University of Massachusetts Amherst

ScholarWorks@UMass Amherst

Doctoral Dissertations Dissertations and Theses

March 2017

Inference from network data in hard-to-reach populations

Isabelle Beaudry
University of Massachusetts - Amherst

INFERENCE FROM NETWORK DATA IN HARD-TO-REACH POPULATIONS

A Dissertation Presented

by

ISABELLE BEAUDRY

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2017

Mathematics and Statistics

INFERENCE FROM NETWORK DATA IN HARD-TO-REACH POPULATIONS

A Dissertation Presented

by

ISABELLE BEAUDRY

Approved as to style and content by:

Krista Gile, Chair

Michael Lavine, Member

John Staudenmayer, Member

Leontine Alkema, Member

Keith Sabin, Member

Farshid Hajir, Department Head
Mathematics and Statistics

DEDICATION

To my parents

ACKNOWLEDGMENTS

I would like to thank my committee members for their numerous questions and comments which have sparked many research questions and interests. Special thanks to Krista Gile who has been an incredible mentor, exceptional academic advisor and who has been very supportive throughout my PhD studies. I am also grateful to all professors who have contributed to my education. Finally, I am also very grateful to be surrounded by friends, family and partner, Javier Campos, who never ceased to encourage me to pursue this graduate degree.

ABSTRACT

INFERENCE FROM NETWORK DATA IN HARD-TO-REACH POPULATIONS

FEBRUARY 2017

ISABELLE BEAUDRY

B.Sc., UNIVERSITÉ LAVAL

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Krista Gile

The objective of this thesis is to develop methods to make inference about the prevalence of an outcome of interest in hard-to-reach populations. The proposed methods address issues specific to the survey strategies employed to access those populations.

One of the common sampling methodology used in this context is respondent-driven sampling (RDS). Under RDS, the network connecting members of the target population is used to uncover the hidden members. Specialized techniques are then used to make inference from the data collected in this fashion. Our first objective is to correct traditional RDS prevalence estimators and their associated uncertainty estimators for misclassification of the outcome variable.

RDS also has the unusual characteristic that the participants are driving the sampling process by recruiting members into the survey. Since the researchers forfeit their control over the sampling process, the estimators are therefore susceptible to

a great extent to participants' behavioral induced biases. Our second objective is therefore to provide a mathematical parametrization for a behavior referred to as differential recruitment and subsequently adjust the inference for potential induced bias.

Finally, a common issue encountered in the application motivating this thesis, that is, HIV prevalence estimation, is the derivation of a national prevalence estimate. Data are often collected at different study sites within a given country. Public health officials however commonly report national prevalence. Therefore, our last objective consists of using Bayesian hierarchical models to derive a national prevalence estimator from regional data.

TABLE OF CONTENTS

Page

CHAPTER

APPENDICES

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1
INTRODUCTION

Public health organizations, such as the Center for Disease Control and Prevention (CDC) and UNAIDS closely monitor the progression of HIV worldwide. HIV surveillance allows, among other things, to efficiently allocate resources to limit the number of new infections and provide care and treatment for people living with HIV.

In concentrated epidemics, HIV disproportionately affects sub-groups of the general population such as people who inject drugs, men who have sex with men and sex workers. Belonging to those key populations is frequently associated with a social stigma. Being a member of those key populations is even considered illegal in some geographies. Therefore, a sampling frame rarely exists for those populations, making the sampling particularly challenging and many traditional sampling methods prohibitively expensive.

The methods discussed in this thesis are developed to address some issues related to the inference about the prevalence of an outcome variable, such as HIV, in the specific context of hard-to-reach populations. In particular, the suggested methods take into account some of the sampling strategies to collect information about those populations.

One of the sampling strategy that may be employed for populations well connected by a social network is link-tracing network sampling. In idealized cases [Goodman, 1961, Handcock and Gile, 2011], the resulting sample is a probability sample, however practical constraints typically interfere, resulting in convenience sampling. For example, an initial probabilistic sample is impractical in most settings [Trow, 1957, Biernacki and Waldorf, 1981] and therefore, a link-tracing or snowball sample collected from that initial convenience sample results in a non-probability sample of the target population [Trow, 1957, Handcock and Gile, 2011].

Respondent-Driven-Sampling (RDS) however, is a specialized form of link-tracing sampling design introduced by Heckathorn [1997] as a practical sampling method to be approximated as a probability sample. Since this sampling process protects participants' confidentiality, it has been widely adopted by public health organizations [Johnston et al., 2008].

Inference from RDS data typically assumes that the outcome variable is measured accurately. The first methodological chapter in this thesis discusses the effect of misclassification on the binary outcome variable of interest. Also, two methods to correct the prevalence estimation are discussed, that is, the matrix method [Barron, 1977] and SIMEX-MC Kuchenhoff et al. [2006], as well as the circumstances under which they may be used with the traditional RDS estimators. Uncertainty estimators are also derived to account for misclassification.

As described in Chapter 4, participants in RDS studies are responsible for selecting most of the survey participants. Researchers conducting RDS surveys have little to no control over the sampling process. Most RDS prevalence estimators however rely on the strong assumption that participants recruit completely at random among their contacts who are members of the target population. In the second methodological chapter of this thesis, we propose extensions to RDS prevalence estimators to correct for bias induced by various recruitment behaviors. A design-based and a model-based approach are proposed to reduce this type of bias.

The third methodological question investigated in this thesis relates to the derivation of a national prevalence estimate and is not specific to RDS data. In many cases, public health practitioners survey key populations at different study cites within a country. The obtained multiple prevalence estimates must subsequently be combined into a national estimate for reporting purposes. The contribution of our work to that research question is to propose a national prevalence estimator based on Bayesian hierarchical models.

In summary, the thesis is organized as follows. Chapter 2 presents the RDS sampling methodology and includes a literature review of the traditional RDS prevalence estimators. It is followed by Chapter 3 which discusses methods to correct RDS prevalence estimators for misclassification on the outcome variable. Proposed methodologies to adjust inference for participants' non random recruitment behaviors are then discussed in Chapter 4. Finally, a Bayesian hierarchical model combining regional prevalence estimates into a national estimate is described in Chapter 5.

CHAPTER 2

RESPONDENT-DRIVEN SAMPLING

Respondent-Driven-Sampling (RDS) [Heckathorn, 1997], is a network based sampling procedure designed to sample hard-to-reach populations when members of such populations are well socially connected. We begin this chapter by describing the RDS methodology. Then, we briefly introduce in Section 2.2 the notation used throughout this thesis. It is followed in Section 2.3 by a description of common simplifying models to represent RDS for inference purposes. Finally, we present a number of RDS prevalence estimators and their associated uncertainty estimators in Section 2.4.

2.1 Sampling Methodology

This section outlines the procedure to collect a respondent-driven sample. Assuming that the studied human population is connected by a social network, the objective of RDS is to leverage this relational structure to reach members who would not otherwise be accessible through a conventional sampling framework. Typically, researchers select the initial participants, the *seeds*, through convenience sampling. Once the seeds are enrolled in the survey, they receive a small number of uniquely identified coupons to distribute among their social ties in the target population. Individuals receiving coupons who return to the survey center are enrolled in the study. The individuals who were recruited from the seeds are said to be part of the first wave of recruitment. The subsequent waves occur in the same fashion, that is, participants in each wave are given the same number of coupons to distribute to their contacts until a desired sample size is achieved. By restricting the number of referrals per participant, a given sample size forces samples many steps away from the initial sample, reducing the dependence of the final sample on the initial convenience sample. The respondents commonly receive a small financial incentive both for their participation and for each successful recruitment. Finally, the coupon mechanism helps diminish serious confidentiality issues related to the recruitment of stigmatized populations, contributing to its wide adoption by public health organizations.

All RDS participants are asked to report on their number of contacts in the target population, their *self-reported degree*. Similarly to other link-tracing samples, RDS allows the recruitment of individuals otherwise unknown to researchers.

2.2 Notation

Suppose a hard-to-reach human population consists of N individuals, also called the *nodes* of the network. We assign the labels $1, 2, ..., N$ to the nodes. This population of N nodes is connected by social ties which may be represented by a sociomatrix $Y \in \{0,1\}^{N \times N}$. Entries in the sociomatrix, y_{ij}, are equal to 1 if nodes i and j are connected or 0 otherwise. Ties are assumed to be reciprocated such that $y_{ij} = y_{ji} \, \forall \, i, j \in \{1, 2, ..., N\}$.

The outcome of interest is represented by a vector $\mathbf{z} \in \{0,1\}^N$. We refer to the outcome of interest as the "infection status" since RDS studies have found many applications in public health settings, such as HIV/AIDS surveillance of at-risk populations [Johnston et al., 2008, Malekinejad et al., 2008, Montealegre et al., 2013]. However, $\mathbf{z}$ may be interpreted as any binary vector of length N. The $i-th$ entry of this vector is such that:

$$z_i = \begin{cases} 1 & \text{person } i \text{ is infected} \\ 0 & \text{otherwise.} \end{cases} \qquad i \in \{1, 2, ..., N\}$$

Note that $\mathbf{z}$ represents the true infection status, typically assumed to be observable. We introduce notation for the misclassification of $\mathbf{z}$ in Chapter 3. Finally, we define the set of infected individuals and uninfected individuals as $\mathcal{Z}^1 = \{i : z_i = 1\}$ and $\mathcal{Z}^0 = \{i : z_i = 0\}$, respectively.

The RDS estimators described in the remainder of this section estimate the prevalence of the infection status in the target population. The actual population prevalence is denoted μ. RDS estimates are based on a sample of n individuals for whom the self-reported degree is observed and is assumed to be equal to the true degree $d_i = \sum_{j=1}^{N} y_{ij}$. The vector $\mathbf{S} \in \{0,1\}^N$ indicates whether the nodes were sampled such that:

$$S_i = \begin{cases} 1 & \text{person } i \text{ has been sampled} \\ 0 & \text{otherwise} \end{cases} \qquad i \in \{1, 2, ..., N\}.$$

Similar to notation for infected individuals, we define the set of sampled nodes as $\mathcal{S}^1 = \{i : S_i = 1\}$.

2.3 Approximating RDS

Respondent-driven sampling is a complex sampling method and estimating the probability of sampling any given individuals from the target population is a challenging problem since a large portion of the network typically remains unobserved. RDS prevalence estimators commonly rely on simplifying models to approximate the RDS mechanism. In this Section, we describe two of these simplifications, that is, a discrete Markov chain on the network nodes and probability proportional to size without replacement sampling (PPSWOR) or equivalently successive sampling (SS) [Yates and Grundy, 1953].

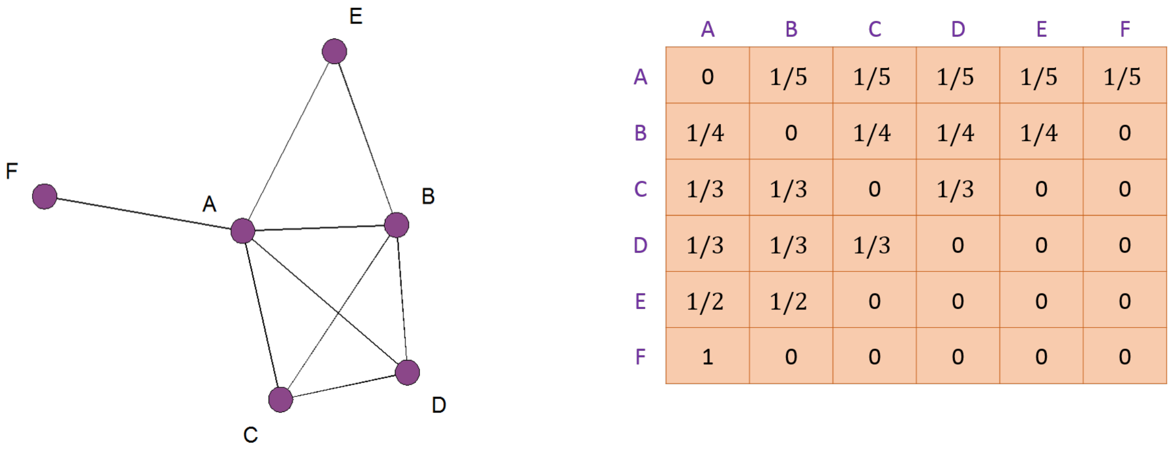
2.3.1 Discrete Markov Chains

A number of RDS prevalence estimators assume that RDS may be well approximated by a discrete Markov chain (MC) on the state space of the network nodes [Salganik and Heckathorn, 2004, Volz and Heckathorn, 2008, Lu, 2013]. Conceptually, the transition from one state (e.g. node i) to another state (e.g. node j) represents peer recruitment (e.g. i recruited j) as if nodes may only recruit one participant. Furthermore, even though in reality members of the target population may only participate once in RDS studies, this model allows for multiple participation. Participants are also assumed to recruit completely at random among all their contacts in the target population, that is, among their *alters*. In addition, these estimators typically assume that the recruitment process occurs on a single component network solely constituted of reciprocated ties. In summary, RDS is represented by a random walk (RW) on a the nodes of a fully connected undirected network.

The probability of node j entering the survey at step t under this RW model strictly depends on the recruiter i at step $t-1$. Let P denote the transition probability matrix of a RW and p_{ij} the entry on the i-th row and j-th column. Since node i is constrained to recruit among its alters, the probability that node j is selected at step t conditional on recruiter i is equal to $p_{ij} = y_{ij}/d_i$ for all i and $j \in \{1, 2, ..., N\}$.

Figure 2.1 illustrates a simple example of a transition probability matrix characterizing the RW on the nodes of the undirected network showed in panel (2.1a). The probability in any given cell p_{ij} is the conditional probability of transitioning to node j (column) given chain's current state i (row).

Under the presumed network structure, the MC is irreducible. Furthermore, the with-replacement assumption effectively leads to the positive recurrence of all states of the MC. The combination of these properties results in the existence of a unique stationary distribution denoted π. Under random recruitment, it may be proven that the stationary distribution of this RW on the network node is as stated in Result 2.1.



	A	B	C	D	E	F
A	0	1/5	1/5	1/5	1/5	1/5
B	1/4	0	1/4	1/4	1/4	0
C	1/3	1/3	0	1/3	0	0
D	1/3	1/3	1/3	0	0	0
E	1/2	1/2	0	0	0	0
F	1	0	0	0	0	0

(a) Small Network **(b)** Transition matrix P

Figure 2.1: Transition probability matrix (b) for a random walk on the nodes of the network depicted in (a) under a random recruitment regime.

Result 2.1. *Let RW_t denotes the state at step t of a MC on the nodes of a fully connected undirected network without self ties. Assume that this MC has the following transition probabilities: $p_{ij} = \frac{y_{ij}}{d_i}$. Then the stationary distribution of this random walk is such that:*

$$\pi_i = \frac{d_i}{\sum_{i=1}^{N} d_i} \propto d_i \quad \textit{for} \; \forall \; i \in \{1, 2, ..., N\}. \tag{2.1}$$

The resulting stationary distribution may be interpreted as the proportion of time the process visits each state in the long run. The RDS estimators developed under this framework assume the sampling starts at stationarity. This implies that the seed is selected with a probability proportional to its degree. If this holds, all participants' sampling probabilities are proportional to their degree. In other words, the more people someone is connected to, the greater the chances this person is recruited and participates in the study.

2.3.2 Successive Sampling

The RW approximation to RDS provides a convenient model to make inference with RDS data. However it over simplifies many features of the RDS process. Recent work has relaxed some of the RW assumptions. The Successive Sampling (SS) representation of RDS proposed by Gile [2011], for example, captures the without replacement nature of the RDS sampling.

Under an SS or PPSWOR process, units in a population are sampled without replacement and with sampling probability proportional to its unit size from among the remaining unsampled units. Let $u = \{u_1, u_2, ..., u_N\}$ denote the sizes of all units in the population and let $G = \{G_1, G_2, ..., G_n\}$ denote the order in which the units are sampled. The transition probabilities of such an SS process are as follows:

$$\begin{aligned} &P(G_i = i | G_1, G_2, ..., G_{i-1} = (g_1, g_2, ..., g_{i-1}), U = u) \\ &\qquad = \begin{cases} \dfrac{u_i}{\sum_{j \notin \{g_1, g_2, ..., g_{i-1}\}}^{N} u_j} & i \notin \{g_1, g_2, ..., g_{i-1}\} \\ 0 & i \in \{g_1, g_2, ..., g_{i-1}\} \end{cases} \end{aligned} \tag{2.2}$$

Determining the unit sizes of all members in the target population is therefore central to the parametrization of an SS process. For the SS approximation to RDS, Gile [2011] argues that the unit sizes are equal to the individuals' degree. This finding assumes that the SS takes place over the nodes of all networks generated from a configuration network model [Molloy and Reed, 1995] with a fixed degree distribution and that participants recruit at random. These unit sizes are then used in a algorithm which jointly estimates the sampling probabilities and the degree distribution.

2.4 Existing Methodology for Respondent-Driven Sampling

The random walk and successive sampling approximation to RDS are used to derive the sampling probabilities for each individuals participating in the RDS survey.

Those probabilities are then used to make inference about the prevalence of an outcome variable such as the prevalence of HIV in the target population. In this section, we describe some of these RDS prevalence estimators and their associated variance estimators.

2.4.1 Hájek Estimator

A number of design-based estimators have been developed for RDS data to estimate the prevalence of an outcome variable, $\mu = \frac{\sum_{i=1}^{N} z_i}{N}$. Several of those estimators are closely related to the Hájek estimator:

$$\hat{\mu}^{Hájek} = \frac{\sum_{i=1}^{N} \frac{\mathbf{S}_i \mathbf{z}_i}{\pi_i}}{\sum_{i=1}^{N} \frac{\mathbf{S}_i}{\pi_i}}, \tag{2.3}$$

where π_i is the sampling probability for individual i.

Due to the complexity of RDS, the sampling probabilities are unknown. A number of methodologies have been proposed to estimate them. We refer to an estimator of the Hájek form but based on estimated sampling probability as an estimator of the Hájek style. Such an estimator is of the form:

$$\tilde{\mu}^{Hájek} = \frac{\sum_{i=1}^{N} \frac{\mathbf{S}_i \mathbf{z}_i}{\hat{\pi}_i}}{\sum_{i=1}^{N} \frac{\mathbf{S}_i}{\hat{\pi}_i}}. \tag{2.4}$$

The sample mean, the Volz-Heckathorn estimator [Volz and Heckathorn, 2008] and the Successive Sampling estimator [Gile, 2011] all are of the Hájek style and rely on distinct methodologies to estimate the sampling probabilities. These methodologies are described in Section 2.4.1.1 - 2.4.1.3. Next, in Section 2.4.2, we present the estimator introduced by Salganik and Heckathorn [2004], which under certain conditions, may also be formulated as an estimator of the Hájek style.

2.4.1.1 Sample Mean

The naive approach to making inference with RDS data is to consider the sample mean as an estimator for the total population mean. This implicitly assumes a common sampling probability for all members in the target population. However, this assumption almost never holds in practice in the context of RDS. Therefore, the sample mean estimator is not expected to perform well in most circumstances. The estimator shown in equation (2.4) with constant sampling probabilities results in the sample mean:

$$\hat{\mu}_{mean} = \frac{\sum_{i=1}^{N} S_i z_i}{\sum_{i=1}^{N} S_i}. \tag{2.5}$$

2.4.1.2 Volz-Heckathorn Estimator

The Volz and Heckathorn [2008] estimator is an estimator of the Hájek style which is based on the RW approximation to RDS as described in Section 2.3.1. The authors therefore argue that the sampling probabilities are proportional to the nodal degrees, d_i and the resulting prevalence estimator is as follows:

$$\hat{\mu}_{VH} = \frac{\sum_{i=1}^{N} S_i \frac{z_i}{d_i}}{\sum_{i=1}^{N} S_i \frac{1}{d_i}}. \tag{2.6}$$

2.4.1.3 Successive Sampling Estimator

The Volz-Heckathorn estimator relies on the strong assumption that the sampling is performed with replacement. However, in practice this assumption is violated as members of the target population are only allowed to participate once in the survey. The contribution of the Successive Sampling estimator [Gile, 2011] is to address this issue. The sampling procedure is instead approximated by a SS process. The resulting $\hat{\mu}_{SS}$ outperforms $\hat{\mu}_{VH}$ for large sampling fractions.

This estimator uses a successive sampling procedure [Yates and Grundy, 1953] with unit size equal to degree to estimate the sampling probabilities jointly with the population degree distribution. The author suggests an algorithm iterating between the estimation of the population degree distribution and the inclusion probabilities. The obtained estimated sampling probabilities are then used in the expression for estimators of Hájek style (2.4).

2.4.2 Salganik-Heckathorn

2.4.2.1 Salganik-Heckathorn Estimator

The estimator introduced by Salganik and Heckathorn [2004] relies on the argument that if all ties are reciprocated, then the total number of ties from infected to uninfected individuals equals the total number of ties from uninfected to infected individuals. This quantity is referred to as the number of cross ties and is denoted $T_{(k,1-k)} = \sum_{i=1}^{N} \sum_{j=1}^{N} z_i(1-z_j)y_{ij}$ for $k \in \{0,1\}$. Multiplying by terms which conveniently cancel out leads to this alternate expression for the number of cross-ties:

$$T_{(k,1-k)} = \ p_{(k,1-k)} \cdot \ \bar{D}_k \cdot \ (\mu k + (1-\mu)(1-k)) \cdot \ N, \tag{2.7}$$

where:

1. $k \in \{0,1\}$,
2. $p_{(k,1-k)} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} z_i(1-z_j)y_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} (kz_i+(1-k)(1-z_i))y_{ij}}$, i.e. the proportion of cross-ties for nodes belonging to Z^k.
3. $\bar{D}_k = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (kz_i+(1-k)(1-z_i))y_{ij}}{|\mathcal{Z}^k|}$, the average degree of nodes belonging to Z^k.

Using the argument that all ties are reciprocated, and thus $T_{(0,1)}$ equals $T_{(1,0)}$, and equation (2.7) the following expression for the actual population proportion is obtained:

$$\mu = \frac{p_{(0,1)}\bar{D}_0}{p_{(1,0)}\bar{D}_1 + p_{(0,1)}\bar{D}_0}. \tag{2.8}$$

The quantities in equation (2.8) are not directly observable from a sample. However, the authors argue that they may be estimated from the collected data. The methodology they proposed assumes that RDS may be reasonably well represented by a with-replacement random walk on the space of network nodes at stationarity. Because of the implied unform distribution of edge sampling, the cross-ties proportions, $p_{(k,1-k)}$, may be estimated from the observed recruitment patterns, such that:

$$\hat{p}_{(k,1-k)} = \frac{r_{(k,1-k)}}{r_{(k,1-k)} + r_{(k,k)}}, \tag{2.9}$$

where $r_{(k,1-k)}$ and $r_{(k,k)}$ are the number of recruitment from nodes belonging to $\{\mathcal{Z}^k, \mathcal{S}^1\}$ to nodes belonging to $\{\mathcal{Z}^{1-k}, \mathcal{S}^1\}$ and $\{\mathcal{Z}^k, \mathcal{S}^1\}$, respectively, for $k \in \{0,1\}$. The random walk assumption also leads to the average degrees, $\bar{D}_0$ and $\bar{D}_1$, to be estimated as follows:

$$\hat{\bar{D}}_k = \frac{n_k}{\sum_{i=1}^N S_i \frac{(kz_i + (1-k)(1-z_i))}{d_i}}, \tag{2.10}$$

where $n_k = |\{\mathcal{Z}^k, \mathcal{S}^1\}|$. The following expression for the estimator $\hat{\mu}_{SH}$ is therefore derived by substituting $p_{(k,1-k)}$'s by $\hat{p}_{(k,1-k)}$'s and $\bar{D}_k$'s by $\hat{\bar{D}}_k$'s in expression (2.8):

$$\hat{\mu}_{SH} = \frac{\hat{p}_{(0,1)} \hat{\bar{D}}_0}{\hat{p}_{(1,0)} \hat{\bar{D}}_1 + \hat{p}_{(0,1)} \hat{\bar{D}}_0}, \tag{2.11}$$

which may be expressed as:

$$\hat{\mu}_{SH} = \frac{\sum_{i=1}^N S_i \frac{z_i}{d_i}}{\sum_{i=1}^N S_i \frac{z_i}{d_i} + c \ \sum_{i=1}^N S_i \frac{(1-z_i)}{d_i}}, \text{ where } c = \left(\frac{n_1}{n_0} \frac{r_{(0,0)} + r_{(0,1)}}{r_{(1,1)} + r_{(1,0)}} \frac{r_{(1,0)}}{r_{(0,1)}} \right). \tag{2.12}$$

2.4.2.2 Relation Between $\hat{\mu}_{SH}$ and $\hat{\mu}_{VH}$

In this section, we establish a relation between $\hat{\mu}_{SH}$ and $\hat{\mu}_{VH}$.

The Salganik-Heckathorn estimator may be formulated as a function of the Volz-Heckathorn estimator:

$$\hat{\mu}_{SH} = \frac{\hat{\mu}_{VH}}{\hat{\mu}_{VH} + c\ (1 - \hat{\mu}_{VH})}. \tag{2.13}$$

The value c in the above relation has a number of important implications. First, we observe that for $c = 1$, $\hat{\mu}_{SH} = \hat{\mu}_{VH}$, or equivalently, the Salganik-Heckathorn estimator is of the Hájek style. Secondly, c approaches 1 under the assumption that the sampling may be approximated by a Markov Chain at stationarity. However, c may significantly differ from 1 in RDS data.

2.4.2.3 SH Estimator With Ego-Network Data

The extension of the SH estimator proposed by Lu [2013] provides an improved estimator for the proportion of cross-ties, that is, $\hat{p}_{(k,1-k)}$. In lieu of estimating this proportion with the observed recruitment patterns as shown in equation (2.9), Lu proposed to estimate this proportion with a generalized Hansen and Hurwitz [1943] estimator. This procedure however requires the collection of ego-network composition data. In the context of this estimator, the ego-network composition data refers to the number of ties in the target population to infected individuals (d_{i1}). More specifically, the introduced estimator is as follows:

$$\hat{p}^{ego}_{(k,1-k)} = \frac{\sum_{i=1}^{N} S_i(kz_i + (1-k)(1-z_i))\dfrac{d_{i1-k}}{d_i}}{\sum_{i=1}^{N} S_i(kz_i + (1-k)(1-z_i)}, \tag{2.14}$$

where $d_{ik} = \sum_{j \neq i} y_{ij}(kz_j + (1-k)(1-z_j))$ and $k \in \{0,1\}$. For instance, the estimated proportion of cross-ties from a non-infected individuals is:

$$\hat{p}^{ego}_{(0,1)} = \frac{\sum_{i=1}^{N} S_i(1-z_i)\dfrac{d_{i1}}{d_i}}{\sum_{i=1}^{N} S_i(1-z_i)} = \frac{\sum_{i=1}^{N} S_i(1-z_i)\dfrac{d_{i1}}{d_i}}{\sum_{i=1}^{N} S_i(1-z_i)\dfrac{d_i}{d_i}}. \tag{2.15}$$

In that expression, the numerator and the denominator represent estimates, up to the same constant of proportionality, of the total degree to infected individuals d_{i1} and total degree to the entire population d_i, respectively, for uninfected individuals.

The form of the estimator is identical to the earlier version of the SH estimator. However $\hat{p}_{(k,1-k)}$ is substituted by $\hat{p}^{ego}_{(k,1-k)}$ in the derivation of equation (2.12). Importantly, both estimators assume that RDS may be approximated by a random walk at stationarity on the space of the network nodes.

Similarly to $\hat{\mu}_{SH}$, $\hat{\mu}^{ego}_{SH}$ may be expressed as a function of the $\hat{\mu}_{VH}$ such that:

$$\hat{\mu}^{ego}_{SH} = \frac{\hat{\mu}_{VH}}{\hat{\mu}_{VH} + c^{ego}\left(1 - \hat{\mu}_{VH}\right)}, \quad \text{where} \quad c^{ego} = \frac{n_1}{n_0}\left(\frac{\sum_{i=1}^{N} S_i z_i d_{i0}/d_i}{\sum_{i=1}^{N} S_i(1-z_i)d_{i1}/d_i}\right). \tag{2.16}$$

2.4.3 Variance Estimation

2.4.3.1 Salganik Bootstrap

In this section, we describe the bootstrap procedure proposed by Salganik [2006] to estimate the variability of RDS estimators. Since RDS does not produce a classic probability sample, Salganik introduced a non-parametric bootstrap that would capture the recruitment dependencies between infected and non-infected nodes. The algorithm consists of the following steps:

1. **Resampling** A new RDS sample is drawn from the observed data:

   (a) A first node is selected at random among all nodes in the observed RDS sample.

   (b) Two vectors are constructed: $\mathbf{w}^0$ and $\mathbf{w}^1 \in \{0,1\}^n$. The i^{th} entry in each vector indicates whether node i was recruited by a non-infected or by an infected node, respectively.

   (c) Nodes are subsequently resampled node-by-node by sampling at random with replacement with weights proportional to $\mathbf{w}^0$ if the infection status of the recruiting node is non-infected or proportional to $\mathbf{w}^1$ otherwise. The resampling is performed with replacement.

   (d) The process stops when n nodes are recruited.

2. **RDS estimates**: A prevalence estimate is calculated based on the resampled data from step 1.

3. **Confidence Interval for μ**: Steps 1 and 2 are repeated a large number of times. For the purpose of this paper, the variability of the resulting resampled estimates is used to construct t-intervals.

2.4.3.2 SH-ego's Bootstrap

The variance estimator for $\hat{\mu}_{SH}^{ego}$ proposed by Lu [2013], that is, the SH-ego Bootstrap, extends the Salganik Bootstrap procedure described above in two ways. First, the author modifies the sampling weights to sample from the revised estimated stationary distribution of the random walk. For instance, after selecting the first node i_1 at random, the random walk transitions to a node of status $1-k$ with probability $\hat{p}_{(k_1,1-k)}^{ego}$ or to a node of status k with probability $1-\hat{p}_{(k_1,1-k)}^{ego}$, where $k_1 = z_{i_1}$. Subsequent re-sampled nodes are selected in a similar manner, where transition probabilities are sequentially updated to appropriately reflect the infection status of the recruiting node. The second extension simply substitutes the prevalence estimator $\hat{\mu}_{SH}$ in the **RDS estimates** step by $\hat{\mu}_{SH}^{ego}$.

2.4.3.3 Successive Sampling Bootstrap

The Successive Sampling Bootstrap (SS Bootstrap) is a procedure that was proposed by Gile [2011] to estimate the variance of $\hat{\mu}_{SS}$, described in Section 2.4.1.3.

The SS Bootstrap procedure is based on a sampling model similar to the one assumed for the Successive Sampling estimator ($\hat{\mu}_{SS}$), but it allows for additional RDS features, such as multiple seeds and a fixed number of recruits per participants. It is also formulated to capture network homophily on the infection status.

In order to simulate sampling under Successive Sampling design [Yates and Grundy, 1953], the unit size of each element in the population is required. Therefore, each SS

Bootstrap replicate is initiated by the simulation of a unit size distribution, i.e. the degree distribution, of a population of N individuals. This distribution is also divided between the infection status classes, i.e. infected or uninfected, so an RDS estimate may be computed.

The author argues however that drawing a successive sample based on these units would likely result in anti-conservative estimates of the variance. Consequently, she extended the proposed methodology to account for network homophily on the infection status. The homophily is represented by an estimated mixing matrix partitioned relative to the infection status, which is estimated based on the observed recruitment patterns.

The resampling process stops when n nodes are sampled. An RDS prevalence estimate based on the bootstrap sample is calculated. This process is repeated a large number of times. The SS Bootstrap variance estimator is the sample variance of the RDS estimates from the replicates.

CHAPTER 3

MISCLASSIFICATION ON NODAL ATTRIBUTE

RDS is a novel sampling mechanism and inference from RDS data relies on a number of strong assumptions regarding the network properties and the sampling process. Due to the great interest in this sampling methodology, the research community has made significant progress in understanding some of the critical RDS assumptions.

Various concerns have been raised regarding the participants' self-reported degree accuracy since current methodology heavily relies on this metric. For instance, researchers have studied whether relationships may safely be assumed to be reciprocated [Mccreesh et al., 2012, Rudolph et al., 2013] and the potential sensitivity of the estimators to directed ties Lu et al. [2012]. Lu et al. [2013] proposed an extension of the Salganik and Heckathorn [2004] which accounts for directed ties. Another assumption related to the degrees is that participants are commonly presumed to report their degree accurately. Several studies have recently assessed the impacts of inaccurately self-reported degrees on RDS estimators [Lu et al., 2012, Rudolph et al., 2013], finding that RDS estimators are robust to many forms of mis-reporting of degrees, but subject to bias in special circumstances such as when mis-reporting patterns are related to the outcome of interest or when respondents report degrees rounded to multiples of five, ten and one hundred [Mills et al., 2014].

To date, however, the assumption that the outcome of interest is measured accurately has not been discussed in the context of RDS data. In this chapter, we show that neglecting such misclassification may lead to biased estimates. This may be a source of concerns for many RDS studies. For instance, dozens of RDS studies have been implemented to estimate HIV prevalence among key populations [Johnston et al., 2008, Malekinejad et al., 2008, Montealegre et al., 2013]. Accuracy of HIV diagnosis is considered crucial in that erroneous results may lead to severe repercussions for misdiagnosed individuals [Smith et al., 2008] and to serious consequences for epidemic prevention [Marks et al., 2005]. As pointed out by the World Health Organization in their recent consolidated guidelines on HIV testing services [World Health Organization, 2015], HIV misdiagnoses have occurred in numerous settings nonetheless.

The main contribution of this chapter is to extend two existing methods for inference in the presence of misclassification to the dependent-sampling weighted-data case of RDS. The first method is an analytical adjustment, also referred to as the matrix method [Barron, 1977], to correct a population proportion.

Despite the fact that it is not possible to assume independence and identical distribution for the sampled units in RDS studies, we demonstrate that this correction is applicable to RDS estimators of the Hájek style such as the sample mean, the Volz-Heckathorn estimator [Volz and Heckathorn, 2008] and the Successive-Sampling estimator [Gile, 2011]. We also introduce a novel formulation for the Salganik-Heckathorn estimator [Salganik and Heckathorn, 2004]. This formulation elucidates the reasons for the suboptimal performance of the analytical adjustment with this estimator. We then discuss the Simulation Extrapolation Misclassification (SIMEX MC) [Kuchenhoff et al., 2006] approach which does not rely on the form of the estimator, but instead requires that the estimator may be expressed as a function of the misclassification error present in the data. Both methods assume a classical misclassification model with known error rates. As the error rates may not be known in practice but instead estimated from external validation studies for instance, we assess the effect of uncertain error rates on the correction methods' ability to reduce misclassification bias in our simulation study. We also extend two RDS Bootstrap uncertainty estimation procedures to account for misclassification.

We have applied the correction methods to RDS surveys conducted in India among people who inject drugs and men who have sex with men. In those studies, the participants were asked to answer questions regarding their knowledge of their HIV infection status. In addition, on-site biological testing was performed to determine their actual HIV infection status. The self-reported data contained substantial false negative rates as participants were largely unaware of their infection status. Their lack of knowledge of their infection status may occur for a number of reasons, such as the fact that they may not have been tested recently. In our application, we address the challenge of inference based on only the self-reported HIV status and known error rates. We compare our results to analysis based on biological test data. We find that inference from self-reported data may be significantly improved when applying the correction methods discussed in this paper.

In Section 3.1 we describe the two correction methods as well as our proposed methodology to estimate the variance of the corrected estimators. In Section 4.5, we present a simulation study illustrating the performance of the proposed methods. Section 3.3 discusses the results from the RDS application in India. Finally, in Section 3.4, we present a discussion of the proposed methods.

3.1 Methods to Correct For Misclassification

In many contexts, it is not possible to directly observe the outcome variable z_i. For example, the medical procedure to determine the infection status of an individual may not be perfectly accurate. Failure to account for misclassification may lead to biased estimates. In this section, we describe two methods to adjust RDS estimators for bias resulting from misclassification on a binary nodal attribute. We first introduce an analytical adjustment for estimators of the Hájek style. Then, we describe the

Simulation-Extrapolation Misclassification algorithm, as it may be applied to RDS prevalence estimators. Finally, we also propose methods to estimate the variance of the corrected estimators.

Before describing adjustments for measurement error, we need to introduce the error-prone binary random variable Z_i^* which takes value one if the observed infection status is positive and zero otherwise. The observed infection status may differ from the actual one. Our approach assumes that the risk of misdiagnosis occurs at known false positive and false negative rates, f^+ and f^-. These probabilities are the conditional probability of observing a positive or negative infection status when the actual status differs:

$$\begin{aligned} f^+ = & \quad P(Z_i^* = 1|z_i = 0) \\ f^- = & \quad P(Z_i^* = 0|z_i = 1). \end{aligned}$$

For simplicity, we refer to these rates as either misdiagnosis or testing error rates interchangeably. We recognize though that in practice more than one tests may be needed to obtain a diagnosis.

An estimate based on taking the observed data, z_i^* at face value, is referred to as the naive estimator. An expression for the naive estimator of Hájek style is given by:

$$\hat{\mu}^{naive} = \frac{\sum_{i=1}^{N} \frac{S_i z_i^*}{\hat{\pi}_i}}{\sum_{i=1}^{N} \frac{S_i}{\hat{\pi}_i}}, \tag{3.1}$$

the same form as equation (2.4) but with z_i replaced by the observed status, z_i^*.

3.1.1 Corrected Prevalence Estimators

3.1.1.1 Analytical Adjustment Estimator

The analytical adjustment, also referred to as the matrix method [Barron, 1977], discussed in this section applies to estimators of the Hájek style (2.4). We denote the resulting adjusted estimator $\hat{\mu}^{adj}$.

Equation (3.1) may be interpreted as a ratio of estimators. The numerator represents an estimate of the number of observed infected individuals, $\widehat{|\mathcal{Z}^{*1}|}$, where $\mathcal{Z}^{*1}$ is the set of individuals for whom a positive infection status would be observed. As for the denominator, it is an estimate of the total number of individuals in the population, $\hat{N}$. Therefore, equation (3.1) may alternatively be expressed as:

$$\hat{\mu}^{naive} = \frac{\widehat{|\mathcal{Z}^{*1}|}}{\hat{N}}.$$

Provided that the $\hat{\pi}_i$'s were true for all i, then $\hat{N}$ would be unbiased for N. Also, under the assumption that the misclassification is the result of a mechanism that is independent of the sampling procedure, we have that $E(\widehat{|\mathcal{Z}^{*1}|}) = N\big[\mu(1-f^-)+(1-\mu)f^+\big]$. Therefore, the ratio of estimators leads to an analytical form for a corrected estimator, $\hat{\mu}^{adj}$, which is approximately unbiased for μ in large samples:

$$\hat{\mu}^{adj} = \frac{\hat{\mu}^{naive} - f^+}{1 - f^+ - f^-}. \tag{3.2}$$

The analytical adjustment may result in a corrected estimate smaller than zero or greater than one. In such cases, the corrected estimate may be set to zero and one, respectively [Buonaccorsi, 2010].

Equation (3.2) provides a general way to correct estimators of the Hájek style for misclassification on the nodal attribute. The specific estimators are denoted $\hat{\mu}^{adj}_{mean}$, $\hat{\mu}^{adj}_{VH}$ and $\hat{\mu}^{adj}_{SS}$ depending on which of the naive estimator is used.

Under the Salganik-Heckathorn estimator assumptions, the term c in equation (2.12) approaches one for large sample size. This implies that $\hat{\mu}_{SH}$ may be close enough to the Hájek style for the analytical correction to apply. Similarly to the estimators of the Hájek style, we denote its corrected estimator $\hat{\mu}_{SH}^{adj}$. Our simulations show that for c significantly departing from 1 or for large discrepancies between c and c^* (i.e. the apparent c-factor based on the observed infection status), the effectiveness of the analytical adjustment in reducing the bias induced by misclassification is diminished.

3.1.1.2 SIMEX MC Estimators

In this section, we present an alternative method to correct for misclassification on the nodal attribute, the Simulation Extrapolation Misclassification (SIMEX MC) introduced by Kuchenhoff et al. [2006]. This method is a discrete version of a the Simulation Extrapolation (SIMEX) procedure [Cook and Stefanski, 1994]. Contrary to the analytical correction discussed in Section 3.1.1.1, this method does not make any assumption on the form of the estimator and therefore is particularly useful when it is not possible to derive a tractable expression for analytical adjustment. However, it requires that the estimator may be expressed as function of the error structure which is presumed to be known.

Cook and Stefanski [1994] describe their simulation-based method SIMEX which corrects estimators for measurement error generated from an additive measurement error model with known variance. The general idea is that if an estimator, say $\hat{\theta}$, may be expressed as a function of measurement error variance then it is possible to extrapolate such function to the theoretical level where such variance is zero.

To illustrate the SIMEX procedure, let's suppose that each observation, X_i^*, comes from an additive measurement error model such that $X_i^* = X_i + \xi_i$, where X_i is the true unobserved data and ξ_i is the random error with known variance σ_ξ^2. Also, we assume that X_i is independent of ξ_i for $i \in \{1 \ldots n\}$. Furthermore, let $g(\cdot)$ be the function mapping the estimator $\hat{\theta}$ to the measurement error variability. Their proposed two-stage algorithm consists of the following steps:

1. **Simulation**: In the simulation step, for each of K levels of perturbation, a large number of data sets, B, are simulated by perturbing the observed data according to a variant of the assumed error model. In our example, this translates into $X^*_{i,b} = X^*_i + \lambda_k \cdot \xi_{i,b}$, where λ_k is a multiplicative scalar that inflates the measurement error variability present in the simulated data and where $\xi_{i,b}$ has the same distribution as ξ_i. For each of the K levels of λ_k, B data sets are simulated which all contain the same measurement error variability. Estimates $\hat{\theta}_b(\lambda_k)$ are computed for each of the data sets at this variability level and are subsequently averaged to obtain $\hat{\theta}(\lambda_k)$.

2. **Extrapolation**: The outcome of the simulation step is a set of K $\hat{\theta}(\lambda_k)$. These $\hat{\theta}(\lambda_k)$ are estimates for the function $g(\cdot)$ at the measurement error variance level $(1 + \lambda_k)\sigma^2_\xi$. The purpose of the extrapolation is to use those points on the estimated curve to derive a function that can be evaluated at $\lambda_k = -1$, that is, the point where the estimate is based on data free of measurement error variability. The choice of the functional form is critical as it may significantly impact the estimate. The resulting extrapolated estimate is referred to as the SIMEX estimate.

Kuchenhoff et al. [2006] have extended the Cook and Stefanski [1994] method to misclassified discrete data, referring to their approach as *SIMEX MC*. The main difference from the continuous version of SIMEX lies in the simulation of the perturbed data sets. Analog to the parametric model for continuous data, SIMEX MC parameterizes the error process with a misclassification matrix, Π. The matrix Π is a matrix of conditional probabilities of observing a specific value of the data given the true value. Each entry of the Π matrix is therefore $\pi_{z_i^*,z_i} = P(Z_i^* = z_i^* | Z_i = z_i)$. As with SIMEX, it is assumed that the Π matrix is known. In the context of misclassification on a binary outcome variable, the Π matrix is:

$$\Pi = \begin{bmatrix} \pi_{0,0} & \pi_{0,1} \\ \pi_{1,0} & \pi_{1,1} \end{bmatrix} = \begin{bmatrix} 1 - f^+ & f^- \\ f^+ & 1 - f^- \end{bmatrix}.$$

A spectral decomposition of the Π matrix is the first step in simulating data at different misclassification magnitudes. The spectral decomposition of Π is $\Pi = E\Lambda E^{-1}$, where Λ is a diagonal matrix with the eigenvalues of Π on the diagonal and where the columns of E are the corresponding eigenvectors. The level of the additional misclassification applied to the observed data is controlled by λ_k. For a given λ_k, data are simulated according to the conditional probabilities specified by the matrix $\Pi_k = E\Lambda^{\lambda_k}E^{-1}$. The simulated data are consequently related to the true unobserved data by the matrix $E\Lambda^{(1+\lambda_k)}E^{-1}$. Extrapolation to $\lambda_k = -1$ gets rid of the misclassification present in the data in principle. Therefore, once the data are simulated, the remainder of the algorithm remains the same as the SIMEX algorithm and the SIMEX MC estimator is the extrapolated estimate at $\lambda_k = -1$.

In the present manuscript, the estimators from the SIMEX MC procedure are denoted $\hat{\mu}^{lin}$ and $\hat{\mu}^{quad}$ when the form for $g(\cdot)$ is assumed linear and quadratic, respectively. Similarly to the analytical adjustment, the specific RDS estimators are indicated in the subscript. For example, the symbol $\hat{\mu}_{VH}^{quad}$ refers to the Volz-Heckathorn estimator corrected for misclassification with the SIMEX MC procedure based on a quadratic functional form.

3.1.2 Uncertainty of the Corrected Estimators

3.1.2.1 Salganik Bootstrap Extensions

A naive approach to estimating the variance of a corrected estimator of the Hájek style would be to perform the Salganik Bootstrap procedure [Salganik, 2006] described in Section 2.4.3.1 based on the observed data without any modifications. However, this fails to take into account the variability from the correction procedure and the fact that the observed infection statuses are measured with uncertainty. In this section, we propose two extensions to the current methodology to address these issues. Alternatively, one could estimate the variance using the methodology proposed by Kuchenhoff et al. [2007]. Here we have nonetheless chosen to extend existing uncertainty estimators to reflect the recruitment structure relevant to the RDS data.

The choice of procedure to correct the naive estimate for misclassification impacts the sampling distribution of the corrected prevalence estimator. The first extension is designed to reflect this source of variability. Simply replacing the naive estimates ($\hat{\mu}^{naive}$) in step (2) of the bootstrap (i.e. "RDS estimates") by the corrected estimates ($\hat{\mu}^{adj}$, $\hat{\mu}^{lin}$, or $\hat{\mu}^{quad}$) using the selected correction procedure accounts for the inherent variability due to the correction method.

The purpose of the second extension is to adjust for the variability associated with the potential misclassification of the recruiters' infection status. The re-sampling weights, $\mathbf{w}^0$ and $\mathbf{w}^1$, defined in step (1) of the bootstrap algorithm (i.e. "Resampling") implicitly assume that the infection statuses are measured accurately. We suggest to substitute those weights with the vectors $\mathbf{w}^{*0}$ and $\mathbf{w}^{*1}$ defined as the conditional probabilities that the recruiter's infection status is negative ($\mathbf{w}^{*0}$) or positive ($\mathbf{w}^{*1}$) given his or her observed status. For instance, let's assume individual i was recruited by j, then:

$$w_i^{*k} = \quad P(Z_j = k | Z_j^* = z_j^*) = \quad (k\mu + (1-k)(1-\mu)) \quad \frac{P(Z_j^* = z_j^* | Z_j = k)}{P(Z_j^* = z_j^*)},$$

where $k \in \{0, 1\}$. One limitation of this method is that these resampling weights require the true population proportion μ and $P(Z_j^* = z_j^*)$. We suggest that μ may be approximated by the selected corrected estimator. Likewise, $P(Z_j^* = 1)$ and $P(Z_j^* = 0)$ may be approximated by $\hat{\mu}^{naive}$ and $1 - \hat{\mu}^{naive}$, respectively.

An additional modification to this algorithm is proposed to incorporate the uncertainty arising from using uncertain misclassification rates, if applicable. The known error rates correcting the naive prevalence estimates are replaced with draws from the error rates' distribution. For the SIMEX MC algorithm, this involves updating Π, the misclassification matrix, used in the **Simulation** step.

3.1.2.2 Successive Sampling Bootstrap Extension

It is possible to adapt the first extension discussed in Section 3.1.2.1 to the successive sampling Bootstrap procedure [Gile, 2011], reflecting the variability associated with the correction procedure. Similarly to the extension for the Salganik Bootstrap algorithm, the naive estimates are substituted for the corrected estimates which are calculated either with the known misclassification rates or with draws from the best estimate distributions. Because the resampling step of the successive sampling bootstrap is more complex, the second extension described in the previous section is not applicable.

3.2 Simulation Study

Because of the inherent complexity of the RDS process, and the inadequacy of any approximating model for it, we use simulation as the primary tool for evaluating the performance of the proposed methods. In the next sections, we describe the design and present the results of a simulation study assessing the performance of the two misclassification correction methods for RDS estimators: the analytical correction and the SIMEX MC, and also assessing the uncertainty estimators. All prevalence and variance estimates based on true or observed data in this simulation study, as well as in the RDS application discussed in Section 3.3, are calculated with functions available in the R package `RDS` [Handcock et al., 2015a].

3.2.1 Simulation Study Design

3.2.1.1 Network, Sampling and Misclassification Rates Simulation Conditions

This simulation study's main objective is to assess the performance of the correction methods under a variety of conditions capturing the main sources of randomness involved in the RDS estimation procedure. These sources include the random process underlying the network structure, the RDS sampling procedure and the misclassification mechanism. The selected scenarios were constructed to capture those sources of uncertainty.

Our first objective was to design a baseline scenario where the effect of misclassification errors could be isolated from other factors. Our second objective consisted in evaluating the robustness of the correction methods to conditions inducing biases in RDS estimators from sources unrelated to misclassification. Under those circumstances, the misclassification correction methods are expected to retrieve the estimate based on the true infection statuses rather than the actual population parameter μ. Our third objective was to assess the ability of the methods to eliminate the misclassification bias for large asymmetric misclassification rates such as those found in the RDS application in India discussion in Section 3.3. Our last objective was to ensure that the performance of the methods is not significantly degraded by uncertain misclassification rates, such as rates obtained from external validation studies. Scenarios' features intended to assess those objectives are summarized in Table 3.1.

Baseline scenario (S1): The purpose of this scenario is to isolate the effect of misclassification. The average prevalence estimates based on the true outcome variable ($z_i's$) approach the true population prevalence so that the bias in the naive prevalence estimates is mainly attributable to misclassification. Methodology to simulate the networks, RDS samples and misclassified infection statuses are outlined below.

1. Network Simulation: One thousand undirected networks are generated at random using the exponential-family random graph model (ERGM) [Frank and Strauss, 1986, Hunter et al., 2008, Hunter and Handcock, 2006]. Networks are simulated such that on average, each individual is connected to 7 members of the population. The total population size is 1000 individuals. Each individual is assigned an infection status at random, with the true infection prevalence maintained at exactly 20% for each network. Networks are simulated using the R package `statnet` [Handcock et al., 2015b].

2. Sampling: One RDS sample is drawn per network with a sample size of 200. A total of 10 seeds are selected completely at random among all nodes. Each respondent recruits 2 participants completely at random among their contacts. The sampling is performed without replacement.

3. Misclassification: One set of misclassified infection statuses is generated for every network. For the baseline case, a false positive rate of 10.3% and false negative rate of 0.5% are assumed. The false positive rate corresponds to the findings of a study conducted in the Democratic Republic of Congo [Shanks et al., 2013].

Sampling and network assumption violations (S2): In S2, network and sampling features are simulated to purposively induce bias in the RDS prevalence estimators. The objective is to assess whether the performance of the correction methods is altered by those biases.

Table 3.1: Network and sampling features included in the simulation study scenarios.

Condition	Parametrization	S1	S2	S3
Homophily	$\dfrac{P(Y_{ij}=1 \mid z_i=1, z_j=1)}{P(Y_{ij}=1 \mid z_i \neq z_j)}$	1.0	5.0	1.0
Seed Selection$^{(1)}$	$P(i \in \mathcal{S}_0 \mid z_i=1)$	$1/N$	$1/\lvert\mathcal{Z}^1\rvert$	$1/N$
	$P(i \in \mathcal{S}_0 \mid z_i=0)$	$1/N$	0	$1/N$
Diff. Recruitment$^{(2)}$	$\dfrac{P(S_{i,t}=1 \mid \ S_{j,t-1}=1,\ z_i=1,\ Y_{ij}=1)}{P(S_{i,t}=1 \mid \ S_{j,t-1}=1,\ z_i=0,\ Y_{ij}=1)}$	1.0	2.0	1.0
Diff. Activity	$\dfrac{\frac{1}{\lvert\mathcal{Z}^1\rvert}\sum_{i\in\mathcal{Z}^1} d_i}{\frac{1}{\lvert\mathcal{Z}^0\rvert}\sum_{i\in\mathcal{Z}^0} d_i}$	1.0	1.0	1.4
f^+ rate (%)		10.3	10.3	1.0
f^- rate (%)		0.5	0.5	57.0

[(1)] $\mathcal{S}_0$: Set of initial participants in the survey, that is, the seeds.
[(2)] $S_{i,t}$: Indicates if i is sampled at step t assuming a random walk on the network nodes.

Networks were simulated with elevated *homophily* and the sampling procedure with *seed bias* and *differential recruitment*. The mathematical parametrization of those terms is given in Table 3.1. Conceptually, homophily is a network feature which represents the propensity of alike nodes to tie more often than expected at random. Networks under S2 were produced with an average homophily of five whereas the ones in S1 displayed no homophily on average. The seed selection regime was also modified in S2 to force initial participants to be selected among the infected nodes. We refer to this notion as seed bias. Gile and Handcock [2010] demonstrate that the selection of the participants starting the referral chains may bias the estimates. Finally, differential recruitment denotes the propensity of participants to recruit individuals with a given characteristic with higher probability. Literature discusses how this form of differential recruitment induces bias in many RDS estimators [Gile and Handcock, 2010, Lu, 2013, Tomas and Gile, 2011, Verdery et al., 2015]. Although one

RDS estimator has shown robustness to this source of bias [Lu, 2013, Verdery et al., 2015] when information about the participants' ego network is available, none of the estimators included in this study adjust for this type of bias. Differential recruitment in S2 is such that infected individuals are twice as likely to be recruited than the non-infected ones.

Large asymmetric misclassification rates (S3): Under S3, the misclassification rates were chosen to replicate the average misclassification rates from the RDS application discussed in Section 3.3, that is, $f^+ = 1\%$ *and* $f^- = 57\%$. Data from this application also suggest an average *differential activity* of approximately 1.4. Differential activity exists when one group has more social connections than the other. More specifically, differential activity is defined as the ratio of mean degree of the infected individuals in the population to the mean degree of the non-infected ones. The baseline scenario was produced with an average differential activity of one, or in other words, without differential activity, while S3 used 1.4.

In the three scenarios, we assumed known misclassification rates. In practice however, researchers may instead have to rely on uncertain error rates such as rates estimated from an external validation study for instance. To assess the performance of the correction methods with uncertain error rates, Scenarios 1 to 3 were repeated with infection statuses (z_i^*'s) simulated with rates generated from Beta distributions. The parameters of the Beta distributions were chosen so the expected values would equal the known error rates. For S1 and S2, the parameters of the Beta generating the false positive rates were also chosen to reproduce the precision of the rate in the work of Shanks et al. [2013]. The 95% confidence interval for the error rates under S1 to S3 are as follows:

- S1 and S2: (.071, .14) for f^+ and (.002, .009) for f^-; and
- S3: (.005, .017) for f^+ and (.52, .62) for f^-.

The naive estimates are subsequently corrected with the best guess misclassification rates, that is, the expected value of the distributions.

3.2.1.2 SIMEX Misclassification Parameters

The objective of SIMEX Misclassification (SIMEX MC) is to express the estimator as a function of the magnitude of misclassification in the data. This procedure relies on a number of tuning parameters, one of which controls the amount of misclassification at which the function $g(\cdot)$ is evaluated. This parameter is λ_k and is described in Section 3.1.1.2. For the simulations, we have used $\lambda_k \in \{0, 0.4, 0.8, 1.2, 1.6, 2\}$ which is a slightly finer grid than what found in the literature related to SIMEX. Our analysis of the RDS application also suggested that in presence of greater misclassification, the optimal choice of λ_k's might differ. As such, we have instead used $\lambda_k \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ for S3. We have simulated $B = 100$ data sets for each levels of λ_k with the exception of $\lambda_k = 0$ for which $\hat{\theta}(\lambda_k) = \hat{\mu}^{naive}$.

For the purpose of our simulation study, and subsequently for the RDS application in India, we have selected two functional forms to extrapolate the simulated estimates to the theoretical level where there is no misclassification, that is, to $\lambda_k = -1$. We have selected the linear and quadratic functional forms based on standard practice in the literature, visual inspection of the functions, and on a comparison of a number of model selection criteria. These two functional forms appear to reasonably fit the data simulated with additional misclassification. However, the objective model-selection criterion favor the quadratic form approximately 80% to 90% of the time under the selected scenarios.

3.2.2 Simulation Study: Point Estimates

Simulation study results for all estimators, under the three scenarios and calculated with known and uncertain misclassification rates are presented in Figure 3.1.

Results in Figure 3.1 are organized in three panels on the horizontal axis corresponding to the three scenarios. In addition, two panels on the vertical axis separate the results produced with known rates from those produced with uncertain error rates. In each of the six sections of the plot, the naive and corrected prevalence estimates are summarized by box plots for each of the four estimators ($\hat{\mu}_{Mean}$, $\hat{\mu}_{VH}$, $\hat{\mu}_{SS}$ and $\hat{\mu}_{SH}$). The average estimates based on the true infection statuses over one thousand simulations for a given estimator and scenario are depicted by the horizontal lines. Those lines represent the best case value to retrieve. Since RDS estimators may be subject to other sources of biases than misclassification and we expect the correction methods to strictly address the misclassification bias, the placement of the blue line may differ from the population prevalence of 20%. Finally, the "*"'s indicate that the method belongs to the set of methods achieving the lowest misclassification bias, for a given scenario and estimator based on a Bonferroni pairwise comparison at a family-wise error rate of 5%.

The first key finding that Figure 3.1 reveals is that the corrected estimates exhibit significantly less misclassification bias than the naive approach. However, the methods do not perform equally well under all circumstances.

For the estimators of the Hájek style, the analytical adjustment is the best method to reduce the misclassification bias in all presented scenarios. For practical purposes though, the SIMEX MC with quadratic extrapolation displays similar performance under S1 and S2. The large false negative rates used in S3 however alters this method's ability to reduce the misclassification bias.

Similar conclusions may be reached for the Salganik-Heckathorn estimator under S1 and S3. However we observe a poorer performance of the analytical adjustment under S2. As demonstrated in Section 2.4.2.2, the Salganik-Heckathorn estimator is exactly of the Hájek style when c in equation (2.12) equals one. Consequently, the analytical adjustment is expected to do reasonably well for a c of one. As discussed

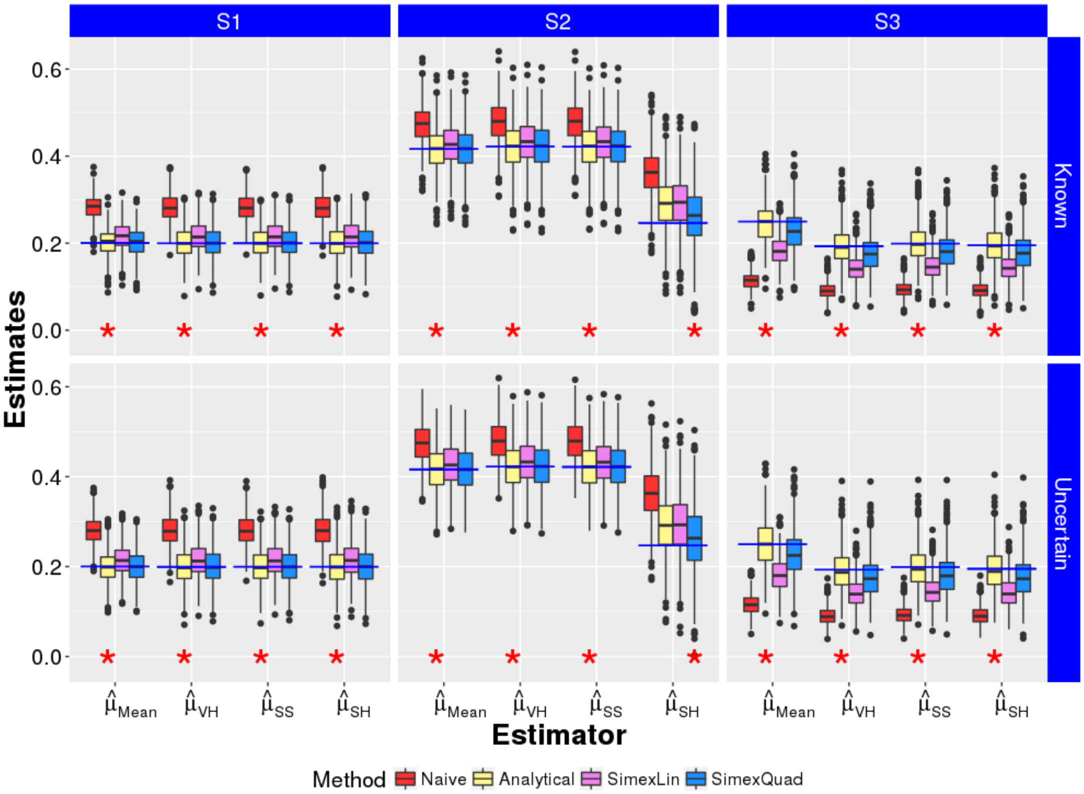


Figure 3.1: Estimates under the three scenarios summarized in Table 3.1 and under known and uncertain misclassification rates. The estimates were calculated based on the observed data ($\hat{\mu}^{naive}$) and on the observed data but adjusted for misclassification with the correction methods ($\hat{\mu}^{adj}$, $\hat{\mu}^{lin}$ and $\hat{\mu}^{quad}$). A "*" on the horizontal axis indicates that the method is in the set of methods producing the least biased estimates based on a Bonferroni pairwise comparison at a family-wise error rate of 5%. The horizontal lines are set at the average estimates based on the true infection statuses.

in Appendix A, discrepancies between c and its analog observed version c^* may also impact the efficiency of the analytical adjustment. The average c and c^* factors over the one thousand simulations under S2 are 2.37 and 1.65, respectively. This discrepancy combined with the magnitude of c explain the inability of the analytical adjustment to eliminate a substantial portion of the misclassification bias in S2. For comparison purposes, those averages were 1.00 and 1.00 for S1 and 0.99 and 0.99 for S3. Lastly, since the SIMEX MC algorithm does not depend on the form of the estimator the performance of this method with quadratic extrapolation is mostly unaffected by the assumption violations simulated under S2.

Although SIMEX MC with linear extrapolation displays significantly less misclassification bias than the naive approach, it consistently results in larger error than the quadratic extrapolation. This agrees with our prior findings which suggested a better fit for the quadratic form.

The distribution of the prevalence estimates with known and uncertain error rates appear similar in Figure 3.1. The main difference is the increased variability of the estimates computed with the uncertain rates. The increase in standard deviation ranges from 9.5% to 27.1% in the selected scenarios. More details regarding the absolute bias, standard deviation and root mean-squared-error (RMSE = $\sqrt{MSE}$) may be found in Appendix B.

The performance of the correction methods have also been assessed at various levels of miclassification. Results are presented in Appendix B. In most instances, the RMSE based on the analytical adjustment is substantially lower than the naive RMSE, with a maximum reduction of approximately 84%. The few exceptions occur when the estimates contain little misclassification bias. In those cases, our analysis suggests that the benefits from the reduction in misclassification bias are offset by the increase in the uncertainty of the corrected prevalence estimates.

The discussed correction methods rely on the knowledge of the misclassification rates f^+ and f^-. In practice however, those rates may be uncertain and possibly contain measurement error. In Appendix B we have evaluated the impact of inaccurate error rates on the correction methods. We found lower misclassification bias in the corrected estimates than in the naive estimates when using moderate departure from the true error rate for either f^+ or f^- for S1 to S3.

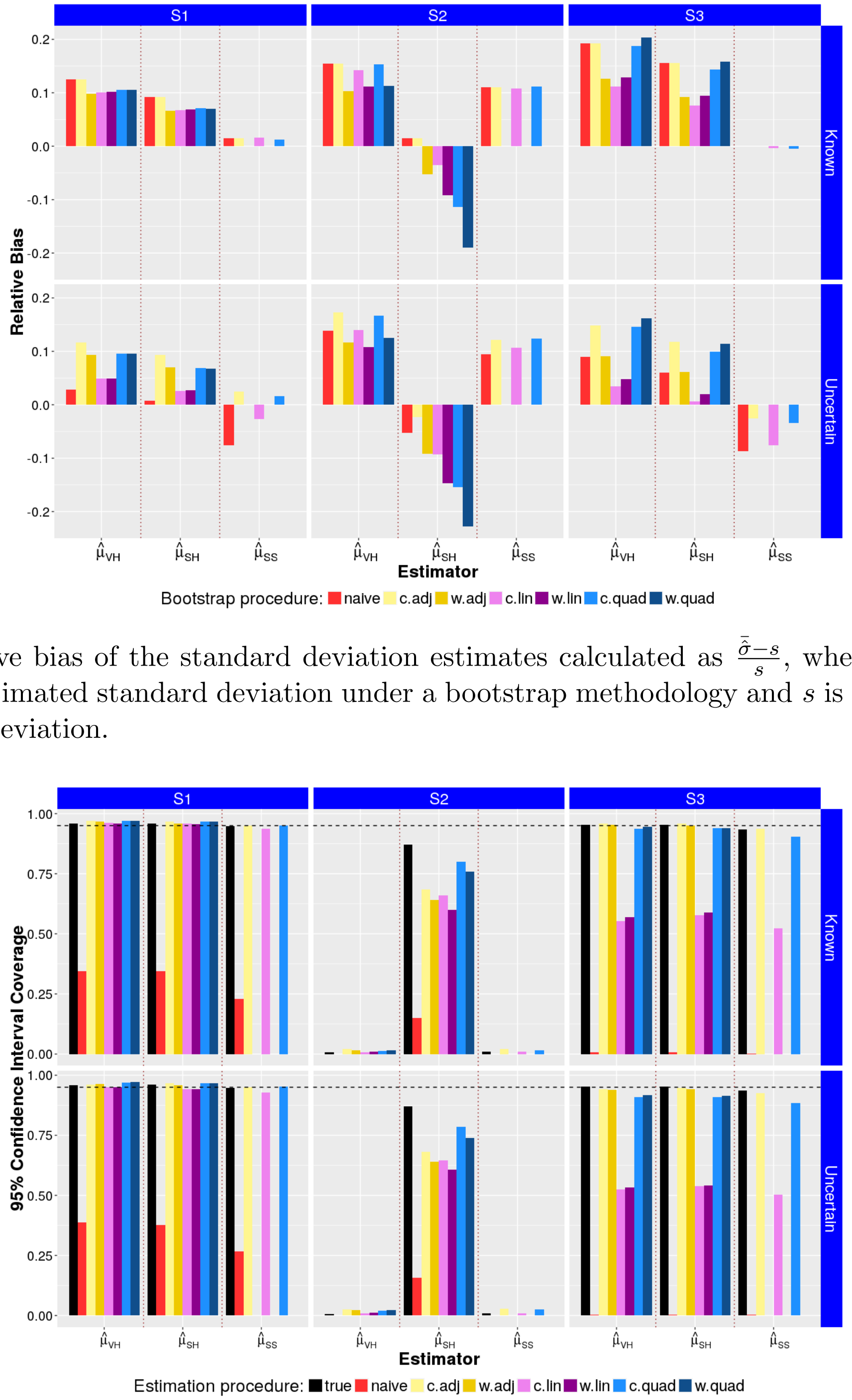
Overall, the correction methods perform better than the naive approach in all scenarios presented in our simulation study. The performance of the analytical adjustment and the SIMEX MC with quadratic extrapolation is similar with two exceptions: when misclassification rates are very large (analytical preferred) and when the analytical adjustment is not suitable for the Salganik-Heckathorn estimator (SIMEX MC preferred).

3.2.3 Simulation Study: Variance Estimates

In Section 3.1.2 we proposed extensions to the existing bootstrap procedures to account for the additional variability of the RDS estimators due to the correction methods, the misclassification on the outcome variable and the uncertainty of the misclassification rates, if applicable. In this section, we evaluate the performance of these extended variance estimation procedures against the naive application of the original method.

Ideally, a bootstrap variance estimator should produce results aligned with the total variance of the stochastic process. Our closest estimate of this total variance is the variability among the estimates in the simulation study for each scenario (s's). Figure 3.2a displays the relative differences between the average estimated standard deviation under the various bootstrap methodologies ($\bar{\hat{\sigma}}$'s) and their respective sample standard deviation (s's). The relative bias is computed as $\frac{\bar{\hat{\sigma}}-s}{s}$.

Figure 3.2a presents, for each of the three scenarios, six versions of the extended Salganik Bootstrap procedure to estimate the variance of $\hat{\mu}_{VH}$ and $\hat{\mu}_{SH}$ and three versions of the extended Successive Sampling Bootstrap procedure to estimate the variance of $\hat{\mu}_{SS}$. For the Salganik Bootstrap procedure, each of the three correction methods produce a set of two variance estimators. The first estimator of that set only accounts for the first extension, i.e. corrected resampled estimates, while the second one also reflects the second extension, i.e. modified resampling weights. Results produced with uncertain misclassification rates include the additional modifications to



(a) Relative bias of the standard deviation estimates calculated as $\frac{\bar{\hat{\sigma}}-s}{s}$, where $\bar{\hat{\sigma}}$ is the average estimated standard deviation under a bootstrap methodology and s is the sample standard deviation.

(b) 95% confidence interval coverage rates, where the coverage rates are the percentage of the intervals including the true population proportion μ of 20%.

Figure 3.2: Standard deviation estimation and 95% confidence interval coverage results for $\hat{\mu}_{VH}$, $\hat{\mu}_{SH}$ and $\hat{\mu}_{SS}$ and for the various versions of the Bootstrap procedures under S1 to S3 with known or uncertain misclassification rates. The notation 'adj", "lin" or "quad" indicates whether the variance is being estimated for $\hat{\mu}^{adj}$, $\hat{\mu}^{lin}$ or $\hat{\mu}^{quad}$ whereas "c." and "w." refers to the first and second bootstrap extensions, respectively.

the algorithm described in Section 3.1.2.1, that is, the known error rates are replaced by draws from the error rates' distribution.

In Figure 3.2a, we observe that including both extensions to the Salganik Bootstrap variance estimator for $\hat{\mu}_{VH}^{adj}$ and $\hat{\mu}_{SH}^{adj}$ reduces the relative bias in most instances. The main exception is under S2 for $\hat{\mu}_{SH}^{adj}$, that is, when $\hat{\mu}_{SH}^{adj}$ is not of the Hájek style. The improvement from the second extension, if any, is negligible when applied to the SIMEX MC correction. Overall though, no methods appear to consistently be the best method across all conditions.

For the variance estimation of $\hat{\mu}_{SS}$, the extended Bootstrap with the three corrected methods perform in a similar fashion. There is a slightly higher relative bias when uncertain error rates are used as opposed to known rates. Again however, none of the methods systematically lead to the best performance under all circumstances.

Figure 3.2a suggests that the naive Bootstrap procedure sometimes outperform the extended Bootstrap estimators with uncertain misclassification rates. However, the decrease in relative bias with uncertain rates is mainly caused by the fact that the uncertainty of the error rates is not accounted for in the naive procedure rather than by superior properties of the procedure. Larger uncertainty around the error rates would deteriorate its performance.

In conclusion, we recommend using the variance estimator corresponding to the appropriate correction method for the problem at hand. For the Salganik Bootstrap, one has to further decide between applying the first extension or both of them. We suggest applying both extensions solely with the analytical adjustment. The two extensions showed smaller relative bias in our simulation study with this correction method, which was not systematically the case when used in combination with the SIMEX-MC algorithm.

Figure 3.2b helps evaluate the combined performance of the point estimation and the variance estimation procedures. The 95% confidence interval coverage rates with respect to the true population proportion of $\mu = 20\%$ for $\hat{\mu}_{VH}$, $\hat{\mu}_{SH}$ and $\hat{\mu}_{SS}$ under three scenarios with known or uncertain error rates and using the different Bootstrap variance estimators are shown in this plot. This figure clearly highlights that the naive approach is either worse than or, at best, equivalent to the correction methods. Also the analytical adjustment and the SIMEX MC with quadratic extrapolation have similar coverage for each scenario. In addition, their coverage rates are comparable to the coverage calculated based on the true infection statuses. For $\hat{\mu}_{SH}$ under S2, since the analytical adjustment does not strictly apply, SIMEX MC with quadratic extrapolation performs better. Similarly, since the analytical correction reduces a larger proportion of the misclassification bias with large error rates, this inference is slightly better with this method under S3. Finally, the SIMEX MC with linear extrapolation tends to do worse than the other two correction methods.

Consequently, we conclude that for the scenarios examined in this simulation study, the methodologies proposed do improve the statistical inference when compared to the naive approach and that unless the Salganik-Heckathorn is far from the Hájek style, the analytical approach is preferred to the other correction methods.

3.3 Application to High Risk Populations in India

RDS has been used extensively in the context of HIV/AIDS surveillance for populations at high risk of infection such as people who inject drugs (PWID), men who have sex with men (MSM) and female sex workers (FSW) [Johnston et al., 2008, Malekinejad et al., 2008, Montealegre et al., 2013]. In this section, we present HIV prevalence estimates for RDS studies conducted in India among two of these key populations, that is, among PWID and MSM. We compare two sets of estimates which are either derived from self-report HIV status or from blood testing. The former is likely an inaccurate measurement of the actual HIV infection status since, as discussed by the gap report [UNAIDS, 2014], around 54% of people living with HIV-positive status are unaware of their status. Therefore, in this section we show that in most cases, it is possible to reduce the misclassification bias present in the estimates based on self-reported status by using the methods proposed in this paper.

The first study on which our analysis is based consists of 15 RDS samples collected in 2013 in multiple cities in India [Lucas et al., 2015]. In that study, a total of 14,481 PWID were surveyed. Two to three seeds were selected to initiate the sampling in each city. Every respondent could recruit up to two individuals. With the exception of one location, all sites recruited approximately one thousand individuals from the target population.

Participants' HIV status was determined based on three rapid HIV testing kits [Lucas et al., 2015]. The results from the on-site HIV test were compared with the self-reported HIV status. This status was determined based on questions regarding their past HIV testing and result history. Participants who answered that their last HIV test was positive are treated as positive HIV self-reports whereas participants who had never been tested or who reported a non-positive test result are treated as negative self-reports. Finally, for the purpose of our analysis, we assume the on-site HIV test is 100% specific and sensitive. All indeterminate results were confirmed using western blot, and this assumption is likely to be quite accurate. Therefore, these values are treated as the truth for estimating error rates and the evaluation of our methods.

The Volz-Heckathorn HIV prevalence estimates without misclassification for the 15 sites range from 5.9% to 44.8% with a weighted average of 18.2%. The Volz-Heckathorn naive estimates are much lower, ranging from 0.9% to 30.2% with a weighted average of 8.9%. The large discrepancy between the two sets of estimates is attributable to large false negative rates (weighted average of 53.9%). These false negative rates may be imputable to non recent testing, for example, and indicate that individuals in the populations are largely unaware of their positive infection status.

The false positive rates (weighted average of 1.3%) are not compensating for the observed unawareness. The weighting is proportional to the sample sizes.

We have applied similar analysis to another RDS study which was conducted among MSM in India [Solomon et al., 2015]. This study covered 12 locations for a total of 12,022 participants. The data collection was performed under nearly the same methodology as the PWID study. The weighted HIV false negative and false positive average rates, 59.3% and 0.2%, are comparable to the ones in the PWID populations.

Figure 3.3 displays the absolute relative bias, as defined as the difference between the corrected or naive estimate and the corresponding estimate based on the true infection status divided by the latter, as a function of the false negative rates. The results are shown for $\hat{\mu}_{VH}$, $\hat{\mu}_{SS}$ and $\hat{\mu}_{SH}$, for all PWID populations. One MSM site is omitted since the analytical adjustment could not be evaluated in that instance. In that sample, no false positives were observed and all HIV positive individuals were unaware of their infection status.

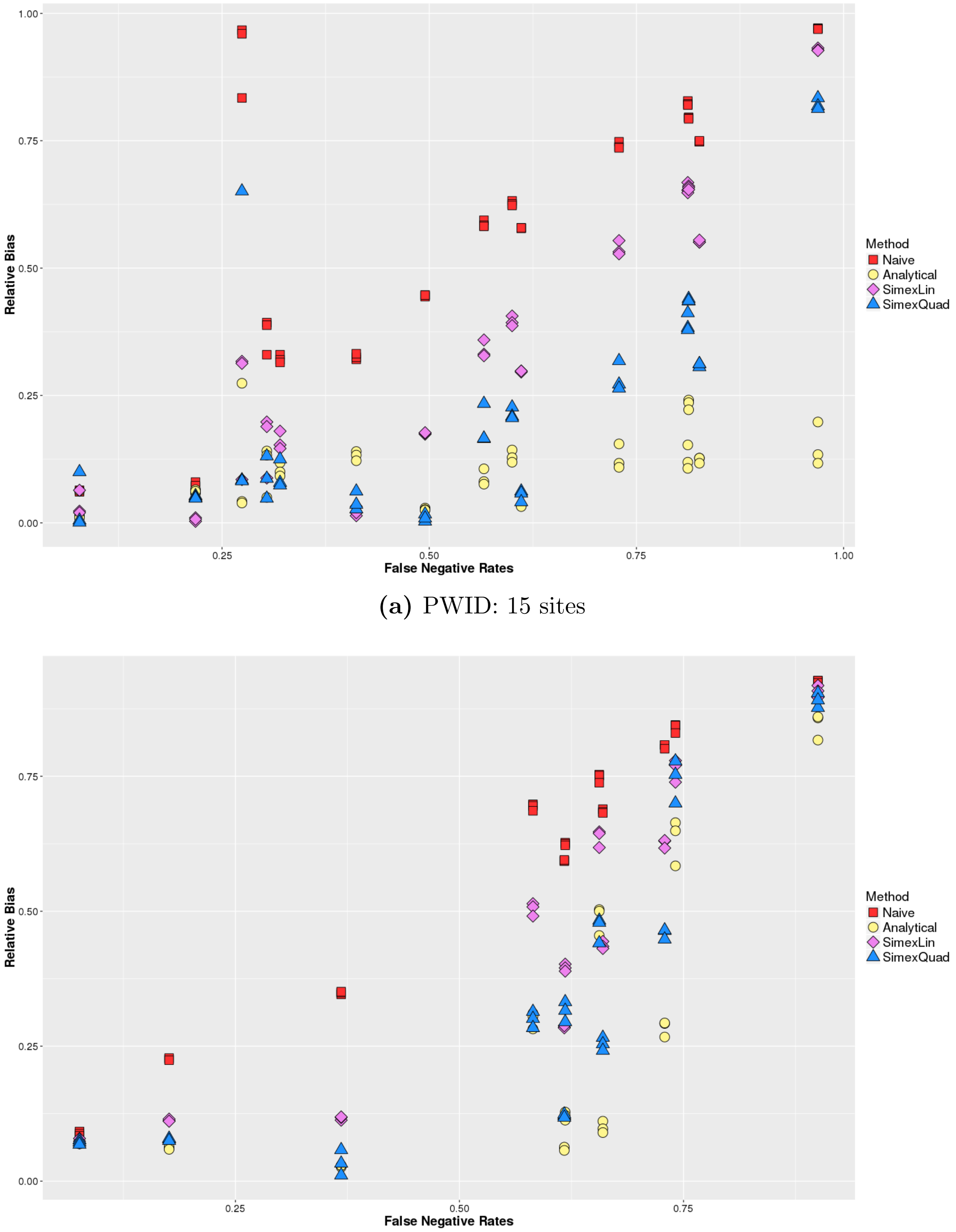
For all data sets, the factor c discussed in Section 2.4.2.2 is close to one and to c^*. This implies that we expect the analytical adjustment to perform well in adjusting the Salganik-Heckathorn estimator. In general, c and c^* may substantially differ from one in RDS studies. They may be close to their theoretical values, as well as close to each other in these examples because of the small number of seeds and the large sample sizes.

A similar analysis to the one performed in the simulation study was conducted to decide on the SIMEX tuning parameters and extrapolation function. We concluded that a larger number of simulated data sets is necessary to improve the model fit. Consequently, $B = 500$ was selected in all but two scenarios where even greater B's were chosen. Also, we established a false negative error rate threshold of 25% to determine whether the lambdas would be $\{0, 0.4, 0.8, 1.2, 1.6, 2\}$ ($f^- < 25\%$) or $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ ($f^- > 25\%$). This choice is justified by improvement to model selection criteria. Finally, the quadratic function appears to be a better choice based on model selection criteria.

Both studies lead to similar methodological findings. For all but one study the naive estimates are more biased than estimates produced by any of the three correction methods. We also observe that the SIMEX procedure tends to perform better for lower false negative rates. This suggests that the functional form fitted with large error rates may not be representative of the functional form at lower error rates. The performance of the analytical correction is also poorer for large error rates, but to a lesser extent. These findings are consistent with results from S3 in our simulation study. Under that scenario, the conditions were purposely chosen to mimic on average some of the conditions in this application.

One of the sites in the PWID study appears to have a greater relative bias than the remaining sites despite the false negative rate being small in comparison to other cities. The noticeable deviation is explained by the larger false positive rate observed at that site ($f^+ = 7.6\%$). The weighted average for the remainder of the sites is 0.8%.

Results from the implementation of the adjusted estimates along with the extended Bootstrap procedures are summarized in Table 3.2. In this table, we compare the number of 95% confidence intervals that include the corresponding "true" value without misclassification for the different sites, treated as a favorable-case for evaluating coverage performance. For comparison purposes, results from the naive point estimates and variance estimates are also presented. As expected, since the false negative rates are so high, very few of the intervals for the 15 PWID and 11 MSM samples based on the naive methodologies include the estimate without misclassification. However, it is clear from this table that the corrected estimates used in combination with the extended versions of the Bootstrap procedures significantly increase the number of confidence intervals including the prevalence estimates based on the true data.



(a) PWID: 15 sites

(b) MSM: 11 sites

Figure 3.3: Point estimate relative bias as a function of the false negative rates for PWID and MSM for a) 15 PWID sites and b) 11 MSM sites of the studies conducted in India. The estimates using the naive and the corrected estimators are shown for the Volz-Heckathorn, the Salganik-Heckathorn and the Successive Sampling estimators.

An additional finding from these results is that, perhaps not surprisingly, the intervals based upon the analytical adjustment produce higher coverage than their SIMEX MC counterparts in all but one case. From Figure 3.3, it is clear that the misclassification bias is smaller for the former method in most instances. Finally, since all correction methods are reasonably applicable to all estimators, the coverage is similar across the three estimators.

Table 3.2: Number of sites for which the estimate without misclassification lies inside the 95% confidence interval, out of a total of 15 PWID and 11 MSM sites.

	Prevalence	Variance Estimators						
Study	Estimator	$\hat{\sigma}_{naive}$	$\hat{\sigma}_{c.adj}$	$\hat{\sigma}_{c.lin}$	$\hat{\sigma}_{c.quad}$	$\hat{\sigma}_{w.adj}$	$\hat{\sigma}_{w.lin}$	$\hat{\sigma}_{w.quad}$
PWID	$\hat{\mu}_{VH}$	2	15	7	11	15	7	11
	$\hat{\mu}_{SH}$	2	15	7	10	15	7	11
	$\hat{\mu}_{SS}$	2	15	5	8	—	—	—
MSM	$\hat{\mu}_{VH}$	2	8	4	6	8	4	6
	$\hat{\mu}_{SH}$	3	8	4	6	8	4	6
	$\hat{\mu}_{SS}$	1	8	6	9	—	—	—

Overall, adjusting for misclassification on the outcome variable in the presented examples improves the inference made from RDS data. The three correction methods all reduce the misclassification bias in the estimates, although the analytical adjustment tends to perform best in the studies discussed in this section.

3.4 Discussion

The main contribution of this article is to introduce approaches to correct existing RDS estimators for the bias introduced by the misclassification on a binary nodal attribute, and associated novel estimators of uncertainty. We also have highlighted circumstances for which the performance of the correction methods is impaired in the specific context of RDS.

The first approach is an analytical adjustment, applicable to estimators of the Hájek style. Under the conditions explored in our simulation studies and with the RDS application, this method has shown to substantially reduce the misclassification bias present in the naive estimates.

In some scenarios the ability of the analytical adjustment to reduce the misclassification bias was compromised. We found this to be the case in particular when the SH estimator, which is not of the Hájek style, diverges most from the Hájek style. This issue has arisen in instances where the observed recruitment patterns could not be used as a proxy to estimate the network mixing matrix partitioned on the infection status. In such cases, the c-factor introduced in Section 2.4.2 is different than one. In practice, since we do not observe this c-factor directly, we have to rely on the related observed c^*-factor to determine whether the analytical adjustment is suitable. Since the c- and c^*-factors are positively correlated (see Appendix A), c^* may be used as a proxy for c to evaluate whether the analytical adjustment is likely to be appropriate.

The second approach we discussed is the SIMEX MC procedure. Although it does not require that the estimators be of the Hájek style, it necessitates that the estimator may be expressed as a function of the measurement error present in the data. In many instances, this method produced comparable results to the analytical adjustment in terms of the reduction of the misclassification bias. However, in cases where large error rates prevailed, this method did not eliminate as much misclassification bias. This suggests that the function mapping the estimates to the measurement error variance at higher error rates may not be representative of the function when little to no misclassification is present. The main advantage of using this method is therefore for situations where the Salganik-Heckathorn estimator is far from the Hájek style, in which case, the SIMEX MC with quadratic extrapolation provided the largest reduction in the misclassification bias.

In this paper, we have also extended procedures to estimate the variance of the corrected estimators. The extensions are intended to capture the variance component attributable to the misclassification on the outcome variable, to the adopted correction methodology and to the uncertain misclassification rates, if applicable. The first extension substitutes the corrected estimates for the naive estimates in the naive bootstrap procedures. The main innovation is the modification to the resampling weights applicable to the Salganik Bootstrap procedure only. We have seen that in most instances, with known error rates, the extended methodology for variance estimation does better or at least similarly to the naive approach for estimators of the Hájek style. The second extension provides only marginal improvements, if any, over the first extension for the SIMEX MC corrected estimator, but does appreciably improve the estimators corrected with the analytical adjustment. All versions of the SS Bootstrap procedure perform similarly and the first extension does not appear to significantly improve the performance of the SS Bootstrap procedure. No method systematically outperformed the other, especially in the case of uncertain error rates.

The application to the RDS data from India led to similar findings. Inference based on the self-reported HIV status displayed large misclassification error as participants were widely unaware of their actual HIV status. The 95% confidence interval coverage rates illustrating the combined performance of the point estimation and variance estimation procedures showed that the naive estimation procedures may severely compromise the validity of the inference from self-reported HIV status. The analytical correction performed best in most instances especially with the largest misclassification rates.

One limitation of the proposed methodology is that it relies on the assumption that f^+ and f^- are known and uniform in the population. In many cases this assumption might not hold. The results from our simulation study however suggest that using uncertain misclassification rates from an external validation study result in nearly unbiased estimates when the uncertain rates are unbiased.

CHAPTER 4

DIFFERENTIAL RECRUITMENT

4.1 Introduction

Inference from RDS data relies on a number of strong assumptions which are often unrealistic in conventional settings. Despite the growing empirical evidence [Frost et al., 2006, Iguchi et al., 2009, Liu et al., 2012, Mccreesh et al., 2012] that participants systematically favor the selection of alters with particular characteristics for instance, random recruitment generally remains the default assumption. Sensitivity analysis performed with simulated and real data demonstrate that non-random recruitment potentially yields large biases in RDS prevalence estimators when the favored characteristics are associated with the outcome variable [Frost et al., 2006, Gile and Handcock, 2010, Tomas and Gile, 2011, Lu et al., 2012, Verdery et al., 2015]. The contribution of this work is to identify and measure recruitment dynamics and correct the prevalence estimators for their induced bias.

Most of the RDS prevalence estimators assume that respondents recruit completely at random among their peers. Many have proposed diagnostics to detect non random recruitment patterns in data [Wejnert and Heckathorn, 2008, Liu et al., 2012, Yamanis et al., 2013, Gile et al., 2015].

Subsequently, some have measured its impact on prevalence estimates [Frost et al., 2006, Tomas and Gile, 2011, Verdery et al., 2015] .

Recent advancements also include an extension of the Salganik and Heckathorn [2004] estimator to reduce the bias introduced by non random recruitment behaviors [Lu, 2013].

Lu [2013] extended Salganik and Heckathorn [2004] estimator to incorporate data on self-reported ego-network composition in the estimation of the recruitment matrix. The resulting estimator is considerably more robust to differential recruitment than its original counterpart and displays significantly lower variability. The extended version of the SH estimator proposed by Lu [2013] relies on an improved estimation of the recruitment matrix. However, the suggested methodology implicitly assumes that differential recruitment does not affect participants' probability of being sampled. In addition, the method in its current form does not allow differential recruitment to take place on any other variables than the outcome variable. This is a major concern for the main application of RDS study, that is, estimation of disease prevalence such as HIV. For instance, UNAIDS recently estimated that only 48% of people living with HIV know their infection status [UNAIDS, 2014]. It is therefore even more unlikely that participants could report their contacts' HIV status accurately. Such level of misclassification would inevitably result in an underestimation of HIV prevalence.

In this chapter, we develop methods to reduce differential recruitment bias. The first set of estimators we propose are design-based estimators. They extend the Volz-Hechatorn and Lu's estimators described in Sections 2.4.1.2 and 2.4.2.3, respectively. Similar to the estimator proposed by Lu, our estimators require ego-network data on the variable over which the differential recruitment takes place. However, our estimators provide greater flexibility in that they allow for additional forms of differential recruitment and the variable inducing differential recruitment may differ from the outcome variable. The proposed design-based prevalence estimators along with the parametrization of the differential recruitment forms are discussed in Section 4.2. It is followed in Section 4.3 by a description of a model-based estimator designed to address one of the limitations of the design-based estimators. A comparison of their performance under various sampling conditions and network features is assessed in a simulation study presented in Section 4.5. Finally, we conclude with a discussion of the proposed methods in Section 4.6.

4.2 Design-Based Inference - Random Walk Approximation

In Section 2.4.1.2 and 2.4.2.3 we described the RDS estimators $\hat{\mu}_{VH}$ and $\hat{\mu}_{SH}^{ego}$. Those two estimators are design-based estimators and consequently, no population model is assumed for the outcome variable. The randomness instead stems from the sampling design and each observed unit is weighted based on their sampling probability to obtain a prevalence estimate representative of the target population.

An exact determination of the sampling probabilities under RDS is not however possible due to the complexity of this sampling method. These estimators resort to a random walk (RW) approximation to the RDS process and the sampling probabilities are presumed equal to the RW stationary distribution ($\pi_i = d_i / \sum_{i=1}^{N} d_i$). Conveniently, the unobserved constant of proportionality ($\sum_{i=1}^{N} d_i$) is eliminated due to the ratio nature of these estimators. As per equations (2.6) and (2.16), $\hat{\mu}_{VH}$ and $\hat{\mu}_{SH}^{ego}$ are as follows:

$$\hat{\mu}_{VH} = \frac{\sum_{i=1}^{N} S_i z_i / d_i}{\sum_{i=1}^{N} S_i / d_i}, \text{and}$$

$$\hat{\mu}_{SH}^{ego} = \frac{\hat{\mu}_{VH}}{\hat{\mu}_{VH} + c^{ego}\,(1 - \hat{\mu}_{VH})}, \quad \text{where} \quad c^{ego} = \frac{n_1}{n_0} \left(\frac{\sum_{i=1}^{N} S_i z_i d_{i,0} / d_i}{\sum_{i=1}^{N} S_i (1 - z_i) d_{i,1} / d_i} \right).$$

In this section, we extend these estimators. We begin by parameterizing the concept of differential recruitment. Then, we specify the transition matrices reflecting three forms of differential recruitment and derive the RW's stationary distributions. A maximum likelihood estimator is subsequently proposed to estimate the differential recruitment parameters factored in the sampling probabilities. Lastly, we present the extended version of $\hat{\mu}_{VH}$ and $\hat{\mu}_{SH}^{ego}$ and discuss how they are derived.

4.2.1 Parametrization

Under a random recruitment regime, participants are assumed to recruit among their alters completely at random. Because recruitment is a social act, it is naive to assume this is always the case. Systematic violations of the random recruitment assumption are referred to as differential recruitment.

Differential recruitment may arise in a variety of ways. For instance, participants may favor the recruitment of individuals based on their characteristics (nodal attributes), or based on the nature of their relationship (tie attributes). The nodal characteristic inducing differential recruitment is represented by the indicator vector $\mathbf{x} \in \{0, 1\}^N$ whereas the tie characteristic is represented by the indicator matrix $W \in \{0, 1\}^{N \times N}$.

Consistent with Tomas and Gile [2011], differential recruitment on the nodal attributes may be partitioned into two categories: within groups and between groups differential recruitment. Within groups differential recruitment occurs when participants select alters similar to themselves, such as contacts of the same ethnic group, whereas between groups differential recruitment results from all classes of respondents preferentially recruiting their contacts with a given characteristic. Gile et al. [2015] find, for example, that respondents in four studies of injecting drug users in the Dominican Republic seem to systematically recruit their employed contacts more often than their unemployed contacts, perhaps due to the recruiters elevated confidence that these more reliable contacts would follow through in participating in the study. Differential recruitment on the tie attribute is the result of participants preferably selecting individuals on the basis of their relationship with them. In an attempt to assess the reciprocity of the network ties, Wang et al. [2005] found that 78.9% of respondents in an MDMA users study reported being recruited by a friend as opposed to 14.9% by an acquaintance and 3.4% by a relative. Participants' actual tie composition differing from those proportions would be evidence of recruitment based on tie characteristic. In this section, we address these three forms of differential recruitment.

The magnitude of those behaviors is quantified by the parameter ϕ. In each case, this parameter represents the ratio of the probability of selecting a member of the target population with the nodal or tie preferred attribute to the probability of recruiting a member without it. For example, survey participants systematically recruiting males with a probability twice as high as other genders translates into a ϕ of two. Also, a recruitment regime completely at random implies that ϕ is equal one.

Our definition for the three parameters are presented in Table 4.1. The subscripts b, w, t indicate the form of differential recruitment, that is, between groups, within groups, and on tie attribute, respectively. Furthermore, the superscript RW specifies that RDS is represented by a RW.

Table 4.1: Parametrization of the three forms of differential recruitment (DR) under the RW scheme. $S_{i,t}$ indicates if node i is sampled at step t of the RW.

DR Form	Parametrization
Between groups	$\phi_b^{RW} = \dfrac{P(S_{i,t}=1\mid S_{j,t-1}=1,\ y_{ij}=1,\ x_i=1)}{P(S_{i,t}=1\mid S_{j,t-1}=1,\ y_{ij}=1,\ x_i=0)}$
Within groups	$\phi_w^{RW} = \dfrac{P(S_{i,t}=1\mid S_{j,t-1}=1,\ y_{ij}=1,\ x_i=x_j)}{P(S_{i,t}=1\mid S_{j,t-1}=1,\ y_{ij}=1,\ x_i\neq x_j)}$
Tie	$\phi_t^{RW} = \dfrac{P(S_{i,t}=1\mid S_{j,t-1}=1,\ y_{ij}=1,\ w_{ij}=1)}{P(S_{i,t}=1\mid S_{j,t-1}=1,\ y_{ij}=1,\ w_{ij}=0)}$

4.2.2 Sampling Probabilities

Deriving the sampling probabilities under this framework is equivalent to obtaining the stationary distributions of the random walks with differential recruitment. Consequently, we define in this Section the transition matrices characterizing the three Markov chains and prove the existence and uniqueness of their stationary distributions contingent on some network features.

The transition matrices, denoted P, specify the conditional probabilities of getting to any states given the previous state visited. The entry in the i-th row and j-th column of that matrix, denoted p_{ij}, for instance, is the probability of getting to node j given that node i is the recruiting node.

Figure 4.1 shows a simple example of transition matrices for each of the three cases of differential recruitment of magnitude two ($\phi = 2$). Let us suppose that the size of the nodes in figure 4.1a and 4.1b represents a nodal attribute inducing differential recruitment. For instance, let the large nodes indicate that the individual resides in neighborhood N_1 as opposed to living in neighborhood N_2 depicted by the smaller size nodes. Under the previously introduced notation, $\mathbf{x} = \{0, 0, 1, 0, 1, 0\}$ where the nodes are arranged in alphabetic order. Figure 4.1a illustrates the case of between group differential recruitment so that all classes of participants favor the recruitment of nodes in N_1. This represents a hypothetical situation where every participant systematically favors the recruitment of their contacts living in the neighborhood where the study is conducted for instance. As the left hand side of Figure 4.1a suggests, when the RW is in state B, the probability of selecting node C or E ($p_{BC} = p_{BE} = 2/6$) is twice as high as the probability of selecting node A or D ($p_{BA} = p_{BD} = 1/6$), that is, $\phi_b^{RW} = 2$. Also, to ensure that the sum of the probabilities equals one, the denominator has to be equal to six, i.e. $\sum_{j=1}^{N}(\phi_b^{RW} x_j + (1 - x_j))y_{ij} = 6$. The full transition matrix P for this small network may be found in the right hand side of Figure 4.1a. More generally,

$$p_{ij} = \frac{(\phi_b^{RW} x_j + (1 - x_j))y_{ij}}{\sum_{j=1}^{N}(\phi_b^{RW} x_j + (1 - x_j))y_{ij}} \tag{4.1}$$

in the presence of between group differential recruitment. In this expression, the summand in both the numerator and denominator includes the term y_{ij}. This ensures that the process is restricted to visiting adjacent nodes to the current state. For example, p_{BF} is equal to zero in the illustration due to the lack of a tie between those two nodes (i.e. $y_{BF} = 0$). Finally, it may be observed that for $\phi_b^{RW} = 1$, that is, for a recruitment regime completely at random, $p_{ij} = \frac{y_{ij}}{\sum_{j=1}^{N} y_{ij}} = \frac{y_{ij}}{d_i}$ as expected.

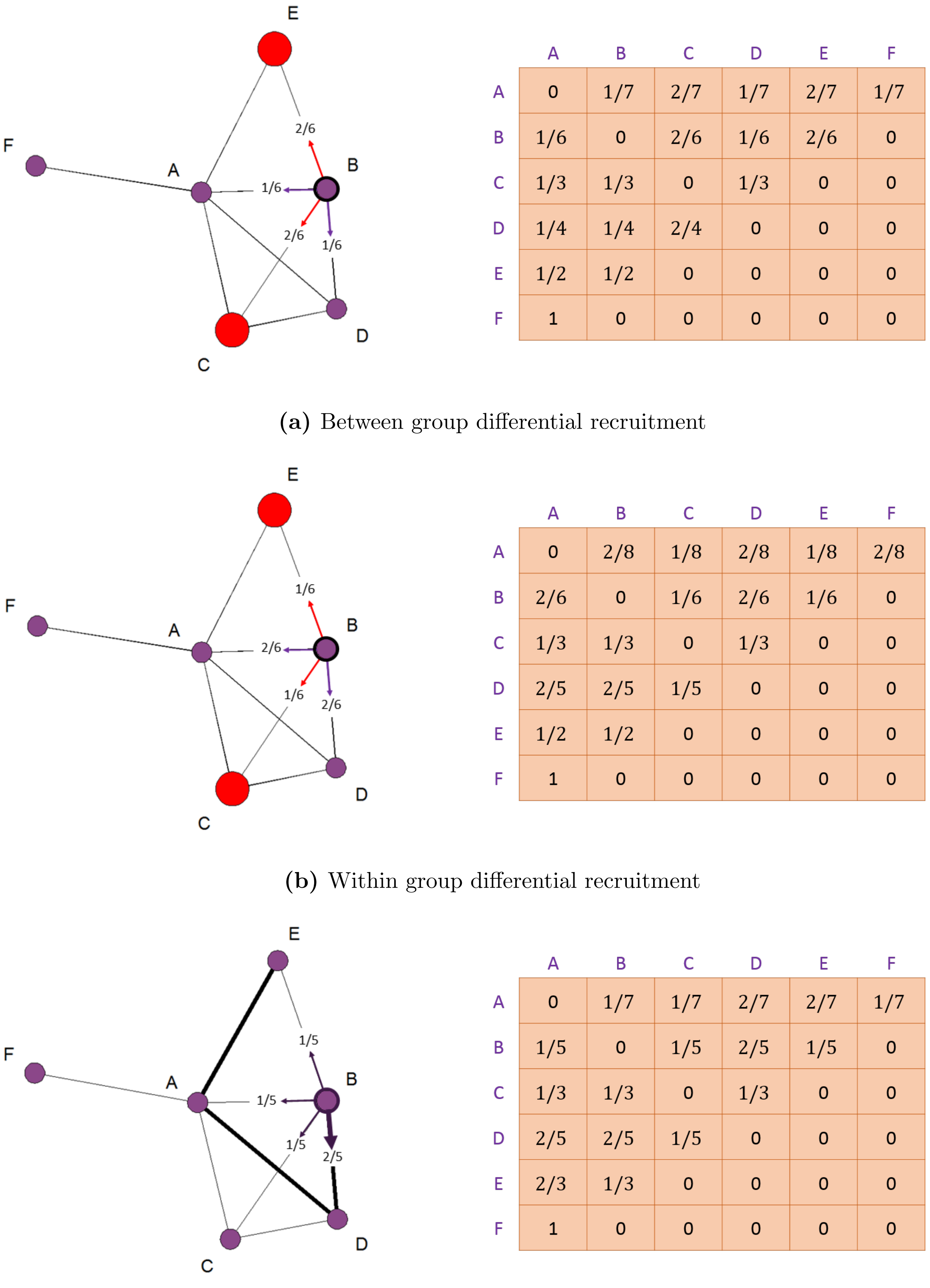
The derivation of within group differential recruitment transition probabilities is similar. However, instead of always favoring individuals residing in N_1, participants recruit more heavily alters living in the same neighborhood as themselves. Node B in Figure 4.1b for instance recruits A or D with a probability twice as large as the probability of selecting node C or E ($\phi_w^{RW} = 2$). Consequently, we obtain the following expression for the within group transition probability between node i and j:

$$p_{ij} = \frac{(\phi_w^{RW} x_j + (1 - x_j))x_i y_{ij} + (\phi_w^{RW}(1 - x_j) + x_j)(1 - x_i)y_{ij}}{\sum_{j=1}^{N}(\phi_w^{RW} x_j + (1 - x_j))x_i y_{ij} + (\phi_w^{RW}(1 - x_j) + x_j)(1 - x_i)y_{ij}}. \tag{4.2}$$

An illustration for transition probabilities for tie attribute differential recruitment is provided in Figure 4.1c. Thicker ties in the plot on the left panel signify that the relationship type induces differential recruitment. Participants may exhibit the tendency to recruit more frequently close friends than acquaintances for example. According to this figure, only six entries in the underlying matrix of tie attributes W are equal to one, w_{AD}, w_{AE}, w_{BD}, and the corresponding reciprocal relationships w_{DA}, w_{EA} and w_{DB}. Under this RW, B is twice as likely to select D over the other nodes. The complete matrix P for this example is provided in the right panel of Figure 4.1c but the expression for any entry p_{ij} is given below:

$$p_{ij} = \frac{(\phi_t^{RW} w_{ij} + (1 - w_{ij}))y_{ij}}{\sum_{j=1}^{N}(\phi_t^{RW} w_{ij} + (1 - w_{ij}))y_{ij}}. \tag{4.3}$$

The three random walks now being fully specified, we may now discuss the associated stationary distributions which are used as sampling weights in the extended version of $\hat{\mu}_{VH}$ and $\hat{\mu}_{SH}^{ego}$. To ensure the Markov chains (MC) are irreducible, we



	A	B	C	D	E	F
A	0	1/7	2/7	1/7	2/7	1/7
B	1/6	0	2/6	1/6	2/6	0
C	1/3	1/3	0	1/3	0	0
D	1/4	1/4	2/4	0	0	0
E	1/2	1/2	0	0	0	0
F	1	0	0	0	0	0

(a) Between group differential recruitment

	A	B	C	D	E	F
A	0	2/8	1/8	2/8	1/8	2/8
B	2/6	0	1/6	2/6	1/6	0
C	1/3	1/3	0	1/3	0	0
D	2/5	2/5	1/5	0	0	0
E	1/2	1/2	0	0	0	0
F	1	0	0	0	0	0

(b) Within group differential recruitment

	A	B	C	D	E	F
A	0	1/7	1/7	2/7	2/7	1/7
B	1/5	0	1/5	2/5	1/5	0
C	1/3	1/3	0	1/3	0	0
D	2/5	2/5	1/5	0	0	0
E	2/3	1/3	0	0	0	0
F	1	0	0	0	0	0

(c) Tie attribute differential recruitment

Figure 4.1: Transition probability matrix (right) for a random walk on the nodes of the networks depicted on the left with three forms of differential recruitment of magnitude two ($\phi = 2$).

strictly consider random walks on fully connected undirected networks where self ties are not permitted, a standard assumption for RDS, and assume all ϕ's are greater than zero. Finally, we assume a finite network to ensure the MC is positive recurrent. If those conditions are met, then there exists a unique stationary distribution for each of those stochastic processes.

Result 4.1. *Let RW_t denote the state at step t of a MC on the nodes of a fully connected undirected network without self ties. Assume that there exists at least one $y_{ij} = 1$ such that $x_i \in \mathcal{X}^1$ and $x_j \in \mathcal{X}^0$ and that this MC has the following transition probabilities (i.e. between group differential recruitment):*

$$p_{ij} = \frac{(\phi_b^{RW} x_j + (1 - x_j))y_{ij}}{\sum_{j=1}^{N}(\phi_b^{RW} x_j + (1 - x_j))y_{ij}}, \tag{4.4}$$

where $\phi_b^{RW} > 0$. Then the stationary distribution of this random walk is such that:

$$\pi_i \propto d_i^b = (\phi_b^{RW} x_i + (1 - x_i))(\phi_b^{RW} d_{i1} + d_{i0}) \quad \textit{for } i \in \{1, 2, ..., N\}. \tag{4.5}$$

Proof.
By assumption, $p_{ij} = \dfrac{(\phi_b^{RW} x_j + (1 - x_j))y_{ij}}{\sum_{j=1}^{N}(\phi_b^{RW} x_j + (1 - x_j))y_{ij}}$.Therefore, we have that:

$$\begin{aligned}
\sum_{i=1}^{N} \pi_i p_{ij} &= \sum_{i=1}^{N} \left[\frac{(\phi_b^{RW} x_i + (1 - x_i))(\phi_b^{RW} d_{i1} + d_{i0})}{K}\right] \left[\frac{(\phi_b^{RW} x_j + (1 - x_j))y_{ij}}{\sum_{j=1}^{N}(\phi_b^{RW} x_j + (1 - x_j))y_{ij}}\right] \\
&= \sum_{i=1}^{N} \left[\frac{(\phi_b^{RW} x_i + (1 - x_i))}{K}\right] (\phi_b^{RW} x_j + (1 - x_j))y_{ij} \\
&= \frac{(\phi_b^{RW} x_j + (1 - x_j))}{K} \sum_{i=1}^{N} (\phi_b^{RW} x_i + (1 - x_i))y_{ij} \\
&= \frac{(\phi_b^{RW} x_j + (1 - x_j))(\phi_b^{RW} d_{j1} + d_{j0})}{K} = \pi_j,
\end{aligned}$$

where K is a normalizing constant such that $\sum_{i=1}^{N} \pi_i = 1$. Therefore, π_i satisfies the global balance equations for all $i \in \{1, 2, ..., N\}$ and $\pi = \{\pi_1, \pi_2, ..., \pi_N\}$ is the stationary distribution for this RW. □

Result 4.2. *Let RW_t denote the state at step t of a MC on the nodes of a fully connected undirected network without self ties. Assume that there exists at least one $y_{ij} = 1$ such that $x_i \in \mathcal{X}^1$ and $x_j \in \mathcal{X}^0$ and that this MC has the following transition probabilities (i.e. within group differential recruitment):*

$$p_{ij} = \frac{(\phi_w^{RW} x_j + (1 - x_j))x_i y_{ij} + (\phi_w^{RW}(1 - x_j) + x_j)(1 - x_i)y_{ij}}{\sum_{j=1}^{N}(\phi_w^{RW} x_j + (1 - x_j))x_i y_{ij} + (\phi_w^{RW}(1 - x_j) + x_j)(1 - x_i)y_{ij}},$$

where $\phi_w^{RW} > 0$. Then the stationary distribution of this random walk is:

$$\pi_i \propto d_i^w = \ (\phi_w^{RW} x_i + (1 - x_i))d_{i1} + (\phi_w^{RW}(1 - x_i) + x_i)d_{i0} \tag{4.6}$$

for $i \in \{1, 2, ..., N\}$.

Proof.
By assumption, $p_{ij} = \dfrac{(\phi_w^{RW} x_j + (1 - x_j))x_i y_{ij} + (\phi_w^{RW}(1 - x_j) + x_j)(1 - x_i)y_{ij}}{\sum_{j=1}^{N}(\phi_w^{RW} x_j + (1 - x_j))x_i y_{ij} + (\phi_w^{RW}(1 - x_j) + x_j)(1 - x_i)y_{ij}}$.

Therefore, we have that:

$$\begin{aligned}
\sum_{i=1}^{N} \pi_i p_{ij} =& \sum_{i=1}^{N} \left[\frac{(\phi_w^{RW} x_i + (1 - x_i))d_{i1} + (\phi_w^{RW}(1 - x_i) + x_i)d_{i0}}{K}\right] \\
&\left[\frac{(\phi_w^{RW} x_j + (1 - x_j))x_i y_{ij} + (\phi_w^{RW}(1 - x_j) + x_j)(1 - x_i)y_{ij}}{\sum_{j=1}^{N}(\phi_w^{RW} x_j + (1 - x_j))x_i y_{ij} + (\phi_w^{RW}(1 - x_j) + x_j)(1 - x_i)y_{ij}}\right] \\
=& \sum_{i=1}^{N} \frac{(\phi_w^{RW} x_j + (1 - x_j))x_i y_{ij} + (\phi_w^{RW}(1 - x_j) + x_j)(1 - x_i)y_{ij}}{K} \\
=& \frac{(\phi_w^{RW} x_j + (1 - x_j))d_{j1} + (\phi_w^{RW}(1 - x_j) + x_j)d_{j0}}{K} = \pi_j,
\end{aligned}$$

where K is a normalizing constant such that $\sum_{i=1}^{N} \pi_i = 1$. Therefore, π_i satisfies the global balance equations for all $i \in \{1, 2, ..., N\}$ and $\pi = \{\pi_1, \pi_2, ..., \pi_N\}$ is the stationary distribution for this RW. □

Result 4.3. *Let RW_t denote the state at step t of a MC on the nodes of a fully connected undirected network without self ties. Assume that the MC has the following transition probabilities (i.e. tie attribute differential recruitment):*

$$p_{ij} = \frac{(\phi_t^{RW} w_{ij} + (1 - w_{ij}))y_{ij}}{\sum_{j=1}^{N}(\phi_t^{RW} w_{ij} + (1 - w_{ij}))y_{ij}},$$

where $\phi_t^{RW} > 0$. Then the stationary distribution of this random walk is:

$$\pi_i \propto d_i^t = \phi_t^{RW} w_{i1} + w_{i0} \quad i \in \{1, 2, ..., N\}. \tag{4.7}$$

Proof.

By assumption, $p_{ij} = \dfrac{(\phi_t^{RW} w_{ij} + (1 - w_{ij}))y_{ij}}{\sum_{j=1}^{N}(\phi_t^{RW} w_{ij} + (1 - w_{ij}))y_{ij}}$.

Therefore, we have that:

$$\begin{aligned}
\sum_{i=1}^{N} \pi_i p_{ij} &= \sum_{i=1}^{N} \left[\frac{(\phi_t^{RW} w_{i1} + w_{i0})}{K}\right] \left[\frac{(\phi_t^{RW} w_{ij} + (1 - w_{ij}))y_{ij}}{\sum_{j=1}^{N}(\phi_t^{RW} w_{ij} + (1 - w_{ij}))y_{ij}}\right] \\
&= \sum_{i=1}^{N} \frac{(\phi_t^{RW} w_{ij} + (1 - w_{ij}))y_{ij}}{K} \\
&= \frac{\phi_t^{RW} w_{j1} + w_{j0}}{K} = \pi_j,
\end{aligned}$$

where K is a normalizing constant such that $\sum_{i=1}^{N} \pi_i = 1$. Therefore, π_i satisfies the global balance equations for all $i \in \{1, 2, ..., N\}$ and $\pi = \{\pi_1, \pi_2, ..., \pi_N\}$ is the stationary distribution for this RW. □

The resulting stationary distributions all involve the ϕ parameters which are generally unknown since the sampling is driven by the respondents. However, these parameters may be estimated by maximizing the following likelihood functions:

$$\begin{aligned}
L(\phi | G = g) &\propto \prod_{i \in S^1 \setminus S_0} p(G_i = g_i | G_{i-1} = g_{i-1}, \phi), \\
&= \prod_{i \in S^1 \setminus S_0} p_{g_{i-1} g_i},
\end{aligned} \tag{4.8}$$

where:

G : n-dimensional vector of random variables specifying nodes' sampling order.

S_0 : set of seeds.

p_{ij} : transition probabilities between node i and node j for the given form of differential recruitment.

The resulting estimate for ϕ's may be replaced in the stationary distributions so that the estimated stationary distributions for node $i \in \{1, 2, ..., N\}$ in equations (4.5), (4.6) and (4.7) respectively become proportional to:

$$\widehat{d_i^b} = (\widehat{\phi_b^{RW}} x_i + (1 - x_i))(\widehat{\phi_b^{RW}} d_{i1} + d_{i0}), \tag{4.9}$$

$$\widehat{d_i^w} = (\widehat{\phi_w^{RW}} x_i + (1 - x_i))d_{i1} + (\widehat{\phi_w^{RW}}(1 - x_i) + x_i)d_{i0}, \text{ and} \tag{4.10}$$

$$\widehat{d_i^t} = \widehat{\phi_t^{RW}} w_{i1} + w_{i0}. \tag{4.11}$$

4.2.3 Extended Design-Based Estimators

Obtaining the extended version of $\hat{\mu}_{VH}$ is straightforward. The only modification to the original estimator consists in replacing the sampling probabilities by the appropriate RW estimated stationary distribution. For instance, in the case of between group differential recruitment, the extended estimator is:

$$\hat{\mu}^b_{VH.dr} = \frac{\sum_{i=1}^N S_i z_i / \widehat{d_i^b}}{\sum_{i=1}^N S_i / \widehat{d_i^b}}. \tag{4.12}$$

In addition, similar to equation (2.16), the extended version of $\hat{\mu}^{ego}_{SH}$ may be expressed as a function of the corresponding $\hat{\mu}_{VH.dr}$. For instance, in the case of between group differential recruitment, we have that:

$$\hat{\mu}^b_{SH.dr} = \frac{\hat{\mu}^b_{VH.dr}}{\hat{\mu}^b_{VH.dr} + c^b\ (1 - \hat{\mu}^b_{VH.dr})}, \quad \text{where} \ \ c^b = \frac{n_1}{n_0} \frac{\sum_{i=1}^N S_i x_i d_{i0} / \widehat{d^b_i}}{\sum_{i=1}^N \widehat{\phi}^{RW}_b S_i (1 - x_i) d_{i1} / \widehat{d^b_i}} \tag{4.13}$$

The extended design-based estimators for all forms of differential recruitment are summarized in Table 4.2. We note that for all estimators, the ego-network composition of every participant i, that is, d_{i1} and d_{i0} for between and within group and w_{i1} and w_{i0} for tie attribute differential recruitment, are necessary to compute the estimate. This information has not traditionally been collected in RDS surveys, but an increasing number of studies now include this information [Liu et al., 2009, 2012]. Furthermore, for the tie attribute differential recruitment, information about $z_j w_{ij}$ and $(1 - z_j) w_{ij}$ also need to be collected for every individual i in the sample. These data represent the allocation of preferred ties by outcome variable.

Table 4.2: Summary of RDS design-based estimators under various recruitment regimes

	Differential Recruitment Form		
	Between Group	Within Group	Tie Attribute
$\pi_i \propto$	$d_i^b = (\phi_b^{RW})^{x_i}(\phi_b^{RW} d_{i1} + d_{i0})$	$d_i^w = d_{i1}(\phi_w^{RW})^{x_i} + d_{i0}(\phi_w^{RW})^{1-x_i}$	$d_i^t = \phi_t^{RW} w_{i1} + w_{i0}$
Estimator	$\hat{\mu}_{VH.dr}^b = \dfrac{\sum_{i=1}^n S_i z_i / \widehat{d_i^b}}{\sum_{i=1}^n S_i / \widehat{d_i^b}}$	$\hat{\mu}_{VH.dr}^w = \dfrac{\sum_{i=1}^n S_i z_i / \widehat{d_i^w}}{\sum_{i=1}^n S_i / \widehat{d_i^w}}$	$\hat{\mu}_{VH.dr}^t \dfrac{\sum_{i=1}^n S_i z_i / \widehat{d_i^t}}{\sum_{i=1}^n S_i / \widehat{d_i^t}}$
Estimator	$\hat{\mu}_{SH.dr}^b = \dfrac{\hat{\mu}_{VH}}{\hat{\mu}_{VH} + c^b(1 - \hat{\mu}_{VH})}$	$\hat{\mu}_{SH.dr}^w = \dfrac{\hat{\mu}_{VH}}{\hat{\mu}_{VH} + c^w(1 - \hat{\mu}_{VH})}$	$\hat{\mu}_{SH.dr}^t = \dfrac{\hat{\mu}_{VH}}{\hat{\mu}_{VH} + c^t(1 - \hat{\mu}_{VH})}$
c^b, c^w and c^t	$\dfrac{n_1}{n_0} \dfrac{\sum_{i=1}^n S_i x_i d_{i0} / \widehat{d_i^b}}{\sum_{i=1}^n \widehat{\phi}_b^{RW} S_i (1 - x_i) d_{i1} / \widehat{d_i^b}}$	$\dfrac{n_1}{n_0} \dfrac{\sum_{i=1}^n S_i x_i d_{i0} / \widehat{d_i^w}}{\sum_{i=1}^n S_i (1 - x_i) d_{i1} / \widehat{d_i^w}}$	$\dfrac{n_1}{n_0} \dfrac{\sum_{i=1}^n S_i z_i \sum_{j=1}^N (1 - z_j)(\widehat{\phi}_t^{RW})^{w_{ij}} y_{ij} / \widehat{d_i^t}}{\sum_{i=1}^n S_i (1 - z_i) \sum_{j=1}^N z_j (\widehat{\phi}_t^{RW})^{w_{ij}} y_{ij} / \widehat{d_i^t}}$

4.3 Bayesian Inference - Successive Sampling Approximation

One of the main limitations of our design-based estimators presented in the previous section is that the estimation of ϕ's does poorly with RDS data in the presence of an important form of dependency in social networks, that is, homophily. Homophily is a network property which arises when members of the target population form ties with alike members more frequently than with other members of the population. For instance, we say that there is homophily on X if the probability of Y_{ij} being equal to one is greater when $x_i = x_j$ than when $x_i \neq x_j$.

In this Section, we propose a model-based framework to estimating the prevalence of an outcome variable Z for RDS samples collected with differential recruitment. Our approach extends the Bayesian methodology proposed by West [1996] and later extended to RDS data by Handcock et al. [2014]. In their work, the authors leverage the information about the order in which the items are sampled as well as the observed unit sizes to make inference about the target population size N when data are collected exactly or approximately from an SS process. Our work is similar in that respect with the exception that N is presumed known and the object of inference is instead the prevalence μ. The key contributions of our work lie in our choice of:

1. definition of unit sizes; and

2. super population model responsible for the distribution of those unit sizes.

Together, these choices allow for both network homophily and differential recruitment to be captured and estimated.

In Section 4.3.1 we describe the likelihood function for the super-population model parameters η as well as the sampling parameter ϕ. It is followed in Section 4.3.2 by a description of the proposed Bayesian framework to estimate those parameters. The methodology is developed assuming between group differential recruitment, which is denoted ϕ throughout this section for simplicity of notation. Also, it is presumed that the characteristic inducing differential recruitment is the outcome of interest Z.

4.3.1 Likelihood For the Network and Sampling Parameters

Under a model-based inference framework, the observations are assumed to be the realization of a super population model. The parameter of that model, η, may be estimated through likelihood inference. In the case where the population variable V is fully observed then the likelihood for η is:

$$L(\eta|V) \ \propto p(V=v|\eta). \tag{4.14}$$

In the situation at hand, a super population model is posited for the participants' degree, $D=\{D_1,D_2,...,D_N\}$ and their outcome variable $Z=\{Z_1,Z_2,...,Z_N\}$, that is, $V=(D,Z)$. The set of parameters for the super population model is $\eta=(\Gamma,\mu)$, where Γ is the vector of parameters for D and μ is the parameter for Z. Therefore, for a fully observed degree distribution and outcome variable, the likelihood function in equation (4.14) becomes:

$$L(\Gamma,\mu|D,Z) \ \propto p(D=d,Z=z|\Gamma,\mu). \tag{4.15}$$

In the current problem however, the participants' degree and outcome variable are only partially observed since RDS studies typically sample a fraction of the entire network. Furthermore, as we discuss in Section 4.3.1.1, the data are not missing at random (NMAR). Consequently, likelihood inference must include a missing data mechanism [Little and Rubin, 2002]. Under the SS approximation to RDS, the SS process is the mechanism responsible for the missing data. Therefore, similarly to Handcock et al. [2014], the likelihood which reflects this missing data process is:

$$
\begin{aligned}
L(\Gamma, \mu, \phi | G = g, V^{obs} = v^{obs}) &\propto \; p(G = g, V^{obs} = v^{obs} | \Gamma, \mu, \phi) \\
&= \sum_{v^{unobs} \in \mathcal{V}(v^{obs})} p(G = g, V = (v^{obs} + v^{unobs}) | \Gamma, \mu, \phi) \\
&= \sum_{v^{unobs} \in \mathcal{V}(v^{obs})} p(G = g | V = (v^{obs} + v^{unobs}), \phi) \; p(V = (v^{obs} + v^{unobs}) | \Gamma, \mu),
\end{aligned}
\tag{4.16}
$$

where

1. $G = (G_1, G_2, ..., G_n)$ denotes the random variable indicating the items' sampling order and $g = (g_1, g_2, ..., g_n)$ is its realized valued. In the remainder of this chapter, to simplify the notation and without loss of generality, we assume that $g = (g_1, g_2, ..., g_n)$ is equal to (1, 2, ..., n).

2. $V^{obs} = (D^{obs}, X^{obs})$ and $V^{unobs} = (D^{unobs}, X^{unobs})$ denote the observed and unobserved portion of the degree distribution and outcome variable, respectively. Also, $V = V^{obs} + V^{unobs} = (D^{obs} + D^{unobs}, X^{obs} + X^{unobs})$.

3. $\mathcal{V}(v^{obs})$ is the set of populations of size N consistent with the observed degrees and outcomes.

As noted by Handcock et al. [2014], the high dimension of $v^{unobs} \in \mathcal{V}(v^{obs}, N)$ often makes it impractical to perform likelihood inference. Therefore, we augment the data and also carry out Bayesian inference to simultaneously estimate the superpopulation model and sampling parameters. All parameters in the model with the exception of μ are nuisance parameters.

4.3.1.1 Successive Sampling With Differential Recruitment

The SS sampling estimator developed by Gile [2011] is based on a network configuration model [Molloy and Reed, 1995]. Under such model, the degree distribution is fixed and pairs of edge-ends are randomly attached. For instance, in Figure 4.2, "A" could be paired at random with any edge-ends in $\{B, C, ..., H\}$ to form a tie.

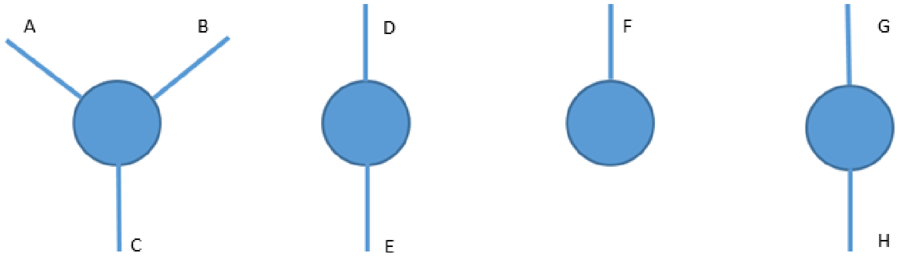


Figure 4.2: Example of edge-ends in a configuration network model.

The author argues that a self-avoiding random walk on the nodes marginalized over all networks generated by this model has the following transition probabilities:

$$\begin{aligned} P(G_i = i|G_1, G_2, ..., G_{i-1} = (1, 2, ..., i-1), D = d) \\ = \begin{cases} \dfrac{d_i}{\sum_{j=i}^{N} d_j} & i \notin \{1, 2, ..., i-1\} \\ 0 & i \in \{1, 2, ..., i-1\} \end{cases} \end{aligned} \tag{4.17}$$

which is equivalent to a successive sampling process with unit size equal to the degree of the individuals in the population. Therefore, this justifies why in absence of differential recruitment the degrees are commonly used for unit size.

Alternative definitions of unit sizes may however be used to reflect various recruitment patterns. Consider instead a self-avoiding RW on the nodes of all networks from a configuration model with between group differential recruitment of magnitude ϕ as defined in the equation below:

$$\phi = \frac{P(G_i = i|\ G_1, G_2, ..., G_{i-1} = (1, 2, ..., i-1), z_i = 1)}{P(G_i = i|\ G_1, G_2, ..., G_{i-1} = (1, 2, ..., i-1), z_i = 0)}. \tag{4.18}$$

This sampling process yields the following transition probabilities:

$$P(G_i = i | G_1, G_2, ..., G_{i-1} = (1, 2, ..., i-1), Z = z, D = d, \phi) \\ = \begin{cases} \dfrac{\phi^{z_i} d_i}{\sum_{j=i}^{N} \phi^{z_j} d_j} & i \notin \{1, 2, ..., i-1\} \\ 0 & i \in \{1, 2, ..., i-1\}, \end{cases} \tag{4.19}$$

which is also equivalent to a SS process. The unit sizes of this SS process may be formulated as a function of the participants' degree, their outcome and the differential recruitment parameter ϕ such that:

$$u_i = h(d_i, z_i, \phi) = \phi^{z_i} d_i \tag{4.20}$$

$$= (\phi z_i + (1 - z_i)) d_i \qquad \forall i \in \{1, 2, ..., N\}. \tag{4.21}$$

Furthermore, the probability of observing a sequence of n units under this SS process is as follows:

$$P(G = (1, 2, ..., n) | Z = z, D = d, \phi) \\ = \frac{N!}{(N-n)!} \prod_{i=1}^{n} P(G_i = i | G_1, G_2, ..., G_{i-1} = (1, 2, ..., i-1), Z = z, D = d, \phi) \\ = \frac{N!}{(N-n)!} \prod_{i=1}^{n} \frac{\phi^{z_i} d_i}{\sum_{j=i}^{N} \phi^{z_j} d_j}, \tag{4.22}$$

which may be rewritten as:

$$P(G = (1, 2, ..., n) | V = v, \phi) = \frac{N!}{(N-n)!} \prod_{i=1}^{n} \frac{\phi^{z_i} d_i}{\sum_{j=i}^{n} \phi^{z_j} d_j + \sum_{j=n+1}^{N} \phi^{z_j} d_j} \tag{4.23}$$

to emphasize that this expression depends on the unobserved unit sizes through $\sum_{j=n+1}^{N} \phi^{z_j} d_j$. This demonstrates that data collected under this sampling design are NMAR since the following condition is not satisfied:

$$P(G = (1, 2, ..., n)|V = v, \phi) = P(G = (1, 2, ..., n)|V^{obs} = v^{obs}, \phi). \tag{4.24}$$

Therefore, the missing data mechanism is nonignorable. This confirms that the sampling process needs to be incorporated in the model when performing inference about the population parameters.

4.3.1.2 Super-population Model

Under a model-based framework, the observations are presumed to be realization of a super-population model with unknown parameters. The purpose of the inference is to estimate those parameters. In the methodology developed by Handcock et al. [2014], the participants' degrees, which are used as unit sizes, are assumed to be generated by a degree distribution. However, in our extension of their methodology, the unit sizes are a function of both the participants' degree D and their outcome variable Z as shown in equation (4.22). Therefore, our super-population model must jointly model these variables. In this section, we justify our selected model for D and Z which is denoted $f(D, Z|\eta)$.

The core objective motivating the development of this prevalence estimator is to account for network homophily in the presence of differential recruitment. Consequently, the super-population model is specifically designed to capture the dependency between D and Z in a way that reflects this network feature. In particular, the degree distribution conditional on the nodal attribute is derived from an exponential-family random graph model (ERGM) [Frank and Strauss, 1986, Hunter et al., 2008, Hunter and Handcock, 2006] including a term for network homophily. This approach contrasts with the work of Handcock et al. [2014] in which the degrees are modeled independently of any nodal characteristics.

Under our model specification, the probability of observing a tie between node i and j is as follows:

$$P(Y_{ij} = 1 | z_i = z_j) = \gamma \tag{4.25}$$

$$P(Y_{ij} = 1 | z_i \neq z_j) = \gamma_{10}, \tag{4.26}$$

where γ and γ_{10} are the rate of ties among alike nodes and the rate of cross-ties, respectively. Equivalently, conditional on the outcome variable Z, the ties are i.i.d. Bernoulli trials with parameters γ or γ_{10}.

Figure 4.3 depicts an empty socio-matrix that has been partitioned into four regions based on the outcome z_i's displayed in the margins. This network contains twelve nodes, out of which seven nodes have a positive outcome. Based on this network model, the probability that any off-diagonal entries in quadrant I and IV are equal to one is γ since those regions are connecting alike members of the population. The entries in the shaded diagonal are always set to zero since this network model does not allow ties to self. Similarly, entries in regions II and III have a probability γ_{10} to be equal to one since they represent ties among nodes with dissimilar outcome.

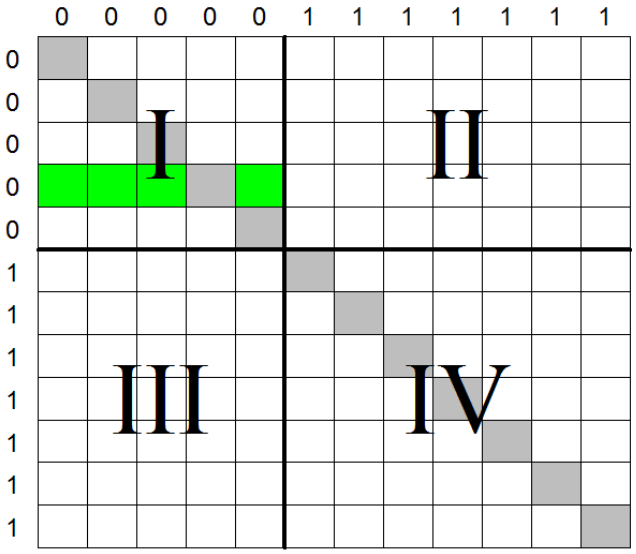


Figure 4.3: Socio-matrix

It is possible to derive the underlying degree distribution for this network model. All ties are presumed independent. Since the rates of ties differ by quadrant, the degree distribution is developed by decomposing the distribution into four pieces which corresponds to the quadrants displayed in Figure 4.3, so that:

$$\begin{aligned} f(D|Z,\Gamma) = \; & p(D_{00} = d_{00}|Z = z, \gamma)\; p(D_{11} = d_{11}|Z = z, \gamma) \\ & p(D_{10} = d_{10}|D_{01} = d_{01}, Z = z, \gamma_{10})\; p(D_{01} = d_{01}|Z = z, \gamma_{10}), \end{aligned} \tag{4.27}$$

where $\Gamma = (\gamma, \gamma_{10})$ and D_{kl}'s are N-dimensional vectors such that for $k, l \in \{0, 1\}$ the i-th element of this vector ($[D_{kl}]_i$) is equal to:

$$D_{i,kl} = \sum_{j=1}^{N} Y_{ij} \left[kZ_i + (1-k)(1-Z_i)\right] \left[lZ_j + (1-l)(1-Z_j)\right]. \tag{4.28}$$

The quantity $D_{i,kl}$ essentially represents the number of ties node i has with any member j in the target population such that $z_j = l$ and $z_i = k$.

Starting with the first quadrant in Figure 4.3, we illustrate how the probability distribution for D_{00} is obtained. The degree of the fourth node, for instance, to nodes with covariate equal to zero, corresponds to the sum of the ties in the green highlighted cells. More specifically, $D_{4,00} = \sum_{j=1}^{12} Y_{4j}(1-Z_4)(1-Z_j) = \sum_{j \in \{1,2,3,5\}} Y_{4j}$. Therefore, conditional on Z, $D_{4,00}$ is the summation of $N_0 - 1$ independent Bernoulli trials with probability γ, where

$$N_k = \sum_{i=1}^{N} kZ_i + (1-k)(1-Z_i) \text{ for } k \in \{0, 1\}. \tag{4.29}$$

Simply put, N_k is the number of nodes in $\mathcal{Z}^k = \{i : z_i = k\}$. It follows that the probability distribution for D_{00} is:

$$p(D_{00} = d_{00}|Z = z, \gamma) = \prod_{i=1}^{N} \left[\binom{N_0 - 1}{d_{i0}} \gamma^{d_{i0}} (1-\gamma)^{N_0 - 1 - d_{i0}} \right]^{1-z_i}, \tag{4.30}$$

where d_{i0} is the number of ties to individuals with $z_j = 0$ regardless of the outcome z_i or more generally, for $k, l \in \{0, 1\}$ and for $i \in \{1, 2, ..., N\}$

$$d_{ik} = \sum_{j=1}^{N} y_{ij}(kz_j + (1-k)(1-z_j)). \tag{4.31}$$

In other words, the probability distribution for D_{00} conditional on Z is simply the product of N_0 Binomial distributions. The same reasoning may be applied to develop the distribution for the vectors D_{11} and D_{01}. However, an additional constraint is imposed on the distribution for D_{10}. In equation (4.27) the distribution of D_{10} is conditional on D_{01}. This constraint ensures that the number of ties in quadrant II of a network is the same as the number of ties in its third quadrant. Although this does not guarantee a symmetric network this constraint preserves some aspect of the symmetry, that is, the total number of ties. This constraint implies that the degrees in the third quadrant follows a multivariate hypergeometric distribution such that:

$$p(D_{10} = d_{10} | D_{01} = d_{01}, Z = z, \gamma_{10}) = \frac{\prod_{i=1}^{N} \binom{N_0}{d_{i0}}^{z_i}}{\binom{N_1 N_0}{t_{10}}}, \text{ where } t_{10} = \sum_{i=1}^{N} [d_{01}]_i \,. \tag{4.32}$$

In summary, the degree distribution conditional on Z is provided below:

$$\begin{aligned} f(D|Z,\Gamma) = \; & p(D_{00} = d_{00}|Z = z, \Gamma)\ p(D_{11} = d_{11}|Z = z, \Gamma) \\ & p(D_{10} = d_{10}|D_{01} = d_{01}, Z = z, \gamma_{10})\ p(D_{01} = d_{01}|Z = z, \gamma_{10}) \\ = \; & \binom{N_1 N_0}{t_{10}}^{-1} \prod_{i=1}^{N} \left[\binom{N_1 - 1}{d_{i1}} \gamma^{d_{i1}} (1-\gamma)^{N_1 - 1 - d_{i1}} \binom{N_0}{d_{i0}} \right]^{z_i} \\ & \left[\binom{N_0 - 1}{d_{i0}} \gamma^{d_{i0}} (1-\gamma)^{N_0 - 1 - d_{i0}} \binom{N_1}{d_{i1}} \gamma_{10}^{d_{i1}} (1-\gamma_{10})^{N_1 - d_{i1}} \right]^{1-z_i}, \end{aligned} \tag{4.33}$$

Thus far we have discussed the probability distribution $f(D|Z, \Gamma)$. However, the complete super-population model is the joint distribution $f(D, Z|\eta) = f(D|Z, \Gamma) f(Z|\mu)$.

Therefore, a model for the outcome variable Z needs to be formulated, where Z is a vector of binary variables. The outcome variables are assumed to be i.i.d. Bernoulli trials with parameter μ such that $Z_i \overset{iid}{\sim} Bernoulli(\mu)$. Although each random variable Z_i is presumed independent, the stochastic mechanism responsible for generating the network captures the tendency of alike nodes to preferentially attach. It is noteworthy to emphasize that the rate μ is the primary object of inference.

4.3.2 Full Conditional Distributions For Gibbs Sampler

The full conditional posterior distributions for the parameters μ, Γ, ϕ, Z^{unobs} and $D^{unobs} = (D_{00}^{unobs}, D_{01}^{unobs}, D_{10}^{unobs}, D_{11}^{unobs})$ are developed in this section. Their distributions rely on the sampling and super-population models discussed in Sections 4.3.1.1 and 4.3.1.2. However, obtaining a straightforward expression for some of the posterior distributions is not possible and therefore, the data are augmented to circumvent this issue. In this section, we begin by describing the data augmentation technique used by West [1996] and Handcock et al. [2014] in similar modeling settings. Then, we present the iterative steps of the Gibbs sampler algorithm and the derivation of the posterior distributions. Finally, we discuss our choice of prior distributions.

4.3.2.1 Data Augmentation

Data augmentation is used in order to simplify the denominator included in the SS model stated in equation (4.22), that is:

$$\prod_{i=1}^{n} r_i = \prod_{i=1}^{n} \sum_{j=i}^{N} \phi^{z_j} d_j = \prod_{i=1}^{n} \sum_{j=i}^{N} (\phi z_j + (1 - z_j)) d_j. \tag{4.34}$$

The r_i terms in this expression represent the remainder of the unsampled units when $i - 1$ units have been sampled. West [1996] observed that augmenting the sampling model by a series of n exponential random variables ψ_i's with parameter r_i having the following density function:

$$f_{\Psi_i}(\psi_i|r_i) = r_i e^{-r_i\psi_i}, \text{ where } r_i > 0 \tag{4.35}$$

yields a simpler model to manipulate. In our proposed methodology, this model is as follows:

$$\begin{aligned} P(\Psi = (\psi_1, \psi_2, ..., \psi_n), G = (1, 2, ..., n) | Z = z, D = d, \phi) \\ = \ & P(\Psi = (\psi_1, \psi_2, ..., \psi_n) | G = (1, 2, ..., n), Z = z, D = d, \phi) \\ & P(G = (1, 2, ..., n) | Z = z, D = d, \phi) \\ = \ & \frac{N!}{(N-n)!} \prod_{i=1}^{n} e^{-r_i\psi_i} \phi^{z_i} d_i. \end{aligned} \tag{4.36}$$

Consequently, the denominator term depending on r_i's is now absent in the resulting augmented sampling model. However, this simplification is at the expense of an additional component for Ψ's in the Gibbs sampler which is described in Section 4.3.2.2.

4.3.2.2 Gibbs Sampler

In this section we describe the Gibbs sampler designed to sample from the augmented joint posterior below:

$$p(\mu, \Gamma, \phi, Z^{unobs}, D^{unobs}, \Psi | V^{obs}, G = g) \tag{4.37}$$

It has a total of six components. The algorithm is similar to the one proposed by West [1996] and extended by Handcock et al. [2014]. The main differences are that:

1. the total population size N is presumed known; and
2. the unit sizes distribution captures network homophily and differential recruitment. Therefore three Gibbs components are necessary for this distribution

instead of one: unobserved degrees, unobserved outcome variable and the differential recruitment parameter ϕ.

3. the main object of inference is the prevalence μ.

The full conditional posterior distributions for each parameter are derived within each step of the Gibbs sampler below. The mean of the posterior distribution for μ is used to estimate the prevalence of the outcome variable and is denoted $\hat{\mu}^b_{SS.dr}$.

1. Initialize z_i^{unobs}, d_{i0}^{unobs} and d_{i1}^{unobs} for all $i \in \{n+1, n+2, ..., N\}$

2. Sample μ from (4.39):

$$\begin{aligned} & p(\mu|Z=z, D=d, G=g, \phi, \Gamma, \Psi) \propto \ \pi(\mu)p(Z=z|\mu) \\ & \propto \ \pi(\mu)\mu^{\sum_{i=1}^N z_i}(1-\mu)^{N-\sum_{i=1}^N z_i} \\ & \therefore \ \mu|Z=z, D=d, G=g, \phi, \Gamma, \Psi \sim \ \text{beta}\left(a_\mu + \sum_{i=1}^N z_i, b_\mu + N - \sum_{i=1}^N z_i\right) \qquad (4.38) \\ & \text{when } \mu \sim \ \text{beta}\,(a_\mu, b_\mu) \end{aligned}$$

3. Sample γ from (4.40) and γ_{10} from (4.41):

$$\begin{aligned} \text{(i)} \ & p(\gamma|Z=z, D=d, G=g, \mu, \phi, \gamma_{10}, \Psi) \propto \ \pi(\gamma) \ p(D=d|Z=z, \gamma) \\ & = \pi(\gamma)\prod_{i=1}^N \left[\gamma^{d_{i1}}(1-\gamma)^{N_1-1-d_{i1}}\right]^{z_i}\left[\gamma^{d_{i0}}(1-\gamma)^{N_0-1-d_{i0}}\right]^{1-z_i} \\ & = \pi(\gamma) \ \gamma^{\sum_{i=1}^N d_{i1}z_i + d_{i0}(1-z_i)} \ (1-\gamma)^{N_1(N_1-1)+N_0(N_0-1)-\sum_{i=1}^N d_{i1}z_i - \sum_{i=1}^N d_{i0}(1-z_i)} \\ & \therefore \gamma|Z=z, D=d, G=g, \mu, \phi, \gamma_{10}, \Psi \sim \text{beta}\,(\alpha_\gamma, \beta_\gamma) \qquad (4.39) \\ & \text{where } \alpha_\gamma = a_\gamma + \sum_{i=1}^N (1-z_i)d_{i0} + \sum_{i=1}^N z_i d_{i1} \\ & \qquad \beta_\gamma = b_\gamma + N_0(N_0-1) + N_1(N_1-1) - \sum_{i=1}^N (1-z_i)d_{i0} - \sum_{i=1}^N z_i d_{i1} \\ & \text{when } \gamma \sim \ \text{beta}\,(a_\gamma, b_\gamma) \end{aligned}$$

(ii) $p(\gamma_{10}|Z=z, D=d, G=g, \mu, \phi, \gamma, \Psi) \propto\ \pi(\gamma_{10})\ p(D=d|Z=z,\gamma)$

$$= \pi(\gamma_{10}) \prod_{i=1}^{N} \left[\gamma_{10}^{d_{i1}}(1-\gamma_{10})^{N_1-d_{i1}}\right]^{1-z_i}$$

$$= \pi(\gamma_{10})\ \gamma_{10}^{\sum_{i=1}^{N} d_{i1}(1-z_i)}\ (1-\gamma_{10})^{N_0N_1-\sum_{i=1}^{N} d_{i1}(1-z_i)}$$

$$\therefore \gamma_{10}|Z=z, D=d, G=g, \mu, \phi, \gamma, \Psi \sim \text{beta}\left(\alpha_{\gamma_{10}}, \beta_{\gamma_{10}}\right) \qquad (4.40)$$

where $\alpha_{\gamma_{10}} = a_{\gamma_{10}} + \sum_{i=1}^{N}(1-z_i)d_{i1}$ and $\beta_{\gamma_{10}} = b_{\gamma_{10}} + N_1N_0 - \sum_{i=1}^{N}(1-z_i)d_{i1}$

when $\gamma_{10} \sim$ beta$\left(a_{\gamma_{10}}, b_{\gamma_{10}}\right)$

4. Sample ψ_i for $i \in \{1, 2, ..., n\}$ from

$$\psi_i|Z=z, D=d, G=g, \mu, \phi, \Gamma \sim\ \exp\big(r_i\big),\ \text{where } r_i = \sum_{j=i}^{N} \phi^{z_j} d_j \qquad (4.41)$$

5. Sample joint unobserved degrees $D_1^{unobs} = (D_{11}^{unobs}, D_{01}^{unobs})$ and $D_0^{unobs} = (D_{00}^{unobs}, D_{10}^{unobs})$ from (4.42),(4.43),(4.44) and (4.45), respectively.

$$p(D_0^{unobs} = d_0^{unobs}, D_1^{unobs} = d_1^{unobs}|Z=z, D^{obs}=d^{obs}, G=g, \phi, \Gamma, \Psi)$$

$$\propto\ p(\Psi|Z=z, D=d, G=g, \phi)\ p(G=g|Z=z, D=d, \phi)$$
$$p(D=d|Z=z,\Gamma)$$

$$\propto\ \prod_{j=1}^{n}\left[r_j e^{-r_j\psi_j}\right]\left[\frac{\phi^{z_i}d_j}{r_j}\right] p(D_{00}=d_{00}|Z=z,\gamma)p(D_{11}=d_{11}|Z=z,\gamma)$$
$$p(D_{10}=d_{10}|D_{01}=d_{01}, Z=z, \gamma_{10})p(D_{01}=d_{01}|Z=z,\gamma_{10})$$

$$\propto\ \prod_{j=1}^{n}\left[e^{-\psi_j\sum_{i=j}^{N}(\phi z_i+(1-z_i))(d_{i1}+d_{i0})}\right]\prod_{i=1}^{N}\left[\binom{N_1-1}{d_{i1}}\gamma^{d_{i1}}(1-\gamma)^{N_1-1-d_{i1}}\binom{N_0}{d_{i0}}\right]^{z_i}$$
$$\left[\binom{N_0-1}{d_{i0}}\gamma^{d_{i0}}(1-\gamma)^{N_0-1-d_{i0}}\binom{N_1}{d_{i1}}\gamma_{10}^{d_{i1}}(1-\gamma_{10})^{N_1-d_{i1}}\right]^{1-z_i}$$

$$\propto\ \prod_{i=n+1}^{N}\left[e^{-(\phi z_i+(1-z_i))(d_{i1}+d_{i0})\sum_{j=1}^{n}\psi_j}\right]\left[\binom{N_1-1}{d_{i1}}\gamma^{d_{i1}}(1-\gamma)^{N_1-1-d_{i1}}\binom{N_0}{d_{i0}}\right]^{z_i}$$
$$\left[\binom{N_0-1}{d_{i0}}\gamma^{d_{i0}}(1-\gamma)^{N_0-1-d_{i0}}\binom{N_1}{d_{i1}}\gamma_{10}^{d_{i1}}(1-\gamma_{10})^{N_1-d_{i1}}\right]^{1-z_i}$$

$$
\begin{aligned}
\propto \prod_{i=n+1}^{N} & \left[\binom{N_1-1}{d_{i1}} (\gamma e^{-\phi\sum_{j=1}^{n} \psi_j})^{d_{i1}} (1-\gamma)^{N_1-1-d_{i1}} \binom{N_0}{d_{i0}} e^{-\phi d_{i0}\sum_{j=1}^{n}\psi_j}\right]^{z_i} \\
& \left[\binom{N_0-1}{d_{i0}}(\gamma e^{-\sum_{j=1}^{n}\psi_j})^{d_{i0}}(1-\gamma)^{N_0-1-d_{i0}}\right]^{1-z_i} \\
& \left[\binom{N_1}{d_{i1}}(\gamma_{10}e^{-\sum_{j=1}^{n} \psi_j})^{d_{i1}}(1-\gamma_{10})^{N_1-d_{i1}}\right]^{1-z_i}
\end{aligned}
$$

Therefore,

- $D_{i11}^{unobs}|Z=z, D=d, G=g, \mu, \phi, \Psi, \Gamma \ \sim \ Bin(N_1-1, \delta_{11}),$ (4.42)
  where $\delta_{11} = \dfrac{\gamma e^{-\phi\sum_{j=1}^{n}\psi_j}}{\gamma e^{-\phi\sum_{j=1}^{n}\psi_j} + (1-\gamma)}$
- $D_{i01}^{unobs}|Z=z, D=d, G=g, \mu, \phi, \Psi, \Gamma \ \sim \ Bin(N_1, \delta_{01}),$ (4.43)
  where $\delta_{01} = \dfrac{\gamma_{01} e^{-\sum_{j=1}^{n}\psi_j}}{\gamma_{01} e^{-\sum_{j=1}^{n}\psi_j} + (1-\gamma_{01})}$
- $D_{i00}^{unobs}|Z=z, D=d, G=g, \mu, \phi, \Psi, \Gamma \ \sim \ Bin(N_0-1, \delta_{00}),$ (4.44)
  where $\delta_{00} = \dfrac{\gamma e^{-\sum_{j=1}^{n}\psi_j}}{\gamma e^{-\sum_{j=1}^{n}\psi_j} + (1-\gamma)}$
- $D_{i10}^{unobs}|Z=z, D=d, G=g, \mu, \phi, \Psi, \Gamma \ \sim$
  $Hypergeometric_{(N_1-n_1)}(m = N_0 \cdot \mathbb{1}^{N_1-n_1}, N = T_{10} - t_{10}^{obs}),$ (4.45)
  where $n_1 = \sum_{i=1}^{n} z_i$ and $t_{10}^{obs} = \sum_{i=1}^{n} d_{i0} z_i$

6. Sample unobserved outcome variable Z^{unobs} from (4.46):

$$
\begin{aligned}
& p(Z^{unobs} = z^{unobs}|Z^{obs} = z^{obs}, D = d, G = g, \mu, \phi, \Gamma, \Psi) \\
& \quad \propto \ p(\Psi|Z = z, D = d, G = g, \phi) \ p(G = g|Z = z, D = d, \phi) \\
& \qquad p(D = d|Z = z, \Gamma) \ p(Z = z|\mu) \\
& \quad \propto \ \prod_{j=1}^{n} \left[r_j e^{-r_j\psi_j}\right] \left[\frac{\phi^{z_i} d_j}{r_j}\right] \prod_{i=1}^{N} \mu^{z_i}(1-\mu)^{1-z_i} \\
& \qquad p(D_{00} = d_{00}|Z = z, \gamma) \ p(D_{11} = d_{11}|Z = z, \gamma) \\
& \qquad p(D_{10} = d_{10}|D_{01} = d_{01}, Z = z, \gamma_{10}) \ p(D_{01} = d_{01}|Z = z, \gamma_{10})
\end{aligned}
$$

$$
\begin{aligned}
&\propto \prod_{j=1}^{n} \left[e^{-\psi_j \sum_{i=j}^{N} (\phi z_i + (1-z_i))(d_{i1}+d_{i0})} \right] \\
&\quad \binom{N_0 N_1}{T_{10}}^{-1} \prod_{i=1}^{N} \left[\mu \binom{N_1 - 1}{d_{i1}} \gamma^{d_{i1}} (1-\gamma)^{N_1 - 1 - d_{i1}} \binom{N_0}{d_{i0}} \right]^{z_i} \\
&\quad \left[(1-\mu) \binom{N_0 - 1}{d_{i0}} \gamma^{d_{i0}} (1-\gamma)^{N_0 - 1 - d_{i0}} \binom{N_1}{d_{i1}} \gamma_{10}^{d_{i1}} (1-\gamma_{10})^{N_1 - d_{i1}} \right]^{1-z_i} \\
&\propto \binom{N_0 N_1}{T_{10}}^{-1} \prod_{i=1}^{N} \left[\mu \binom{N_1 - 1}{d_{i1}} (\gamma e^{-\phi \sum_{j=1}^{n} \psi_j})^{d_{i1}} (1-\gamma)^{N_1 - 1 - d_{i1}} \binom{N_0}{d_{i0}} e^{-\phi d_{i0} \sum_{j=1}^{n} \psi_j} \right]^{z_i} \\
&\quad \left[(1-\mu) \binom{N_0 - 1}{d_{i0}} (\gamma e^{-\sum_{j=1}^{n} \psi_j})^{d_{i0}} (1-\gamma)^{N_0 - 1 - d_{i0}} \right]^{1-z_i} \\
&\quad \left[\binom{N_1}{d_{i1}} (\gamma_{10} e^{-\sum_{j=1}^{n} \psi_j})^{d_{i1}} (1-\gamma_{10})^{N_1 - d_{i1}} \right]^{1-z_i} \qquad (4.46)
\end{aligned}
$$

Therefore, since there is no closed form distribution for the joint distribution of the unobserved outcome, a Gibbs sampler is used to sample from this distribution one unobserved nodal outcome at a time. Contrary to traditional Gibbs samplers which sequentially draw from the posterior distribution of each variable, the outcome variable to be sampled are instead sequentially selected at random.

7. Sample ϕ from (4.47):

$$
\begin{aligned}
&p(\phi | Z = z, D = d, G = g, \mu, \Gamma, \Psi) \\
&\quad \propto\ p(\Psi | Z = z, D = d, G = g, \phi)\ p(G = g | Z = z, D = d, \phi)\ \pi(\phi) \\
&\quad \propto\ \prod_{i=1}^{n} \left[r_i e^{-\psi_i \sum_{j=i}^{N} z_j d_j \phi} \right] \left[\frac{\phi^{z_i}}{r_i} \right]\ \pi(\phi) \\
&\quad \propto\ \pi(\phi)\ \phi^{\sum_{i=1}^{n} z_i}\ \exp\left[-\sum_{i=1}^{n} \psi_i \sum_{j=i}^{N} z_j d_j \phi \right] \\
&\therefore\ \phi | Z = z, D = d, G = g, \phi, \Gamma, \Psi \sim \text{gamma}\Big(\alpha_\phi, \beta_\phi \Big) \qquad (4.47) \\
&\quad \text{where } \alpha_\phi = a_\phi + \sum_{i=1}^{n} x_i \text{ and } \beta_\phi = b_\phi + \sum_{i=1}^{n} \psi_i \sum_{j=1}^{N} x_j d_j \\
&\quad \text{when } \phi \sim\ \text{gamma}\, (a_\phi, b_\phi)
\end{aligned}
$$

8. Repeat steps (2) to (7) until convergence.

4.3.2.3 Prior Distributions

In the previous section, the selected prior distributions are shown in the derivation of the full conditional distributions. The prior distributions for μ, γ, γ_{10} and ϕ are chosen so that they are conjugate to their respective data models. In this section we describe how the prior parameters for these prior distributions were selected.

First, a vague prior is assumed for the parameter ϕ. The prior for that parameter is simply: $\phi \sim \text{ gamma}\,(a_\phi = 0.001, b_\phi = 0.001)$ which reflects the lack of prior knowledge of the participants' sampling preferences.

Second, a common approach is used in the selection of the prior parameters for μ, γ and γ_{10} since these three parameters all have a beta prior distribution and a binomial data model. An empirical Bayesian procedure is utilized to determine the prior means of the distributions. The prior mean is chosen so that it is equal to the design based estimate for the parameter in equations (4.13), (4.48) and (4.49). Furthermore, the prior variance is calculated to obtain a coefficient of variation of 10 for all priors. This prior selection has helped improve the stability and convergence of the algorithm.

$$\hat{\gamma} = \frac{\hat{\hat{d}}_{11}\hat{\theta} + \hat{\hat{d}}_{00}(1-\hat{\theta})}{\hat{\theta}(\hat{\theta}N - 1) + (1-\hat{\theta})((1-\hat{\theta})N - 1)} \tag{4.48}$$

$$\hat{\gamma}_{10} = \frac{\hat{\hat{d}}_{10}\hat{\theta} + \hat{\hat{d}}_{01}(1-\hat{\theta})}{2N\hat{\theta}(1-\hat{\theta})} \tag{4.49}$$

where $\hat{\theta} = \hat{\mu}^b_{SH.dr}$ and where $\hat{\hat{d}}_{lk} = \sum_{i=1}^n \dfrac{(kz_i + (1-k)(1-z_i))d_{i,kl}}{d_i^b}$.

4.4 Uncertainty of The Estimators

So far, we have discussed methodology to estimate the prevalence of an outcome of variable Z with RDS data when participants preferentially recruit individuals based on their characteristic or on their relationship with them. In this section we develop methodology to assess the uncertainty of the proposed estimators $\hat{\mu}_{VH.dr}$, $\hat{\mu}_{SH.dr}$ under all three forms of differential recruitment and of $\hat{\mu}^{b}_{SS.dr}$.

4.4.1 Design-Based Estimators

The variance estimator described in this section extends the SH-ego bootstrap procedure proposed by Lu [2013] summarized in Section 2.4.3.2. The revised methodology is designed to estimate the uncertainty of $\hat{\mu}_{VH.dr}$ or $\hat{\mu}_{SH.dr}$ under any of the three forms of differential recruitment discussed in this chapter. The two modifications to the existing methodology are as follows:

- **Sampling weights**. Similarly to the SH-ego bootstrap procedure, with the exception of the first node being selected completely at random, every resampled node is selected according to the estimated probability of transitioning from the group to which the recruiting node belongs to any other groups. For instance, suppose that the characteristic of the resampled node at step t is zero ($x_{i_t} = 0$). Furthermore, assume that the recruitment displays between group differential recruitment at an estimated rate of $\widehat{\phi}^{RW}_{b}$. Then, the probability of selecting a node such that $x_{i_{t+1}} = 1$ is equal to the average estimated proportion of cross recruitment given by:

$$\hat{p}^{b}_{(0,1)} = \frac{\sum_{i=1}^{N} S_i(1 - x_i)\widehat{\phi}^{RW}_{b} d_{i1}/\widehat{d^{b}_{i}}}{\sum_{i=1}^{N} S_i(1 - x_i)}. \tag{4.50}$$

  Table 4.3 summarizes all transition probabilities from any group k to any group l used in our proposed bootstrap variance estimator.

- **Estimates**. As usual, prevalence estimates are computed for all replicates. In the present case, the prevalence is determined based on the extended prevalence estimator for which the variability is estimated. For example, if the objective is to estimate the variability of $\hat{\mu}^b_{VH.dr}$, then the prevalence estimates are calculated with equation (4.12). Since the extended design-based estimators all depend on the estimated ϕ's, this quantity is re-evaluated for each replicate prior to calculating the resampled estimates. The determination of the resampled ϕ's is based on the characteristics of the resampled nodes. Therefore, information about the nodes' differential recruitment variable and ego-network compositions is recorded for every resampled nodes.

Table 4.3: Transition probabilities used in the bootstrap procedure for the extended design-based estimators under various recruitment regimes

DR Form	Transition Probability
Between Group	$p^b_{(k,l)} = \dfrac{\sum_{i=1}^N S_i \mathbb{1}_{(x_i=k)} (\widehat{\phi}_b^{RW})^l d_{il}/\widehat{d}_i^b}{\sum_{i=1}^N S_i \mathbb{1}_{(x_i=k)}}$
Within Group	$p^w_{(k,l)} = \dfrac{\sum_{i=1}^N S_i \mathbb{1}_{(x_i=k)} d_{il} (\phi_w^{RW})^{1-\lvert k-l\rvert}/\widehat{d_i^w}}{\sum_{i=1}^N S_i \mathbb{1}_{(x_i=k)}}$
Tie Attribute	$p^t_{(k,l)} = \dfrac{\sum_{i=1}^N S_i \mathbb{1}_{(z_i=k)} \sum_{j=1}^N \mathbb{1}_{(z_j=l)} (\widehat{\phi}_t^{RW})^{w_{ij}} y_{ij}/\widehat{d}_i^t}{\sum_{i=1}^N S_i \mathbb{1}_{(z_i=k)}}$

The resulting bootstrap procedure is intended to capture the uncertainty pertaining to the sampling process assuming a random walk approximation to RDS. Also, by recalculating ϕ for each replicate, we adjust for the variability of this parameter. However, neither the variability due to a super-population model nor the variability induced by other RDS-specific characteristics is reflected in this bootstrap estimator.

4.4.2 Bayesian Estimator

Deriving a variance estimator for $\hat{\mu}^b_{SS.dr}$ is straightforward. The posterior distribution for the parameter μ is obtained as part of the Bayesian estimation procedure. Therefore, measures of uncertainty may be obtained directly from a summary statistic of the posterior samples, such as the standard deviation. The standard deviation of the posterior samples reflects the sources of uncertainty modeled in the estimation framework, such as the uncertainty due to the super-population model and the successive sampling procedure. However, again, variability induced by other RDS-specific features is not reflected in this uncertainty estimator.

4.5 Simulation Study

4.5.1 Simulation Study Design

The complexity of the RDS sampling method prevents an analytical assessment of the performance of the proposed prevalence estimators. Therefore, we have designed a simulation study to compare their performance under a variety of sampling conditions and network features. In this section, we present the design and results from the simulation study consisting of the scenarios intended to capture randomness due to the population model and to the sampling. We also describe the tuning parameters of the MCMC. The simulation study was performed with the statistical software R and the packages `statnet` [Handcock et al., 2015b] and `RDS` [Handcock et al., 2015a].

4.5.1.1 Network Features

There is a vast body of literature studying the tendency of people to form ties with individuals with whom they share common attributes [Kandel, 1978, McPherson et al., 2001, Currarini et al., 2009]. Therefore, one of the primary objectives of this simulation study it to evaluate the sensitivity of the proposed methodology to this social behavior.

Exponential-family random graph model (ERGM) provides the flexibility to incorporate this feature. This may be done for instance by adding a homophily term which differentiates between the rate of ties among alike members, γ in equation (4.25), from the rate of ties among members belonging to different groups, γ_{10} in equation (4.26). This parametrization of homophily is then given by:

$$\text{homophily} = \tau = \gamma/\gamma_{10}. \tag{4.51}$$

The simulated networks were generated using $\tau = 1$ (no homophily) and $\tau = 5$ (elevated homophily) with respect to the outcome variable Z. The rate of ties were also chosen so to produce an average degree of ten and all ties were reciprocated. Furthermore, the number of positive outcomes were randomly drawn from a Binomial distribution with probability $\mu = 0.35$ or $\mu = 0.30$ when comparing the three forms of differential recruitments.

A total of one thousand networks were generated with functions in the R packages `statnet` [Handcock et al., 2015b] for each of the two attachment regimes. Each one of those populations comprises a thousand members.

4.5.1.2 Sampling

The simulated RDS process in this study is intended to exhibit features approaching those of actual RDS studies. For instance, the nodes are sampled without replacement. Also, a set of ten seeds initiate the sample instead of one as assumed by the proposed design-based and Bayesian estimators. Such seeds are selected completely at random. Each node subsequently recruits a maximum of two participants. A smaller number of recruits is allowed when there are less than two unsampled alters connected to the recruiting node. Nodes are presumed to recruit under one of the three recruitment regimes:

- recruitment completely at random (i.e. $\phi = 1$),

- moderate differential recruitment (i.e. $\phi = 2$), or
- elevated differential recruitment (i.e. $\phi = 4$)

with respect to the outcome variable Z or the tie attribute matrix W. Nodes receiving an invitation to participate into the survey are presumed to systematically accept the invitation. Finally, the sampling process stops when the target sample size of two hundred is attained. One RDS sample is drawn from each network.

In summary, the six basic scenarios correspond to the sampling conditions and network features described above simulated with one of the three levels of differential recruitment and one of the two homophily levels.

4.5.1.3 MCMC parameters

The Gibbs sampler algorithm requires a number of tuning parameters. For instance, a set of starting values for the unobserved outcome variable and degrees has to be generated. For the purpose of the simulation study, the unobserved outcomes are sampled from a Bernouilli distribution. The probability of a positive outcome of such distribution is assumed to be equal to the estimated prevalence (based on $\hat{\mu}_{SH.dr}$) among the remaining nodes. As for the unobserved degrees, they are simulated from the prior distributions described in Section 4.3.2.3.

4.5.2 Results: Point Estimates

Results from the simulation study for between group differential recruitment are presented in Figure 4.4. This figure displays results for the six scenarios described in the previous section. The two levels of network homophily are shown on the horizontal panels, $\tau \in \{1, 5\}$ and the three levels of differential recruitment are shown on the vertical panels, $\phi \in \{1, 2, 4\}$. Estimates from seven estimators are summarized by box plots which appear in the following order for each scenario: $\hat{\mu}_{VH}$, $\hat{\mu}^{b}_{VH.dr}$, $\hat{\mu}_{SH}$, $\hat{\mu}^{ego}_{SH}$, $\hat{\mu}^{b}_{SH.dr}$, $\hat{\mu}_{SS}$ and $\hat{\mu}^{b}_{SS.dr}$. Estimators are grouped into three categories $\hat{\mu}_{VH}$, $\hat{\mu}_{SH}$ and

$\hat{\mu}_{SS}$ on the x-axis of each scenario and the box plot color within each category indicates the specific version of the estimator: original estimators (green), Lu's extension (purple) and finally, the extended estimators for differential recruitment proposed in this chapter (red). The Bayesian estimator is grouped with the SS estimator even though it is not an extension of the original SS design-based estimator per se. However, both estimators in this category are based on a SS approximation to RDS. The true population parameter μ is represented by the horizontal blue line on that figure. Finally, we note that the last scenario ($\tau = 5$ and $\phi = 4$) is based on 979 populations as opposed to 1000 for all the other scenarios. The algorithm failed to converge for twenty one populations under those simulated conditions.

We first observe from this figure that all estimators have little to no bias in the two scenarios in which no differential recruitment is simulated. In addition, most extended estimators under those two scenarios have reduced variability. This reduction in the uncertainty is partly attributable to the fact that although ϕ is approximately equal to one on average, it slightly varies from this value in any particular simulated sample. These small departures from recruitment completely at random are corrected for in the extended estimators and therefore, produce estimates with smaller errors. For $\hat{\mu}^{b}_{VH.dr}$, the bias introduced by the network homophily offset this effect. For estimators in the SH category, the decrease in variability is also explained by the improved estimation of the c-factors.

Our simulation corroborates the findings discussed in various studies that differential recruitment induces strong biases [Frost et al., 2006, Gile and Handcock, 2010, Tomas and Gile, 2011, Lu et al., 2012, Verdery et al., 2015]. This holds even for a moderate value for ϕ. A between group differential recruitment of magnitude two for instance yields an average bias of roughly 11% in the original estimators in scenarios without homophily and an average bias of 18% with homophily.

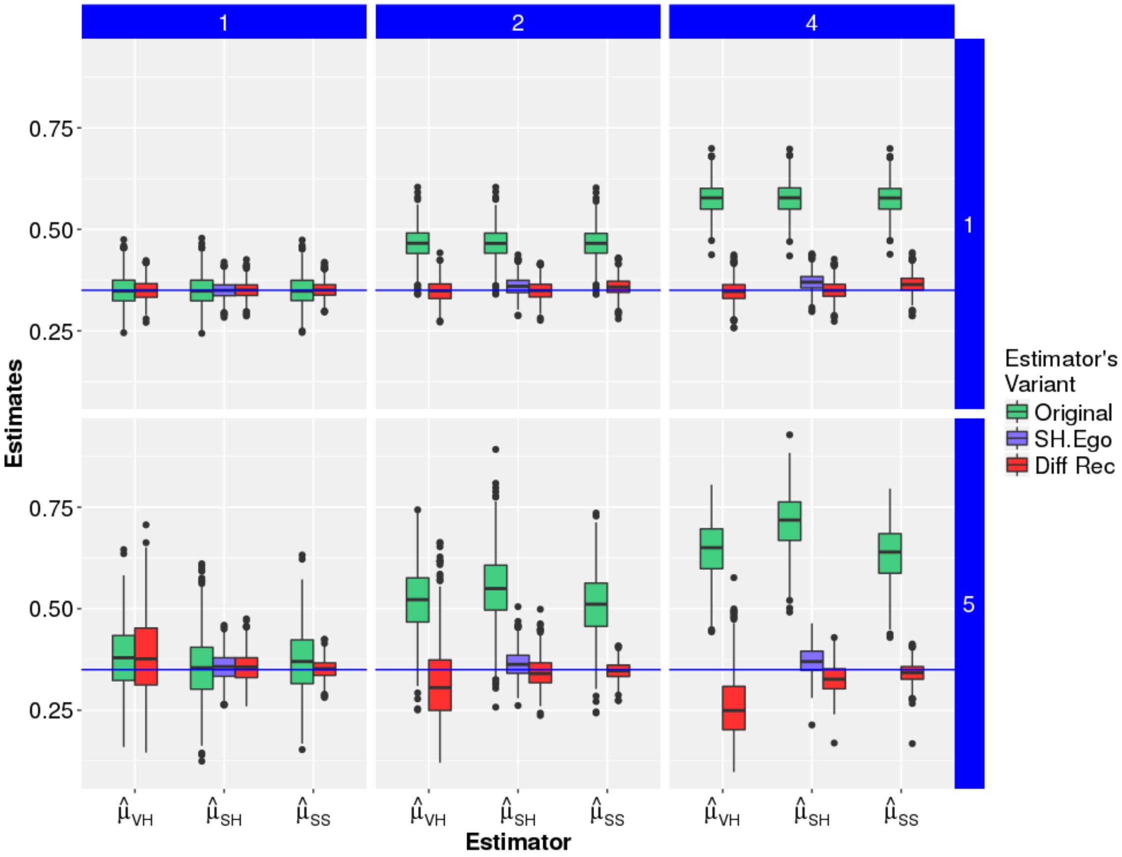


Figure 4.4: Estimates produced with varying level of network homophily, that is, $\tau \in \{1, 5\}$, (horizontal panels) and between group differential recruitment, that is, $\phi \in \{1, 2, 4\}$ (vertical panels). Estimators are presented in the following order: $\hat{\mu}_{VH}$, $\hat{\mu}^{b}_{VH.dr}$, $\hat{\mu}_{SH}$, $\hat{\mu}^{ego}_{SH}$, $\hat{\mu}^{b}_{SH.dr}$, $\hat{\mu}_{SS}$ and $\hat{\mu}^{b}_{SS.dr}$. The blue horizontal line represents the true population prevalence.

As observed in Figure 4.4 all discussed extended estimators reduce substantially the differential recruitment bias under all scenarios. All estimators however do not perform equally well under all circumstances. The estimator proposed by Lu for instance, $\hat{\mu}^{ego}_{SH}$, is far more robust to differential recruitment than the original estimator under all assessed scenarios, but a residual bias remain when $\phi \neq 1$. This bias is explained by the fact that the sampling probabilities on which the estimator relies do not account for differential recruitment. The estimators $\hat{\mu}_{VH.dr}$ and $\hat{\mu}_{SH.dr}$ which modify the sampling weights are therefore outperforming the $\hat{\mu}^{ego}_{SH}$ in the scenarios where there is no homophily. In the presence of homophily however, ϕ is overestimated due to the branching and without replacement nature of RDS and the resulting estimates are therefore overcorrected. The Bayesian estimator which specifically accounts for network homophily reduces a greater portion of the bias under those circumstances.

Table 4.4 displays the performance of the estimators with respect to the root mean-squared-error (RMSE = $\sqrt{MSE}$) for all six scenarios. For each scenario, the RMSE for the four estimators are compared using a Bonferroni comparison at a family-wise error rate of 5%. Results displayed in bold characters indicate that the estimator is in the set of best estimator for the particular scenario. Those results suggest that $\hat{\mu}^{b}_{SH.dr}$ systematically appears in the best set of estimators when there is no network homophily and $\hat{\mu}^{b}_{SS.dr}$ outperforms all estimators otherwise.

Table 4.4: RMSE for the extended estimators under S1-S6. The RMSE's in bold indicates that the method is in the best set of estimators for a particular scenario based on Bonferroni pairwise comparison at a family-wise error rate of 5%.

	Parameters		Estimators			
Scenario	τ	ϕ	$\hat{\mu}^{b}_{VH.dr}$	$\hat{\mu}^{ego}_{SH}$	$\hat{\mu}^{b}_{SH.dr}$	$\hat{\mu}^{b}_{SS.dr}$
S1	1	1	.0249	**.0210**	**.0213**	**.0202**
S2	1	2	.0265	.0247	**.0227**	**.0225**
S3	1	4	.0260	.0297	**.0227**	.0273
S4	5	1	.1025	.0355	.0375	**.0225**
S5	5	2	.0985	.0365	.0358	**.0214**
S6	5	4	.1206	.0391	.0415	**.0258**

For the extended design-based estimators, the ability of the estimators to reduce differential recruitment bias has also been assessed for the other forms of differential recruitment. Figure 4.5 presents the results of this analysis using a $\phi = 2$ and $\tau = 1$. The findings are similar to the results for between group differential recruitment. In other words, extended estimators decrease the differential recruitment bias and Lu's estimator appears to have a residual bias except for the within group differential recruitment. Also, $\hat{\mu}_{SH.dr}$ tend to display less variability than $\hat{\mu}_{VH.dr}$.

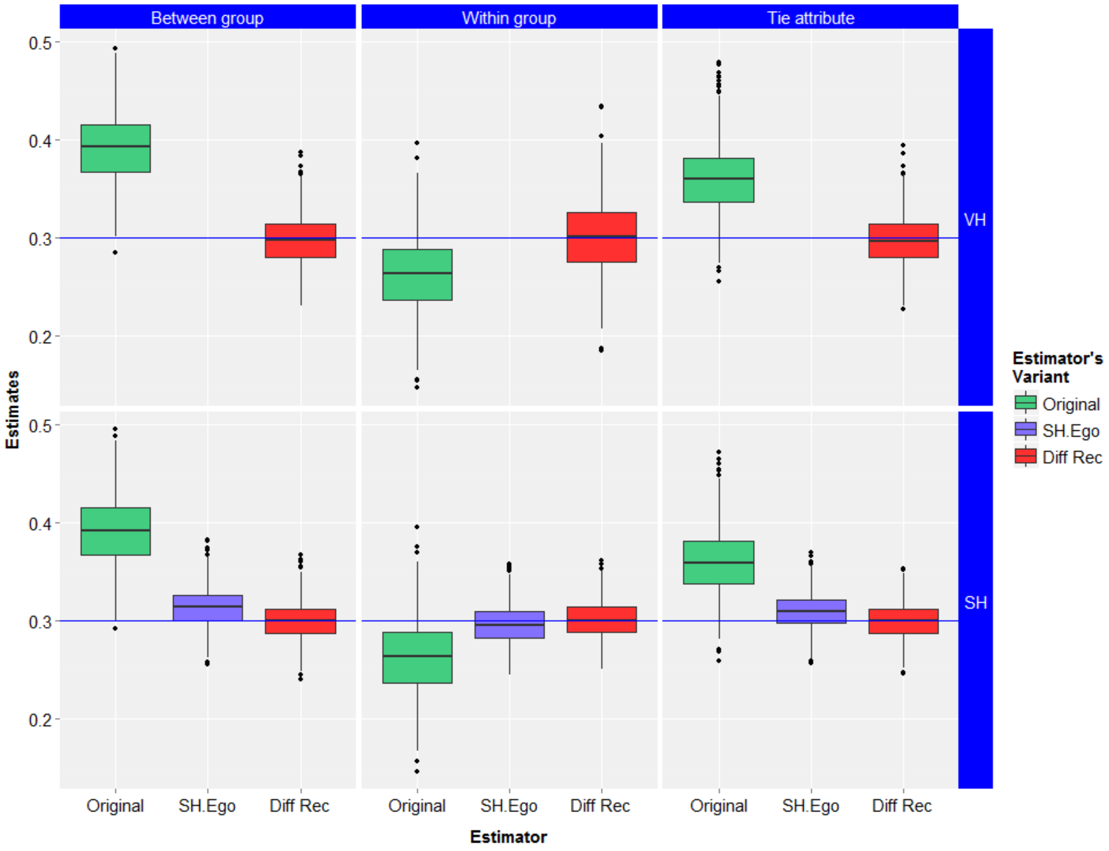


Figure 4.5: Design-based estimates produced under three forms of differential recruitment (vertical panels). Networks are simulated with $\tau = 1$ and samples with $\phi = 2$. $\hat{\mu}_{VH}$ is compared with the corresponding $\hat{\mu}_{VH.dr}$ in the upper horizontal panel and $\hat{\mu}_{SH}$ with $\hat{\mu}_{SH}^{ego}$, and $\hat{\mu}_{SH.dr}$ in the lower panel. The blue horizontal line represents the true population prevalence.

The analysis presented above suppose that the variable inducing differential recruitment is the outcome variable Z. However, a simulation study has also been performed to assess the performance of the design-based estimators when the variable inducing is an arbitrary variable nodal X. The results were similar to the ones presented. However, the differential recruitment bias is smaller in instances where the variable X is not closely related to the outcome variable Z.

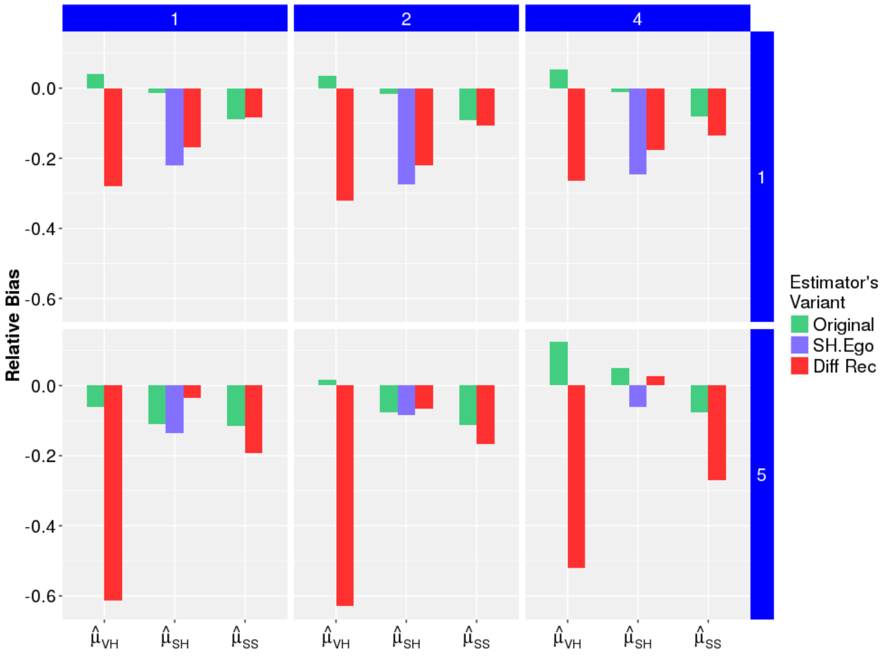
In addition, the convergence of the Gibbs Sampler was assessed using standard MCMC diagnostics. A total of 1500 samples were drawn from the full posterior distribution of which five hundred were discarded for the burnin. Trace plots showed slow mixing of the chains but they appeared to converge. The effective sample size [Heidelberger and Welch, 1981, Geweke, 1992] for some parameters was small and therefore, we will increase the number of posterior draws in our future work.

4.5.3 Results: Variance Estimates

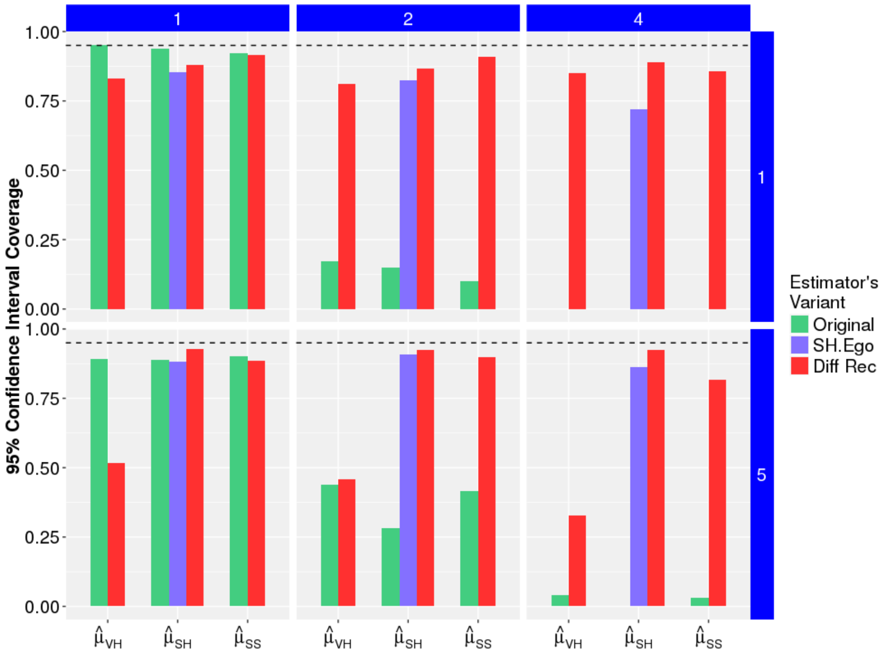
In this section, we assess the performance of the proposed bootstrap variance estimators described in Section 2.4.3 and Section 4.4 at various levels of between group differential recruitment and network homophily. We also evaluate the impact of differential recruitment on the overall inference by comparing coverage rates of the 95% confidence intervals for the traditional RDS estimators and their extended versions.

Similarly to the approach in Section 3.2.3, the performance of the uncertainty estimators is evaluated by comparing the estimated standard deviation ($\hat{\sigma}$) to our best estimates of the true variability which consists of the standard deviation of the simulated prevalence estimates under each scenario (s's).

Figure 4.6a presents the relative differences between the average estimated variability and the variability of the simulated estimated such that the relative bias $= \frac{\bar{\hat{\sigma}}-s}{s}$. The results are shown for the seven prevalence estimators discussed in this section and for between group differential recruitment only. This figure is organized in the same way as Figure 4.4, that is, the two horizontal blocks display the results for the two levels of network homophily ($\tau \in \{1,5\}$) and the vertical panels are divided according to the differential recruitment parameter $\phi \in \{1,2,4\}$. The estimators are presented in the following order within each scenario: $\hat{\sigma}(\hat{\mu}_{VH})$, $\hat{\sigma}(\hat{\mu}^{b}_{VH.dr})$, $\hat{\sigma}(\hat{\mu}_{SH})$, $\hat{\sigma}(\hat{\mu}^{ego}_{SH})$, $\hat{\sigma}(\hat{\mu}^{b}_{SH.dr})$, $\hat{\sigma}(\hat{\mu}_{SS})$ and $\hat{\sigma}(\hat{\mu}^{b}_{SS.dr})$.



(a) Relative bias of the standard deviation estimates calculated as $\frac{\bar{\hat{\sigma}}-s}{s}$, where $\bar{\hat{\sigma}}$ is the average estimated standard deviation under a bootstrap methodology and s is the sample standard deviation.



(b) 95% confidence interval coverage rates, where the coverage rates are the percentage of the intervals including the true population proportion μ of 35%. The dashed line is set at 95%.

Figure 4.6: Standard deviation estimation and 95% confidence interval coverage results for using the bootstrap procedures for the various versions of $\hat{\mu}_{VH}$, $\hat{\mu}_{SH}$ and $\hat{\mu}_{SS}$.

As suggested by this figure, the estimated uncertainty of the extended prevalence estimators is most often underestimating the variance. The relative bias ranges from approximately 3% to -63%. These large negative biases may be explained by the fact that the RW and the successive sampling processes ignore some RDS-specific sampling features, such as the branching. The variability associated with those features is not captured in the Bootstrap procedures.

We note however that the overall inference with the extended estimators is not severely impaired by the underestimation of the variability. As seen in Figure 4.6b coverage rates for the 95% confidence intervals are either rather comparable ($\phi = 1$) or much higher ($\phi > 1$) than the coverage rates for the traditional estimators. However, inference using $\hat{\mu}^b_{VH.dr}$ in the presence of network homophily is an exception. Under those circumstances, the prevalence estimator is largely biased and therefore, the true population parameter falls within the constructed intervals less often.

In summary, despite the underestimation of the variance of the extended prevalence estimators, the inference from RDS data is improved by the extensions proposed in this section in the presence of differential recruitment. As seen in Figure 4.6, inference with $\hat{\mu}^b_{SH.dr}$ and $\hat{\mu}^b_{SS.dr}$ appear to have the best performance in most scenarios.

4.6 Discussion

Sampling hard-to-reach populations is a challenging problem. RDS has provided ways to circumvent some of the issues specific to those populations that make the use of traditional sampling methods unpractical. However, the sampling process under RDS is out of the control of the researchers conducting the studies and therefore, this sampling method is highly susceptible to biases induced by participants' behaviors. The main contribution of this work is to introduce inferential methodologies correcting existing RDS prevalence estimators and their uncertainty estimators for biases induced by various forms of differential recruitment.

Our first approach to correct for differential recruitment extends the traditional design-based RDS estimators. Conventional estimators under this framework suppose that participants' sampling probabilities may be estimated from the stationary distribution of a random walk (RW) on the state space of the network nodes. The derivation of the stationary distribution assumes that participants recruit completely at random among their contacts in the target population. Our approach modifies this assumption and instead proposes three sampling schemes under which participants systematically recruit individuals based on one of their nodal characteristics or based on their relationship nature with them. By explicitly defining those sampling schemes we were able to derive the RW characterizing those behaviors and their associated stationary distributions. The revised estimators rely on the stationary distributions of the modified RW. Results from the simulation study show that this methodology greatly reduces biases induced by the various forms of differential recruitment. However, these methods require additional data about participants' ego-network compositions.

One of the important limitation of the proposed design-based approach is its poor performance with networks featuring homophily. To address this issue, we have extended a model-based approach which allows us to simultaneously estimate network homophily and between group differential recruitment on the outcome variable. Under this framework, Bayesian inference is performed about the parameters of the super-population model and the sampling model. Since the super-population model explicitly allows for network homophily, this method has shown to substantially reduce the traditional estimators' differential recruitment bias present in homophilous networks. However, similarly to the design-based inference, this model-based estimation framework also requires ego-network compositions data to be collected.

The comparison of the root mean-squared-error (RMSE) in our simulation study suggests that the design-based estimator $\hat{\mu}_{SH.dr}$ generally outperforms alternative estimators for networks simulated with random attachment. The same criteria favors the model-based estimator $\hat{\mu}_{SS.dr}$ in the presence of network homophily.

We have also proposed uncertainty estimators in this section. For design-based prevalence estimators, the uncertainty is estimated through a bootstrap procedure capturing the variability associated with the RW sampling as well as with the estimation of the magnitude of the differential recruitment ϕ. For the model-based approach, the standard deviation is simply calculated from the posterior draws of the prevalence parameter μ. In addition to reflecting the variability induced by the successive sampling model, this variance also takes into account the variability of the super-population model. Results from the simulation study show that the variance estimators tend to underestimate variability. This may be explained by the fact that those procedures do not reflect some of the RDS specific features. Although the underestimation of the variance affects the width of the 95% confidence intervals, the coverage rates for $\hat{\mu}_{SH.dr}$ and $\hat{\mu}_{SS.dr}$ are significantly better than those produced by the conventional estimators when with $\phi = 2$ or 4. We conclude that the proposed extended methods improve the inference in the presence of differential recruitment despite the underestimation of the variance.

Additional analysis not presented in this section shows that one of the limitations of the model-based estimator is its sensitivity to the mispecification of the network model. To address this issue, we intend to examine in our future work alternative and more flexible ways to formulate this model.

The model-based approach also assumes a known target population size N. This assumption is often unrealistic in most RDS studies. One possible extension to our work would be to treat this quantity as a parameter to be estimated along with the other model parameters. Similar methodology has been developed by West [1996] and Handcock et al. [2014], but sensitivity to an additional parameter in the problem at hand has not been evaluated yet.

One of the major advantage of the design-based framework is its ability to correct for differential recruitment on any nodal or tie characteristics. Although our preliminary work to allow this feature to be incorporated in the model-based framework has not been conclusive, we intend to pursue this objective in our future work.

A number of questions could be further investigated to ensure that the proposed methodologies are both sounded and practical for practitioners. For instance, since we have identified homophily as a key factor in determining which of the two approaches is the most suitable, providing RDS users with measures of network homophily based on RDS data would represent a useful addition to our work. Similarly, in the event that the model-based approach remains sensitive to model choices, providing diagnostic tools to assess the fit of the data to those choices would represent a critical future contribution. Besides, since RDS surveys relies on self-reported ego-network data, studying the sensitivity of the methods to misclassified data remains a key objective of our future work. Finally, we hope to work with practitioners to develop guidance on prior determination of variables that could lead to differential recruitment so that ego-network information about those variables may be included in RDS questionnaires.

Overall, we believe that the proposed methodologies are promising and could significantly improve traditional estimators when participants do not recruit at random.

CHAPTER 5

NATIONAL PREVALENCE ESTIMATION

5.1 Introduction

Public Health organizations studying concentrated HIV epidemics commonly conduct a series of surveys within a country among its populations at elevated risk of infection, such as men who have sex with men (MSM), sex workers (SW) and people who inject drugs (PWID). Collecting information from those hard-to-reach populations is however often challenging and expensive and specialized sampling techniques, such as RDS [Heckathorn, 1997], are used. Consequently, it is not uncommon that only a subset of the key populations of a given country are sampled. For instance, samples may be collected in twenty of the thirty major cities of a country. Estimates of quantities of interest, such as disease prevalence and key population size, are therefore often available only for a subset of the country's key populations. This specific nature of the data collection poses a challenge when national estimates are sought.

Recent methodological advances have allowed the derivation of national estimates from local estimates. For instance, Bao et al. [2015] developed a Bayesian hierarchical model to estimate national key population size from regional estimates. Their method incorporates many of the data sources typically available in the context of HIV surveillance. It also reflects the uncertainty and some biases inherent to the conventional data sources.

National prevalence estimates, denoted $\hat{\pi}$, are commonly derived by computing the average of the regional prevalence estimates (y_j's) weighted by their target population size estimates (n_j's):

$$\hat{\pi} = \frac{\sum_{j=1}^{\mathrm{J}} y_j n_j}{\sum_{j=1}^{\mathrm{J}} n_j}, \tag{5.1}$$

where J is the number of regions where data were collected. This methodology however raises two main concerns related to (1) the estimation of uncertainty of national prevalence estimate, and (2) the treatment of regions for which no survey data are available. We discuss each of these in turn.

(1) Estimation of uncertainty of national prevalence estimates A commonly used approach to constructing confidence intervals is to ignore the uncertainty of the population size estimates. The bounds of the confidence interval are calculated in a similar fashion to the prevalence point estimate, that is, as weighted average of the regional confidence interval bounds. This would in principle be an appropriate procedure if the regional population sizes were known with certainty, which is rarely the case in the context of hard-to-reach populations. Therefore, this approach often underestimates the uncertainty of the national prevalence estimate.

(2) Treatment of regions for which no survey data are available Often regional estimates are not available for all the regions of the studied country, although national estimates are desired. The national estimator in its current form does not explicitly model the missing regions. It instead assumes that these regions have the same disease prevalence as the national estimate. Although it may sometimes be a reasonable assumption, this could potentially be problematic in other instances. In resource limited settings for example, studies may be conducted only in regions with the most susceptible populations. By design, the prevalence estimates in the selected areas may be substantially higher than the prevalence in the unobserved areas. Consequently, the overall national prevalence would be overestimated. Additionally, if a large number of regions are missing, this could translate into an underestimation of the national prevalence estimate uncertainty.

The contribution of our research is to address some of these issues. For instance, using a similar approach to Bao et al. [2015] for population sizes, our proposed method naturally incorporates uncertainty in regional population sizes. Our proposed Bayesian approach also allows for direct modeling of prevalence in regions where no data are available.

In this chapter, we discuss a proposed approach to estimating the national prevalence. The chapter is organized as follows. Section 5.2 describes the data used in our research. The prevalence and population size models are subsequently presented in Section 5.3. Assessment of the models' fit is discussed in Section 5.4 along with the derivation of the national prevalence estimate. We conclude in Section 5.5 with a brief overview of the methodology discussed, its current limitations and thoughts for future research.

5.2 Data

Data from two target populations of a given country are used for this study. The country name is however not identified for confidentiality purposes. Data from the country may be divided into two categories: prevalence data and key population size data. These two types of data are described in this section followed by a discussion about additional data.

We denote observed data with lower case letters and parameters to be estimated from the Bayesian models presented in Section 5.3 with Greek letters. It should therefore be clear from the context if, for example, we refer to the prevalence point estimates from the RDS surveys, which are treated as observed data, or whether we refer to the prevalence estimates obtained from the Bayesian model.

5.2.1 Survey Prevalence Estimates

Prevalence estimates and their respective variance were derived from survey data. In particular, data were collected with RDS surveys, which were conducted in five of the country's regions ($J = 5$). For two of these regions, the surveys included participants residing in various cities. Our analysis demonstrated that participants recruited almost exclusively individuals living in their city. Consequently, the regional surveys for these two regions were subsequently divided into two or three surveys to reflect the fact that the samples were obtained from different populations. This yielded a total of eight surveys ($I = 8$) from five regions for each of the two key populations, denoted KP_1 and KP_2. Therefore, the data contain a total of sixteen point estimates.

The prevalence estimates are derived from the Volz and Heckathorn [2008] estimator described in Section 2.4.1.2 and are denoted y_i for $i \in \{1, 2, ...8\}$. Although the variance of the prevalence estimates, v_i, are determined by a bootstrap procedure [Salganik, 2006], they assumed to be known with certainty for this study.

The prevalence data are presented in Figure 5.1. This plot shows the sixteen point estimates along with their respective 95% confidence intervals. The estimates are presented separately for the two key populations and are grouped by region, when applicable. The colors pink and blue represent KP_1 and KP_2, respectively, and the vertical lines delimit the five regions. We observe that the prevalence estimates vary significantly across regions both in magnitude and in variability.

5.2.2 Population Size Data

Our model incorporates three sources of information regarding the size of the target population: object multiplier estimates [Archibald and Sutherland, 2001], wisdom of the crowd estimates and experts' estimates of the proportion of the reference pop-

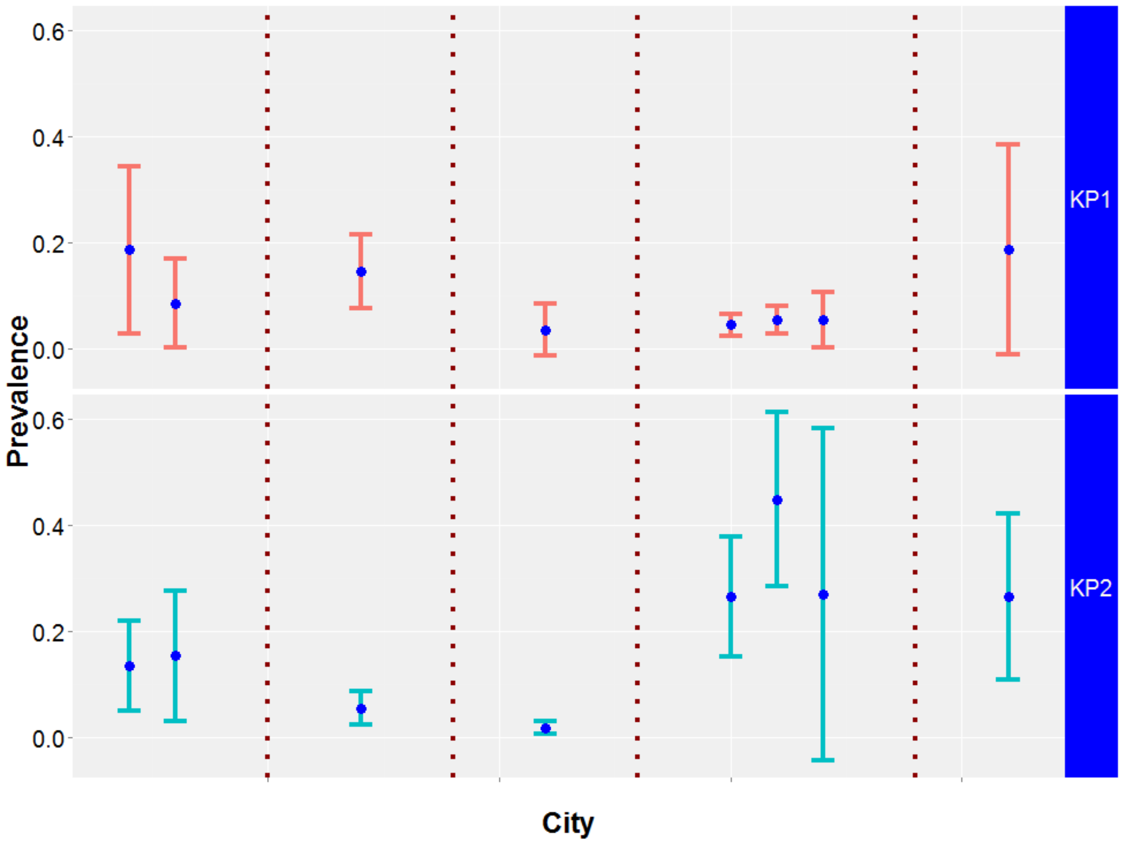


Figure 5.1: Survey prevalence estimates y_i (data) along with their 95% confidence intervals. The prevalence are grouped by regions and shown separately for the two key populations.

ulation who belongs to the key population. All these data sources are also used in the work of Bao et al. [2015].

The unique object multiplier method is equivalent to a capture-recapture method. The first step of this method consists in distributing characteristic objects to members of the target population. The objects are sometimes distributed at venues typically attended by members of the target population for example. Secondly, a survey is performed shortly after and the number of participants having received the object are counted.

It is possible to estimate the size of the target population with this collected information with a method of moments estimator. The method of moments estimator commonly used with this type of data assumes that the proportion of participants who received the objects in the survey is approximately equal to the proportion of people who were given the objects in the overall target population such that:

$$\frac{o_j}{r_j} = \frac{d_j}{n_j} \quad \Leftrightarrow \quad n_j = \frac{d_j}{o_j/r_j}, \tag{5.2}$$

where

o_j number of objects observed in the surveys in region j,

r_j number of participants in the survey in region j,

d_j number of objects distributed in region j,

n_j target population size estimate for region j.

This estimator is unbiased as long as the two sources of data are independent and as long as the survey is representative of the target population. However, as pointed out by the WHO and UNAIDS in their guideline on population size estimation [UNAIDS/WHO, 2010], the properties of this estimator heavily depend on the quality of the collected data. In the present study, unique object multiplier estimates of the population sizes are available at the regional level, that is, for the five regions and for the two key populations.

In addition to the unique object multiplier estimates, wisdom of the crowd estimates were also collected for the five regions. Under this method, the population size estimate is simply the average of the survey participants' best estimate of the target population size. This method however often leads to large biases and does not provide any measure of uncertainty.

Finally, field experts also provided their best guess estimate of the proportion of the reference population, p_e, belonging to the target populations. This proportion is a global estimate for the entire country.

5.2.3 Additional Data

In addition to prevalence and key population size estimates, the number of individuals in the general population, i.e. the reference population size, is also used as a predictor in the model to estimate the size of the key populations. Other predictive variables have been considered from the Demographic Health Survey (https://dhsprogram.com/) and from UNAIDS Key Populations Atlas (http://www.aidsinfoonline.org/kpatlas). None of the tested variables have shown strong predictive power. Therefore they are not discussed in this chapter.

5.3 Methods

The main objective of this study is to determine the national prevalence of HIV among two susceptible key populations from a given country. In this Section, we describe our proposed approach to obtain a national prevalence estimate. The methodology relies on hierarchical Bayesian models for both the prevalence and the population size estimates. The hierarchical structure of the models is designed to reflect both the variability within cities or regions as well as across them. Sections 5.3.1 and 5.3.2 describe two possible models for either the prevalence or the target population size. In those Sections, the two key populations are treated as two different groups. In other words, the models are the same for the two key populations but they are fitted separately for each dataset. It is then followed in Section 5.3.3 by a description of the national prevalence estimator.

5.3.1 Prevalence Model

Multiple models were considered and evaluated in our analysis. In this Section however, we only describe two models. A description of alternative models along with results may be found in Appendix C.

Since RDS is a complicated sampling process, the prevalence estimators do not follow a known distribution. The two models below assume that the logarithm of the prevalence follows a Normal distribution. Also, the variance of the distributions (v_{logy_i}) is parameterized based on the known variability for y_i which is estimated by a standard RDS Bootstrap procedure [Salganik, 2006].

Model 1 - Partial Pooling by City

$$\begin{aligned} log(y_i)|\tau_i^{m1}, v_{logy_i} &\sim N(\tau_i^{m1}, v_{logy_i}) \\ \tau_i^{m1}|\mu_{\tau^{m1}}, \sigma^2_{\tau^{m1}} &\sim N(\mu_{\tau^{m1}}, \sigma^2_{\tau^{m1}}). \end{aligned} \tag{5.3}$$

Model 2 - Partial Pooling by Region

$$\begin{aligned} log(y_i)|\tau_{j[i]}^{m2}, v_{logy_i} &\sim N(\tau_{j[i]}^{m2}, v_{logy_i}) \\ \tau_j^{m2}|\mu_{\tau^{m2}}, \sigma^2_{\tau^{m2}} &\sim N(\mu_{\tau^{m2}}, \sigma^2_{\tau^{m2}}). \end{aligned} \tag{5.4}$$

In the expressions above, y_i is the HIV prevalence for city $i \in \{1, 2, ..., 8\}$, and v_{logy_i} is determined based on the known variance of y_i, that is, v_i. In particular,

$$v_{logy_i} = log\left[\frac{1}{2}\left(1 + \sqrt{1 + \frac{4v_i}{e^{2\tau_i}}}\right)\right], \tag{5.5}$$

where τ_i represents τ_i^{m1} or τ_j^{m2} for model 1 or 2, respectively.

The parameters $\mu_{\tau^{m1}}$ and $\mu_{\tau^{m2}}$ represent the overall mean prevalence across all cities (model 1) or regions (model 2). As for $\sigma^2_{\tau^{m1}}$ and $\sigma^2_{\tau^{m2}}$, they represent the variability across cities and across regions, respectively. Vague prior distributions for those parameters are assumed to reflect our lack of prior knowledge of the national HIV prevalence and the variability across cities or regions:

$$\begin{aligned}
\mu_{\tau^{m1}} &\sim Normal(0, 1000)\\
\mu_{\tau^{m2}} &\sim Normal(0, 1000)\\
\sigma_{\tau^{m1}} &\sim U(0,\ 5)\\
\sigma_{\tau^{m2}} &\sim U(0,\ 5).
\end{aligned} \tag{5.6}$$

The two models differ in their treatment of the surveys within a region. Model 2 implies that there exists a unique mean τ_j^{m2}'s for each region. This assumption is appropriate when individuals from the various cities of a given region may be viewed as belonging to a common target population. Should this not be the case, then model 1 would be more suitable since under model 1, each prevalence point estimate has its own mean.

In the present case, for example, some regions have multiple prevalence estimates y_i. However, the data were truly obtained from one survey per region. The results were subsequently divided into two and three estimates for two of the regions. The reason behind this decision is related to the recruitment chains which were highly clustered on the city variable. In other words, except a few exceptions, individuals only recruited participants living in the same city as themselves. This indicates that the network is not very well connected between cities and therefore, the data in fact represent samples from different populations and therefore, model 1 is more appropriate for the data.

5.3.2 Population Size Estimates

We also propose two models for the the population size parameters (η's): complete pooling and partial pooling models. The method closest to the work of Bao et al. [2015] is the partial pooling model. Both models account for the three sources of data:

- **Service Multiplier**: Since the service multiplier follows a capture-recapture approach, the number of objects observed in the survey, o_j, are modeled with a hypergeometric distribution.
- **Wisdom of the crowd**: Similar to Bao et al. [2015], the wisdom of the crowd estimates, i.e. z_j is modeled on a log scale with a bias component. The bias is expressed as a proportion of the regional reference population size, q_j.
- **Expert's opinion**: The expert's opinion, i.e. the proportion of the reference population who belongs to the target population (p_e), has been accounted for in the specification of the prior distributions.

Model 1, the complete pooling model:

$$\begin{aligned} o_j|d_j, \eta_j, r_j &\sim Hypergeometric(d_j, \eta_j, r_j) \\ \eta_j|\theta, q_j &\sim Bin(q_j, \theta) \\ log(z_j)|\eta_j, q_j, \sigma_z^2 &\sim N(log(\eta_j) + \beta log(q_j), \sigma_z^2) \end{aligned}$$

with the following prior distributions:

$$\begin{aligned} \theta &\sim N(p_e, \sigma_\theta = 0.0005) \\ \beta &\sim N(0, \sigma_\beta = 10) \\ \sigma_z &\sim U(0, 100) \end{aligned}$$

and where,

η_j Parameter for the target population size in region j;

o_j Number of objects retrieved in the sample of region $j \in \{1, 2, ..., J\}$;

d_j Number of objects distributed in region j;

r_j Survey sample size in region j;

q_j Reference population size in region j;

z_j Wisdom of the crowd estimate of η_j; and

p_e Expert's opinion of the proportion of reference population who belong to the target population.

This model is referred to as the "complete pooling" model since the parameter θ, which represents the proportion of the reference population who belongs to the target population, is the same for all regions. In other words, the information of all regions is completely pooled into a single estimate. **Model 2** differs in that respect such that regional proportions are instead modeled with a hierarchical structure. The hierarchical structure allows information to be shared across regions, i.e. to be "partially pooled". **Model 2**, the partial pooling model is given by:

$$\begin{aligned} o_j|d_j, \eta_j, r_j &\sim Hypergeometric(d_j, \eta_j, r_j) \\ \eta_j|\theta_j, q_j &\sim Bin(\theta_j, q_j) \\ logit(\theta_j) &\sim N(\mu_\theta, \sigma_\theta^2) \\ log(z_j)|\eta_j, q_j, \sigma_z^2 &\sim N(log(\eta_j) + \beta log(q_j), \sigma_z^2) \end{aligned}$$

with the following prior distributions:

$$\begin{aligned} \mu_\theta &\sim N(logit(p_e), 4) \\ \beta &\sim N(0, \sigma_\beta = 10) \\ \sigma_z &\sim U(0, 100). \end{aligned}$$

The prior distribution for θ and μ_θ were chosen to be centered at the experts' guess. As no measure of uncertainty is provided with the experts' opinion, we had to choose this prior parameter. We chose it so that the prior was informative. To verify that the prior distributions were indeed informative, we plotted the posterior and the prior distributions for those parameters. Figure 5.2 illustrates our findings for KP_1. In that figure, the histograms represent the posterior distribution whereas the red lines represent the prior distributions. We conclude that the prior distributions are informative. Similar results were obtained for KP_2.

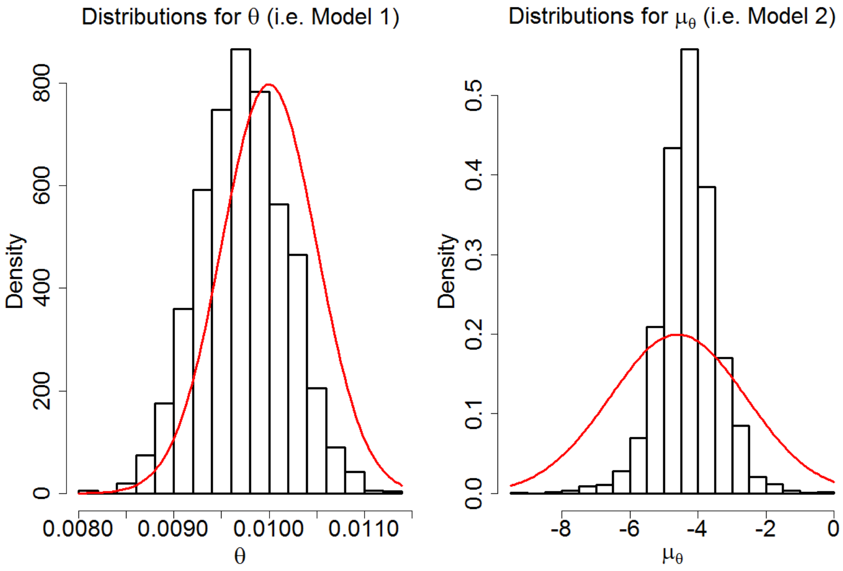


Figure 5.2: KP_1 prior and posterior distributions (θ and μ_θ)

5.3.3 National Prevalence Estimator

Current national prevalence estimator is in the form of a weighted average as described in equation (5.1). Our proposed Bayesian estimator has a similar form but takes into account different sources of data and uncertainty and also reflects the fact that data might be missing for a number of regions or cities. In this section, we describe our proposed approach and highlight the main differences with the current estimator.

The proposed prevalence estimators for the prevalence models 1 and 2, respectively, are as follows:

$$\pi_{m1}^{(s)} = \frac{\sum_{i=1}^{\mathrm{I}_{all}} \pi_{m1,i}^{(s)} \eta_i^{(s)}}{\sum_{i=1}^{\mathrm{I}_{all}} \eta_i^{(s)}} \qquad \text{and} \qquad \pi_{m2}^{(s)} = \frac{\sum_{j=1}^{\mathrm{J}_{all}} \pi_{m2,j}^{(s)} \eta_j^{(s)}}{\sum_{j=1}^{\mathrm{J}_{all}} \eta_j^{(s)}}, \tag{5.7}$$

where "s" indicates the "s"-th posterior draw and where I_{all} and J_{all} refer to the total number of cities (model 1) or regions (model 2), including cities and regions where no data were collected.

Firstly, we notice that the observed prevalence estimates y_i shown in equation (5.1) are substituted by $\pi_{m1,i}^{(s)}$ and $\pi_{m2,i}^{(s)}$, respectively. These values are derived from the posterior draws for τ^{m1}'s and τ^{m2}'s. Since the prevalence models are on a log-scale, the posterior prevalence estimates are obtained by taking the exponential of the posterior parameters τ^{m1}'s and τ^{m2}'s such that, $\pi_{m1,i}^{(s)} = exp\left[(\tau_i^{m1})^{(s)}\right]$ and $\pi_{m2,j}^{(s)} = exp\left[(\tau_j^{m2})^{(s)}\right]$. For cities and regions with observed prevalence estimates, the posterior draws for τ^{m1}'s and τ^{m2}'s are available directly from the model fit. However, for cities and regions with missing data, obtaining posterior draws for τ^{m1}'s and τ^{m2}'s requires two steps. For example, for model 1, the two steps are as follows:

1. sample $(\mu_{\tau^{m1}}^{(s)}, \sigma_{\tau^{m1}}^{(s)})$ from their posterior distribution $p(\mu_{\tau^{m1}}, \sigma_{\tau^{m1}}|y)$
2. sample $(\tau^{m1})^{(s)}$ from its posterior distribution $p(\tau^{m1}|\mu_{\tau^{m1}}^{(s)}, \sigma_{\tau^{m1}}^{(s)})$.

These two steps are equivalent to sampling from the posterior distribution of τ_i^{m1} since its posterior distribution may be expressed as follows:

$$p(\tau_i^{m1}|y) = \int p(\tau_i^{m1}|y, \alpha)p(\alpha|y)d\alpha, \text{ where } \alpha = (\mu_{\tau^{m1}}, \sigma_{\tau^{m1}}). \tag{5.8}$$

As such, since we have posterior draws for all cities, it is possible to sum over I_{all} in equation (5.7). A similar approach may be adopted for model 2.

Secondly, the observed population size estimates n_j shown in equation (5.1) are substituted by $\eta_i^{(s)}$ and $\eta_j^{(s)}$, respectively. The posterior draws are obtained from the population size model 2 described in Section 5.3.1. For regions without observed population size estimates, the posterior draws are obtained using a 2-step procedure similar to the one used for the prevalence parameters. It is worth noting however that posterior draws for the population size estimates η's are only available at the regional level. Therefore, for model 1, to evaluate the sum in equation (5.7), we need to formulate an assumption regarding the allocation of the regional population size estimates between their respective cities. Results presented in Section 5.4 assume that cities in a given region are of equal size. As such, for region j containing c_j cities, the estimated target population size for a city i in that region is $\eta_i = \eta_j / c_j$. This approach was adopted due to the lack of sufficient data to obtain city specific estimates. Our model could however easily be extended should the necessary data become available.

Finally, the Bayesian national prevalence estimate is obtained by taking the mean of the posterior distribution for $\pi_{m1}^{(s)}$ or $\pi_{m2}^{(s)}$. Similarly, the uncertainty of this estimator may be estimated directly from the posterior draws. Results in Section 5.4 were produced from 3000 draws. This therefore led to 3000 $\pi_{m1}^{(s)}$ and $\pi_{m2}^{(s)}$ samples from which we could derive the posterior mean (i.e. the Bayesian estimate) as well as the 2.5% and 97.5% quantiles to obtain the credible interval.

5.4 Results

In this Section, we present results from fitting the different models to the data. Section 5.4.1 discusses results for the two prevalence models, Section 5.4.2 discusses results for the population size models and finally, Section 5.4.3 describes the results for the national prevalence estimates.

5.4.1 Prevalence Estimates

The first validation to assess whether the models provide a reasonable fit to the data is to examine the residuals. The residuals are calculated as follows:

$$\begin{aligned} res_i &= y_i - \pi_{m1,i} && \text{for model 1} \\ res_i &= y_i - \pi_{m2,j[i]} && \text{for model 2,} \end{aligned} \tag{5.9}$$

where $\pi_{m1,i} = \sum_{s=1}^{S} \pi_{m1,i}^{(s)}/S$, $\pi_{m2,i} = \sum_{s=1}^{S} \pi_{m2,j[i]}^{(s)}/S$ and S is the number of posterior draws. Under the two models and for the two key populations, the average residuals are nearly zero, which is ideal.

Figure 5.3 also provides a visualisation of the model fit. This graph shows the 95% predictive intervals (PI's) for each y_i as well as these observed estimates (red dots). The results are displayed for the two key populations (horizontal panels) and the two prevalence models (vertical panels). All data points fall inside the PIs for both models. The model fit again appears to be reasonable.

We have also compared the two models in terms of Relative Mean Absolute Error (RMAE), Root Square Mean Error (RSME) and Deviance Information Criterion (DIC) [Spiegelhalter et al.]. Results are shown in Table 5.1. Lower values indicate a better model fit to the data. In the present case, model 2 achieves the lowest values for almost all criteria. However, as discussed in Section 5.3.1, model 1 is more appropriate in our case due to the nature of the prevalence data.

Table 5.1: In-sample predictive accuracy and DIC for prevalence models

Key Population	Relative Mean Absolute Error		Root Square Mean Error		DIC	
	M1	M2	M1	M2	M1	M2
KP_1	0.247	0.205	0.052	0.044	19.2	15.4
KP_2	0.138	0.123	0.044	0.061	22.3	13.7

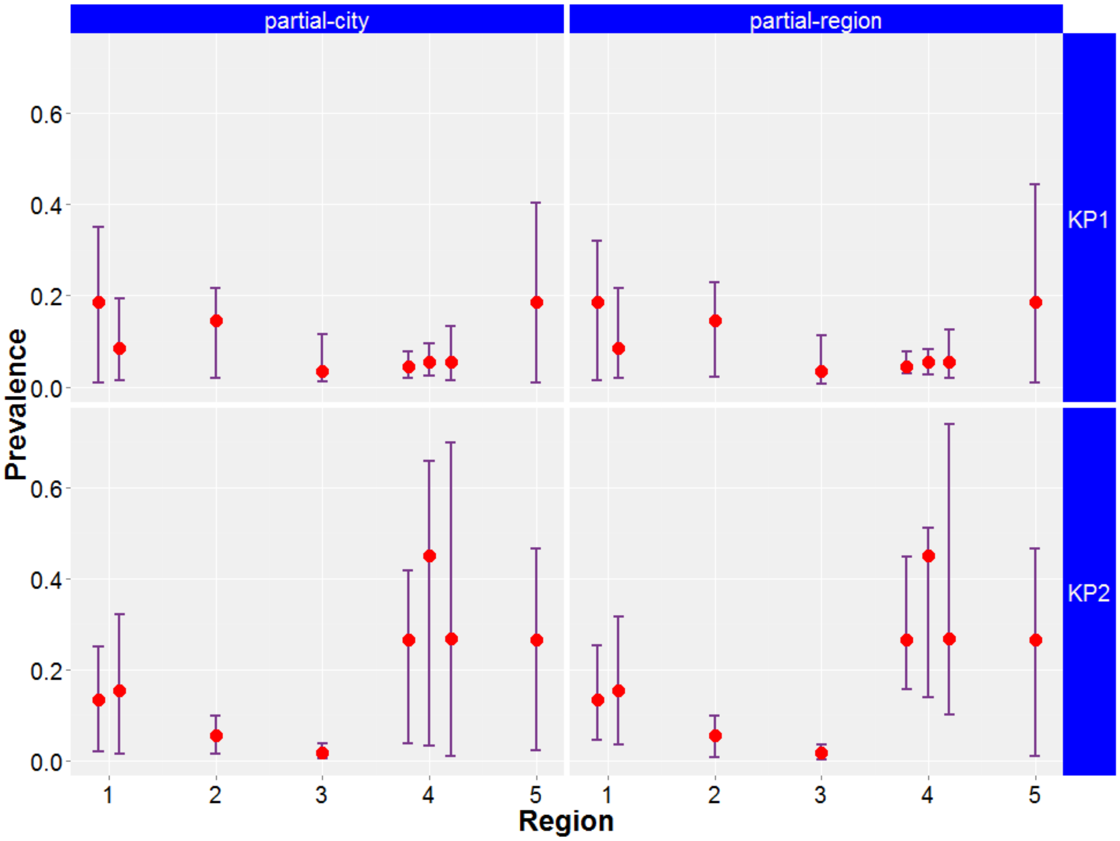


Figure 5.3: 95% predictive intervals (PI) for the prevalence of the two key populations along with the observed prevalence estimates y_i depicted by red dots.

Finally, the convergence of the models has been assessed by visual inspection of trace plots as well as with other traditional MCMC diagnostic statistics, such as the $\hat{R}$ [Gelman and Rubin, 1992] and the effective sample size [Heidelberger and Welch, 1981, Geweke, 1992]. All measures appear to indicate convergence of the chains to the target distributions.

5.4.2 Population Size Estimates

We have also verified the model fit for the two target population size models. All validations suggest that model 2 strongly outperforms model 1. For instance Figure 5.4 displays the 95% predictive intervals (PI's) for the observed population size estimates as a proportion of the reference population with dashed lines (n_j/q_j).

The PI's for the first model are depicted with the lighter lines and with thicker lines for model 2. These PI's are compared with the observations (n_j/q_j) represented by dashed lines. As expected, for model 1, the proportions are fairly constant across regions. We note that the observed data always lie inside the 95% PI's for model 2, whereas it is generally not true for the first model. Model 2 has a much better fit.

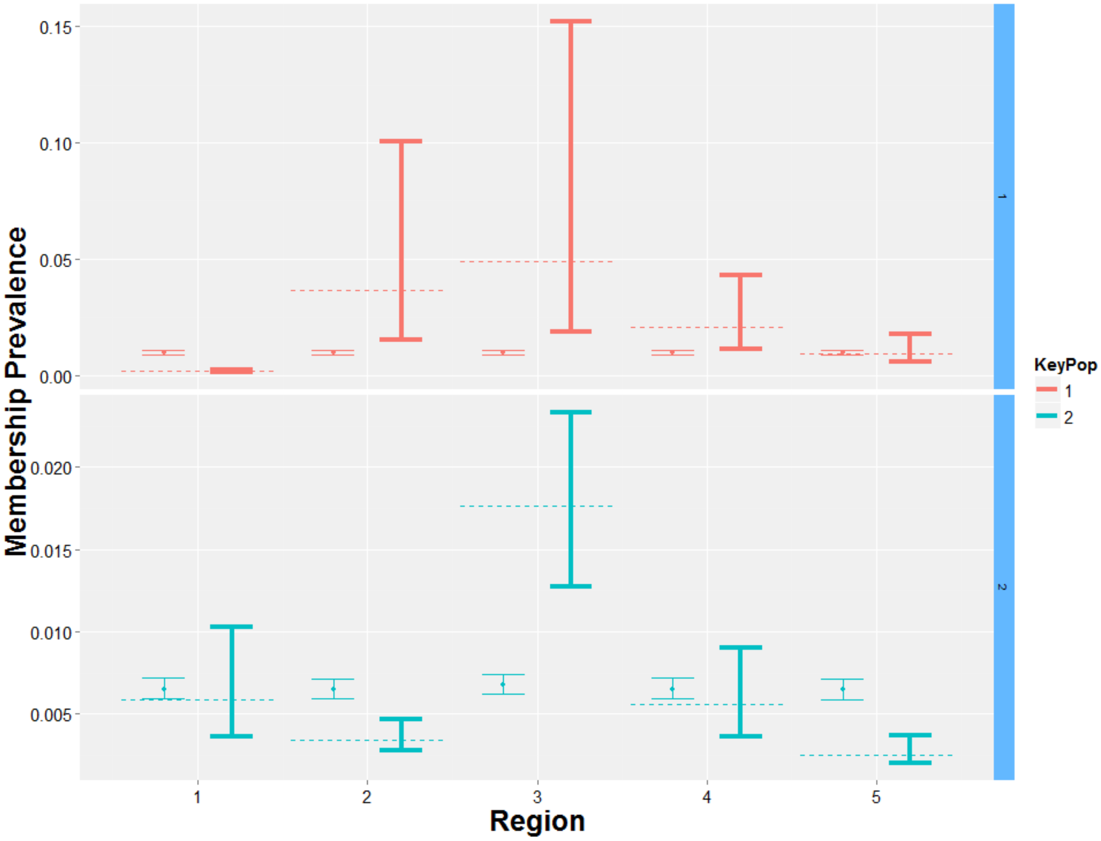


Figure 5.4: 95% predictive intervals (PI) for the population size estimates expressed as a proportion of the reference population. Model 1 PI's are depicted with the lighter lines and model 2 PI's with thicker lines. Dashed lines represents the observed data.

We have also compared the models using in-sample predictive accuracy and DIC information criteria. Table 5.2 shows the results. Again, all metrics indicate model 2 vastly outperforms model 1.

Table 5.2: In-sample predictive accuracy and DIC for population size models

Model	Key Population	Relative MSE	Relative Mean Absolute Error	DIC
1	1	25.452	1.743	177.4
	2	3.255	.590	157.1
2	1	.035	.079	55.6
	2	.011	.043	55.5

We have also verified that the inclusion of the variable q_j as a predictor of the bias for the wisdom of the crowd estimate was appropriate. Based on the posterior estimates for β (model 2) shown in Table 5.3, we conclude that q_j explains some of the variability in these estimates.

Table 5.3: Posterior estimates and 95% CI for β

Key Population	Posterior Estimates
1	-0.25 (-0.40, -0.10)
2	-0.13 (-0.24, -0.03)

Finally, similarly to the prevalence models, convergence to the target distribution was assessed with trace plots and MCMC diagnostic tools. No convergence issues were diagnosed.

5.4.3 National Prevalence Estimates

The national prevalence estimates that are presented in this section for the two key populations are based on model 1 for the prevalence and on model 2 for the population size.

Table 5.4 displays estimates based on the current methodology and compares them to the proposed revised estimates. We first note that both methods produce point estimates that are lower than the current methodology. This is due to the fact that the larger prevalence point estimates have larger variability than the smaller point estimates. Therefore, the larger estimates are pulled towards the smaller estimates.

The 95% CI is substantially wider for KP_2 than the one produced from current methodology. This is due to the fact that the revised estimator accounts for the target population size estimates uncertainty as well as the additional uncertainty for including regions for which no data were collected. This effect is however not observed for KP_1 since the prevalence point estimates do not vary as much across cities.

Table 5.4: Comparison of national prevalence estimates

Key Population	Current Method	Bayes Estimates
1	.082 (.034, .128)	.060 (.037, .117)
2	.222 (.146, .298)	.163 (.070, .346)

5.5 Conclusion and discussion

In summary, we have developed Bayesian models to improve the estimation of the national HIV prevalence among high risk populations. The developed methodology overcomes some of the issues encountered with the current practice. First, the developed estimator incorporates the target population size estimate uncertainty. Also, it accounts for regions where no data have been collected. Finally, similar to the work of Bao et al. [2015], it incorporates multiple sources of data about the target population size which are often available.

The main limitations of the proposed methodology are that, first, the model assumes that the unobserved regions are similar to the observed ones. However, this assumption may not always be reasonable in practice. As discussed in the Section 5.1, in resource limited settings, the regions might not be missing at random. In our future work, we would like to include predictive variables for the prevalence and the target population size estimates that would help factor potential dissimilarities across regions or cities.

A second limitation to our work is that our analysis depends on only five surveys in the entire country. Ideally, it would be better to have a larger number of surveys to fit the models.

Finally, the choice of prior for the target population size estimates is rather informative. Determining the sensitivity of the results to the choice of prior will be assessed in future work.

APPENDIX A

PERFORMANCE OF THE ANALYTICAL ADJUSTMENT WITH THE SALGANIK-HECKATHORN ESTIMATOR

We discuss here why the c-factor and its observed version c^* both play a role in whether or not the linear adjustment applies to the Salganik-Heckathorn estimator. The argument is based on the fact that, if we assume a random walk at stationarity, this implies that $c \to 1$ and $c^* \to 1$. As such, we also have that:

1. $\lim_{c\to 1} \frac{\hat{\mu}_{VH}^{adj}}{\hat{\mu}_{VH}^{adj}+c(1-\hat{\mu}_{VH}^{adj})} = \hat{\mu}_{VH}^{adj}$ and

2. $\lim_{c^*\to 1} \hat{\mu}_{VH}^{naive} = \hat{\mu}_{SH}^{naive}$ or equivalently, $\lim_{c^*\to 1} \hat{\mu}_{VH}^{adj} = \hat{\mu}_{SH}^{adj}$

Therefore, under those conditions,

$$\hat{\mu}_{SH}^{adj} \approx \frac{\hat{\mu}_{VH}^{adj}}{\hat{\mu}_{VH}^{adj} + c(1-\hat{\mu}_{VH}^{adj})}.$$

By definition of the analytical adjustment given by equation (3.2), we also have that:

$$\hat{\mu}_{SH}^{adj} = \frac{\hat{\mu}_{SH}^{naive}-f^+}{1-f^+-f^-}.$$

By relating the right hand side of the two equations above and by using the following relations:

1. $\hat{\mu}_{SH}^{naive} = \frac{\hat{\mu}_{VH}^{naive}}{\hat{\mu}_{VH}^{naive}+c^*(1-\hat{\mu}_{VH}^{naive})}$

2. $\hat{\mu}_{VH}^{adj} = \frac{\hat{\mu}_{VH}^{naive}-f^+}{1-f^+-f^-}$,

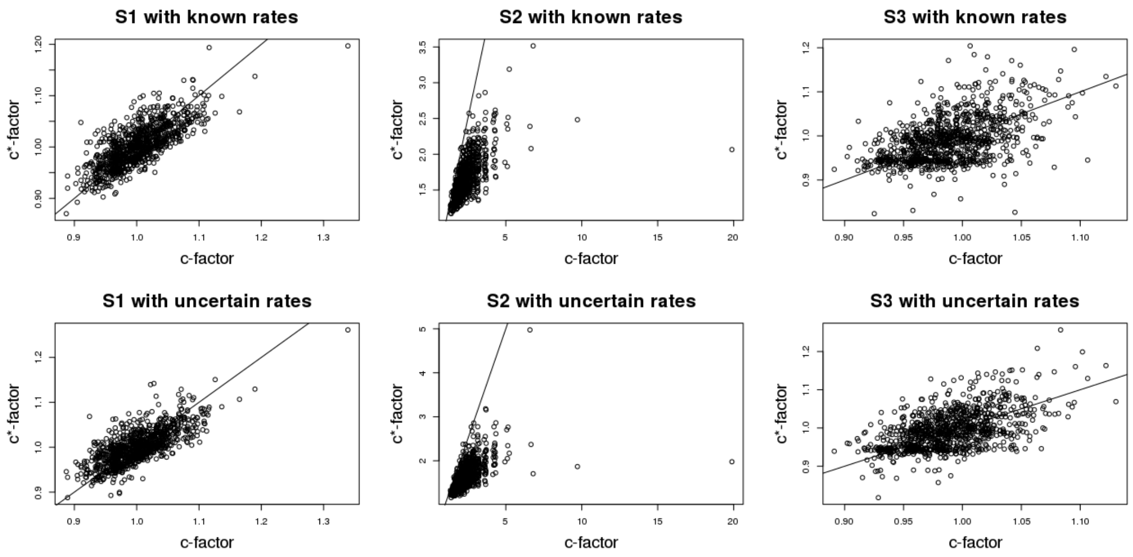


Figure A.1: Relation between c^* and c under the three scenarios of the simulation study.

we obtain:

$$\begin{aligned}
&\frac{\hat{\mu}_{VH}^{naive}}{\hat{\mu}_{VH}^{naive} + c^*(1-\hat{\mu}_{VH}^{naive})} \\
&\qquad\approx \frac{(\hat{\mu}_{VH}^{naive} - f^+)(1-f^+-f^-) + f^+(\hat{\mu}_{VH}^{naive} - f^+ + c(1-f^- - \hat{\mu}_{VH}^{naive}))}{\hat{\mu}_{VH}^{naive} - f^+ + c(1-f^- - \hat{\mu}_{VH}^{naive})}.
\end{aligned}$$

Therefore, when the random walk at stationarity assumption is met, that is, when $c \to 1$ and $c^* \to 1$, the limit on each side of the equation when $c^* \to 1$ and $c \to 1$, respectively, are equal. However, other values for c and/or c^* may create a discrepancy between the two sides of the equation thus indicating that the analytical adjustment will have a poor performance. This has been found to indeed create biases in the simulations. In practice, it is not possible to calculate c in presence of misclassification. Although no exact linear relationship exists between c and c^*, as seen in Figure from our simulations, they tend to be positively correlated. As such, a high c^* may imply an elevated c and should serve as an indicator that the analytical adjustment might not be the best correction method for the Salganik-Heckathorn estimator.

APPENDIX B

ADDITIONAL RESULTS FROM MISCLASSIFICATION SIMULATION STUDY

Root Mean-Squared-Error

Table B.1: Absolute average bias, standard deviation and RMSE for the naive and corrected estimators under S1-S3 with fixed and uncertain misclassification rates. The RMSE's in bold indicates that the method is in the best set of correction methods for a particular scenario and estimator based on Bonferroni pairwise comparison at a family-wise error rate of 5%.

			Absolute Bias				Standard Deviation				RMSE			
			$\hat{\mu}^{naive}$	$\hat{\mu}^{adj}$	$\hat{\mu}^{lin}$	$\hat{\mu}^{quad}$	$\hat{\mu}^{naive}$	$\hat{\mu}^{adj}$	$\hat{\mu}^{lin}$	$\hat{\mu}^{quad}$	$\hat{\mu}^{naive}$	$\hat{\mu}^{adj}$	$\hat{\mu}^{lin}$	$\hat{\mu}^{quad}$
Fixed	$\hat{\mu}_{mean}$	S1	.0822	.0013	.0161	.0021	.0174	.0193	.0190	.0198	.0840	**.0193**	.0249	.0199
		S2	.0576	.0002	.0103	.0006	.0163	.0176	.0173	.0182	.0598	**.0175**	.0201	.0182
		S3	.1345	.0006	.0668	.0209	.0215	.0369	.0273	.0403	.1362	**.0369**	.0721	.0454
	$\hat{\mu}_{VH}$	S1	.0822	.0013	.0161	.0022	.0195	.0214	.0211	.0221	.0845	**.0215**	.0266	.0222
		S2	.0570	.0002	.0102	.0007	.0178	.0193	.0190	.0200	.0597	**.0193**	.0216	.0201
		S3	.1019	.0001	.0507	.0160	.0197	.0333	.0247	.0365	.1037	**.0333**	.0564	.0398
	$\hat{\mu}_{SS}$	S1	.0822	.0013	.0161	.0022	.0191	.0210	.0207	.0216	.0844	**.0211**	.0262	.0217
		S2	.0570	.0002	.0102	.0007	.0175	.0190	.0187	.0197	.0596	**.0190**	.0213	.0197
		S3	.1054	.0001	.0525	.0166	.0196	.0333	.0247	.0365	.1072	**.0333**	.0580	.0401
	$\hat{\mu}_{SH}$	S1	.0822	.0013	.0161	.0021	.0201	.0221	.0218	.0228	.0846	**.0222**	.0271	.0229
		S2	.1156	.0444	.0455	.0141	.0384	.0410	.0415	.0469	.1218	.0604	.0615	**.0490**
		S3	.1029	.0006	.0510	.0155	.0205	.0349	.0259	.0386	.1049	**.0349**	.0572	.0416
Uncertain	$\hat{\mu}_{mean}$	S1	.0806	.0004	.0143	.0005	.0222	.0246	.0242	.0251	.0836	**.0246**	.0281	.0251
		S2	.0575	.0002	.0103	.0006	.0193	.0211	.0207	.0217	.0607	**.0211**	.0231	.0217
		S3	.1356	.0019	.0686	.0230	.0242	.0445	.0325	.0460	.1377	**.0445**	.0759	.0514
	$\hat{\mu}_{VH}$	S1	.0808	.0003	.0145	.0007	.0239	.0264	.0260	.0271	.0842	**.0264**	.0298	.0271
		S2	.0571	.0000	.0104	.0008	.0207	.0224	.0221	.0231	.0607	**.0224**	.0244	.0231
		S3	.1031	.0027	.0528	.0185	.0218	.0385	.0285	.0403	.1054	**.0386**	.0600	.0443
	$\hat{\mu}_{SS}$	S1	.0808	.0003	.0145	.0007	.0236	.0261	.0257	.0267	.0841	**.0261**	.0295	.0267
		S2	.0571	.0001	.0103	.0008	.0205	.0222	.0218	.0229	.0607	**.0222**	.0241	.0229
		S3	.1066	.0027	.0546	.0190	.0219	.0388	.0287	.0405	.1089	**.0389**	.0617	.0447
	$\hat{\mu}_{SH}$	S1	.0809	.0001	.0147	.0009	.0243	.0270	.0265	.0278	.0845	**.0269**	.0303	.0278
		S2	.1163	.0452	.0463	.0150	.0429	.0463	.0474	.0534	.1240	.0647	.0662	**.0554**
		S3	.1044	.0030	.0537	.0191	.0225	.0397	.0295	.0423	.1068	**.0398**	.0613	.0464

RMSE at Various Levels of Misclassification Rates

Figure B.1 presents the average RMSE improvement when using the analytical adjustment compared to the naive approach, that is:

$$\frac{\overline{RMSE}_{naive} - \overline{RMSE}_{adj}}{\overline{RMSE}_{naive}}.$$

The calculations were performed with false positive and negative rates varying from 0 to 0.4 by 0.04 increments and under scenario 1 for the Volz-Heckathorn estimator. The average improvement is expressed as a function of the average misclassification bias present in the estimates. In most instances, the RMSE is significantly lower than under the naive approach. The limited instances where the average RMSE with the analytical adjustment is higher than the naive one occur when the estimates contain little misclassification bias. In those cases, the benefits from the reduction in misclassification bias are offset by the increase in the uncertainty of the estimates.

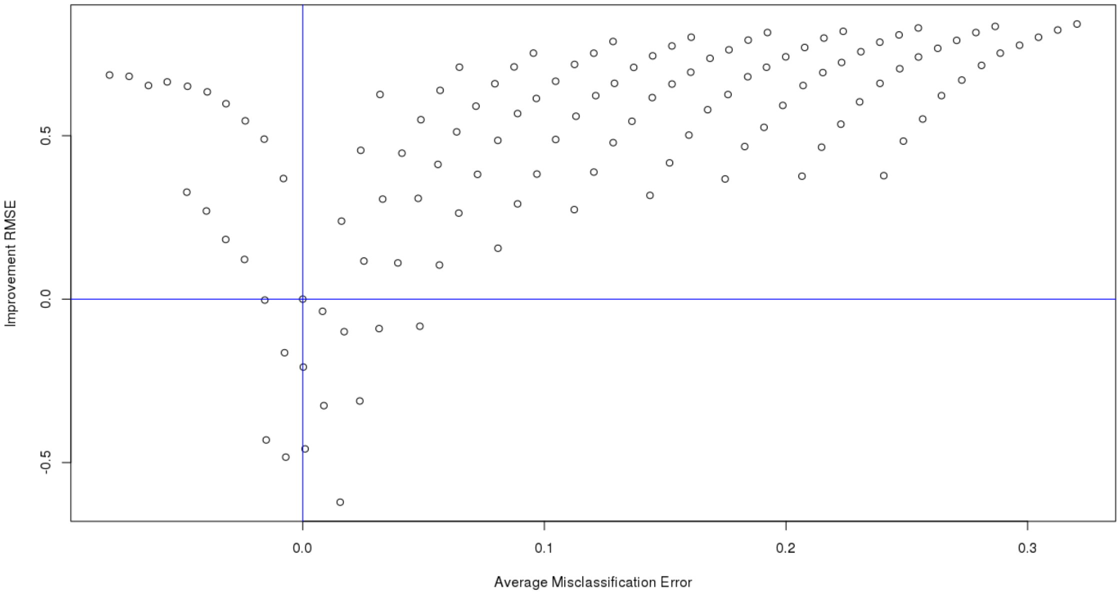
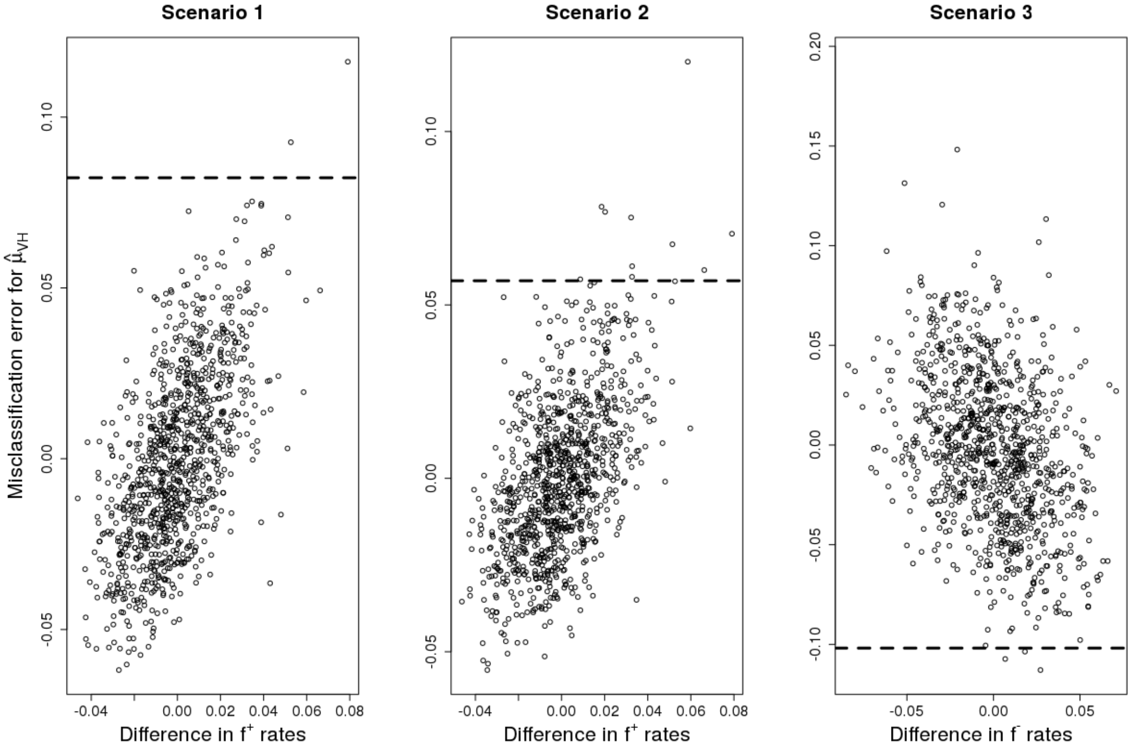


Figure B.1: Relative decrease in the average RMSE for the Volz-Heckathorn estimator under S1 as a function of the average misclassification bias in the estimates.

Sensitivity to Erroneous Error Rates

Figure B.2 shows the misclassification bias still present in the Volz-Heckathorn estimates after applying the analytical adjustment when inaccurate misclassification error rates (i.e. f^+ and f^-) are used in equation (3.2). The impact of inaccurate rates is presented for S1 to S3 at various levels of inaccuracy in f^+ and f^-. The relation is shown in terms of f^+ for S1 and S2 and f^- for S3 since those rates significantly deviate from the true rates under the corresponding scenarios. As for the dash line, it represents the average misclassification bias in the naive point estimate. Very few point estimates in either of the three scenarios contain more misclassification bias than the one present in the average naive point estimate. This suggests that for moderate departure from the true misclassification rates, the correction methods may still result in less misclassification bias than the naive approach. Although Figure B.2 is based on inaccurate f^+ and f^- for all scenarios, the uncertain f^-'s in S1 and S2 and f^+'s in S3 are fairly close to the true rate.



Scenario 1

Scenario 2

Scenario 3

Figure B.2: Misclassification error remaining in the VH estimates ($\hat{\mu}_{VH}$) after applying the analytical adjustment for S1 to S3 as a function of the inaccuracy in the error rates (either f^+ or f^-). The dash line represents the average misclassification bias in the naive point estimate.

APPENDIX C

NATIONAL PREVALENCE ESTIMATION SUPPLEMENT

C.1 Alternative Prevalence Models and Results

In this section, we describe alternative models that were assessed to model prevalence. Similarly to the log-normal model proposed in Section 5.3.1, each data model has two versions: one partial pooling by city and one partial pooling by region.

Model 1 (Normal) - Partial Pooling by City

$$y_i|\pi_i^{m1}, v_i \sim N(\pi_i^{m1}, v_i)$$
$$\pi_i^{m1}|\mu_{\pi^{m1}}, \sigma^2_{\pi^{m1}} \sim N(\mu_{\pi^{m1}}, \sigma^2_{\pi^{m1}}).$$

Model 2 (Normal) - Partial Pooling by Region

$$y_i|\pi_{j[i]}^{m1}, v_i \sim N(\pi_{j[i]}^{m2}, v_i)$$
$$\pi_j^{m2}|\mu_{\pi^{m2}}, \sigma^2_{\pi^{m2}} \sim N(\mu_{\pi^{m2}}, \sigma^2_{\pi^{m2}}).$$

The prior distributions are as follows:

$$\pi^{m1} \sim Uniform(0, 1)$$
$$\sigma_{\pi^{m1}} \sim Uniform(0, 1)$$
$$\pi^{m2} \sim Uniform(0, 1)$$
$$\sigma_{\pi^{m2}} \sim Uniform(0, 1).$$

Model 1 (Beta) - Partial Pooling by City

$$y_i|a_i, b_i \sim beta(a_i, b_i)$$

$$\pi_i \sim beta(\alpha_\pi, \beta_\pi), \text{where}$$

$$a_i = \ \pi_i^2((1-\pi_i)/v_i - 1/\pi_i) \text{ and } b_i = \ a_i(1/\pi_i - 1)$$

Model 2 (Beta) - Partial Pooling by Region

$$y_i|a_i, b_i \sim beta(a_i, b_i)$$

$$\pi_j \sim beta(\alpha_\pi, \beta_\pi), \text{where}$$

$$a_i = \ \pi_{j[i]}^2((1-\pi_{j[i]})/v_i - 1/\pi_{j[i]}) \text{ and } b_i = \ a_i(1/\pi_{j[i]} - 1)$$

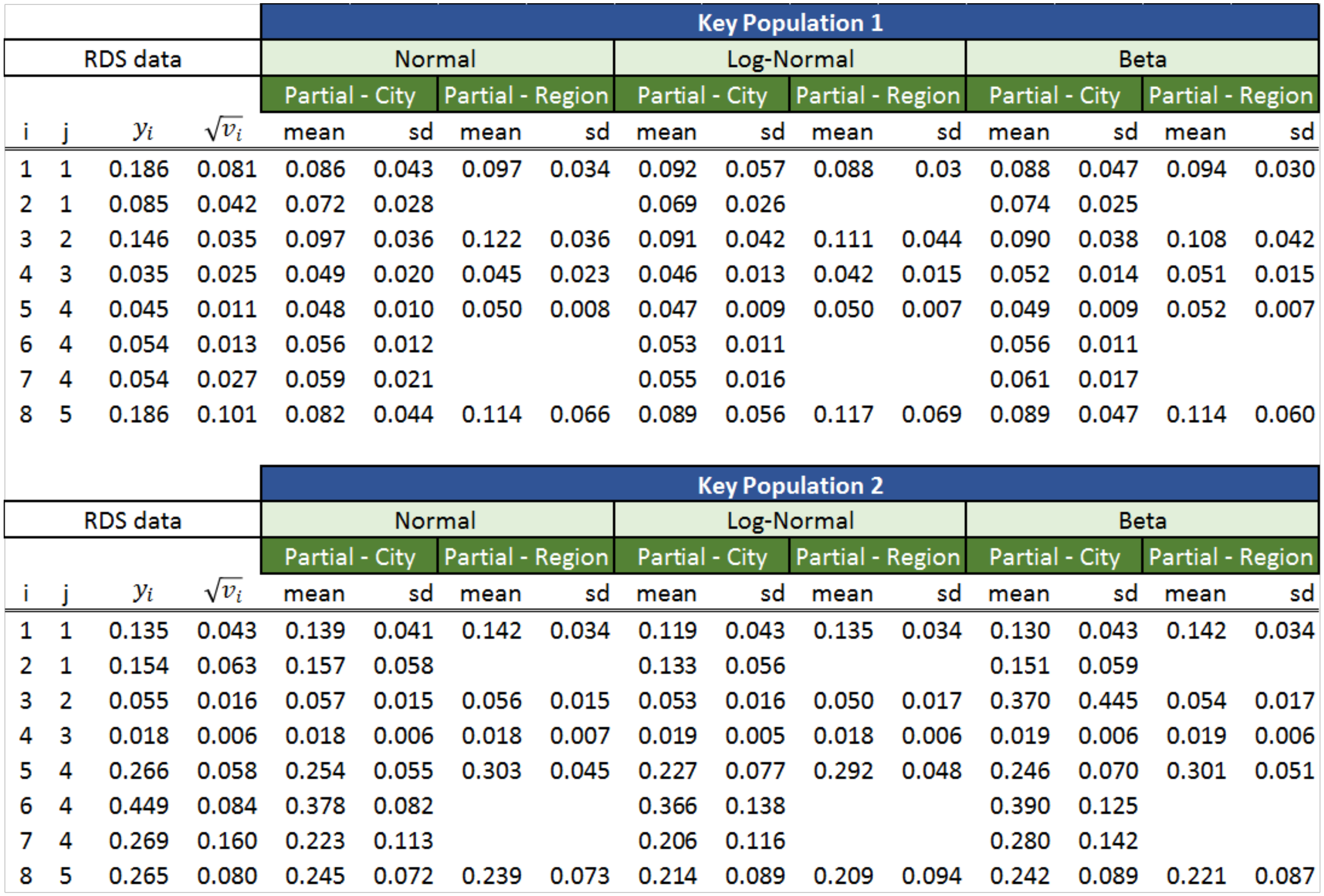
The prior distributions are as follows:

$$\alpha_\pi \sim gamma(0.001, 0.001)$$

$$\beta_\pi \sim gamma(0.001, 0.001)$$

The priors distributions were chosen to be vague. Also, in all models, the variance of the data models is assumed known.

Figure C.1: Observed prevalence estimates and their standard deviation along with posterior prevalence estimates under six models for two key populations



				Key Population 1											
RDS data				Normal				Log-Normal				Beta			
				Partial - City		Partial - Region		Partial - City		Partial - Region		Partial - City		Partial - Region	
i	j	y_i	$\sqrt{v_i}$	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
1	1	0.186	0.081	0.086	0.043	0.097	0.034	0.092	0.057	0.088	0.03	0.088	0.047	0.094	0.030
2	1	0.085	0.042	0.072	0.028			0.069	0.026			0.074	0.025		
3	2	0.146	0.035	0.097	0.036	0.122	0.036	0.091	0.042	0.111	0.044	0.090	0.038	0.108	0.042
4	3	0.035	0.025	0.049	0.020	0.045	0.023	0.046	0.013	0.042	0.015	0.052	0.014	0.051	0.015
5	4	0.045	0.011	0.048	0.010	0.050	0.008	0.047	0.009	0.050	0.007	0.049	0.009	0.052	0.007
6	4	0.054	0.013	0.056	0.012			0.053	0.011			0.056	0.011		
7	4	0.054	0.027	0.059	0.021			0.055	0.016			0.061	0.017		
8	5	0.186	0.101	0.082	0.044	0.114	0.066	0.089	0.056	0.117	0.069	0.089	0.047	0.114	0.060

				Key Population 2											
RDS data				Normal				Log-Normal				Beta			
				Partial - City		Partial - Region		Partial - City		Partial - Region		Partial - City		Partial - Region	
i	j	y_i	$\sqrt{v_i}$	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
1	1	0.135	0.043	0.139	0.041	0.142	0.034	0.119	0.043	0.135	0.034	0.130	0.043	0.142	0.034
2	1	0.154	0.063	0.157	0.058			0.133	0.056			0.151	0.059		
3	2	0.055	0.016	0.057	0.015	0.056	0.015	0.053	0.016	0.050	0.017	0.370	0.445	0.054	0.017
4	3	0.018	0.006	0.018	0.006	0.018	0.007	0.019	0.005	0.018	0.006	0.019	0.006	0.019	0.006
5	4	0.266	0.058	0.254	0.055	0.303	0.045	0.227	0.077	0.292	0.048	0.246	0.070	0.301	0.051
6	4	0.449	0.084	0.378	0.082			0.366	0.138			0.390	0.125		
7	4	0.269	0.160	0.223	0.113			0.206	0.116			0.280	0.142		
8	5	0.265	0.080	0.245	0.072	0.239	0.073	0.214	0.089	0.209	0.094	0.242	0.089	0.221	0.087

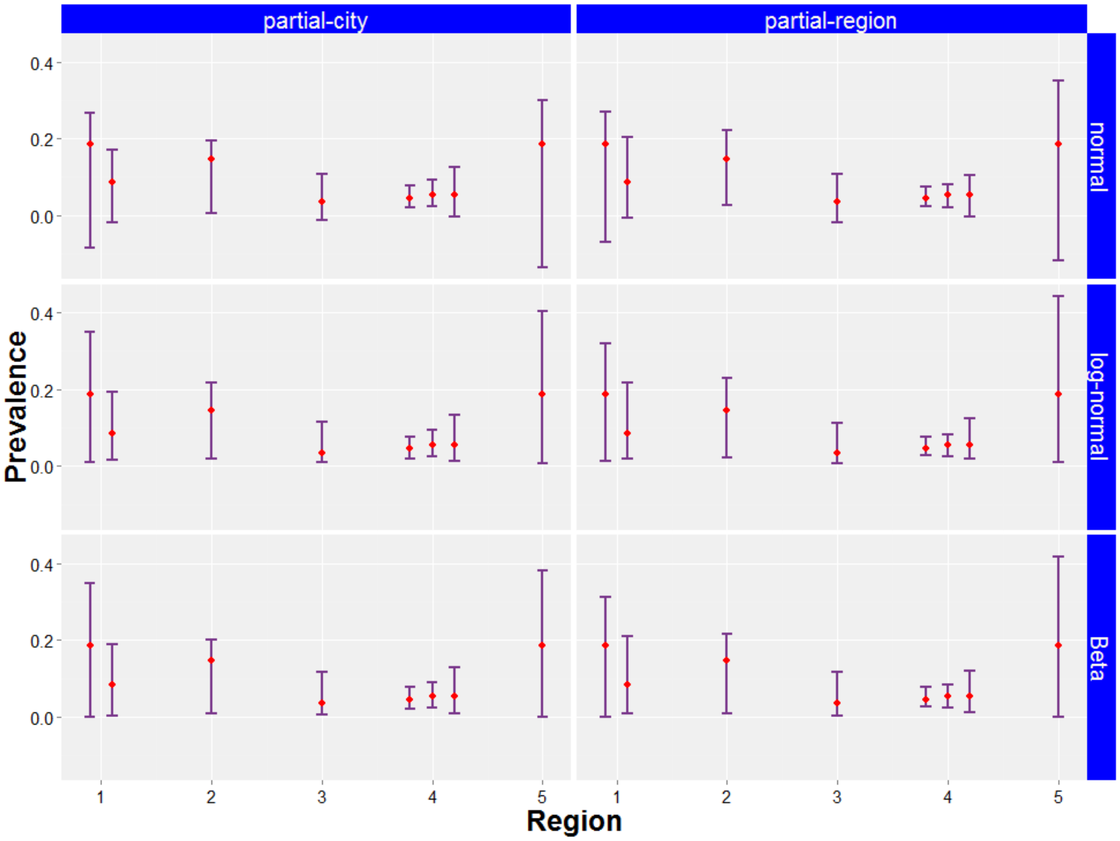


Figure C.2: 95% predictive intervals (PI) for the prevalence of key population 1 along with the observed prevalence estimates y_i depicted by red dots under six models.

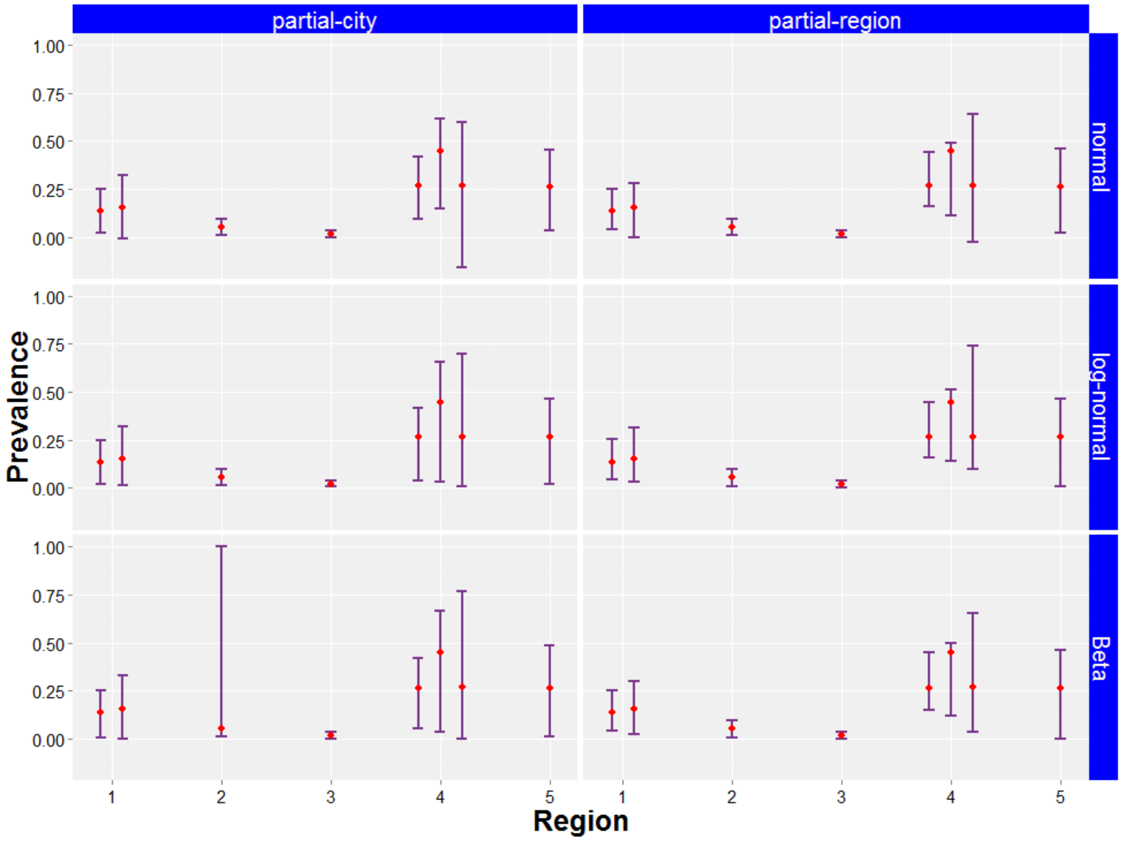


Figure C.3: 95% predictive intervals (PI) for the prevalence of key population 2 along with the observed prevalence estimates y_i depicted by red dots under six models.

BIBLIOGRAPHY

GC Jayaraman C. Major DM Patrick SM Houston Archibald, CP and D. Sutherland. Estimating the Size of Hard-to-reach Populations: A Novel Method Using HIV Testing Data Compared to Other Methods. *AIDS (London, England)*, 15:41–48, 2001.

Le Bao, Adrian E Raftery, and Amala Reddy. Estimating the Sizes of Populations At Risk of HIV Infection From Multiple Data Sources Using a Bayesian Hierarchical Model. *Statistics and its interface*, 8(2):125–136, 2015.

Bruce A. Barron. The effects of misclassification on the estimation of relative risk. *Biometrics*, 33(2):414–418, 1977.

P. Biernacki and D. Waldorf. Snowball sampling: problem and techniques of chain referral sampling. *Sociological Methods and Research*, 10:141–163, 1981.

John P. Buonaccorsi. *Measurement error: models, methods, and applications*. Chapman & Hall, New York, 2010.

J.R. Cook and L. A. Stefanski. Simulation extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89 (428):1314–1328, 1994.

Sergio Currarini, Matthew O. Jackson, and Paolo Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009. ISSN 1468-0262.

Ove Frank and David Strauss. Markov graphs. *Journal of the American Statistical Association*, 81:832–842, 1986.

Simon D. W. Frost, Kimberly C. Brouwer, Michelle A. Firestone Cruz, Rebeca Ramos, Maria Elena Ramos, Remedios M. Lozada, Carols Magis-Rodriguez, and Steffanie A. Strathdee. Respondent-driven sampling of injection drug users in two U.S.-Mexico border cities: Recruitment dynamics and impact on estimates of HIV and syphilis prevalence. *Journal of Urban Health*, 83:83–97, 2006.

Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 1992.

John Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics 4 : Proceedings of the Fourth Valencia International Meeting, April 15-20, 1991 (Edited by J.M. Bernardo Et Al.)*, 1992.

Krista J. Gile. Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *Journal of the American Statistical Association*, 106:135–146, 2011.

Krista J. Gile and Mark S. Handcock. Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology*, 40:285–327, 2010.

Krista J. Gile, Lisa G. Johnston, and Matthew J. Salganik. Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):241–269, 2015. ISSN 1467-985X.

Leo A Goodman. Snowball sampling. *Annals of Mathematical Statistics*, 32:148–170, 1961.

Mark S. Handcock and Krista J. Gile. Comment: On the concept of snowball sampling. *Sociological Methodology*, 41(1):367–371, 2011.

Mark S. Handcock, Krista J. Gile, and Corinne M. Mar. Estimating hidden population size using respondent-driven sampling data. *Electronic Journal of Statistics*, 8: 1491–1521, 2014.

Mark S. Handcock, Ian E. Fellows, and Krista J. Gile. *RDS: Respondent-Driven Sampling*. Los Angeles, CA, 2015a. URL `http://CRAN.R-project.org/package=RDS`. R package version 0.7-2.

Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, Skye Bender-deMoll, and Martina Morris. *statnet: Software Tools for the Statistical Analysis of Network Data*. The Statnet Project (`http://www.statnet.org`), 2015b. URL `CRAN.R-project.org/package=statnet`. R package version 2015.6.2.

Morris H. Hansen and William N. Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362, 1943.

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

Philip Heidelberger and Peter Welch. A spectral method for confidence interval generation and run length control in simulations: a frame of reference. *Communications of the ACM*, 24(4):233–245, 1981.

David R. Hunter and Mark S. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15:565–583, 2006.

David R. Hunter, Steven M. Goodreau, and Mark S. Handcock. Goodness of fit for social network models. *Journal of the American Statistical Association*, 103: 248–258, 2008.

Martin Y. Iguchi, Allison J. Ober, Sandra H. Berry, Terry Fain, Douglas D. Heckathorn, Pamina M. Gorbach, Robert Heimer, Andrei Kozlov, Lawrence J. Ouellet, Steven Shoptaw, and William A. Zule. Simultaneous Recruitment of Drug Users and Men Who Have Sex with Men in the United States and Russia Using Respondent-Driven Sampling: Sampling Methods and Implications. *Journal of Urban Health*, 86(S1):5–31, 2009.

Lisa G. Johnston, Moshen Malekinejad, Carl Kendall, Irene M. Iuppa, and George W. Rutherford. Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: Field experiences in international settings. *AIDS and Behavior*, 12:131–141, 2008.

Denise B. Kandel. Homophily, selection, and socialization in adolescent friendships. *American Journal of Sociology*, 84(2):427–436, 1978.

Helmut Kuchenhoff, Samuel M. Mwalili, and Emmanuel Lesaffre. A general method for dealing with misclassification in regression: The misclassification simex. *Biometrics*, 62(1):85–96, 2006.

Helmut Kuchenhoff, Wolfgang Lederer, and Emmanuel Lesaffre. Asymptotic variance estimation for the misclassification SIMEX. *Computational Statistics and Data Analysis*, 51(12):6197–6211, 2007.

Roderick J. A. Little and Donald B. Rubin. *Nonignorable Missing-Data Models*, pages 312–348. John Wiley and Sons, Inc., 2002.

Hongjie Liu, Tiejian Feng, Hui Liu, Hucang Feng, Yumao Cai, Anne G. Rhodes, and Oscar Grusky. Egocentric Networks of Chinese Men Who Have Sex with Men: Network Components, Condom Use Norms, and Safer Sex. *AIDS Patient Care and STDs*, 23(10):885–893, oct 2009. ISSN 1087-2914. doi: 10.1089/apc.2009.0043. URL `http://www.liebertonline.com/doi/abs/10.1089/apc.2009.0043`.

Hongjie Liu, Jianhua Li, Toan Ha, and Jian Li. Assessment of Random Recruitment Assumption in Respondent-Driven Sampling in Egocentric Network Data. *Social networking*, 1(2):13–21, 2012. ISSN 2169-3285.

Xin Lu. Linked Ego Networks: Improving estimate reliability and validity with respondent-driven sampling. *Social Networks*, 35(4):669–685, 2013. ISSN 03788733.

Xin Lu, Linus Bengtsson, Tom Britton, Martin Camitz, Beom Jun Kim, Anna Thorson, and Fredrik Liljeros. The sensitivity of respondent-driven sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(1):191–216, 2012. ISSN 1467-985X.

Xin Lu, Jens Malmros, Fredrik Liljeros, and Tom Britton. Respondent-driven Sampling on Directed Networks. *Electronic Journal of Statistics*, 7:292–322, 2013.

Gregory M. Lucas, Sunil S. Solomon, Aylur K. Srikrishnan, Alok Agrawal, Syed Iqbal, Oliver Laeyendecker, Allison M. McFall, Muniratnam S. Kumar, Elizabeth L. Ogburn, David D. Celentano, Suniti Solomon, and Shruti H. Mehta. High HIV burden among people who inject drugs in 15 Indian cities. *AIDS*, page 1, 2015. ISSN 0269-9370.

Mohsen Malekinejad, Lisa Johnston, Carl Kendall, Ligia Kerr, Marina Rifkin, and George Rutherford. Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review. *AIDS and Behavior*, 12:105–130, 2008.

Gary Marks, Nicole Crepaz, J Walton Senterfitt, and Robert S Janssen. Meta-analysis of high-risk sexual behavior in persons aware and unaware they are infected with hiv in the united states: implications for hiv prevention programs. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 39(4):446–453, 2005.

Nicky Mccreesh, Simon D W Frost, Janet Seeley, Joseph Katongole, Matilda N Tarsh, Richard Ndunguse, Fatima Jichi, Natasha L Lunel, Dermot Maher, Lisa G Johnston, Pam Sonnenberg, Andrew J Copas, Richard J Hayes, and Richard G White. Evaluation of Respondent-driven Sampling. *Epidemiology (Cambridge, Mass.)*, 23 (1):138–47, 2012.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

Harriet L. Mills, Samuel Johnson, Matthew Hickman, Nick S. Jones, and Caroline Colijn. Errors in reported degrees and respondent driven sampling: Implications for bias. *Drug and Alcohol Dependence*, 142(0):120 – 126, 2014. ISSN 0376-8716.

Michael S. Molloy and Bruce A. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6:161–179, 1995.

JaneR. Montealegre, LisaG. Johnston, Christopher Murrill, and Edgar Monterroso. Respondent driven sampling for hiv biological and behavioral surveillance in latin america and the caribbean. *AIDS and Behavior*, 17(7):2313–2340, 2013.

AbbyE. Rudolph, CrystalM. Fuller, and Carl Latkin. The importance of measuring and accounting for potential biases in respondent-driven samples. *AIDS and Behavior*, 17(6):2244–2252, 2013. ISSN 1090-7165.

Matthew J. Salganik. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *Journal of Urban Health*, 83, 2006.

Matthew J. Salganik and Douglas D. Heckathorn. Sampling and estimation in hidden populations using respondent-drive sampling. *Sociological Methodology*, 34:193–239, 2004.

Leslie Shanks, Derryck Klarkowski, and Daniel P. O'Brien. False Positive HIV Diagnoses in Resource Limited Settings: Operational Lessons Learned for HIV Programmes. *PLoS ONE*, 8(3):8–13, 2013.

R Smith, K Rossetto, and BL Peterson. A meta-analysis of disclosure of one's hiv-positive status, stigma and social support. *AIDS Care*, 20(10):1266 – 1275, 2008. ISSN 0954-0121.

Sunil S. Solomon, Shruti H. Mehta, Aylur K. Srikrishnan, Canjeevaram K. Vasudevan, Allison M. Mcfall, Pachamuthu Balakrishnan, Santhanam Anand, Panneerselvam Nandagopal, Elizabeth L. Ogburn, Oliver Laeyendecker, Gregory M. Lucas, Suniti Solomon, and David D. Celentano. High HIV prevalence and incidence among MSM across 12 cities in India. *AIDS*, 29:723–731, 2015. ISSN 0269-9370.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and pages = 583–639 title = Bayesian measures of model complexity and fit volume = 64 year = 2002 van der journal = Journal of the Royal Statistical Society: Series B (Statistical Methodology), number = 4.

Amber Tomas and Krista J. Gile. The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electronic Journal of Statistics*, 5:899–934, 2011.

Martin Trow. *Right-Wing Radicalism and Political Intolerance.* Arno Press, New York, 1957. Reprinted 1980.

UNAIDS. The gap report. 2014.

UNAIDS/WHO. Guidelines on estimating the size of populations most at risk to hiv. Technical report, UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance, 2010. URL `http://files.unaids.org/en/media/unaids/contentassets/documents/epidemiology/2011/2011_Estimating_Populations_en.pdf`.

Ashton M. Verdery, M. Giovanna Merli, James Moody, Jeffrey A. Smith, and Jacob C. Fisher. Respondent-driven sampling estimators under real and theoretical recruitment conditions of female sex workers in China. *Epidemiology*, 26(5):661–665, 2015.

Erik Volz and Douglas D. Heckathorn. Probability based estimation theory for Respondent Driven Sampling. *The Journal of Official Statistics*, 24(1):79–97, 2008.

Jichuan Wang, Robert G. Carlson, Russel S. Falck, Harvey A. Siegal, Ahmmed Rahman, and Linna Li. Respondent-driven sampling to recruit MDMA users: a methodological assessment. *Drug and Alcohol Dependence*, 78:147–157, 2005.

Cyprian Wejnert and Douglas D. Heckathorn. Web-based network sampling: Efficiency and efficacy of respondent-driven sampling for online research. *Sociological Methods and Research*, 37:105–134, 2008.

Mike West. Inference in successive sampling discovery models. *Journal of Econometrics*, 75:217–238, 1996.

World Health Organization. Consolidated guidelines on HIV testing services 2015. Technical report, World Health Organization, 2015. URL `http://www.who.int/hiv/pub/guidelines/hiv-testing-services/en/`.

Thespina J. Yamanis, M Giovanna Merli, William Whipple Neely, Felicia Feng Tian, James Moody, Xiaowen Tu, and Ersheng Gao. An Empirical Analysis of the Impact of Recruitment Patterns on RDS Estimates Among a Socially Ordered Population of Female Sex Workers in China. *Sociological Methods & Research*, 42(3):392–425, 2013.

Frank Yates and P. Michael Grundy. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):253–261, 1953.