11-2019

Reduced Bias for Respondent Driven Sampling: Accounting for Non-Uniform Edge Sampling Probabilities in People Who Inject Drugs in Mauritius

Miles Q. Ott
Smith College, mott@smith.edu

Krista J. Gile
University of Massachusetts Amherst

Matthew T. Harrison
Brown University

Lisa G. Johnston

Joseph W. Hogan
Brown University

Recommended Citation

Ott, Miles Q.; Gile, Krista J.; Harrison, Matthew T.; Johnston, Lisa G.; and Hogan, Joseph W., "Reduced Bias for Respondent Driven Sampling: Accounting for Non-Uniform Edge Sampling Probabilities in People Who Inject Drugs in Mauritius" (2019). Statistical and Data Sciences: Faculty Publications, Smith College, Northampton, MA.
https://scholarworks.smith.edu/sds_facpubs/12

Reduced Bias for respondent driven sampling: accounting for non-uniform edge sampling probabilities in people who inject drugs in Mauritius

Abstract

People who inject drugs are an important population to study in order to reduce transmission of blood-borne illnesses including HIV and Hepatitis. In this paper we estimate the HIV and Hepatitis C prevalence among people who inject drugs, as well as the proportion of people who inject drugs who are female in Mauritius. Respondent driven sampling (RDS), a widely adopted link-tracing sampling design used to collect samples from hard-to-reach human populations, was used to collect this sample. The random walk approximation underlying many common RDS estimators assumes that each social relation (edge) in the underlying social network has an equal probability of being traced in the collection of the sample. This assumption does not hold in practice. We show that certain RDS estimators are sensitive to the violation of this assumption. In order to address this limitation in current methodology, and the impact it may have on prevalence estimates, we present a new method for improving RDS prevalence estimators using estimated edge inclusion probabilities, and apply this to data from Mauritius.

KEY WORDS: Respondent Driven Sampling, Link-Tracing, Network Sampling, Edge Inclusion, HIV, AIDS, Hepatitis

Authors: Miles Q. Ott, Krista J. Gile, Matthew T. Harrison, Lisa G. Johnston, Joseph W. Hogan

Corresponding Author: Miles Q. Ott, mott@smith.edu, Statistical and Data Sciences Program, Smith College, 7 College Lane, Northampton MA 01063

1 Introduction

1.1 Injection Drug Use in Mauritius

Mauritius is estimated to have one of the highest per capita percentages of people who inject drugs (PWID) of all African countries (Johnston et al., 2013; National AIDS Secretariat, 2014). This high rate of injection drug use has seriously impacted public health, as it is the primary mode of HIV transmission within Mauritius, and accounts for 44% of all HIV transmissions in the country (Mau, 2015). In order to measure HIV and other infections' prevalence and associated risk factors in Mauritius, a sample of 500 PWID was collected using respondent driven sampling (RDS, Heckathorn (1997)) as part of a biological and behavioral surveillance survey in 2011(Johnston et al., 2011). In that survey, PWID were defined as males or females, of at least fifteen years in age who injected drugs in the previous three months and were living in Mauritius. In this paper we estimate the HIV and Hepatitis C prevalence, and the proportion of PWID in Mauritius who are female. As the data were collected using RDS, we next describe the RDS recruitment process, as well as the accompanying estimation methods and the assumptions that they require to produce valid inference.

1.2 Respondent Driven Sampling Background

RDS is a network sampling method typically used to infer population proportions of binary traits in hard-to-reach human populations. RDS has been widely adopted to estimate the prevalence of disease or risk behaviors within high-risk hard-to-reach human populations, including PWID, sex workers, and men who have sex with men. It has been used in hundreds of studies around the world (Johnston et al., 2008, 2016; Montealegre et al., 2013), for surveys of biological behavioral surveillance funded by the Global Fund to monitor HIV prevalence, assess risk and program coverage and to measure trends over time (Lansky et al., 2007). Despite its wide use in settings of public health importance, the statistical properties and optimal inferential strategies for data resulting from RDS still require much additional study.

In RDS, the sampling is a variant of a link-tracing network sampling procedure (Handcock and Gile, 2011). Link-tracing sampling has been widely used in hard-to-reach populations (Goodman, 1961; Faugier and Sargeant, 1997; Sheil et al., 1968). In link-tracing, a number of individuals from the target population are enrolled into the study as 'seeds', and subsequent samples are selected based on their network connections with previous sample members.

Networks are used to represent systems of inter-related entities. In social networks, people (or groups of people) are represented by nodes, and their inter-relationships are represented by edges. Frank (1977) presented an overview of network concepts. Critically, for our work, edges may be either directed (relationships may or may not be reciprocated) or undirected (every relationship is reciprocated). Two nodes are considered incident to each other if they are connected by an edge. RDS draws its name from the fact that respondents are responsible for recruitment by distributing uniquely identified coupons to population members known to them, who are then asked to enroll those they know into the sample, and so on.

Because the sampling process depends on the network structure (Crawford, 2014; Verdery et al., 2015a), the sample mean (or proportion) from a link-tracing sample is typically a biased estimator of the population mean (or proportion). Staying within the design-based frame, existing RDS prevalence estimators utilize estimates of sampling probabilities π, which are typically a function of a respondent's reported number of social ties in a population, called their *degrees*. It should be noted that there are many possible ways of defining the sampling probabilities π. Here we define π as marginal without-replacement sampling probabilities, marginal over all selections of seeds. In practice, RDS diverges from its theoretical approximations in several ways, as discussed in Gile and Handock (2010); Lu et al. (2012); Goel and Salganik (2010); Gile (2011); Tomas and Gile (2011); Lu et al. (2013); Rocha et al. (2016); Aronow and Crawford (2015).

1.3 Edge Inclusion Probabilities

Social networks tend to have complex structure, and are often difficult to observe in their entirety. Typical mechanisms for sampling networks can either rely on pre-specified global rules determining sampling probabilities (i.e. simple random sampling on nodes), or rely on local decision procedures for growing a sample, such as by tracing network edges from previously-observed nodes. Examples of the latter include snowball sampling (Goodman, 1961; Handcock and Gile, 2011), adaptive web sampling (Thompson, 2006a), targeted random walk sampling (Thompson, 2006b), Bayesian adaptive link tracing (St Clair and O'Connell, 2012; Chow and Thompson, 2003), and RDS (Heckathorn, 1997, 2002; Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008). In each of these network sampling strategies, initial nodes are chosen in some fashion, and then some subset of the nodes incident to the initial nodes are sampled. This procedure of sampled nodes recruiting a certain number of their neighbors is repeated until the desired sample size is achieved. In this way, both nodes and edges are observed. Note that in typical RDS practice, only edges traversed by the sampling process are observed.

Often these network sampling approaches build on the theory of random-walks (Lovasz, 1993), where the random walk forms a first-order Markov chain on the space of nodes (Goel and Salganik, 2009). Less-commonly considered is the related implied distribution on traversed network edges. Consider an idealized random walk of the following form:

1. The network is undirected and connected, (i.e. consisting of a single connected component), without self-ties (loops)
2. An initial node is selected with probability proportional to degree: $p_1(i) = \frac{d_i}{\sum_{i=1}^{N} d_i}$
3. Subsequent nodes are selected completely at random, with replacement from among the contacts of the prior sampled node: $P(S_{k+1} = j) = \begin{cases} \frac{1}{d_{S_k}} & Y_{S_k j} = 1 \\ 0 & \text{else} \end{cases}$,

where N is the population size, d_i is the degree of node i, $p_k(i)$ is the probability of sampling the i^{th} node at the k^{th} step, S_k is the index of the node sampled at the k^{th} step, and the $N \times N$ matrix $\mathbf{Y}$ represents the sociomatrix of network ties, such that $Y_{ij} = Y_{ji} = 1$ if there is an edge between i and j, and $Y_{ij} = Y_{ji} = 0$ otherwise. Then the draw-wise edge sampling probabilities are uniform (Salganik and Heckathorn, 2004; Ott and Gile, 2016). Several methods for RDS data, including the estimator in Salganik and Heckathorn (2004) rely on treating edge sampling probabilities as equal. However, Ott and Gile (2016), show that in without-replacement link-tracing sampling, such as RDS, edge sampling probabilities are not uniform.

Our purpose in this paper is to find the proportion of PWID in Mauritius who are HIV-positive, Hepatitis-C positive, and who are female with a new estimator that improves upon the RDS estimator in Salganik and Heckathorn (2004), (also referred to as the SH, or RDS-I estimator) by adjusting for the bias induced by without-replacement sampling in both the estimation of average degree and accounting for non-uniform edge inclusion probabilities. In Section 2 we introduce the most commonly used RDS prevalence estimators and explain how these estimators are subject to bias when there are non-uniform edge inclusion probabilities. In Section 3, we propose a method for estimating edge inclusion probabilities, and in Section 4 we address limitations in current methodology by presenting a prevalence estimator which utilizes estimated edge inclusion probabilities that is particularly suited to the Mauritius data. In Section 5 we compare this new RDS prevalence estimator to existing estimators through simulation studies. In Section 6 we apply this novel method to PWID in Mauritius and estimate the prevalence of HIV and hepatitis C, as well as the proportion of PWID who are female. In Section 7 we present a brief discussion.

2 RDS prevalence estimators

Most RDS inference is aimed at estimating the population proportion of a binary covariate, or the population prevalence:

$$\mu = \frac{1}{N}\sum_{k=1}^{N} z_k, \tag{1}$$

where N is the total population size, and z_i is a binary quantity of interest for the i^{th} unit. For example, in the Mauritius data, which motivates this paper, the binary covariates are HIV positive status, Hepatitis C positive status, and female gender. Several RDS prevalence estimators have been proposed and implemented (Heckathorn, 1997, 2002; Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008; Gile, 2011; Gile and Handcock, 2015), however it has been demonstrated that no one estimator is superior in all cases (Tomas and Gile, 2011). We briefly review the three most commonly implemented estimators of μ: the Volz-Heckathorn estimator (VH), the successive sampling estimator (SS), and the Salganik-Heckathorn estimator (SH), and their relationships to the proposed estimator (Johnston et al., 2016). The VH and SS estimators estimate the probability of observing individuals based on their degree, and use this probability to perform inverse-probability weighting. The SH estimator relies on the assumption that each edge has an equal probability of being included in the sample, and leverages edge-wise information to estimate prevalence. The SH out-performs the alternatives in the presence of homophily, the tendency for individuals with similar attributes to be connected with each other, and differential recruitment effectiveness, that is, when the target populations forms ties preferentially among similar people, and when one group tends to have more successful recruitments per recruiter (Tomas and Gile, 2011). It also performs especially well when the initial sample is highly unrepresentative of the overall population, such as when all individuals selected in the initial sample are HIV positive (Gile and Handcock, 2015). However, it can be severely biased when the sample fraction is large (Gile and Handock, 2010). Note that the the improvement offered by the SS over the VH is the adjustment for without-replacement sampling. In this paper, we use an approximation to the sampling process similar to that in the SS to create a new estimator similar to the SH, but adjusting for without-replacement sampling. Other methods follow the RDS sampling procedure, but then collect additional information about each sampled individual's ego network (Lu, 2013) which is then utilized in the estimation process. In this work we focus on how to improve RDS estimation without collecting additional information. First, we describe the VH, SS, and SH estimators.

The VH estimator (Volz and Heckathorn, 2008) assumes that the RDS sample can be treated as an independent sample from the stationary distribution of a random walk on the space of network nodes. Because the stationary distribution is proportional to the nodal degrees (d_i's), this estimator inverse-weights the observed nodal values of the quantity of interest (z_i's), by the degrees, in a generalized Horvitz-Thompson estimator (Thompson, 2002) ratio format:

$$\widehat{\mu_{VH}} = \frac{\sum_{i=1}^{n} \frac{z_i}{d_i}}{\sum_{i=1}^{n} \frac{1}{d_i}},$$

where n is the sample size and nodes are ordered such that the sampled nodes appear first.

While VH performs well in many settings it is subject to bias under several sampling conditions including differential recruitment effectiveness (individuals passing coupons in one group are likely to disperse more of their coupons than individuals in the other group), and differential activity (individuals in one group tend to have a higher degree than individuals in the other group) in the presence of a large sample fraction(Tomas and Gile, 2011).

The SS estimator (Gile, 2011) has a form very similar to the VH estimator. While the VH estimator assumes that sampling probabilities are proportional to degree, the SS estimator approximates these probabilities based on a without replacement process (Gile, 2011):

$$\widehat{\mu_{SS}} = \frac{\sum_{i=1}^{n} \frac{z_i}{\hat{\pi}_i(\mathbf{d})}}{\sum_{i=1}^{n} \frac{1}{\hat{\pi}_i(\mathbf{d})}}.$$

The formulation of the SS estimator differs from the formula for the VH estimator as the estimated sampling probability of i is a function of both d_i, as well as the degree sequence of the entire sample (noted as $\mathbf{d}$). The SS estimator is not subject to bias when there are large sampling proportions, though it is still subject to bias resulting from other conditions including differential recruitment effectiveness in combination with homophily. Its finite population correction also relies on a working estimate of the population size.

In contrast to the VH and SS estimators, the SH estimator (Salganik and Heckathorn, 2004) relies heavily on the number of within and between group edges (i.e. recruitments) in an RDS sample, rather than a weighted proportion of the sample that has the attribute of interest. For this reason, the SH estimator is especially sensitive to unequal edge sampling probabilities. For example, if the RDS recruitment process has a disproportionate number of recruitments from someone who has HIV to someone who does not have HIV (relative to the number of social ties between people in these two groups) the SH estimator will be heavily biased towards underestimating the proportion of people with HIV. However, it is also because of this very different formulation that uses the number of between group edges that the SH performs well in circumstances where the alternative VH and SS perform poorly, in particular in the face of the combination of differential recruitment effectiveness, differential activity, and homophily. Because our proposed estimator builds on the SH, we describe its form in greater detail.

While we assume that the underlying network is undirected, in the RDS sample we observe edges as they are traversed in a directed manner, so we treat equivalence classes of observed directed edges. For instance, consider the case where we are interested estimating the proportion μ of a networked-population that is HIV positive, also equal to the population average of a binary nodal variable $z_i \in \{0, 1\}$, where $z_i = 1$ if node i is infected. Define

$$T_{(k,1-k)} = \frac{\sum_{i:z_i=k} \sum_{j:z_j=1-k} Y_{ij}}{N_k},$$

to be the average number of ties a single type $z_i = k$ node has to type $z_i = 1 - k$ nodes, $k \in \{0, 1\}$, where N_k is the population number of nodes of type k. Since the network is undirected, $N_0 T_{(0,1)} = N_1 T_{(1,0)}$, so $\frac{T_{(1,0)}}{T_{(0,1)}} = \frac{N_0}{N_1}$, and

$$\mu = \frac{N_1}{N_0 + N_1} = \frac{T_{(0,1)}}{T_{(0,1)} + T_{(1,0)}}.$$

We can express $T_{(k,1-k)}$ in terms of $D_{(k)}$ and $C_{(k,1-k)}$, where $D_{(k)} = \frac{1}{N_k} \sum_{i:z_i=k} d_i$ is the average degree for those nodes of type k and

$$C_{(k,1-k)} = \frac{\sum_{i:z_i=k} \sum_{j:z_j=1-k} Y_{ij}}{\sum_{i:z_i=k} d_i}$$

is the proportion of cross-group ties among the ties of type k nodes. Then the proportion of proportion with the characteristic of interest is calculated as:

$$\mu = \frac{D_{(0)} C_{(0,1)}}{D_{(0)} C_{(0,1)} + D_{(1)} C_{(1,0)}}. \tag{2}$$

The Salganik and Heckathorn (2004) method takes the form of the above equation and utilizes the observed between and within group referrals and the degree for each participant sampled through RDS. They estimate:

$$\widehat{C}_{(k,1-k)} = \frac{r_{(k,1-k)}}{r_{(k,1-k)} + r_{(k,k)}}, \quad k \in \{0,1\}, \tag{3}$$

where $r_{(k,1-k)}$ is the number of referrals from an k node to a $1 - k$ node, $k, \in \{0,1\}$. This is based on the assumption that the sampling process can be treated as from the stationary distribution of a Markov chain on the network nodes, leading to uniform edge-traversal probabilities. $D_{(1)}$ and $D_{(0)}$ are also unknown and need to be estimated in order to compute prevalence estimates. Like the VH estimator, the SH estimator makes use of the generalized Hansen-Hurwitz estimator (Hansen and Hurwitz, 1943; Thompson, 2002), but rather than estimating the prevalence, the SH estimator uses this form to estimate the average degree for each group:

$$\widehat{D}_{(1)} = \frac{\sum_{i=1}^{n} d_i z_i}{\sum_{i=1}^{n} \frac{z_i}{d_i}}, \; \widehat{D}_{(0)} \quad = \frac{\sum_{i=1}^{n} d_i(1-z_i)}{\sum_{i=1}^{n} \frac{1-z_i}{d_i}}. \tag{4}$$

Substituting these estimates directly into (2) gives the form of the SH estimator:

$$\widehat{\mu_{SH}} = \frac{\widehat{D}_{(0)}\widehat{C}_{(0,1)}}{\widehat{D}_{(0)}\widehat{C}_{(0,1)} + \widehat{D}_{(1)}\widehat{C}_{(1,0)}}. \tag{5}$$

The SH estimator out-performs other estimators in certain situations, particularly when there is differential recruitment effectiveness (Tomas and Gile, 2011). However, the SH estimator has been noted to perform poorly in the presence of differential activity, and when there is a large sample fraction (Gile and Handock, 2010).

2.1 Implication for Inference on Nodal Characteristics: Salganik-Heckathorn Estimator

In equation 5 we see that the SH estimator relies on intermediate estimates (3) and (4) to estimate μ. While the form of (4) is based on weighting the nodal sample and does not depend directly on assumptions about edge sampling probabilities, (3) relies heavily on the assumption of uniform edge sampling probabilities. When these probabilities are far from uniform, as in without-replacement sampling with substantially large sample fractions, the estimation in (3) may be quite inaccurate, leading to biased estimates produced by the SH estimator. These biases will be exacerbated in the presence of differential activity. Based on the analysis in Ott and Gile (2016), we expect that $\widehat{C}_{(0,1)}$ will have positive bias, and $\widehat{C}_{(1,0)}$ will have negative bias when group 0 has higher mean degree than group 1, leading to positive bias in $\widehat{\mu_{SH}}$.

3 Estimating Edge Inclusion Probabilities

The without replacement sampling design in RDS results in unequal edge sampling probabilities which have previously been unaccounted for (Ott and Gile, 2016). In order to address the estimation problem resulting from unequal edge inclusion probabilities, we propose to estimate edge inclusion probabilities and use inverse-probability weighting with these weights to improve the SH estimator.

Because of the complexity of the RDS sampling process, we must estimate these probabilities under an approximation to the sampling process. In particular, we follow Gile (2011) and use a successive sampling approximation to the sampling process. Successive sampling, also known as probability proportional to size without replacement (PPSWOR) sampling, is a sampling mechanism used to draw a sample of size n from a population of size N of units with *unit sizes* $\mathbf{u} = (u_i),\ i \in \{1 \ldots N\}$ (Raj, 1956; Rao et al., 1991; Gile, 2011). It proceeds as follows:

1. The first unit is sampled with probability proportional to $\mathbf{u}$.

2. Each subsequent sample is drawn with probability proportional to $\mathbf{u}$ *from the previously unsampled units*, resulting in step-wise sampling probabilities:

$$P(S_k = i | S_1, \ldots, S_{k-1}) = \begin{cases} \\ \frac{u_i}{\sum_{j=1}^{N} u_j - \sum_{j=1}^{k-1} u_{S_j}} & i \notin \{S_1, \ldots, S_{k-1}\} \\ 0 & \text{else.} \end{cases} \tag{6}$$

3. Sampling ends when sample size n is attained.

Gile (2011) notes that treating nodal degrees $\mathbf{d}$ as unit sizes ($\mathbf{u}$ above), the probabilities (6) are equivalent to the step-wise sampling probabilities of a without-replacement random walk on a network drawn from a Molloy-Reed distribution (Molloy and Reed, 1995) conditional on the population degree distribution $\mathbb{N} = \mathbb{N}_1, \mathbb{N}_2, \ldots, \mathbb{N}_K$ where $\mathbb{N}_j$ is the number of nodes with degree j, and K is the maximum degree. Recall that $\mathbf{d}$ is the vector of degrees in the sample, whereas $\mathbb{N}$ is the distribution of degrees in the population.

The approximation in Gile (2011) has proved highly effective in adjusting the VH estimator. Therefore, it stands to reason that using the corresponding directed edge weights to adjust the SH estimator will be similarly effective in accounting for finite population effects. We therefore present an approach to estimating directed edge sampling weights based on a successive sampling approximation to the RDS process.

Assuming that there is an underlying network which we cannot fully observe, we will estimate the probability that we observe in the RDS sample the directed edge between node i and node j, given that i and j are connected in the underlying graph: $P(i \to j)$.

We specify that $S_{i,j} = 1$ if node i is sampled as a non-terminal node while node j is still unsampled. Then:

$$P(i \to j) = P(i \to j, S_{i,j} = 1) = P(i \to j | S_{i,j} = 1) P(S_{i,j} = 1). \tag{7}$$

Because successive sampling acts on equivalence classes of units (nodes) based on nodal degrees, we treat equivalence classes of directed edges based on ordered pairs of nodal degrees. We make the approximation:

$$P(i \to j | S_{i,j} = 1) \approx \min\left(\frac{n_c}{h(d_i)}, 1\right) \approx \min\left(\frac{n_c}{g(d_i)}, 1\right), \tag{8}$$

where $h(d)$ is the average number of connections incident to a node with degree d that are unsampled when such a node is sampled, and, n_c is the maximum number of coupons that each participant may pass on. We further approximate $h(d)$ as $g(d) = d(1 - \mu(d)/N)$, where $\mu(d)$ is the average number of nodes that have been sampled when a node with degree d is sampled. The intuition for this is that if a node i is chosen as a seed, then $\mu(d_i) = 0$ and $P(i \to j | S_{i,j} = 1) = n_c/d_i$. As the sampling process continues, more nodes are included, and thus are not available to be sampled again. If a node i is chosen after 10% of nodes from the population are included in the sample, then we approximate the number of unsampled nodes that i could then recruit into the sample as $0.9d_i$, and $P(i \to j|S_{i,j} = 1) = n_c/(0.9d_i)$. Here since we are marginalizing over all nodes with degree d_i, we find the average sampling order of nodes with degree d_i as $\mu(d_i)$. We use these approximations so that we can efficiently estimate them from a PPSWOR sample without instantiating an unknown network.

The more challenging aspects of estimating $P(i \to j)$ are estimating $P(S_{i,j} = 1)$ and $\mu(d)$. Gile (2011) faces a similar challenge of estimating nodal inclusion probabilities in RDS in the presence of large sample fractions. In Gile (2011)'s successive sampling approximation, the nodal inclusion probabilities are not available in closed form, and must be estimated by an iterated simulated sampling process. This procedure converges to sampling probabilities $\hat{\pi}_k = \mathrm{f}(k, n, \mathbb{N})$, for equivalence classes of nodes according to degree k, and dependent on sample size n and population distribution of degrees, $\mathbb{N}$.

Following Gile (2011), we estimate the nodal sampling probabilities $\mathrm{f}(k, n, \mathbb{N})$ and the degree distribution $\mathbb{N}$ with successive sampling. We then conduct an additional round of simulated resampling to estimate $P(S_{i,j} = 1)$ and $\mu(d)$. Given the estimated degree distribution $\mathbb{N}$, we simulate M resamplings according to the following procedure:

1. For t in $1 \dots M$:

   (a) Draw a sample $S_1, \dots, S_n$ of size n from $\mathbb{N}$ using the successive sampling method treating nodal degrees as unit sizes, with initial node chosen with probability proportional to degree and subsequent nodes drawn according to (6).

   (b) For all ordered pairs k, l such that $k, l \in 1, \dots, K$, record the number of times a node with degree k is sampled before a node with degree l forming a $K \times K$ matrix V^t (where K is the largest degree in the degree distribution). For example, suppose that $\mathbb{N}_4 = 3$ and $\mathbb{N}_5 = 2$, and the simulated sample includes two nodes of degree 4 and one of degree 5 in the order $4, 4, 5$. Then $V^t_{44} = 3, V^t_{55} = 1, V^t_{45} = 4, V^t_{54} = 1$, all other entries of V^t are zero, and we would record this in a 5×5 matrix V^t:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 4 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

(c) Let $\hat{\mu}^t(k)$ be the average ordered index of all nodes of degree k observed in the simulated sample, and $g^t(k) = k(1 - \frac{\hat{\mu}^t(k)}{N})$. Treat $g^t(k)$ as null whenever no such nodes are sampled.

2. Let W be a $K \times K$ matrix in which W_{ij} is the proportion of instances in the simulation in which a node with degree i was sampled before a node with degree j where:

$$W_{kl} = \frac{\sum_{t=1}^{M} V_{kl}^t + 1}{\mathbb{N}_k \times \mathbb{N}_l \times M + 1}, \quad k \neq l, \text{ and } W_{kk} = \frac{\sum_{t=1}^{M} V_{kk}^t + 1}{(\mathbb{N}_k - 1) \times \mathbb{N}_k \times M + 1}, \quad k \in 1, \ldots, K.$$

3. Estimate $P(S_{i,j} = 1)$ with $W_{d_i d_j}$, for $i, j \in 1, \ldots N$.

4. Estimate $\hat{g}(k)$ as the average of the non-null values of the M realizations of $g^t(k)$, for $k \in 1 \ldots K$.

5. Estimate $P(i \to j)$ for equivalence classes of d_i and d_j as:

$$\hat{q}_{d_i,d_j} = \min\left(\frac{n_c}{\hat{g}(d_i)}, 1\right) W_{d_i d_j}.$$

4 The weighted SH estimator

Recall that SH estimates prevalence by estimating the average number of cross-ties from each group to the other group. Here we use the new estimated inclusion probabilities ($\hat{q}_{d_i,d_j}$) to improve upon the SH estimator by weighting observed edges and nodes. We refer to this new estimator as the *Weighted SH Estimator*. Let $r_{i,j}$ be the indicator that person i passed a coupon to person j, and recall that $z_i = 1$ if i is HIV positive, and $z_i = 0$ if person i is HIV negative. Then we have:

$$\widehat{C}_{W(1,0)} = \frac{\sum_{i:z_i=1}\sum_{j:z_j=0} r_{ij}/\hat{q}_{d_i,d_j}}{\sum_{i:z_i=1}\sum_{j:z_j=0} r_{ij}/\hat{q}_{d_i,d_j} + \sum_{i:z_i=1}\sum_{j:z_j=1,i\neq j} r_{ij}/\hat{q}_{d_i,d_j}},$$

$$\widehat{C}_{W(0,1)} = \frac{\sum_{i:z_i=0}\sum_{j:z_j=1} r_{ij}/\hat{q}_{d_i,d_j}}{\sum_{i:z_i=0}\sum_{j:z_j=1} r_{ij}/\hat{q}_{d_i,d_j} + \sum_{i:z_i=0}\sum_{j:z_j=0,i\neq j} r_{ij}/\hat{q}_{d_i,d_j}}.$$

The SH estimator relies on the assumption that individuals are sampled in proportion to their degree in order to estimate the average degree for each group. Here we follow Gile (2011) and make use of the SS estimator to estimate $D_{(1)}$ and $D_{(0)}$ using $\hat{\pi}_{d_i}$, the estimated probability that someone with degree d_i is included in the sample:

$$\widehat{D}_{W(1)} = \frac{\sum_i z_i d_i \hat{\pi}_{d_i}^{-1}}{\sum_i z_i \hat{\pi}_{d_i}^{-1}},$$

$$\widehat{D}_{W(0)} = \frac{\sum_i (1-z_i) d_i \hat{\pi}_{d_i}^{-1}}{\sum_i (1-z_i) \hat{\pi}_{d_i}^{-1}}.$$

Now we have

$$\widehat{\mu}_{WSH} = \frac{\widehat{D}_{W(0)}\widehat{C}_{W(0,1)}}{\widehat{D}_{W(0)}\widehat{C}_{W(0,1)} + \widehat{D}_{W(1)}\widehat{C}_{W(1,0)}}, \tag{9}$$

which we use to estimate prevalence. Note that the weighted SH prevalence estimator is in the same form as the original SH estimator, but includes adjustments for unequal edge sampling probabilities.

4.1 Variance Estimation of Weighted SH Estimator

Uncertainty estimation for RDS estimators is typically conducted using a bootstrap procedure, most often the procedure introduced in Salganik (2006). Little is known about the performance of this bootstrap, although the studies that do exist suggest that it is anti-conservative (Goel and Salganik, 2010; Nesterko and Blitzstein, 2015; Wejnert, 2009; Verdery et al., 2015b). Creating an uncertainty estimator that corrects the under-estimation of the Salganik bootstrap is a separate line of inquiry beyond the scope of this project. We therefore propose to apply a version of the Salganik bootstrap, and suggest that users remember that this estimate should be regarded with all the caveats applied to other RDS uncertainty estimators. The resulting bootstrap estimation procedure proceeds as follows:

1. Categorize nodes in the model as either referred by infected, or referred by uninfected.
2. Sample uniformly and with replacement from the RDS sample n_s seeds, where $n_s =$ the number of seeds in the RDS sample. These become our bootstrap seeds.
3. For each new member in our bootstrap sample that is infected, sample n_c with replacement from the observed RDS sample that were referred by someone who is infected.
4. For each new member in our bootstrap sample that is uninfected, sample n_c with replacement from the observed RDS sample that were referred by someone who is uninfected.
5. From steps 3 and 4, we have now collected our new wave of sample collection. Repeat steps 3 and 4 with the newest wave, until the sample size in the original RDS sample is reached.
6. Calculate the RDS estimate of disease prevalence $\hat{P}_{bs}$.
7. Repeat steps 2-6 many times to estimate the bootstrap distribution of $\hat{P}_{bs}$.
8. Use the distribution of $\hat{P}_{bs}$ to form confidence intervals or calculate standard errors using a Normal approximation.

In this way we seek to estimate the variance of the the RDS estimate of the population proportion of the characteristic(s) of interest, explicitly, the variance of an RDS estimator if repeated RDS samples of the same sample size, with the same number of seeds, using the same maximum number of coupons, were collected from the same population.

5 Simulation Studies

The goal of this project is to create an improved estimator for RDS to better estimate the HIV prevalence, Hepatitis C prevalence, and proportion female of PWID in Mauritius. To do so, we propose an inferential method relying on estimation of edge-sampling probabilities. Here, we present simulations to illustrate the performance of our proposed RDS estimator.

5.1 Simulations to compare RDS prevalence estimators

To evaluate the performance of the proposed estimator, we simulated RDS on networks and compared five prevalence estimators: the mean of the sample, the SS, the VH, the original SH, and the weighted SH. To allow for comparisons with previous RDS work, we used simulated networks that were used in Tomas and Gile (2011) and Gile and Handock (2010), where a detailed description of the methods used to generate these networks can be found. These networks are each composed of 1000 nodes composed of two groups: *infected* (200 nodes) and *uninfected* (800 nodes). In the networks used in the simulations, the networks either have a homophily value of one or two and differential activity of one or two. Additionally, as we simulate RDS on these different networks, we vary the sampling proportion (20%, 50%, 70%), and the differential recruitment effectiveness (1 and 1, or .9 and .6). These terms are defined mathematically in Table 1 for a network Y with N nodes where N^1 is the number of infected nodes in the population and $N^0 = N - N^1$. We use these parameters to determine the parameters of an exponential random graph model, then sample networks from these models using the `statnet` R package (Handcock et al., 2016). The parameters network statistics specified in each condition are then the expected values under the network-generating model.

Differential recruitment effectiveness occurs when individuals in one group are more likely to successfully recruit people into the sample. In these simulations differential recruitment effectiveness was set to either (1,1) or (.9,.6). Differential recruitment effectiveness = (1,1) when individuals in both the infected and uninfected groups would recruit as many people into the sample as they were allowed. Differential recruitment effectiveness = (.9, .6) when individuals in the infected group have a 90% chance of successfully recruiting someone for each coupon they are given, and those in the uninfected group have a 60% chance of successfully recruiting someone for each coupon they are given. In these simulations we begin the sampling process with 10 seeds, and the number of coupons, $n_c = 2$, so each

Table 1: Network Parameters for Simulations

Parameter	Definition
Number of Nodes	N
Prevalence	$\mu = \sum_{i=1}^{N} z_i/N$
Mean degree	$\bar{d} = 1/N \sum_{i=1}^{N} d_i/2$
Homophily	$H = \frac{2/(N^1(N^1-1))\sum_{i,j} z_i z_j \mathbb{E}(y_{ij})}{1/(N^1 N^0)\sum_{i,j} z_i(1-z_j)\mathbb{E}(y_{ij})}$
Differential Activity	$DA = \bar{d}^1/\bar{d}^0$

person in the sample recruited up to two others given the sample size had not yet been attained.

In the first set of simulations, we demonstrate how the weighted SH estimator compared to the other RDS prevalence estimators in the absence of differential activity, differential recruitment effectiveness, and homophily effects. We simulated RDS on 1,000 networks of size 1,000, with prevalence 0.20, and a sampling proportion of 20%. In these networks, the average degree was 7.07. The prevalence estimates from these 1,000 simulations are presented in Figure 1, and the MSE is displayed in the first panel of Table 2. All five estimators, including the naive mean perform comparably when there is no differential activity, differential recruitment effectiveness or homophily present, and the sampling proportion is relatively small.

In the second set of simulations, we used 1000 networks where homophily was set to 1 and differential activity was held constant at 2. In the simulated RDS sampling, the differential recruitment effectiveness was set to (1,1), and we varied the sampling proportion (20%, 50%, 70%). In these networks, the average degree for those with the characteristic of interest was 11.7, and 5.83 among those who did not have the characteristic of interest. Overall, the average degree was 6.98. On each of the 1,000 networks we drew three RDS samples of sizes 200, 500, and 700, to produce the desired sampling proportions. Boxplots of the estimated prevalence are displayed in Figure 2, and the second panel of Table 2 contains the MSE of the various estimators. In these simulations, regardless of the sampling proportion, the SS

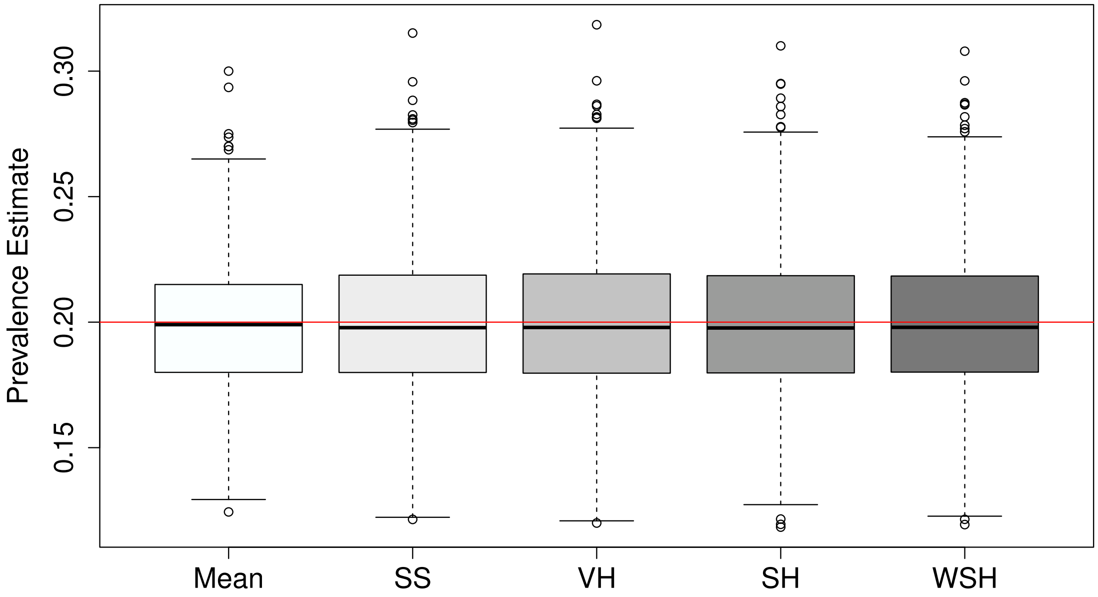

Figure 1: Prevalence estimates using the sample mean, SS, VH, SH, and weighted SH estimators without differential recruitment effectiveness, differential activity, or homophily effects. Sampling fraction is 20%. The horizontal red line represents the true prevalence.

estimator performs the best in terms of MSE, and the Weighted SH performing comparably to the SS when the sampling proportion is 20% or 50%. Because the VH and SH estimators would perform best when sampling is conducted with replacement, we would expect that larger sampling proportions would incur greater bias for these estimators. Indeed, the VH and SH estimators display increasing bias as the sampling proportion increases, and perform similarly in terms of MSE. The naive mean performs very poorly relative to all the other estimators when the sample size is small, but has a smaller MSE than the VH and the SH estimators when the sampling proportion is 70%.

The next set of simulations were performed on networks where homophily was varied to be either 1 or 2 and differential activity was held constant at 2. In each homophily condition, RDS was carried out on each of the networks (1,000 networks where homophily =1, 1,000 networks where homophily=2) with differential recruitment effectiveness set to (1,1) and (.9,.6). In these networks, the average degree for those with the characteristic of interest was 11.7, and 5.83 among those who did not have the characteristic of interest. Overall, the average degree was 6.98. In this set of simulations, the sampling proportion was held constant at 20%. We compare the performance of the estimators from these simulations in

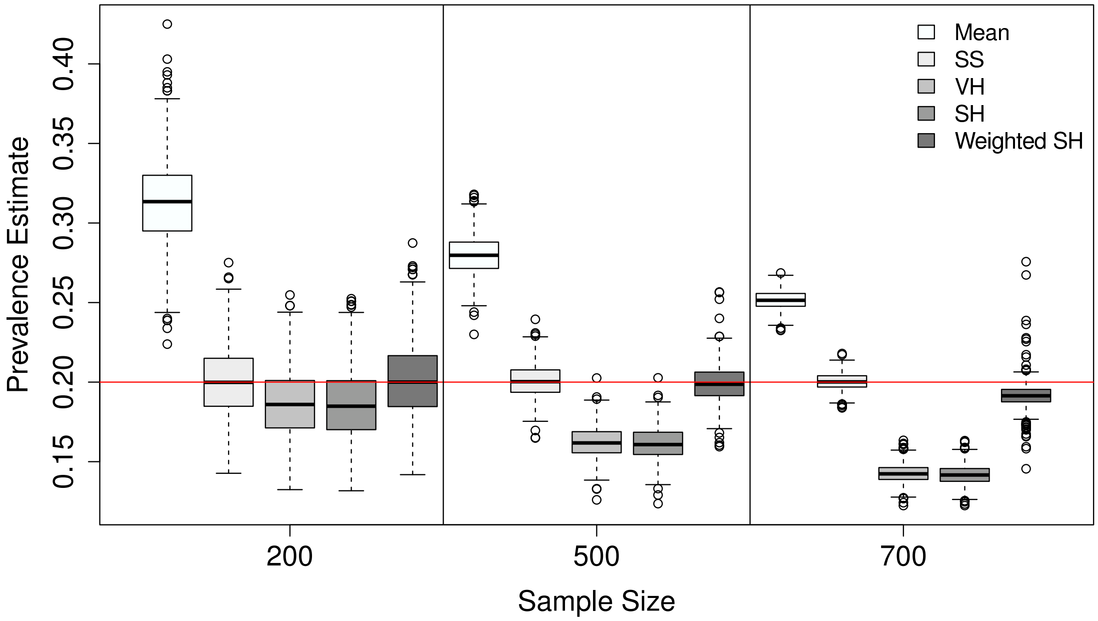

Figure 2: Prevalence estimates using the sample mean, SS, VH, SH, and weighted SH estimators for samples of size 200, 500, and 700 from simulated networks of 1000. Here differential activity =2, homophily=1, and differential recruitment effectiveness =(1,1). The horizontal red line represents the true prevalence.

Figure 3, and in third panel of Table 2. Regardless of homophily level, when the differential recruitment effectiveness is set to (1,1) all five estimators perform similarly in terms of MSE, except the mean which is positively biased. Similarly, when homophily is set to 1, and differential recruitment effectiveness is set to (.9,.6) the mean has the highest MSE, and the other four estimators perform comparably. However, when differential recruitment effectiveness is set to (.9,.6) and homophily is set to 2, the weighted SH estimator has the lowest MSE, followed by the SH. This is consistent with our expectation that the revised estimator corrects for finite population biases (as does the SS), while also maintaining the insensitivity to differential recruitment effectiveness of the SH. In fact, for all simulation conditions,the weighted SH estimator has lower MSE than the SH.

In each of the previous simulations, it is assumed that the population size is known in the estimation for both the SS and the weighted SH. In the final set of simulations, we focus exclusively on the weighted SH estimator, and investigate the impact of misspecifying the population size. In these simulations, we again use 1000 networks where the population size is 1000, homophily and differential activity are both present. We simulate respondent driven

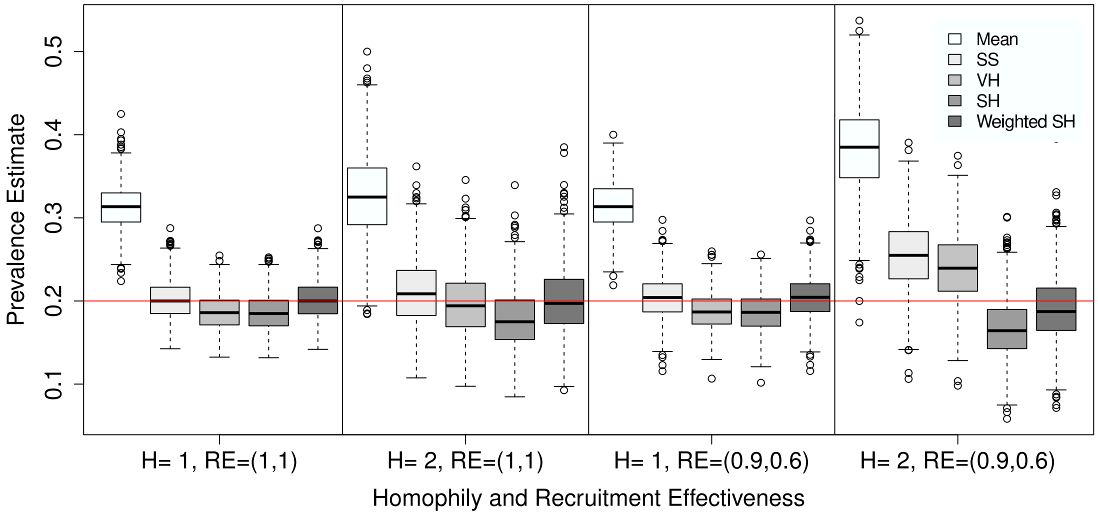

Figure 3: Prevalence estimates using the sample mean, SS, VH, SH, and weighted SH estimators for varying homophily $\in (1, 2)$, recruitment effectiveness $\in [(1, 1); (.9, .6)]$, and differential activity =2. Sampling fraction is 20%. The horizontal red line represents the true prevalence.

Table 2: Mean Squared Error (times 10^3) of Simulated Prevalence Estimates on 1000 Networks with 1000 Nodes

Recruitment Effectiveness:	1,1	1,1	1,1	1,1	1,1	1,1	0.9,0.6	0.9,0.6
Homophily:	1	1	1	1	1	2	1	2
Differential Activity:	1	2	2	2	2	2	2	2
Sampling Fraction:	0.2	0.2	0.5	0.7	0.2	0.2	0.2	0.2
Estimator								
Mean	0.69	13.65	6.47	2.72	13.65	18.40	13.95	36.00
SS	0.84	0.54	0.10	0.03	0.54	1.68	0.63	4.85
VH	0.88	0.62	1.52	3.33	0.62	1.50	0.62	3.30
SH	0.93	0.69	1.59	3.44	0.69	1.79	0.70	2.36
Weighted SH	0.90	0.54	0.13	0.14	0.54	1.62	0.63	1.71

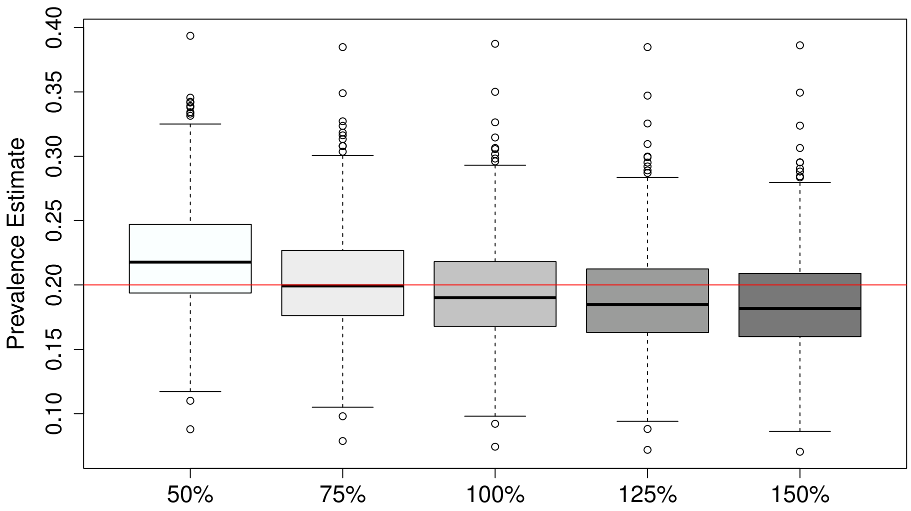

Figure 4: Prevalence estimates weighted SH estimators with homophily=2, recruitment effectiveness =(.9,.6), differential activity =2, with a sampling fraction of 20% and varying the specification of the population size to be 50%, 75%, 100%, 125%, and 150% of the original sample size of 1000. The horizontal red line represents the true prevalence.

sampling on each network, using a sample size of 200, and induce differential recruitment effectiveness of (.9,.6). We then calculate the weighted SH in each of these simulations varying the the specified population size from 50% of the true population size to 150% of the true population size by increments of 25% which we display in Figure 4. We found that the MSE for each of these estimates was less than the MSE for any of the other estimators under the same sampling and network conditions (see the last column in the in third panel of Table 2), and that the MSE was lowest for the weighted SH with the correct sample size specified during the estimation process.

6 Analysis of Mauritius Data

We applied this method to data sampled using RDS in 2011 among PWID in Mauritius. Using six seeds, each participant was provided with up to three recruitment coupons resulting in a sample size of 500 and a maximum sample wave of 12. Here we focus on estimating the prevalence of HIV and Hepatitis C, and the proportion of female PWID in Mauritius.

6.1 HIV Prevalence

In the sample, 219 respondents were HIV positive, 279 were HIV negative, and 2 had missing values. Here all missing values were coded as HIV negative. The average degree in the sample was 12.79, and the average degree among those who were HIV positive was 10.38, while the average degree among those who were HIV negative was 14.67. HIV positive participants in the sample successfully recruited 1.05 on average, while HIV negative participants successfully recruited 0.93 on average (recruitment effectiveness ratio=1.14) (Gile et al., 2015). As there seems to be both differential activity and differential recruitment effectiveness and assuming that homophily is present in this sample, the weighted SH estimator developed in this work may be of use. In the presence of differential recruitment effectiveness, homophily, and differential activity, we would expect the VH, SS, and SH to overestimate the proportion of the population that is a member of the group with smaller average degree. In particular, since those who were HIV positive had lower average degree than those who were HIV negative, we would expect that the weighted SH estimator would provide a more accurate estimate of HIV prevalence, and that the other estimators would tend to overestimate HIV prevalence.

In order to apply the weighted SH estimator, we must use an estimate of the total number of PWID in Mauritius. We used the estimated population size of 9253 at the time of the study (Johnston et al., 2011). Using this data set and the SH method we estimate: $\widehat{C}(AB) = 0.48$, $\widehat{C}(BA) = 0.36$, $\widehat{D}(A) = 4.46$,$\widehat{D}(B) = 6.31$. For the weighted SH method, we estimate $\widehat{C}(AB) = 0.56$, $\widehat{C}(BA) = 0.26$, $\widehat{D}(A) = 4.31$, $\widehat{D}(B) = 6.02$. The estimated proportions of HIV positive PWID are 51.35% (SH), 39.65% (Weighted SH), 51.89% (SS), 52.10% (VH), and 43.80% (naive mean).

Figure 5 displays the estimated proportions of HIV positive PWID and a bootstrap 95% confidence interval (CI) (from 10,000 bootstrap samples) for all five estimators that we discuss here. In the figure we can see that the weighted SH estimator has both the lowest estimate of the proportion who are HIV positive and the narrowest 95% CI. The VH has the highest estimate.

6.2 Hepatitis C Prevalence

In the sample, 92.2% respondents tested positive for the Hepatitis C virus. The average degree among those who are Hepatitis C positive was 12.90, while the average degree among those who were Hepatitis C negative was 11.46. Hepatitis C positive participants in the sample successfully recruited 1.00 on average, while Hepatitis C negative participants successfully recruited 0.74 on average (recruitment effectiveness ratio=1.35). Since those who tested positive for Hepatitis C had higher average degree than those who tested negative for Hepatitis C, we would expect that the weighted SH estimator would provide a more accurate estimate of Hepatitis C prevalence, and that the other estimators would tend to underestimate Hepatitis C prevalence.

As before, we used the estimated population size of 9253 and the SH method to estimate: $\widehat{C}(AB) = 0.08$, $\widehat{C}(BA) = 0.93$, $\widehat{D}(A) = 5.16$,$\widehat{D}(B) = 4.84$. For the weighted SH method, we estimate $\widehat{C}(AB) = .04$, $\widehat{C}(BA) = 0.98$,$\widehat{D}(A) = 5.39$,$\widehat{D}(B) = 5.04$. The estimated prevalence of PWID who are Hepatitis C positive are 91.63% (SH), 95.79% (Weighted SH), 91.75% (SS), 91.74% (VH), and 92.20% (naive mean).

Figure 6 displays the estimated proportions of Hepatitis C positive PWID and a bootstrap 95% CIs (from 10,000 bootstrap samples) for the five estimators. In the figure we can see that the weighted SH estimator has both the highest estimate of the proportion who are Hepatitis C positive and the narrowest 95% CI. The VH, SS, and SH all have comparable estimates. when there is differential recruitment effectiveness and differential activity with those who are Hepatitis C positive having higher average degree.

6.3 Proportion Female

In the sample, 6% of respondents were female. The average degree among those who were female was 13.77, while the average degree among those who were not female was 12.73. Female participants in the sample successfully recruited 1 person on average, while non-female participants successfully recruited .98 people on average (recruitment effectiveness ratio=1.02).

Contrasting the SH estimator and the weighted SH (as well as the other three estimators)

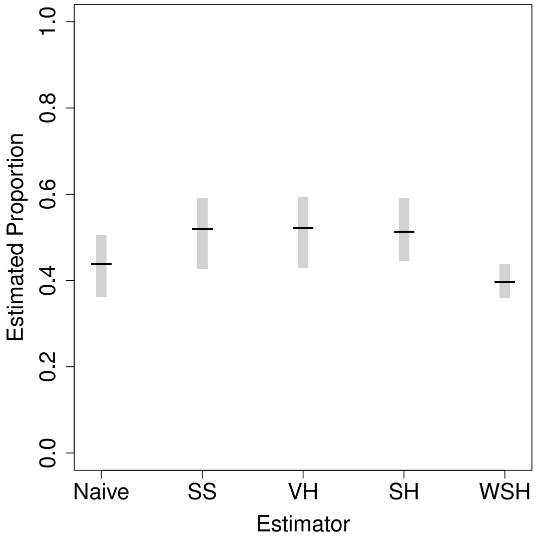

Figure 5: Estimated proportion of HIV positive PWID with 95% CIs for five different estimators.

when estimating the proportion of the population that is female, with the SH method we find: $\widehat{C}(AB) = 0.83$, $\widehat{C}(BA) = 0.05$, $\widehat{D}(A) = 4.46$, $\widehat{D}(B) = 5.18$. For the weighted SH method, we estimate $\widehat{C}(AB) = 0.83$, $\widehat{C}(BA) = 0.03$, $\widehat{D}(A) = 4.70$,$\widehat{D}(B) = 5.42$. The estimated prevalence of PWID who are female are 6.76% (SH), 3.53% (Weighted SH), 6.89% (SS), 6.91% (VH), and 6.00% (naive mean).

Since those who were female had higher estimated average degree (both with the SH method and the weighted SH method) than those who were not female, we would expect that the weighted SH estimator would provide a more accurate estimate of the proportion female, and that the other estimators would tend to overestimate the proportion female, which is what we concluded above.

Figure 7 displays the estimated proportions of PWID who are female and a bootstrap 95% CIs (from 10,000 bootstrap samples) for the five estimators.

7 Discussion

In this paper, we have estimated features of the population of PWID in Mauritius. To do so, we have introduced a new estimator which improves upon existing RDS prevalence estimation by accounting for the unequal edge sampling probabilities that result from the

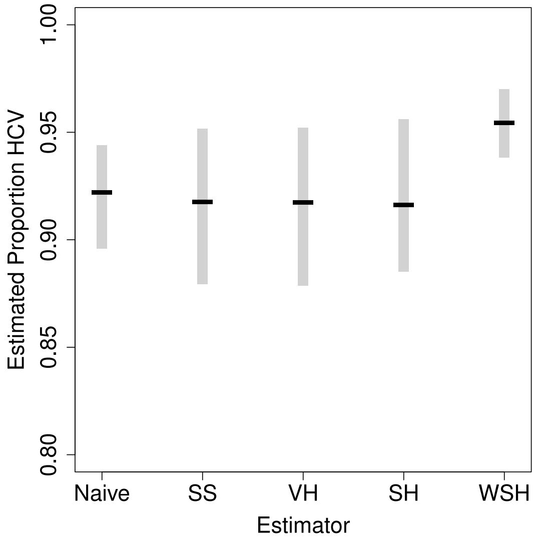

Figure 6: Estimated proportion of Hepatitis C positive PWID with 95% CIs for five different estimators.

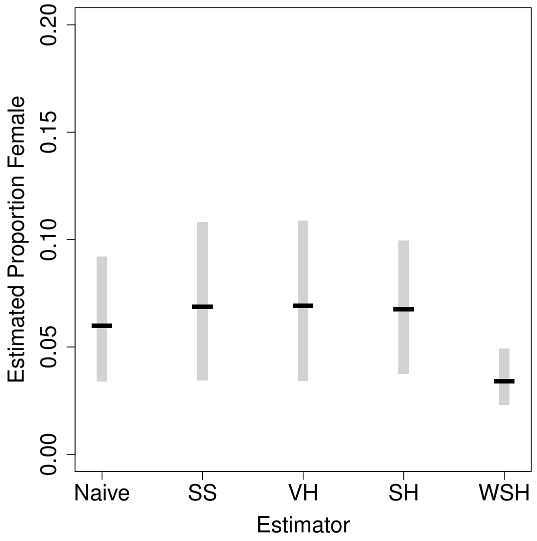

Figure 7: Estimated proportion of Hepatitis C positive PWID with 95% CIs for five different estimators.

violation of the with-replacement sampling assumption. Our new estimator, which we call the "weighted SH" weights edges inversely to an estimated edge sampling probability. While this estimator follows the general form of the SH estimator (Salganik and Heckathorn, 2004), it differs from the SH estimator in two important ways. First, when edge characteristics are needed for the method of moments approach, rather than assume that all edges have an equal chance of being included in the RDS sample, we estimate the edge inclusion probabilities of each edge taking into account the without-replacement sampling inherent in RDS. Estimating these edge inclusion probabilities is a difficult endeavor, and we have provided an approximation that allows significant improvement over assuming all edge inclusion probabilities are equal. Secondly, we also estimate the average degrees again accounting for the without-replacement sampling. The weighted SH estimator will be particularly useful when the degree distributions and recruitment effectiveness vary by the outcome of interest, and homophily is present, conditions that are common in settings where RDS is typically used, including in each area of concern in our application in Mauritius (HIV, Hepatitis C status, and gender).

We have shown here that the weighted SH estimator that we propose has uniformly smaller MSE than the SH estimator in simulation, under several different conditions both of the network and of the sampling process. Most notably, in the simulations which we think most closely represent real-world conditions (with differential recruitment effectiveness, homophily, and differential activity) the SH estimator has 1.38 times the MSE of the weighted SH, the VH has 1.93 the MSE of the weighted SH, and the SS has 2.83 times the MSE of the weighted SH. Ideally, we would derive analytic results to compare estimators, however since the RDS is so complex, simulations provide the best option for comparing estimators to find the estimator that will perform the best in practice.

While the weighted SH estimator that we propose here improves over the SH in many ways, the SH estimator does not require that the underlying population size is known. Estimating the population size of hidden populations is a difficult task, though new statistical methodology can aid in this estimation (Handcock et al., 2014). However, in simulation we found that even with large misspecification (50% greater or 50% less) of the population size, the weighted SH has lower MSE than all of the other estimators that we considered, including the SH and VH which do not assume that population size is known. Like the SH estimator, the weighted SH estimator may be subject to biases introduced by biased seeds, preferential recruitment, or other sampling and network anomalies. The weighted SH estimator, like the original SH estimator tends to exhibit less bias than estimators that weight the nodal attributes (SS, VH, mean) in the presence of differential recruitment effectiveness coupled with homophily. We also note that the SH estimator assumes that the underlying network is undirected, and our newly proposed estimator also makes this assumption.

We also propose an adaption of the commonly used bootstrap-based variance estimator (Salganik, 2006). However other variance estimators have recently been proposed such as in Baraff et al. (2016), which utilizes a tree bootstrap method, and shows promising results. Future research should investigate the performance of variance estimation methods when using different RDS estimation methods (Spiller et al., 2017).

RDS estimation is commonly applied to estimate HIV prevalence of traditionally higher-risk and hard to reach networked populations, such as men who have sex with men, sex workers, and PWID. Here, we have applied this new estimator, as well as the most commonly used RDS estimators, to estimate HIV and Hepatitis C prevalence, and the proportion females among PWID in Mauritius. This is an excellent case study, as the population is well-defined (Mauritius is an island country), and is one of the traditional hard to reach populations to which RDS prevalence estimations are applied. In this study, males, those who were HIV positive, and those who were Hepatitis C negative had higher average degree than females, those who were HIV negative, and those who were Hepatitis C positive, respectively. As a result of this differential activity, we would expect that the weighted SH would correct for a bias that previous estimators exhibit to overestimate the proportion of groups with lower average degree. Indeed, we have found that the most commonly used prevalence estimators over-estimate HIV prevalence and proportion female, and underestimate Hepatitis C prevalence in this population of PWID in Mauritius relative to the weighted SH. The differences in prevalence estimates between the weighted SH and the other most commonly used RDS prevalence estimators could have substantial implications for disease prevention surveillance and policy.

One of the main contributions of this work is to improve upon a widely adopted prevalence estimator for hidden and networked populations. There are many different RDS prevalence estimators that are now in use. No one estimator is superior in all settings. We conclude that the new weighted SH estimator is best when there is differential activity, differential recruitment effectiveness, and homophily effects, which is what we would expect to see in a realistic network setting, and what we believe is present in the network of PWID in Mauritius.

8 Acknowledgments

Research reported in this publication was supported by grants from NSF(SES-1230081), including support from the National Agricultural Statistics Service. We are grateful to Ahmed Saumtally, Sewraz Corceal, Indrasen Mahadoo, and Farida Oodally, for allowing us to use these data.

References

Country progress report, Republic of Mauritius, 2015. URL http://www.unaids.org/sites/default/files/country/documents/MUS_narrative_report_2015.

Peter M. Aronow and Forrest W. Crawford. Nonparametric identification for respondent-driven sampling. *Statistics & Probability Letters*, 106:100 – 102, 2015. ISSN 0167-7152. doi: http://dx.doi.org/10.1016/j.spl.2015.07.003. URL http://www.sciencedirect.com/science/article/pii/S0167715215002357.

Aaron J Baraff, Tyler H McCormick, and Adrian E Raftery. Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. *Proceedings of the National Academy of Sciences*, page 201617258, 2016.

Mosuk Chow and S. K. Thompson. A Bayesian approach to estimation with link-tracing sampling designs. *Survey Methodology*, 29:197–205, 2003.

Forrest W Crawford. The graphical structure of respondent-driven sampling. *arXiv preprint arXiv:1406.0721*, 2014.

Jean Faugier and Mary Sargeant. Sampling hard to reach populations. *Journal of Advanced Nursing*, 26(4):790–797, 1997. ISSN 1365-2648. doi: 10.1046/j.1365-2648.1997.00371.x. URL http://dx.doi.org/10.1046/j.1365-2648.1997.00371.x.

Ove Frank. Estimation of graph totals. *Scandinavian Journal of Statistics*, 4:81–89, 1977.

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. Network model-assisted inference from respondent-driven sampling data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):619–639, 2015. ISSN 1467-985X. doi: 10.1111/rssa.12091. URL http://dx.doi.org/10.1111/rssa.12091.

Krista J. Gile and Mark S. Handock. 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*, 178:241–269, 2015.

Sharad Goel and Matthew J. Salganik. Respondent-driven sampling as Markov chain Monte Carlo. *Stat Med*, 28(17):2202–2229, Jul 2009. doi: 10.1002/sim.3613. URL `http://dx.doi.org/10.1002/sim.3613`.

Sharad Goel and Matthew J. Salganik. Assessing respondent-driven sampling. *Proc Natl Acad Sci U S A*, 107(15):6743–6747, Apr 2010. doi: 10.1073/pnas.1000261107. URL `http://dx.doi.org/10.1073/pnas.1000261107`.

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

M. S. Handcock and K Gile. Comment: on the concept of snowball sampling. *Social Methodology*, 41:367–371, 2011.

Mark S. Handcock, Krista J. Gile, and Corinne M. Mar. Estimating hidden population size using respondent-driven sampling data. *Electron. J. Statist.*, 8(1):1491–1521, 2014. doi: 10.1214/14-EJS923. URL `http://dx.doi.org/10.1214/14-EJS923`.

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`), 2016. URL `CRAN.R-project.org/package=statnet`. R package version 2016.9.

M H Hansen and W N Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14:333–362, 1943.

Douglas D. Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(2):pp. 174–199, 1997. ISSN 00377791. URL `http://www.jstor.org/stable/3096941`.

Douglas D. Heckathorn. Respondent-driven sampling II: Deriving valid population estimates from chain referral samples of hidden populations. *Social Problems*, 49:11–34, 2002.

L Johnston, A Saumtally, S Corceal, I Mahadoo, and Oodally F. High HIV and hepatitis C prevalence amongst injecting drug users in Mauritius: findings from a population size estimation and respondent driven sampling survey. *Int J Drug Policy*, 22:252–258, 2011.

L. G. Johnston, D. Prybylski, H.F. Raymond, A. Mirzazadeh, C. Manopaiboon, and McFar. Incorporating the service multiplier method in respondent-driven sampling surveys to estimate the size of hidden and hard-to-reach populations: case studies from around the world. *Sex Transm Dis*, 40:303–310, 2013.

Lisa G. Johnston, Avi J. Hakim, Samantha Dittrich, Janet Burnett, Evelyn Kim, and Richard G. White. A systematic review of published respondent-driven sampling surveys collecting behavioral and biologic data. *AIDS and Behavior*, pages 1–23, 2016. ISSN 1573-3254. doi: 10.1007/s10461-016-1346-5. URL `http://dx.doi.org/10.1007/s10461-016-1346-5`.

Lisa Grazina Johnston, Mohsen 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 Behav*, 12(4 Suppl):S131–S141, Jul 2008. doi: 10.1007/s10461-008-9413-1. URL `http://dx.doi.org/10.1007/s10461-008-9413-1`.

Amy Lansky, Abu S. Abdul-Quader, Melissa Cribbin, Tricia Hall, Teresa J. Finlayson, Richard S. Garfein, Lillian S. Lin, and Patrick S. Sullivan. Developing an HIV behavioral surveillance system for injecting drug users: The National HIV Behavioral Surveillance System. *Public Health Reports (1974-)*, 122:pp. 48–55, 2007. ISSN 00333549. URL `http://www.jstor.org/stable/20057230`.

L Lovasz. Random walks on graphs: A survey. *Combinatorics: Paul Erdos is Eighty*, 2: 1–46, 1993.

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

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. doi: 10.1111/j.1467-985X.2011.00711.x. URL http://dx.doi.org/10.1111/j.1467-985X.2011.00711.x.

Xin Lu, Jens Malmros, Fredrik Liljeros, and Tom Britton. Respondent-driven sampling on directed networks. *Electron. J. Statist.*, 7:292–322, 2013. doi: 10.1214/13-EJS772. URL http://dx.doi.org/10.1214/13-EJS772.

Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms*, 6(2/3):161–179, March 1995. ISSN 1042-9832. URL http://dl.acm.org/citation.cfm?id=259573.259582.

Jane R. Montealegre, Lisa G. 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. ISSN 1573-3254. doi: 10.1007/s10461-013-0466-4. URL http://dx.doi.org/10.1007/s10461-013-0466-4.

National AIDS Secretariat. Programmatic mapping & size estimation of key populations in Mauritius. Technical report, Office of the Prime minister, MAURITIUS, http://actogether.mu/photo

Sergiy Nesterko and Joseph Blitzstein. Bias-variance and breadth-depth tradeoffs in respondent-driven sampling. *Journal of Statistical Computation and Simulation*, 85(1):89–102, 2015. doi: 10.1080/00949655.2013.804078. URL http://dx.doi.org/10.1080/00949655.2013.804078.

Miles Q. Ott and Krista J. Gile. Unequal edge inclusion probabilities in link-tracing network sampling with implications for respondent-driven sampling. *Elec-*

tron. *J. Statist.*, 10(1):1109–1132, 2016. doi: 10.1214/16-EJS1138. URL `http://dx.doi.org/10.1214/16-EJS1138`.

Des Raj. Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association*, 51 (274):269–284, 1956. doi: 10.1080/01621459.1956.10501326. URL `http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1956.10501326`.

T. J. Rao, S. Sengupta, and B. K. Sinha. Some order relations between selection and inclusion probabilities for ppswor sampling scheme. *Metrika*, 38(1):335–343, 1991. ISSN 1435-926X. doi: 10.1007/BF02613629. URL `http://dx.doi.org/10.1007/BF02613629`.

Luis E. C. Rocha, Anna E. Thorson, Renaud Lambiotte, and Fredrik Liljeros. Respondent-driven sampling bias induced by community structure and response rates in social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, pages n/a–n/a, 2016. ISSN 1467-985X. doi: 10.1111/rssa.12180. URL `http://dx.doi.org/10.1111/rssa.12180`.

Matthew J. Salganik. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *J Urban Health*, 83(6 Suppl):i98–112, Nov 2006. doi: 10.1007/s11524-006-9106-x. URL `http://dx.doi.org/10.1007/s11524-006-9106-x`.

Matthew J. Salganik and Douglas D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:pp. 193–239, 2004. ISSN 00811750. URL `http://www.jstor.org/stable/3649374`.

Helen Sheil, Ivan Zwart, Nicola Brackertz, and Denise Meredyth. Community consultation and the hard to reach. 1968.

Michael W Spiller, Krista J Gile, Mark S Handcock, Corinne M Mar, and Cyprian Wejnert. Evaluating variance estimators for respondent-driven sampling. *Journal of Survey Statistics and Methodology*, 2017.

K St Clair and D O'Connell. A Bayesian model for estimating population means using a link-tracing sampling design. *Biometrics*, 68:165–173, 2012.

S. K. Thompson. *Sampling.* Wiley, 2002.

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

Steven K. Thompson. Targeted random walk designs. *Survey Methodology*, 32(1):11–24, 2006b.

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. ISSN 1935-7524. doi: {10.1214/11-EJS630}.

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

Ashton M. Verdery, Ted Mouw, Shawn Bauldry, and Peter J. Mucha. Network structure and biased variance estimation in respondent driven sampling. *PLoS ONE*, 10(12):1–27, 12 2015b. doi: 10.1371/journal.pone.0145296. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0145296`.

Erik Volz and Douglas D. Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24:79–97, 2008.

Cyprian Wejnert. An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociol Methodol*, 39(1):73–116, Aug 2009. doi: 10.1111/j.1467-9531.2009.01216.x. URL `http://dx.doi.org/10.1111/j.1467-9531.2009.01216.x`.