N}_0^{(k)} - \hat{N}_{RB}^{(k)})^2 \tag{5.13}$$

to aid in comparing the performance of the estimators (notice that $\hat{N}_0^{(k)}$ can also be considered an unbiased estimator for $\hat{N}_{RB}^{(k)}$ where $\hat{N}_0^{(k)}$ is the preliminary estimate obtained with sample reordering k). We have replaced the values of $\hat{N}_{RB}^{(k)}$ with the estimated values that are obtained with an MCMC chain of length 5000 where the proposal distribution interchanged one unit from those hypothetical members that are selected for the initial wave with one unit from those that are selected for wave one while working over the most recently accepted reordering, as outlined in Section 4.4. With a chain of length 5000 we can consider these values to be very good approximations to the true values and hence are suitable substitutes for $\hat{N}_{RB}^{(k)}$. Also recall that each neighbourhood defined in case one is a subset of a neighbourhood defined in case two. We shall note here that with $m = 500$ all of the estimates came out with very little to no bias. That is, $\sum_{k=1}^{500} \tilde{N}_{RB,k}^{0} \approx \sum_{k=1}^{500} \tilde{N}_{RB,k}^{r} \approx \sum_{k=1}^{500} \hat{N}_{RB}^{(k)}$ where $N = 300$ is the population size.

Table 5.1: The observed standardized output that corresponds to the 10-1 importance sampler for approximating the Rao-Blackwellized estimates. Case 1 corresponds with the finer definition of the neighbourhoods and Case 2 corresponds with the coarser definition of the neighbourhoods. The preliminary scores correspond with the estimator presented in expression (5.11) and the improved scores correspond with the estimator presented in expression (5.12) where both scores are standardized by expression (5.13).

Estimator	Preliminary	Improved - Case 1	Improved - Case 2
$\hat{N}_{A,RB,0}$	0.995	0.970	0.951
$\hat{N}_{B,RB,0}$	1.011	0.981	0.968

Table 5.2: The observed standardized output that corresponds to the 25-1 importance sampler for approximating the Rao-Blackwellized estimates. Case 1 corresponds with the finer definition of the neighbourhoods and Case 2 corresponds with the coarser definition of the neighbourhoods. The preliminary scores correspond with the estimator presented in expression (5.11) and the improved scores correspond with the estimator presented in expression (5.12) where both scores are standardized by expression (5.13).

Estimator	Preliminary	Improved - Case 1	Improved - Case 2
$\hat{N}_{A,RB,0}$	0.878	0.837	0.635
$\hat{N}_{B,RB,0}$	0.900	0.833	0.709

Notice that under the 10-1 importance sampler, the reported values of the preliminary importance sampler are close to 1. It was determined that the 10-1 importance sampler selected very few sample reorderings that had positive probability of being selected under the target distribution and hence a large majority of the contributions to the estimator came from sample reorderings that registered a value of zero.

Even though there was a relatively small difference in the scores, the improved approach gave rise to estimators that reported corresponding values smaller than 1, as expected.

With respect to the 25-1 importance sampler, it is apparent that this importance sampler approximates the true target distribution better than the 10-1 importance sampler as the reported values from the preliminary estimator are significantly smaller than 1. Using the improved importance sampling procedure has resulted in significantly more efficient estimates. Notice the additional improvement in Case 2 where we defined neighbourhoods which are coarser than those defined in Case 1.

In summary, we could expect a value of zero in the standardized output if $\tilde{N}_{RB} \equiv \hat{N}_{RB}$, which would typically require $m \to \infty$ (since we are sampling with replacement). Notice that the output provided in the simulation study has exemplified that using the improved method will result in more efficient estimates, relative to the preliminary method, as these values are closer to zero.

5.4 Discussion

The improved importance sampling inference procedure outlined in this chapter demonstrated that improvements over the well-known preliminary importance sampling inference procedure are certain. One of the most attractive features about this inferential method is that the neighbourhoods can be defined in any manner desired by the analyst.

The methods that are outlined in this chapter can be extended, if necessary, to approximate any of the Rao-Blackwellized estimators that are outlined in this thesis. All that may be required is a redefining of the neighbourhoods. For example, for the Rao-Blackwellized estimators that are based on the adaptive web sampling designs outlined in chapter 3, one may define the neighbourhoods of the sample reorderings to be based on interchanging one of the units selected at random for the initial wave and one unit that is selected after the initial wave.

Significant improvements using the improved method, perhaps in cases where the choice of an ideal importance sampler is not readily available, can always be

obtained via defining coarser neighbourhoods. For example, in our study we could expect more efficient improved approximations of the Rao-Blackwellized estimates of the population size if we chose neighbourhoods to consist of reorderings that have at least $n_0 - 3$ units in the initial wave in common with the importance sample reordering that is selected on each draw, given that we observe all of the units in the neighbourhoods corresponding with the sampled units.

In our study we took a simple random sample of the units in each neighbourhood to estimate the desired response value from each neighbourhood (that is, y_k and q'_k as outlined in expressions (5.3) and (5.4), respectively), and with a coarser set of neighbourhoods it is likely that a larger simple random sample of elements from the sampled neighbourhoods will be required to obtain reliable approximations of the response values from these neighbourhoods. Hence, there is a trade-off between using coarser neighbourhoods and a greater amount of computational effort that may be required to obtain truly reliable estimates of the Rao-Blackwellized estimates of population unknowns.

The improved importance sampling method can be extended to work over a continuous target distribution of infinite range where neighbourhoods are defined over the domain of the sample space. The methods outlined in this chapter should facilitate in serving as a foundation for making this possible. Furthermore, extending the improved importance sampling procedure to work with an MCMC inference procedure is a topic that is deserving of future attention.

Chapter 6

Discussion

The goal of this thesis was to develop new methods for estimating the size and distribution of networked populations through the use of adaptive sampling methods. Three novel methods were introduced with one method utilizing a model-based approach to inference and two methods utilizing a design-based approach to inference. We have considered cases that cover both a single-sample and a multi-sample study.

In the first project, we introduced an elaborate graph model and developed an extended Bayesian data augmentation routine to make inference for the population size and model parameters. The Bayesian approach facilitated developing estimators as we were able to take advantage of working with the complete data likelihood, as opposed to the observed likelihood, based on a hypothetical full graph realization, which gave rise to suitable posterior distributions that were fairly straightforward to sample from. The inference procedure is somewhat restricted to the one-wave snowball sampling design. However, if sampling were to continue beyond the first wave then a combination of a model-based and design-based approach to inference via obtaining Rao-Blackwellized estimates of the population size and model parameters, perhaps by incorporating the inferential method outlined in Chapter 4 into the Bayesian analysis, can be achieved. We checked this method via a simulation study that consisted of an analysis based on the stochastic block model that was applied to the data from samples obtained from a small simulated networked population and found that the results came out as anticipated; the Rao-Blackwellized estimates

retained the same expectation as their preliminary estimator counterparts and had smaller variance. We should note that one immediate drawback when using this strategy for inference is presented in the computational effort that may be required to obtain the Rao-Blackwellized versions of the preliminary estimates as obtaining the preliminary estimate from each sample reordering may require the computation of an MCMC chain.

The second project introduced a design-based approach to inference of a population size with the use of a multi-sample study. Rao-Blackwellized versions of the preliminary estimates were shown to be obtainable as the unknown population size factors out of the Rao-Blackwellization expression. Obtaining design-based estimates of population unknowns, like the population size and average node degree directly based on the full adaptive samples, is likely to be a complicated task if the the ability to identify the sampled units' neighbours is unavailable and/or if the true population size is not known. The reason for this is that these two restrictions will not permit for the inclusion probabilities of the sampled units to be observed and therefore can not be incorporated into the inferential procedure. Hence the method outlined in this project may prove to be highly useful in an empirical setting. In future work, amalgamating the methods outlined in this project with some of the common capture-recapture models would be highly useful.

The third project extended the methods developed by Frank and Snijders (1994) for estimating population sizes with a design-based approach to inference based on selecting one sample. We proposed new estimates of the population size based on one wave that is selected after the initial wave is obtained. We also developed a Rao-Blackwellized version of these preliminary estimates based on a sufficient statistic in sampling in a manner similar to that which was introduced in the second project. In future work, developing similar inferential methods for samples obtained from a respondent-driven sampling design (Heckathorn, 1997, 2002), where identification of nominated members of the hard-to-reach populations is a challenge and will likely be aided by a model-assisted approach (for example, see Gile and Handcock (2011)), would be extremely beneficial as this sampling method is currently being employed in some empirical studies (Abdul-Quader et al., 2006).

The fourth project introduced a new method for estimating a distribution based on a strategy termed *improved importance sampling*. We showed that, with a single-stage cluster sampling style of design, this method will produce more reliable approximations of the characteristics of a distribution relative to the existing preliminary procedure. Typically, with importance sampling there will be a trade-off between computational effort and the efficiency of the estimates. With our approach we demonstrated that in the event that incorporating the neighbourhood responses into the inference procedure does not result in a significant increase in computational effort, either significantly more efficient estimates can be obtained or significantly less computational effort will be required to obtain approximations comparable in efficiency relative to those obtained with the preliminary procedure. Extending this method to work with adaptive MCMC strategies will be an interesting future challenge.

Bibliography

Abdul-Quader, A., Heckathorn, D., McKnight, C., Bramson, H., Nemeth, C., Sabin, K., Gallagher, K., and Des Jarlais, D. (2006). Effectiveness of respondent-driven sampling for recruiting drug users in new york city: Findings from a pilot study. *Journal of Urban Health* **83,** 459–476. 10.1007/s11524-006-9052-7.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive markov chain monte carlo algorithms. *Bernoulli* **11,** 815–828.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43,** pp. 783–791.

Chao, A., Tsay, P. K., Lin, S.-H., Shau, W.-Y., and Chao, D.-Y. (2001). The applications of capture-recapture models to epidemiological data. *Statistics in Medicine* **20,** 3123–3157.

Chapman, D. (1951). Some properties of the hypergeometric distribution with applications to zoological sample census. *University of California Publications in Statistics* **1,** 131–160.

Felix-Medina, M. and Monjardin, P. (2006). Combining link-tracing 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.

Fienberg, S. E. (2010a). Introduction to papers on the modeling and analysis of network data—I. *The Annals of Applied Statistics* **4,** 1–4.

Fienberg, S. E. (2010b). Introduction to papers on the modeling and analysis of network data—II. *ArXiv e-prints* .

Frank, O. and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics* **10,** 53–67.

Frischer, M., Leyland, A., Cormack, R., Goldberg, D. J., Bloor, M., Green, S. T., Taylor, A., Covell, R., McKeganey, N., and Platt, S. (1993). Estimating the population prevalence of injection drug use and infection with human immunodeficiency virus among injection drug users in glasgow, scotland. *American Journal of Epidemiology* **138,** 170–181.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, London.

Gile, K. and Handcock, M. (2011). Network model-assisted inference from respondent-driven sampling data. *Arxiv preprint arXiv:1108.0298* .

Handcock, M. S. and Gile, K. J. (2010). Modeling social networks from sampled data. *Annals of applied statistics* **4,** 5–25.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika* **57,** 97–109.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems* **44,** 174–199.

Heckathorn, D. D. (2002). Respondent-driven sampling ii: Deriving valid population estimates from chain-referral samples. *Social Problems* **49,** 11–34.

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.

Hook, E. B. and Regal, R. R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews* **17,** 243–264.

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

Mastro, T. D., Kitayaporn, D., Weniger, B. G., Vanichseni, S., Laosunthorn, V., Uneklabh, T., Uneklabh, C., and Choopanya, K. (1994). Estimating the number of HIV-infected injection drug users in Bangkok: a capture–recapture method. *American Journal of Public Health* **84,** 1094–1099.

Petersen, C. (1896). The yearly immigration of young plaice into the limfjord from the german sea. *Report of the Danish Biological Station* **6,** 5–84.

Schwarz, C. J. and Seber, G. A. F. (1999). Estimating animal abundance: Review iii. *Statistical Science* **14,** 427–456.

Seber, G. A. F. (1970). The effects of trap response on tag recapture estimates. *Biometrics* **26,** 13–22.

Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters.* London Blackburn Press, 2nd edition.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin de Méthodologie Sociologique* **36,** 34–58.

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. (2002). *Sampling*. Wiley Series in Probability and Statistics, New Jersey.

Thompson, S. K. (2006a). Adaptive web sampling. *Biometrics* **62,** 1224–1234.

Thompson, S. K. (2006b). Targeted random walk designs. *Survey Methodology* **32,** 11–24.

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.

Appendix A

Development of Stochastic Cluster Model

This appendix provides the derivations of the probability mass/density functions of the missing data and the development of posterior distributions for the model parameters that were outlined in the data augmentation procedure corresponding with the use of the stochastic cluster model, as described in subsection 2.3.3.

A.1 Probability Mass/Density Functions of the Missing Values

A.1.1 The probability mass function of group memberships

After sampling an N from the binomial style posterior distribution (as presented in expression (2.20)), we continue the data augmentation process with the observed graph data

$$d_0 = \{\mathrm{S}, \underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U}\} \tag{A.1}$$

where U is a hypothetical population of size N.

Now, the probability of obtaining a realization of $\underline{C}_{\bar{\mathrm{S}}}$ given d_0 is obtained as

$$\begin{aligned}
\mathrm{P}(\underline{C}_{\bar{\mathrm{S}}}|d_0) &= \mathrm{P}(\underline{C}_{\bar{\mathrm{S}}}|\mathrm{S}, \underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U}) \\
&= \frac{\mathrm{P}(\mathrm{S}, \underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U}|\underline{C}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{C}_{\bar{\mathrm{S}}})}{\mathrm{P}(d_0)} \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U}, \underline{C}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U}|\underline{C}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{C}_{\bar{\mathrm{S}}})}{\mathrm{P}(d_0)} \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_0)} \cdot \mathrm{P}(Y_{\mathrm{S}_0,U}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, \underline{C}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, |\underline{C}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{C}_{\bar{\mathrm{S}}}) \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_0)} \cdot \mathrm{P}(Y_{\mathrm{S}_0,U}|\underline{C}, \underline{Z}_{\mathrm{S}}) \cdot \mathrm{P}(\underline{Z}_{\mathrm{S}}|\underline{C}_{\mathrm{S}}, \underline{C}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{C}_{\mathrm{S}}, |\underline{C}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{C}_{\bar{\mathrm{S}}}) \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_0)} \cdot \mathrm{P}(Y_{\mathrm{S}_0,U}|\underline{C}, \underline{Z}_{\mathrm{S}}) \cdot \mathrm{P}(\underline{Z}_{\mathrm{S}}|\underline{C}) \cdot \mathrm{P}(\underline{C}). \qquad \text{(A.2)}
\end{aligned}$$

Note that $\underline{C}_{\bar{\mathrm{S}}}$ 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,

$$\begin{aligned}
\mathrm{P}(\underline{C}_{\bar{\mathrm{S}}}|d_0) &= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_0)} \cdot \mathrm{P}(Y_{\mathrm{S}_0,U}|\underline{C}, \underline{Z}_{\mathrm{S}}) \cdot \mathrm{P}(\underline{Z}_{\mathrm{S}}|\underline{C}) \cdot \mathrm{P}(\underline{C}) \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_0)} \cdot \mathrm{P}(Y_{\mathrm{S}_0,U}|\underline{C}, \underline{Z}_{\mathrm{S}}) \cdot \mathrm{P}(\underline{Z}_{\mathrm{S}}|\underline{C}) \cdot \prod_{i=1}^{N} \mathrm{P}(C_i) \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_0)} \cdot \mathrm{P}(Y_{\mathrm{S}_0,U}|\underline{C}, \underline{Z}_{\mathrm{S}}) \cdot \prod_{i=1}^{n} \mathrm{BVN}(Z_i; \mu_{C_i}, \sigma^2_{C_i}\mathrm{I}_d) \cdot \prod_{i=1}^{N} \mathrm{P}(C_i) \ , \qquad \text{(A.3)}
\end{aligned}$$

where

$$P(Y_{\mathrm{S}_0,U}|\underline{C},\underline{Z}_{\mathrm{S}}) =$$
$$\prod_{i=1}^{n_0}\prod_{\substack{i,j=1:\\ i<j}}^{n}\mathrm{P}(Y_{ij}|C_i,C_j,Z_i,Z_j)\cdot\prod_{k=n+1}^{N}\int_{Z_k}\prod_{i=1}^{n_0}\mathrm{P}(Y_{ij}|C_i,C_k,Z_i,Z_k)\mathrm{BVN}(Z_k;\mu_k,\sigma_k^2\mathrm{I}_d)\ \mathrm{d}Z_k\ . \tag{A.4}$$

Therefore, by the factorization theorem we have that

$$\begin{aligned} P(\underline{C}_{\bar{\mathrm{S}}}|d_0) &= \prod_{i\in\bar{\mathrm{S}}}\mathrm{P}(C_i|d_0) \\ &= \prod_{i\in\bar{\mathrm{S}}}\mathrm{P}(C_i|\mathrm{S},\underline{C}_{\mathrm{S}},\underline{Z}_{\mathrm{S}},Y_{\mathrm{S}_0,U}) \\ &= \prod_{i\in\bar{\mathrm{S}}}\mathrm{P}(C_i|\mathrm{S},\underline{C}_{\mathrm{S}_0},\underline{Z}_{\mathrm{S}_0},Y_{\mathrm{S}_0,i})\ . \end{aligned} \tag{A.5}$$

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

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

$$
\begin{aligned}
&P(C_i = k|\mathrm{S}, \underline{C}_{\mathrm{S}_0}, \underline{Z}_{\mathrm{S}_0}, Y_{\mathrm{S}_0,i}) \\
&= \frac{\mathrm{P}(C_i = k, \mathrm{S}, \underline{C}_{\mathrm{S}_0}, \underline{Z}_{\mathrm{S}_0}, Y_{\mathrm{S}_0,i})}{\sum_{\ell=1}^{G} \mathrm{P}(C_i = \ell, \mathrm{S}, \underline{C}_{\mathrm{S}_0}, \underline{Z}_{\mathrm{S}_0}, Y_{\mathrm{S}_0,i})} \\
&= \frac{\mathrm{P}(C_i = k) \cdot \mathrm{P}(Y_{\mathrm{S}_0,i}|\mathrm{S}, \underline{C}_{\mathrm{S}_0}, \underline{Z}_{\mathrm{S}_0}, C_i = k)}{\sum_{\ell=1}^{G} \left[\mathrm{P}(C_i = \ell) \cdot \mathrm{P}(Y_{\mathrm{S}_0,i}|\mathrm{S}, \underline{C}_{\mathrm{S}_0}, \underline{Z}_{\mathrm{S}_0}, C_i = \ell)\right]} \\
&= \frac{\mathrm{P}(C_i = k) \cdot \prod_{j=1}^{n_0} \mathrm{P}(Y_{ij} = 0|\mathrm{S}, C_j, Z_j, C_i = k)}{\sum_{\ell=1}^{G} \left[\mathrm{P}(C_j = \ell) \cdot \prod_{j=1}^{n_0} \mathrm{P}(Y_{ij} = 0|\mathrm{S}, C_j, Z_j, C_i = \ell)\right]} \\
&= \frac{\lambda_k \cdot \int_{-\infty}^{\infty} \prod_{j=1}^{n_0} \left(1 - \frac{\exp(\beta_{C_j,k} + \alpha_{C_j,k}||Z_j - Z||)}{1+\exp(\beta_{C_j,k} + \alpha_{C_j,k}||Z_j - Z||)}\right) \mathrm{BVN}(Z; \mu_k, \sigma_k^2 \mathrm{I}_d) \ \mathrm{d}Z}{\sum_{\ell=1}^{G} \left[\lambda_\ell \cdot \int_{-\infty}^{\infty} \prod_{j=1}^{n_0} \left(1 - \frac{\exp(\beta_{C_j,\ell} + \alpha_{C_j,\ell}||Z_j - Z||)}{1+\exp(\beta_{C_j,\ell} + \alpha_{C_j,\ell}||Z_j - Z||)}\right) \mathrm{BVN}(Z; \mu_\ell, \sigma_\ell^2 \mathrm{I}_d) \ \mathrm{d}Z\right]}.
\end{aligned} \tag{A.6}
$$

A.1.2 The probability density function of the parameters corresponding to the covariate information

For a specific realization $\underline{Z}_{\bar{S}}$, the density given $d_1 = \{S, \underline{C}, \underline{Z}_S, Y_{S_0,U}\}$ is evaluated as

$$
\begin{aligned}
&P(\underline{Z}_{\bar{\mathrm{S}}}|d_1) = P(\underline{Z}_{\bar{\mathrm{S}}}|\mathrm{S}, \underline{C}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U}) \\
&= \frac{\mathrm{P}(\mathrm{S}, \underline{C}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U}|\underline{Z}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{Z}_{\bar{\mathrm{S}}})}{\mathrm{P}(d_1)} \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U}, \underline{Z}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{C}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U}|\underline{Z}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{Z}_{\bar{\mathrm{S}}})}{\mathrm{P}(d_1)} \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_1)} \cdot \mathrm{P}(Y_{\mathrm{S}_0,U}|\underline{C}, \underline{Z}_{\mathrm{S}}, \underline{Z}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{C}, \underline{Z}_{\mathrm{S}}|\underline{Z}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{Z}_{\bar{\mathrm{S}}}) \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_1)} \cdot \mathrm{P}(Y_{\mathrm{S}_0,U}|\underline{C}, \underline{Z}) \cdot \mathrm{P}(\underline{C}|\underline{Z}_{\mathrm{S}}, \underline{Z}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{Z}_{\mathrm{S}}|\underline{Z}_{\bar{\mathrm{S}}}) \cdot \mathrm{P}(\underline{Z}_{\bar{\mathrm{S}}}) \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_1)} \cdot \mathrm{P}(Y_{\mathrm{S}_0,U}|\underline{C}, \underline{Z}) \cdot \mathrm{P}(\underline{C}|\underline{Z}) \cdot \mathrm{P}(\underline{Z}) \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_1)} \cdot \mathrm{P}(Y_{\mathrm{S}_0,U}|\underline{C}, \underline{Z}) \cdot \mathrm{P}(\underline{Z}|\underline{C}) \cdot \mathrm{P}(\underline{C}) \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_1)} \cdot \mathrm{P}(Y_{\mathrm{S}_0,U}|\underline{C}, \underline{Z}) \cdot \prod_{i=1}^{N} \mathrm{BVN}(Z_i; \mu_{C_i}, \sigma^2_{C_i}\mathrm{I}_d) \cdot \prod_{i=1}^{N} \mathrm{P}(C_i) \\
&= \frac{\mathrm{P}(\mathrm{S}|\underline{C}_{\mathrm{S}}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U})}{\mathrm{P}(d_1)} \cdot \prod_{j=1}^{N} \prod_{i=1}^{n_0} \mathrm{P}(Y_{ij}|C_i, C_j, Z_i, Z_j) \cdot \prod_{i=1}^{N} \mathrm{BVN}(Z_i; \mu_{C_i}, \sigma^2_{C_i}\mathrm{I}_d) \cdot \prod_{i=1}^{N} \mathrm{P}(C_i).
\end{aligned}
\tag{A.7}
$$

Note that $\underline{Z}_{\bar{S}}$ is dropped in the first term by use of the adaptive sampling design (Thompson and Seber, 1996). Once again, by the factorization theorem, we have that

$$
\begin{aligned}
P(\underline{Z}_{\bar{\mathrm{S}}}|d_1) &= \prod_{i \in \bar{\mathrm{S}}} \mathrm{P}(Z_i|d_1) \\
&= \prod_{i \in \bar{\mathrm{S}}} \mathrm{P}(Z_i|\mathrm{S}, \underline{C}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,U}) \ .
\end{aligned}
\tag{A.8}
$$

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

$$
\begin{aligned}
\mathrm{P}(Z_i = z^*|d_1) &= \mathrm{P}(Z_i = z^*|\mathrm{S}, \underline{C}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,i}) \\
&= \frac{\mathrm{P}(Z_i = z^*, \mathrm{S}, \underline{C}, \underline{Z}_{\mathrm{S}}, Y_{\mathrm{S}_0,i})}{\int_{-\infty}^{\infty} \mathrm{P}(Z_i = Z, \mathrm{S}, \underline{C}, \underline{Z}_{\mathrm{S}}, Y_{S_0,i})\ \mathrm{d}Z} \\
&= \frac{\mathrm{P}(Z_i = z^*|\underline{C}) \cdot \mathrm{P}(Y_{\mathrm{S}_0,i}|\mathrm{S}, \underline{C}, \underline{Z}_{\mathrm{S}_0}, Z_i = z^*)}{\int_{-\infty}^{\infty} \left(\mathrm{P}(Z_i = Z|\underline{C}) \cdot \mathrm{P}(Y_{\mathrm{S}_0,i}|\mathrm{S}, \underline{C}, \underline{Z}_{\mathrm{S}_0}, Z_i = Z)\right)\ \mathrm{d}Z} \\
&= \frac{\prod_{j=1}^{n_0} \mathrm{P}(Y_{ij}|C_j, C_i, Z_j, Z_i = z^*)\mathrm{BVN}(z^*; \mu_{C_i}, \sigma^2_{C_i})\mathrm{I}_d}{\int_{-\infty}^{\infty} \prod_{j=1}^{n_0} \mathrm{P}(Y_{ij}|C_j, C_i, Z_j, Z_i = Z)\mathrm{BVN}(Z; \mu_{C_i}, \sigma^2_{C_i}\mathrm{I}_d)\ \mathrm{d}Z} \\
&= \frac{\prod_{j=1}^{n_0} \left(1 - \frac{\exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i}||Z_j - z^*||}{1+\exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i}||Z_j - z^*||}\right) \mathrm{BVN}(z^*; \mu_{C_i}, \sigma^2_{C_i}\mathrm{I}_d)}{\int_{-\infty}^{\infty} \prod_{j=1}^{n_0} \left(1 - \frac{\exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i}||Z_j - Z||}{1+\exp(\beta_{C_j,C_i} + \alpha_{C_j,C_i}||Z_j - Z||}\right) \mathrm{BVN}(Z; \mu_{C_i}, \sigma^2_{C_i}\mathrm{I}_d)\ \mathrm{d}Z}.
\end{aligned}
\tag{A.9}
$$

A.2 The Posterior Distributions of the Model Parameters

A.2.1 The joint posterior distribution of (σ_k^2, μ_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 $\pi(\sigma_k^2|\underline{Z}_k)$ and $\pi(\mu_k|\sigma_k^2, \underline{Z}_k)$. Recall that we have that

$$
\begin{aligned}
Z_{1,1}, Z_{1,2}, \ldots, Z_{1,N_k} &\overset{\text{iid}}{\sim} \mathrm{N}(\mu_{1,k}, \sigma_k^2), \\
Z_{2,1}, Z_{2,2}, \ldots, Z_{2,N_k} &\overset{\text{iid}}{\sim} \mathrm{N}(\mu_{2,k}, \sigma_k^2),
\end{aligned}
\tag{A.10}
$$

all of which arise independently given the group memberships.

In order to evaluate the posterior distributions we will need

$$
\begin{aligned}
&f(\underline{Z}_k|\mu_{1,k},\mu_{2,k},\sigma_k^2) \\
&= \prod_{i=1}^{N_k}\left[\frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left\{-\frac{1}{2\sigma_k^2}(Z_{1,i}-\mu_{1,k})^2\right\}\right]\cdot\prod_{i=1}^{N_k}\left[\frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left\{-\frac{1}{2\sigma_k^2}(Z_{2,i}-\mu_{2,k})^2\right\}\right] \\
&= \frac{1}{(2\pi\sigma_k^2)^{N_k}}\cdot\exp\left\{-\frac{1}{2\sigma_k^2}\cdot\sum_{i=1}^{N_k}(Z_{1,i}-\mu_{1,k})^2\right\}\cdot\exp\left\{-\frac{1}{2\sigma_k^2}\cdot\sum_{i=1}^{N_k}(Z_{2,i}-\mu_{2,k})^2\right\}\ .
\end{aligned}
\tag{A.11}
$$

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

$$
\pi(\sigma_k^2)=\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma_k^2)^{-\alpha-1}\exp\left\{\frac{-\beta}{\sigma_k^2}\right\}\ . \tag{A.12}
$$

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

$$
\begin{aligned}
&\pi(\mu_{1,k},\mu_{2,k}|\sigma_k^2)=\pi(\mu_{1,k}|\sigma_k^2)\pi(\mu_{2,k}|\sigma_k^2) \\
&= \frac{1}{\sqrt{2\pi\frac{\sigma_k^2}{\nu_1}}}\exp\left\{-\frac{1}{2}\frac{(\mu_{1,k}-\gamma_1)^2}{\frac{\sigma_k^2}{\nu_1}}\right\}\cdot\frac{1}{\sqrt{2\pi\frac{\sigma_k^2}{\nu_2}}}\exp\left\{-\frac{1}{2}\frac{(\mu_{2,k}-\gamma_2)^2}{\frac{\sigma_k^2}{\nu_2}}\right\}\ .
\end{aligned}
\tag{A.13}
$$

Therefore we have the following posterior distribution,

$$
\begin{aligned}
&\pi(\sigma_k^2, \mu_{1,k}, \mu_{2,k}|\underline{Z}_k) \propto \pi(\sigma_k^2) \cdot \pi(\mu_{1,k}, \mu_{2,k}|\sigma_k^2) \cdot f(\underline{Z}_k|\mu_{1,k}, \mu_{2,k}, \sigma_k^2) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma_k^2)^{-\alpha-1} \exp\left\{-\frac{\beta}{\sigma_k^2}\right\} \\
&\quad \cdot \frac{1}{\sqrt{2\pi}\frac{\sigma_k}{\sqrt{\nu_1}}} \exp\left\{-\frac{1}{2}\frac{(\mu_{1,k}-\gamma_1)^2}{\frac{\sigma_k^2}{\nu_1}}\right\} \cdot \frac{1}{\sqrt{2\pi}\frac{\sigma_k}{\sqrt{\nu_2}}} \exp\left\{-\frac{1}{2}\frac{(\mu_{2,k}-\gamma_2)^2}{\frac{\sigma_k^2}{\nu_2}}\right\} \\
&\quad \cdot \frac{1}{(2\pi\sigma_k^2)^{N_k}} \exp\left\{-\frac{1}{2\sigma_k^2}\left[\sum_{i=1}^{N_k}(Z_{1,i}-\mu_{1,k})^2 + \sum_{i=1}^{N_k}(Z_{2,i}-\mu_{2,k})^2\right]\right\} \\
&\propto (\sigma_k^2)^{-\alpha-1} \cdot \exp\left\{-\frac{\beta}{\sigma_k^2}\right\} \\
&\quad \cdot \frac{1}{\frac{\sigma_k}{\sqrt{\nu_1}}} \exp\left\{-\frac{1}{2}\frac{(\mu_{1,k}-\gamma_1)^2}{\frac{\sigma_k^2}{\nu_1}}\right\} \cdot \frac{1}{\frac{\sigma_k}{\sqrt{\nu_2}}} \exp\left\{-\frac{1}{2}\frac{(\mu_{2,k}-\gamma_2)^2}{\frac{\sigma_k^2}{\nu_2}}\right\} \\
&\quad \cdot \frac{1}{(\sigma_k^2)^{N_k}} \exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i=1}^{N_k}(Z_{1,i}-\mu_{1,k})^2\right\} \cdot \exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i=1}^{N_k}(Z_{2,i}-\mu_{2,k})^2\right\} \ .
\end{aligned}
\tag{A.14}
$$

Notice that, by the factorization theorem, $\mu_{1,k}|\sigma_k^2, \underline{Z}_k$ is independent of $\mu_{2,k}|\sigma_k^2, \underline{Z}_k$.

A.2.2 The posterior distribution of σ_k^2

We wish to determine $\pi(\sigma_k^2|\underline{Z}_k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \pi(\sigma_k^2, \mu_{1,k}, \mu_{2,k}|\underline{Z}_k)\ \mathrm{d}\mu_{1,k}\ \mathrm{d}\mu_{2,k}$. Integrating over expression (A.14), we have

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \pi(\sigma_k^2, \mu_{1,k}, \mu_{2,k}|\underline{Z}_k)\ \mathrm{d}\mu_{1,k}\ \mathrm{d}\mu_{2,k} \\
&= (\sigma_k^2)^{-\alpha-1} \exp\left\{-\frac{\beta}{\sigma_k^2}\right\} \cdot \frac{1}{\frac{\sigma_k}{\sqrt{\nu_1}}} \cdot \frac{1}{\frac{\sigma_k}{\sqrt{\nu_2}}} \cdot \frac{1}{(\sigma_k^2)^{N_k}} \\
&\quad \cdot \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\frac{(\mu_{1,k}-\gamma_1)^2}{\frac{\sigma_k^2}{\nu_1}} - \frac{1}{2\sigma_k^2}\sum_{i=1}^{N_k}(Z_{1,i}-\mu_{1,k})^2\right\}\ \mathrm{d}\mu_{1,k} \\
&\quad \cdot \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\frac{(\mu_{2,k}-\gamma_2)^2}{\frac{\sigma_k^2}{\nu_2}} - \frac{1}{2\sigma_k^2}\sum_{i=1}^{N_k}(Z_{2,i}-\mu_{2,k})^2\right\}\ \mathrm{d}\mu_{2,k}\ . \qquad \text{(A.15)}
\end{aligned}
$$

By completing the square over $\mu_{1,k}$ and $\mu_{2,k}$, it can be shown that

$$
\begin{aligned}
\pi(\sigma_k^2|\underline{Z}_k) &\propto (\sigma_k^2)^{-\alpha-1} \exp\left\{-\frac{\beta}{\sigma_k^2}\right\} \cdot \frac{1}{\sigma_k} \cdot \frac{1}{\sigma_k} \cdot \frac{1}{(\sigma_k^2)^{N_k/2}} \cdot \frac{1}{(\sigma_k^2)^{N_k/2}} \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma_k^2}\left[\nu_1\gamma_1^2 + \sum_{i=1}^{N_k} Z_{1,i}^2 - \frac{(\nu_1\gamma_1 + \sum_{i=1}^{N_k} Z_{1,i})^2}{\nu_1 + N_k}\right]\right\} \cdot \frac{1}{\sigma_k} \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma_k^2}\left[\nu_2\gamma_2^2 + \sum_{i=1}^{N_k} Z_{2,i}^2 - \frac{(\nu_2\gamma_2 + \sum_{i=1}^{N_k} Z_{2,i})^2}{\nu_2 + N_k}\right]\right\} \cdot \frac{1}{\sigma_k} \\
&= (\sigma_k^2)^{-\alpha-1-1-N_k-1} \cdot \exp\left\{-\frac{1}{\sigma_k^2} \cdot \left[\beta + \frac{1}{2}\left[\nu_1\gamma_1^2 + \sum_{i=1}^{N_k} Z_{1,i}^2 - \frac{(\nu_1\gamma_1 + \sum_{i=1}^{N_k} Z_{1,i})^2}{\nu_1 + N_k}\right.\right.\right. \\
&\quad \left.\left.\left. + \nu_2\gamma_2^2 + \sum_{i=1}^{N_k} Z_{2,i}^2 - \frac{(\nu_2\gamma_2 + \sum_{i=1}^{N_k} Z_{2,i})^2}{\nu_2 + N_k}\right]\right]\right\}\ . \qquad \text{(A.16)}
\end{aligned}
$$

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

$$A = \alpha + N_k, \text{ and} \tag{A.17}$$

$$B = \beta + \frac{1}{2}\left[\nu_1\gamma_1^2 + \sum_{i=1}^{N_k} Z_{1,i}^2 - \frac{(\nu_1\gamma_1 + \sum_{i=1}^{N_k} Z_{1,i})^2}{\nu_1 + N_k} + \nu_2\gamma_2^2 + \sum_{i=1}^{N_k} Z_{2,i}^2 - \frac{(\nu_2\gamma_2 + \sum_{i=1}^{N_k} Z_{2,i})^2}{\nu_2 + N_k}\right]. \tag{A.18}$$

A.2.3 The posterior distribution of $\mu_k|\sigma_k^2$

To find the posterior distribution of $\mu_{1,k}$, we condition on the σ_k^2 sampled from the distribution found in (A.16), and hence

$$\begin{aligned}
\pi(\mu_{1,k}|\underline{Z}_k, \sigma_k^2) &\propto \exp\left\{-\frac{1}{2}\frac{(\mu_{1,k}-\gamma_1)^2}{\frac{\sigma_k^2}{\nu_1}}\right\}\exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^{N_k}(Z_{1,i}-\mu_{1,k})^2}{\sigma_k^2}\right\} \\
&= \exp\left\{-\frac{1}{2\sigma_k^2}\left[\nu_1\cdot(\mu_{1,k}^2 - 2\gamma_1\mu_{1,k} + \gamma_1^2) + \sum_{i=1}^{N_k}(\mu_{1,k}^2 - 2Z_{1,i}\mu_{1,k} + Z_{1,i}^2)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_k^2}\left[\nu_1\mu_{1,k}^2 - 2\gamma_1\nu_1\mu_{1,k} + N_k\mu_{1,k}^2 - \mu_{1,k}\sum_{i=1}^{N_k} 2Z_{1,i}\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma_k^2}\left[(N_k+\nu_1)\mu_{1,k}^2 + \mu_{1,k}\cdot(-2\gamma_1\nu_1 - \sum_{i=1}^{N_k} 2Z_{1,k})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma_k^2}\left[(N_k+\nu_1)\left(\mu_{1,k} - \frac{\gamma_1\nu_1 + \sum_{i=1}^{N_k} Z_{1,i}}{\nu_1 + N_k}\right)^2\right]\right\} \\
&= \exp\left\{-\frac{1}{2\frac{\sigma_k^2}{N_k+\nu_1}}\cdot\left(\mu_{1,k} - \frac{\gamma_1\nu_1 + \sum_{i=1}^{N_k} Z_{1,i}}{\nu_1 + N_k}\right)^2\right\}
\end{aligned} \tag{A.19}$$

Therefore,

$$\mu_{1,k}|\sigma_k^2, \underline{Z}_k \sim \mathrm{N}\left(\frac{\gamma_1\nu_1 + \sum_{i=1}^{N_k} Z_{1,i}}{\nu_1 + N_k}, \frac{\sigma_k^2}{\nu_1 + N_k}\right) , \tag{A.20}$$

and similarly,

$$\mu_{2,k}|\sigma_k^2, \underline{Z}_k \sim \mathrm{N}\left(\frac{\gamma_2\nu_2 + \sum_{i=1}^{N_k} Z_{2,i}}{\nu_2 + N_k}, \frac{\sigma_k^2}{\nu_2 + N_k}\right) . \tag{A.21}$$

Appendix B

Sufficient Statistic when N is Unknown

This appendix provides an illustration to help clarify the use of the notation and adaptive web sampling designs that are outlined in chapter 3. We also prove that $d_r = \{(i, w_{ij}, w_i+), \underline{\mathcal{J}} : i, j \, \epsilon \, s_k, \; k = 1, 2, ..., K\}$ is a sufficient statistic (as mentioned in Section 3.2) when the population size is unknown and an adaptive web sampling strategy is employed.

B.1 An Illustration to Clarify the use of the Notation and Adaptive Web Sampling Designs

To help clarify the notation used in chapter 3, Figure B.1 provides an example of two samples that are selected under the original adaptive web sampling design where $a_{t_k} = s_{t_k}$ for each step $t_k = 1, 2$ and for each sample $k = 1, 2$ in the sample selection procedure. The size of the initial random samples are $n_{01} = n_{02} = 1$ and the number of members added after the initial samples is two to bring the final sample sizes up to $n_1 = n_2 = 3$. The original order the samples are selected in shall be assumed to be $s_{(0_1,0_2)} = ((A, B, C), (A, D, E))$.

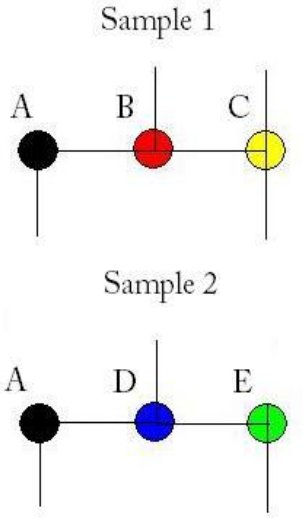

Figure B.1: Two adaptive web samples selected via the original adaptive web sampling design where the initial sample sizes are one and the final sample sizes are three.

In the event that $0 < d < 1$ and both members in s_1 are added via taking a random jump and both members in s_2 are added via tracing a link out of the active set then $\underline{J}_1 = (0, 0, 0)$ and $\underline{J}_2 = (0, 1, 1)$ are the original jump vectors (and hence $\underline{\mathcal{J}} = (0, 1, 1)$). One probable pair of sample reorderings is $s_{(x_1,x_2)} = ((C, A, B), (D, A, E))$ if we allow for a random jump to be taken when unit A is added to the corresponding reordering for sample 1, for some pair of reorderings (x_1, x_2) where $x_1, x_2 = 1, 2, .., 3!$. Notice that this requires unit A to be added via tracing a link out of sample 2 since there is a jump that is taken at this point in the sample selection procedure. In this case, $\underline{J}_1^{(x_1)} = (0, 1, 0)$ and $\underline{J}_2^{(x_2)} = (0, 0, 1)$ so that $\underline{\mathcal{J}}^{(x_1,x_2)} = (0, 1, 1)$ is consistent with $\underline{\mathcal{J}}$. In the event that $d = 1$, then no random jumps are taken and hence $s_{(x_1,x_2)} = ((C, A, B), (D, A, E)$ is not a pair of probable sample reorderings as there is no link to trace from unit C to unit A in the first sample (recall that this requires $\underline{\mathcal{J}} \equiv 0$).

We shall also clarify here that under the nearest neighbours adaptive web sampling design, unless random jumps are permitted (that is $0 < d < 1$) neither of the samples $s_{0_1} = (A, B, C)$ and $s_{0_2} = (A, D, E)$ can be selected since the active set is restricted to the initial sample. That is, if $d = 1$ only members linked to unit A can be added for the samples.

B.2 Sufficient Statistic in Adaptive Web Sampling when the Population Size is Unknown

We shall commence with a review of the adaptive web sampling selection procedure.

The selection of an adaptive web sampling design consists of two stages. For each sample $k = 1, 2, ..., K$, the sample selection procedure commences with the selection of n_{0k} members completely at random and then $n_k - n_{0k}$ members are added adaptively. The adaptively selected members are added as follows. For each step t_k, $t_k = 1, 2, ..., n_k - n_{0k}$, any member i not yet chosen is selected with probability $q_{t_k,i} = d\frac{w_{a_{t_k},i}}{w_{a_{t_k},+}} + (1-d)\frac{1}{N-(n_{0k}+t_k-1)}$ where $w_{a_{t_k},i}$ is the number of links from the current active set $a_{t_k} \subseteq s_{t_k}$ (where s_{t_k} is the current sample at time t_k) out to unit i at step t_k and $w_{a_{t_k},+}$ is the number of links out of the current active set to members not yet selected at step t_k. Hence, with probability $0 \leq d \leq 1$ a unit is added via tracing a link from the active set and with probability $1 - d$ a unit is added completely at random (that is, a random jump is taken), given that $w_{a_{t_k},+} > 0$. In the event that $w_{a_{t_k},+} = 0$, a member is selected completely at random (a random jump is taken) amongst those not yet selected with probability $\frac{1}{N-(n_{0k}+t_k-1)}$.

The observed data is $d_0 = \{(i, w_{ij}, w_i^+, t_{i,k}), \underline{J}_k : i, j \in s_k,\ k = 1, 2, ..., K\}$ where s_k refers to sample k for $k = 1, 2, ..., K$; w_i^+ is the out-degree of individual i (that is, the number of members acknowledged by individual i); $t_{i,k}$ is the time (or step) in the sampling sequence that unit i is selected for sample k; $\underline{J}_k$ is an indicator vector of length $L = \max_{j=1,2,...,K}\{n_j\}$ that records the sequence of jumps after the initial sample is selected for sample $k, k = 1, 2, ..., K$. It shall be understood that for all $k = 1, 2, ..., K$, $J_{1,k}, ..., J_{n_{0k},k} = 0$ and if $n_k < \max_{j=1,2,...,K}\{n_j\}$ then $J_{n_k+1,k}, ..., J_{L,k} = 0$. We

shall let the *reduced data* be $d_r = \{(i, w_{ij}, w_i+), \underline{\mathcal{J}} : i, j \in s_k,\ k = 1, 2, ..., K\}$ where $\underline{\mathcal{J}} = (\sum_{k=1}^{K} J_{1,k}, \sum_{k=1}^{K} J_{2,k}, ..., \sum_{k=1}^{K} J_{L,k}) = (\mathcal{J}_1, \mathcal{J}_2, ..., \mathcal{J}_L)$.

We shall define $\underline{\theta} = (N, \underline{w}_N, \underline{w}_N^+)$ to be the parameter of interest where $\underline{w}_N$ is the adjacency matrix (of size $N \times N$) of the population graph, $\underline{w}_N^+$ is a vector (of length N) which displays the out-degree of the members of the population, and N is the population size. We will make the definition that $\underline{\theta}$ is *consistent* with the reduced data d_r if there exists a subset $d' \subseteq N$ such that $\underline{w}_{d'} \equiv \underline{w}_{d_r}$ and $\underline{w}_{d'}^+ \equiv \underline{w}_{d_r}^+$. The set of all $\underline{\theta}$ that are consistent with the reduced data d_r shall be labeled as Θ_{d_r}. Notice that since the population size is unknown, N (and hence the corresponding values $\underline{w}_N$ and $\underline{w}_N^+$) is permitted to range over all values in the natural number set $\mathbb{N}$.

Claim: d_r is a sufficient statistic.

Proof:

First we will show sufficiency.

For $k = 1, 2, ..., K$ and step $t_k = 1, 2, ..., n_k - n_{0k}$, let $H_{t_k+n_{0k}} = 1$ if $w_{a_{t_k}}^+ = 0$ and 0 otherwise (that is, a random jump is forced at this step in the sample selection procedure as there are no links to trace at selection step t_k in sample k out of the current active set a_{t_k}). We will also let q_{t_k} be the probability of adaptively adding that unit which was selected at time t_k to sample k.

Now, let d_0 be any data point where $P(D_0 = d_0) > 0$. Then,

$$
\begin{aligned}
P_{\underline{\theta}}(D_0 = d_0) &= P(D_0 = d_0 | N, \underline{w}_N, \underline{w}_N^+) I[\underline{\theta} \epsilon \Theta_{d_r}] \\
&= P(D_0 = d_0 | N, \underline{w}_{d_r}, \underline{w}_{d_r}^+) I[\underline{\theta} \epsilon \Theta_{d_r}] \\
&= \prod_{k=1}^{K} \left[\frac{1}{\binom{N}{n_{0k}}} \prod_{\substack{t_k=1: \\ J_{t_k+n_{0k},k}=0}}^{n_k-n_{0k}} dq_{t_k}^{s_k} \prod_{\substack{t_k=1: \\ J_{t_k+n_{0k},k}=1, \\ H_{t_k+n_{0k},k}=0}}^{n_k-n_{0k}} (1-d) \frac{1}{N-(n_{0k}+t_k-1)} \times \right. \\
&\quad \left. \prod_{\substack{t_k=1: \\ J_{t_k+n_{0k},k}=1, \\ H_{t_k+n_{0k},k}=1}}^{n_k-n_{0k}} \frac{1}{N-(n_{0k}+t_k-1)} \right] I[\underline{\theta} \epsilon \Theta_{d_r}] \\
&= \prod_{k=1}^{K} \left[\left(\prod_{\substack{t_k=1: \\ J_{t_k+n_{0k},k}=0}}^{n_k-n_{0k}} dq_{t_k}^{s_k} \right) (1-d)^{\sum_{t_k=1}^{n_k-n_{0k}} J_{t_k+n_{0k},k}(1-H_{t_k+n_{0k},k})} \right] \times \\
&\quad \prod_{k=1}^{K} \frac{1}{\binom{N}{n_{0k}}} \prod_{i=1}^{L} \left(\frac{1}{N-(i-1)} \right)^{\mathcal{J}_i} I[\underline{\theta} \epsilon \Theta_{d_r}] \\
&= h(d_0) \cdot g(d_r, \underline{\theta}) \qquad \text{(B.1)}
\end{aligned}
$$

Therefore, by the Fisher-Neyman Factorization Theorem, d_r is a sufficient statistic. □