Population Size Estimation Using Multiple Respondent-Driven Sampling Surveys

Brian Kim* Mark S. Handcock†

Abstract

Respondent-driven sampling (RDS) is commonly used to study hard-to-reach populations since traditional methods are unable to efficiently survey members due to the typically highly stigmatized nature of the population. The number of people in these populations is of primary global health and demographic interest and is usually hard to estimate. However, due to the nature of RDS, current methods of population size estimation are insufficient. We introduce a new method of estimating population size that uses concepts from capture-recapture methods while modeling RDS as a successive sampling process. We assess its statistical validity using information from the CDCs National HIV Behavioral Surveillance system in 2009 and 2012.

Key Words: Hard-to-reach population sampling, Network sampling, Model-based survey sampling, Capture-Recapture, Probability proportional to size without replacement sampling

1. Introduction

In some populations, such as people at high risk for HIV — such as female sex workers (FSW), men who have sex with men (MSM), or people who inject drugs (PWID) — or recent migrants, obtaining a probability sample can be practically impossible. This can be for a variety of reasons, including social stigma, unwillingness to self-identify, or simply because the people in the group may have engaged in an illegal activity.

In many cases, finding the size of these hard-to-reach populations is of great interest. When deciding how much aid to send for HIV prevention, getting an accurate count of people at high risk for HIV is crucial for efficient allocation of resources (UNAIDS and Organization, 2010). Since these populations are hard to reach, traditional methods of sampling, such as random digit dialing telephone numbers or household surveys, are unfeasible. In addition, due to the stigma attached to many hidden populations, individuals may refuse to release information to protect theirs or others' privacy (Heckathorn, 1997).

An alternative way of surveying these hidden populations involves exploiting their highly connected nature. For example, PWID are much more likely to know someone else who injects drugs. In addition, it is much more likely that a person in the population would even know that someone else is also in that population (e.g. it is more likely for a PWID to know who among people they know is also a PWID). Therefore, researchers have developed various link-tracing sampling methods, in which the network links from sampled members of the population are traced out to unsampled members in the population in order to grow the sample (Spreen, 1992; Handcock and Gile, 2011; Gile and Handcock, 2010).

One link-tracing method that has recently become very popular with researchers for studying hidden populations is respondent-driven sampling, or RDS. Introduced by Heckathorn (1997) as an alternative to traditional snowball sampling and time-location sampling techniques, RDS employs a link-tracing design in the following manner:

1. Start with a small initial sample, usually a convenience sample.

*University of Maryland, 1218 LeFrak Hall, 7251 Preinkert Dr., College Park, MD 20742

†University of California, Los Angeles , UCLA Department of Statistics, Student Services, 8117 Math Sciences Bldg., Los Angeles, CA 90095-1554

2. Give each respondent a few coupons to recruit others, with incentives for both the recruiter and the new recruit.

3. Include each recruit in the study and give them a limited number of coupons, typically approximately 3 (Malekinejad et al., 2008), to hand out to other members of the population.

4. People who receive a coupon may choose to come in to join the study.

5. Each new participant is included in the study and given coupons to recruit others.

6. The process continues until a stopping condition is reached, such as a target sample size. If the chain stops before the stopping condition is reached (e.g. due to unfruitful referrals), new seeds may be chosen to start new chains.

RDS has several benefits over more conventional methods. First, and most prominent, is that it enables researchers to survey a population for which a sampling frame does not exist. Since respondents know others in the population (a reasonable assumption for populations like MSM and FSW), it is much easier to ask them to find more people for the study rather than for researchers to try to find them. In addition, as opposed to other network sampling designs, RDS involves a dual incentive system, with incentives for both the recruiter for recruiting others as well as participants for joining the study. In addition, the long chains generated by RDS result in samples that are not as prone to bias by the initial convenience sample, whereas other recruitment sampling methods may be (Heckathorn, 1997; Gile, 2011). In particular, RDS is able to reach the less visible members of the population that may be missed by other link-tracing designs (Kendall et al., 2008). Further, RDS does not require participants to give up any information about others. Instead, they are simply asked to recruit them into the study, giving agency to the possible recruits. This avoids the issue of asking respondents to reveal information about their friends or acquaintances (Heckathorn, 1997).

Due to its many benefits, respondent-driven sampling has increasingly been the method of choice when surveying these hard-to-reach populations (UNAIDS and Organization, 2010; Bengtsson et al., 2012; Platt et al., 2006). A review in 2008 found that there had been over 120 studies that have used RDS to sample most-at-risk populations for HIV (Malekinejad et al., 2008). Much of the previous work has been done using respondent-driven sampling to find proportions or other statistics (Volz and Heckathorn, 2008; Gile, 2011). Many studies comparing RDS with other methods have found that RDS is an effective and efficient method for sampling hard-to-reach populations (Abdul-Quader et al., 2006; Semaan, 2010; Magnani et al., 2005; Platt et al., 2006).

However, the current literature on estimating the population size using only RDS data is quite limited, and has mostly focused on using a multiplier method (Paz-Bailey et al., 2011; Wattana et al., 2007). Some research has been done regarding using other forms of network sampling, such as incorporating general link-tracing design to capture-recapture (Vincent and Thompson, 2016), but very few model-based methods have been developed specifically for RDS data (Handcock et al., 2014; Crawford et al., 2018). In addition, the current model-based methods for RDS do not take full advantage of the multiple capture population size estimation literature from animal abundance. Because RDS is so popular among researchers of hidden populations, methods must be developed to estimate population size using RDS data.

In this paper, we develop a new method to estimate population size using two RDS surveys. In Section 2, we discuss traditional capture-recapture methods for population size

		list 2		
		1	0	
list 1	1	$a_{1,1}$	$a_{1,0}$	$a_{1,+}$
	0	$a_{0,1}$	$a_{0,0}$	$a_{0,+}$
		$a_{+,1}$	$a_{+,0}$	$a_{+,+}$

Figure 1: Contingency table for a simple capture-recapture design.

estimation, as well as methods developed specifically for RDS data. In Section 3, we introduce our new population size estimation method that uses concepts from both traditional capture-recapture and RDS-based methods. Then, in Section 4, we use simulation studies to assess our new model and compare it to existing methods. Finally, we provide concluding remarks in Section 5.

2. Population Size Estimation

Even though using RDS data for population size estimation has not been studied thoroughly, there is a rich literature on size estimation in general. The basis of these methods come from animal abundance. Researchers interested in finding the number of animals in a certain area developed methods to count them. These methods were applied to surveys of human populations, and extensions were developed to relax some of the assumptions.

2.1 Capture-Recapture

A classic method of estimating the abundance of animals is using capture-recapture. A basic capture-recapture design as it applies to animal abundance can be described as follows:

1. Capture a certain number of animals.
2. Tag them, then release them back into the wild.
3. Perform a second capture (recapture).
4. Count how many of the recaptured animals are tagged.
5. Use the overlap to estimate the abundance.

We can express the simple capture-recapture data in the form of a collapsed 2 by 2 table, as in Figure 1. Then, we see the data as capture histories. That is, we know the capture history of "1, 1", "1, 0", and "0, 1", while we do not observe the capture history "0, 0". This is a condensed version of the complete data, and multiple list methods typically approach the problem in this way, estimating the frequency of capture history of all "0"s.

If we assume independence of the two lists and equal capture probabilities for each unit in the population, we get what Paz-Bailey et al. (2011) and Berchenko and Frost (2011) refer to as the naive estimator,

$$\hat{N}_{\text{naive}} = \frac{(a_{1,0} + a_{1,1})(a_{0,1} + a_{1,1})}{a_{1,1}}, \tag{1}$$

where $a_{1,0}$ is the number of people who were captured in only the first sample and $a_{0,1}$ is the number of people who were captured in only the second sample. This is also called the Lincoln-Petersen estimator. An alternative formulation of this capture-recapture design

uses the Hypergeometric distribution. That is, we model the second capture as a Hypergeometric process, with a "success" defined as a unit that was already observed in the first list (e.g. a tagged animal) and a "failure" defined as one that was not observed (e.g. an untagged animal). A Bayesian implementation of this is provided by Cosenza et al. (2014).

2.2 Multiple List Methods

The basic capture-recapture method of estimating population size can be generalized to include more lists. For example, one might consider using three lists. This would result in a three-dimensional version of the contingency table in Figure 1, and the aim would once again be to estimate every unit that was not captured in any list. In addition, various methods have been developed to try to account for heterogeneity in capture probabilities, as well as heterogeneity in lists (that is, a different propensity to capture animals) or time effects (for example, if an animal that is tagged is more likely to be captured because of the tagging) (Rivest and Baillargeon, 2007; Fienberg et al., 1999; Manrique-Vallier, 2016).

2.3 Network-Based Population Size Estimation Methods

One method to find the size of a population without directly sampling that population is the network scale-up method (Bernard et al., 2010; Salganik et al., 2011). The network scale-up method uses information about personal networks sizes of respondents in the general population and known population proportions to make size estimates. For example, suppose we want to know how many men who have sex with other men (MSM) in a particular group of one million people. If a respondent knows 200 people (i.e. their personal network size is 200) and knows 2 people who are MSM, then we can conclude that 1% of the population are MSM. Known populations are used to estimate the size of each respondent's personal network — if a person reports knowing 2 people named Joe and there are 10,000 Joes out of 1,000,000 people in the population, then that respondent's personal network size would be 200.

The network scale-up method provides an advantage in that it does not need to sample the population of interest directly, but it does require an assumption that the unobserved and observed members have the same distribution of characteristics. In addition, the population of interest is hidden, which means that the general population most likely does not know whether their acquaintances really are in that hidden population. Therefore, respondents might report knowing much fewer members of the hidden population even if they actually know many more members.

Successive Sampling - Population Size Estimation (SS-PSE) was developed specifically to use RDS data, estimating population size using a single RDS sample and modeling the RDS process rather than treating it as a probability sample (Handcock et al., 2014). While performing the RDS, information on each respondent's degree (i.e. how many people they know in the population) is collected, along with the order of observation. The RDS process is treated as sampling with probability proportional to size without replacement (PPSWOR) (Gile, 2011).

Intuitively, if we are sampling with PPSWOR, we would expect the people with higher degrees to be sampled first and the people with lower degrees to be sampled later on. SS-PSE leverages this information to estimate the population size, modeling RDS with a successive sampling approximation, which accounts for the without-replacement nature of RDS and has been found to be effective in estimation of population means (Gile, 2011).

Crawford et al. (2018) introduce another population size estimation method specifically developed for a single RDS survey. They assume that the population social network follows an Erdős-Rényi distribution, and use the timing of recruitment and network degree of

recruits to gain information about the unsampled members of the population. As opposed to SS-PSE, the RDS process is more exactly modeled (rather than treating it as sampling with PPSWOR), but it also requires a stronger assumption that the population network is an Erdős-Rényi graph, such that ties are independent and that there is an equal probability of each tie.

One big limitation of both SS-PSE and the method described by Crawford et al. (2018) is that they both only use one RDS sample. More conventional size estimation methods use multiple samples, so while the RDS process is actually modeled, there is potential for improvement since adding a subsequent RDS survey adds valuable recapture information. In the next section, we will build on the network-based approaches by estimating population size using two RDS surveys.

3. Capture-Recapture with Successive Sampling for Population Size Estimation

One of the biggest benefits of using RDS data is the ability to collect multiple samples on a hidden population relatively easily. In a review of RDS studies related to HIV surveillance, Malekinejad et al. (2008) found that RDS studies took, on average, 9 weeks to complete. Because of this, we want to adapt capture-recapture methods to work with RDS data so that we can use as much information as possible. Currently, capture-recapture and other multiple list methods have not been tailored for RDS data. Paz-Bailey et al. (2011) has used RDS in the recapture stage, but there have been no published studies using RDS for each stage of multiple captures in population size estimation. In addition, Paz-Bailey et al. (2011) only used an adjusted ratio estimator. We aim to take a model-based approach, which will not only give us better estimates, but also better measures of variance.

We aim to improve on current methods by utilizing the degrees of each respondent and the order in which they were sampled to develop a method specifically for multiple RDS lists. In this section, we introduce Capture-Recapture SS-PSE, a method to estimate population size using two RDS surveys.

We assume the existence of a population (for example, FSW in a city) with an associated unit size. This can generally be anything about the individual that affects their catchability, but for our purposes in RDS surveys, the unit sizes will be the personal network size, or degree. Our observed data consists of two RDS surveys which serve as our lists, or captures, in which the personal network size of each observation is recorded in order of observation. That is, we are tracking both the unit size (degree) as well as the sequential order of the when units were including into the study for both RDS surveys. We note that in practice, this is done in the order that respondents come into the research center to be included in the study, regardless of which wave or which seed's chain they are a part of. In addition, the second RDS survey includes a question about whether the respondent had been recruited and participated in the first RDS survey. Notably, in these methods, we do not assume that we have unique matching between the first and second list for those captured in both. Instead, we only know that the unit was part of the first sample. This is similar to the information collected in previous RDS studies that have tried to use capture-recapture methodologies, as the researchers would ask whether the respondent had previously received a unique keychain (Paz-Bailey et al., 2011).

3.1 Likelihood Formulation

We start by describing the likelihood. Let the total population size be N, with each individual person in the population indexed $1, \cdots, N$. Since we want to estimate the population size, N is unknown. Each person has an associated unit size representing the number of

people they know in the population (e.g. the number of FSW they know). These unit sizes are treated as an i.i.d. sample from a superpopulation model. Let $U_1, \cdots, U_N$ be the random variables representing the unit sizes. Let n' and n'' be the sample size of the first list and second list, respectively, and n_0 refer to the size of the overlap while n refers to the overall unique sample size (so that $n' + n'' - n_0 = n$).

We have an ordered sampling design, with $G' = (G'_1, \cdots, G'_{n'})$ representing the random indices of the sequentially sampled units with realization $g' = (g'_1, \cdots, g'_{n'})$ in the first list, and $G'' = (G''_1, \cdots, G''_{n''})$ the random indices of the sequentially sampled units with realization $g'' = (g''_1, \cdots, g''_{n''})$ in the second list. $U'_{obs} = (U_{g'_1}, \cdots, U_{g'_{n'}})$ is the unit sizes, in order of the first list, and $U''_{obs} = (U_{g''_1}, \cdots, U_{g''_{n''}})$ the ordered unit sizes in order of the second list, with realizations $u'_{obs} = (u_{g'_1}, \cdots, u_{g'_{n'}})$ and $u''_{obs} = (u_{g''_1}, \cdots, u_{g''_{n''}})$, respectively. $Y''_{obs} = (Y''_{g''_1}, \cdots, Y''_{g''_{n''}})$ with realizations $y''_{obs} = (y''_{g''_1}, \cdots, y''_{g''_{n''}})$ represent the recapture information. In other words, $y''_{g_i} = 1$ if unit g_i was observed in the first list and 0 otherwise. $U = (U_1, \cdots, U_N)$ and $u = (u_1, \cdots, u_N)$ are the unit sizes of the population. We will assume that the unit sizes are independent and identically distributed by some probability mass function $f(\cdot|\eta)$. The choice of parametric model for the unit size distribution is discussed further in Section 3.7. Finally, for simplicity, we will use $\mathbf{U_{obs}} = \{U'_{obs}, U''_{obs}, Y''_{obs}\}$, with realizations $\mathbf{u_{obs}} = \{u'_{obs}, u''_{obs}, y''_{obs}\}$, to represent all observed data, including both lists as well as information about their overlap.

$$
\begin{aligned}
&L(N, \eta | \mathbf{U_{obs}} = \mathbf{u_{obs}}) \\
&\propto p(\mathbf{U_{obs}} = \mathbf{u_{obs}} | N, \eta) \\
&= p(U'_{obs} = u'_{obs} | N, \eta) \cdot p(U''_{obs} = u''_{obs}, Y''_{obs} = y''_{obs} | U'_{obs} = u'_{obs}, N, \eta) \\
&= p(U'_{obs} = u'_{obs} | N, \eta) \cdot p(U''_{obs} = u''_{obs} | Y''_{obs} = y''_{obs}, U'_{obs} = u'_{obs}, N, \eta) \cdot \\
&\quad p(Y''_{obs} = y''_{obs} | U'_{obs} = u'_{obs}, N, \eta) \\
&= \sum_u \Bigg[\Bigg(\sum_{g'} p(U'_{obs} = u'_{obs} | U = u, G' = g', \eta) p(G' = g' | U = u, \eta) \Bigg) \cdot \\
&\qquad \Bigg(\sum_{g''} p(U''_{obs} = u''_{obs} | U'_{obs} = u'_{obs}, Y''_{obs} = y''_{obs}, U = U, G'' = g'', \eta) \cdot \\
&\qquad p(G'' = g'' | Y''_{obs} = y''_{obs}, U'_{obs} = u'_{obs}, U = u) \\
&\qquad p(Y''_{obs} = y''_{obs} | U'_{obs} = u'_{obs}, U = u, \eta) \Bigg) \cdot p(U = u | \eta) \Bigg] \\
&= \frac{N!}{(N-n)!} \sum_{v \in \mathcal{U}} \Bigg[p(G' = (1 \cdots n') | U = v) \\
&\qquad\qquad p(G'' = g^* | U'_{obs} = u'_{obs}, Y''_{obs} = y''_{obs}, U = v, \eta) \cdot \\
&\qquad\qquad p(Y''_{obs} = y''_{obs} | U'_{obs}, U = v, \eta) \prod_{j=1}^{N} f(u_j | \eta) \Bigg].
\end{aligned}
$$

Here, $\mathcal{U}$ is the set of equivalence classes of unit sizes possible for the N units given that the observed data was $\mathbf{u_{obs}}$. Intuitively, the elements of $\mathcal{U}$ include all possible unit sizes for each of the N units, except n of them are constrained to be the observed unit sizes. Since the likelihood is equivalent for all values of g' and g'' as long as they have the same unit sizes, we assign the labels sequentially starting from 1 and incrementing up when we sample a previously unobserved unit, then multiply by the number of permutations outside the sum. So, in the first list, we have $g' = \{1, \cdots, n'\}$ and in the second list, we have $g'' = g^*$, where the values of g^* takes on the original label from the first list if it was already observed in

the first list, and the next available sequential value if it was not observed in the first list. In other words, the newly-observed units in g^* are in order from $n' + 1$ to n (recall that n refers to the combined sample size, or the number of unique units sampled in the two lists), while the previously-observed units retain their original labeling. Since we are choosing n' indices from N possible in the first list and $n' - n_0$ indices from $N - n'$ possible in the second list, the multiplicative factor is

$$\frac{N!}{(N-n')!} \cdot \frac{(N-n')!}{(N-n'-(n''-n_0))!} = \frac{N!}{(N-n)!}. \tag{2}$$

We note that even though n is not fixed by the study design as n' and n'' are, it is determined by the observed data (specifically, the overlap information), and we are able to apply the multiplicative factor outside the summation due to how we have constructed $\mathcal{U}$.

3.2 Modeling RDS as PPSWOR

RDS is a complex process that is extremely difficult to find a statistical representation for because it relies on the network structure of the population and is not fully controlled by surveyors. There have been many attempts at approximating the RDS process (Gile, 2011; Heckathorn, 1997; Volz and Heckathorn, 2008). Gile (2011) provides theoretical and empirical justification for treating the RDS process as a successive sampling process, using a probability proportional to size without replacement (PPSWOR) sampling scheme to approximate RDS, showing that it reduces finite population biases for RDS estimates of population characteristics.

As such, we model the RDS process as a successive sampling procedure, following Gile (2011). Specifically, our model for the first list is the same as in SS-PSE (Handcock et al., 2014):

$$\begin{aligned} &L(N, \eta | \mathbf{U_{obs}} = \mathbf{u_{obs}}) \\ &\propto \frac{N!}{(N-n)!} \sum_{v \in \mathcal{U}} \Bigg[\Bigg(\prod_{k=1}^{n'} \frac{u_{g'_k}}{r'_k} \Bigg) p(G'' = g^* | U'_{obs} = u'_{obs}, Y''_{obs} = y''_{obs}, U = v, \eta) \cdot \\ &\qquad p(Y''_{obs} = y''_{obs} | U'_{obs}, U = v, \eta) \prod_{j=1}^{N} f(u_j | \eta) \Bigg], \end{aligned}$$

where

$$r'_k = \sum_{i=1}^{n'} u_{g'_i} - \sum_{j=1}^{k-1} u_{g'_j}. \tag{3}$$

We can think of r'_k as representing the remaining total degree (that is, the sum of the degrees of everyone in the population who has not yet been sampled). For the second list, we split it up into two parts: whether the units in second list were in the first list or not, and the order in which they were captured given the information about whether they were in the first list. We start with the latter.

Given we know whether the unit was in the first list or not, we can treat the sampling process as PPSWOR out of the two groups: captured in first list and not captured in first list. Let $g_k^+, k \in \{1, \cdots, n_0\}$ refer the indices of units caught in the first list and $g_k^-, k \in \{1, \cdots, N - n'\}$ refer to the indices of units not caught in the first list. Then, we obtain

$$p(G'' = g^* | U'_{obs} = u'_{obs}, Y''_{obs} = y''_{obs}, U = v, \eta) = \prod_{k=1}^{n''-n_0} \frac{u_{g_k^+}}{r_k^+} \prod_{k=1}^{n_0} \frac{u_{g_k^-}}{r_k^-}, \tag{4}$$

where

$$r_k^+ = \sum_{i=1}^{n''} u_{g_i''} - \sum_{j=1}^{k-1} u_{g_j^+} \qquad \text{and} \qquad r_k^- = \sum_{i=1}^{n''} u_{g_i''} - \sum_{j=1}^{k-1} u_{g_j^-}. \tag{5}$$

So, our full likelihood becomes

$$L(N, \eta | \mathbf{U_{obs}} = \mathbf{u_{obs}})$$

$$\propto \frac{N!}{(N-n)!} \sum_{v \in \mathcal{U}} \left[\prod_{k=1}^{n'} \frac{u_{g_k'}}{r_k'} \prod_{k=1}^{n''-n_0} \frac{u_{g_k^+}}{r_k^+} \prod_{k=1}^{n_0} \frac{u_{g_k^-}}{r_k^-} \cdot p(Y_{obs}'' = y_{obs}'' | U_{obs}', U = v, \eta) \prod_{j=1}^{N} f(u_j|\eta) \right]. \tag{6}$$

(7)

Recall that Y_{obs}'' represents a vector of indicator variables for whether the units in the second list were captured in the first list. Again, we model the process as PPSWOR, so that

$$p(Y_{obs}'' = y_{obs}'' | U_{obs}' = u_{obs}', U = v, \eta) = \prod_{k=1}^{n''} \frac{1}{r_k''} \prod_{k=1}^{n''-n_0} r_k^+ \prod_{k=1}^{n_0} r_k^- \tag{8}$$

where

$$r_k'' = \sum_{i=1}^{n''} u_{g_i''} - \sum_{j=1}^{k-1} u_{g_j''}. \tag{9}$$

After simplification, we obtain

$$L(N, \eta | \mathbf{U_{obs}} = \mathbf{u_{obs}}) \propto \frac{N!}{(N-n)!} \sum_{v \in \mathcal{U}} \left[\prod_{k=1}^{n'} \frac{u_{g_k'}}{r_k'} \prod_{k=1}^{n''-n_0} u_{g_k^+} \prod_{k=1}^{n_0} u_{g_k^-} \prod_{k=1}^{n''} \frac{1}{r_k''} \prod_{j=1}^{N} f(u_j|\eta) \right]. \tag{10}$$

3.3 Bayesian Inference for unit size distribution and population size

The joint posterior is given by

$$p(\eta, N | \mathbf{U_{obs}} = \mathbf{u_{obs}}) \propto \pi(\eta, N) \cdot L[\eta, N | \mathbf{U_{obs}} = \mathbf{u_{obs}}], \tag{11}$$

where $\pi(\eta, N)$ is the joint prior for the unit size distribution parameter and the population size. West (1996) and Handcock et al. (2014) note that the likelihood is difficult to compute due the complexity of $\mathcal{U}$. They develop an ancillary variable to finesse this. We use a variant of this method to sample from the augmented posterior,

$$p(N, \eta, U_{unobs} = u_{unobs} \Psi | \mathbf{U_{obs}} = \mathbf{u_{obs}}), \tag{12}$$

using a four component Gibbs sampler. We start by developing inference for the unit size distribution assuming the population size is known in Section 3.4, then adjusting the method to treat the population size N as a parameter in Section 3.5.

3.4 Bayesian Inference for the unit size distribution

We start by developing inference for the unit size distribution conditional on known N. The posterior is given by

$$p(\eta | \mathbf{U_{obs}} = \mathbf{u_{obs}}) \propto \pi(\eta) \cdot L[\eta | \mathbf{U_{obs}} = \mathbf{u_{obs}}], \tag{13}$$

where $\pi(\eta)$ is the prior for the unit size distribution parameter. West (1996) and Handcock et al. (2014) note that the likelihood is difficult to deal with and use

$$p(U = u|G' = g', G'' = g'', \eta) = \frac{N!}{(N-n)!} \prod_{k=1}^{n'} \frac{u_{g'_k}}{r'_k} \prod_{h=1}^{n''} \frac{u_{g''_h}}{r''_h} \prod_{j=1}^{N} f(u_j|\eta). \tag{14}$$

So, from (14),

$$p(U_{unobs} = u_{unobs}|\eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) \propto \prod_{k=1}^{n'} \frac{1}{r'_k} \prod_{h=1}^{n''} \frac{1}{r''_h} \prod_{j=n+1}^{N} f(u_j|\eta). \tag{15}$$

West (1996) and Handcock et al. (2014) note that the r'_k and r''_k terms are difficult to deal with and use a method involving augmenting the data. We adapt the method to include multiple lists. For $k \in \{1, \cdots, n'\}$, $h \in \{1, \cdots, n''\}$, let ψ'_k and ψ''_h have the exponential distribution with rate parameter r'_k and r''_h, respectively. Then,

$$\int_0^\infty r'_k e^{-r'_k \psi'_k} d\psi'_k = 1 \implies \int_0^\infty e^{-r'_k \psi'_k} d\psi'_k = \frac{1}{r'_k}$$

and

$$\int_0^\infty r''_h e^{-r''_h \psi''_h} d\psi''_h = 1 \implies \int_0^\infty e^{-r''_h \psi''_h} d\psi''_h = \frac{1}{r''_h}.$$

In other words,

$$p(\psi'_k = \psi'|\eta, U_{unobs} = u_{unobs}, \mathbf{U_{obs}} = \mathbf{u_{obs}}) = r'_k \exp(-r'_k \psi') \tag{16}$$

and

$$p(\psi''_h = \psi''|\eta, U_{unobs} = u_{unobs}, \eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) = r''_h \exp(-r''_h \psi''). \tag{17}$$

We can then augment the data with $\Psi' = (\psi'_1, \cdots, \psi'_{n'})$ and $\Psi'' = (\psi''_1, \cdots, \psi''_{n''})$, where the components of Ψ' and Ψ'' are all conditionally independent of one another. Let $\Psi = (\Psi', \Psi'')$. Then,

$$\begin{aligned}
&p(U_{unobs} = u_{unobs}, \Psi|\eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) \\
&= p(\Psi' = \psi'|\eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) p(\Psi'' = \psi''|\eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) \cdot \\
&\quad p(U_{unobs} = u_{unobs}|\eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) \\
&\propto \prod_{j=1}^{n'} e^{-r'_j \psi'_j} \prod_{j=1}^{n''} e^{-r''_j \psi''_j} \prod_{j=n+1}^{N} f(u_j|\eta).
\end{aligned} \tag{18}$$

Using (3), (5), and (18),

$$
\begin{aligned}
&p(U_{unobs} = u_{unobs} | \Psi', \Psi'', \eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) \\
&\propto \prod_{j=1}^{n'} e^{-r_j'' \psi_j'} \prod_{k=1}^{n''} e^{-r_j'' \psi_j''} \prod_{j=n+1}^{N} f(u_j|\eta) \\
&\propto \prod_{i=1}^{n'} e^{-\psi_i' \sum_{j=n'+1}^{N} u_{g_j'}} \prod_{i=1}^{n'} e^{-\psi_i' \sum_{j=i}^{n'} u_{g_j'}} \prod_{i=1}^{n''} e^{-\psi_i'' \sum_{j=n''+1}^{N} u_{g_j''}} \\
&\quad \prod_{i=1}^{n''} e^{-\psi_i'' \sum_{j=i}^{n''} u_{g_j''}} \prod_{j=n+1}^{N} f(u_j|\eta) \\
&\propto \prod_{j=n+1}^{N} \exp\left(- u_j \sum_{i=1}^{n'} \psi_i' \right) \exp\left(- u_j \sum_{i=1}^{n''} \psi_i'' \right) f(u_{g_j}|\eta) \\
&= \prod_{j=n+1}^{N} \exp\left(- u_j \left(\sum_{i=1}^{n'} \psi_i' + \sum_{i=1}^{n''} \psi_i'' \right) \right) f(u_{g_j}|\eta). \qquad (19)
\end{aligned}
$$

We see that the unobserved units are conditionally independent from the unnormalized PMF $\exp(-u_j(\sum_{i=1}^{n'} \psi_i' + \sum_{i=1}^{n''} \psi_i''))f(u_{g_j}|\eta)$. In addition, they are independent of all observed information. Thus, we can get draws from the augmented posterior,

$$
p(\eta, U_{unobs} = u_{unobs}, \Psi | \mathbf{U_{obs}} = \mathbf{u_{obs}}), \qquad (20)
$$

using a three component Gibbs sampler.

3.5 Bayesian Inference for the population size

In the previous section, we assumed known N. However, when estimating the population size, we do not know N and want to estimate it. To do this, we adjust the method in the previous section to treat N as a parameter.

We derive the conditional for N.

$$
\begin{aligned}
&p(N|\eta, \Psi', \Psi'', \mathbf{U_{obs}} = \mathbf{u_{obs}}) \propto \pi(N) p(\mathbf{U_{obs}} = \mathbf{u_{obs}} | N, \eta, \Psi', \Psi'') \\
&= \frac{N!}{(N-n)!} \pi(N) \sum_{v \in \mathcal{U}} \left[\prod_{j=n+1}^{N} \exp\left(- u_j \left(\sum_{i=1}^{n'} \psi_i' + \sum_{i=1}^{n''} \psi_i'' \right) \right) f(u_{g_j}|\eta) \right] \\
&= \frac{N!}{(N-n)!} \pi(N) \prod_{j=n+1}^{N} \left[\sum_{v_j=1}^{\infty} \exp\left(- v_j \left(\sum_{i=1}^{n'} \psi_i' + \sum_{i=1}^{n''} \psi_i'' \right) \right) f(j|\eta) \right] \\
&= \frac{N!}{(N-n)!} \pi(N) \left[\gamma(\sum_{i=1}^{n'} \psi_i' + \sum_{i=1}^{n''} \psi_i'', \eta) \right]^{N-n}, \text{where } \gamma(\alpha, \eta) = \sum_{j=1}^{\infty} e^{-\alpha j} f(j|\eta).
\end{aligned} \qquad (21)
$$

We can use (21) to obtain samples from the joint augmented posterior,

$$
p(N, \eta, U_{unobs} = u_{unobs} \Psi | \mathbf{U_{obs}} = \mathbf{u_{obs}}), \qquad (22)
$$

which we can then use to obtain the marginal posterior distribution of N and η. The full details of the MCMC algorithm are given in the next section.

3.6 Algorithmic Details

Here, we describe in the detail the algorithm for drawing from the joint posterior.

1. Initialize N at a point estimate and U_{unobs} at a set of unit sizes.

2. Sample η from

$$p(\eta|U_{unobs}, \mathbf{U_{obs}} = \mathbf{u_{obs}}.\Psi', \Psi'', N) = \pi(N) \cdot \prod_{j=1}^{N} f(u_j|\eta) \tag{23}$$

   This is done using a Metropolis-Hastings algorithm. In our applications, we used the Conway-Maxwell-Poisson distribution as our unit size distribution, so we had two parameters $\eta = ((\log(\mu), \sigma^2)$. We used a Gaussian proposal for the log mean and an Inverse-χ^2 for the variance.

3. Sample Ψ', Ψ'' from

$$p(\psi'_k = \psi'|\eta, U_{unobs} = u_{unobs}, \mathbf{U_{obs}} = \mathbf{u_{obs}}) = r'_k \exp(-r'_k\psi')$$

   and

$$p(\psi''_k = \psi''|\eta, U_{unobs} = u_{unobs}, \eta, \mathbf{U_{obs}} = \mathbf{u_{obs}}) = r''_k \exp(-r''_k\psi'').$$

   These are independent standard Exponential draws.

4. Sample N from equation (21). In order to make computation easier, we set N_{max} a maximum value for N. We compute (21) for each value between n and N_{max} and use this to sample a value between n and N_{max} directly.

5. Sample U_{unobs} from equation (19). This is done using a rejection sampling method, similar to the one described in West (1996) and used in SS-PSE by Handcock et al. (2014).

   The rejection sampling process is

   (i) Draw d from $f(\cdot|\eta)$ and, independently, $u \sim U(0, 1)$.

   (ii) If $\log(u) > -(\sum_{i=1}^{n'} \psi'_i + \sum_{i=1}^{n''} \psi''_i) \cdot d$, reject d and return to (i). Otherwise, save d and repeat until $N - n$ elements of u_{unobs} have been sampled.

6. Repeat until convergence.

3.7 Unit Size Distribution Model

For our purposes, we need a super-population model for unit size. Here, we focus on the cases in which the unit sizes are the personal network sizes, or degrees. There has been a considerable amount of work done on modeling the degree distribution of a network. Handcock et al. (2014) notes that certain long-tailed distributions such as the Poisson-log-normal and the Waring and Yule distributions, which allow for power-law over-dispersion (Handcock and Jones, 2006), are not able to represent the under-dispersion in degree counts, suggesting the Conway-Maxwell-Poisson distribution (Shmueli et al., 2005) as an alternative.

For the applications in this paper, we chose to use the Conway-Maxwell-Poisson distribution as it offers greater flexibility over similar distributions such as the Poisson while using only one additional parameter.

3.8 Prior specification

We can parametrize the Conway-Maxwell-Poisson distribution in terms of its mean and standard deviation. We then put priors on the log mean and variance parameters, using the Normal distribution for the prior log mean, μ, given the prior standard deviation, σ, and scaled Inverse Chi-squared for the variance, σ^2, so

$$\log(\mu)|\sigma \sim N(\mu_0, \sigma/df_{mean}) \quad \text{and} \quad \sigma^2 \sim \text{Inv}\chi^2(\sigma_0^2; df_{sigma}). \tag{24}$$

In our applications, we use diffuse priors with $df_{mean} = 1$ and $df_{sigma} = 5$.

For the population size, Handcock et al. (2014) uses a two parameter class of priors,

$$\pi(N) = \frac{\beta n(N-n)^{\beta-1}}{N^{\alpha+\beta}} \qquad \text{for } N > n, \alpha > 0, \beta > 0. \tag{25}$$

This prior can be thought of specifying knowledge about the sample fraction (n/N) as a Beta(α, β) distribution. This class of priors was chosen after consultation with field researchers the hyperparameters can be specified by them based on budget and logistic considerations for their choice of sample size. For more information about this prior choice, see Handcock et al. (2014).

4. Assessment of CR-SS-PSE

In this section we compare our method, Capture-Recapture Successive Sampling-Population Size Estimation (CR-SS-PSE), with other population size estimation methods. One approach to assessment would be using asymptotic approximations of its statistical properties. However, in the case of population size estimation, there are a number of different asymptotic frameworks involving the relative sizes of n and N, and the real-world relevance of each proposed asymptotic approximation would need to be carefully considered. Instead, we assess the performance of CR-SS-PSE via simulation studies using ranges of n and N that are commonly met in practice. We can also precisely specify the statistical properties of the networked populations and RDS schemes. This allow us to see how CR-SS-PSE performs under known conditions and the dimensions under which it breaks down.

4.1 Simulation of Networks with Known Statistical Properties

For our simulation studies, we use a real network from the Add Health data set. The Add Health data set was constructed from a series of questionnaires given to students in school. The students were asked to nominate friends either in the same school or in a "sister" school, and a friendship network was generated based on these responses (Harris et al., 2009). We used a network with $N = 2587$ for our simulations. Though the networks generated were directed, we treated them as undirected by turning each tie into an undirected one.

When we apply traditional capture-recapture models, we implicitly make assumptions about the RDS process. That is, we may be treating the RDS as a simple random sample as in the basic Lincoln-Petersen estimator. However, we have reason to believe that the RDS process is actually very different from this (Gile, 2011). Therefore, in these simulations, we try to get as close to real RDS surveys as possible.

In the simulations performed in this chapter, we will define a *trial* as consisting of collecting two respondent-driven samples and recording the order of observation and personal network size (unit size) as well as recapture information. The latter is only collected in the second sample as the answer to, "Were you in the first sample?" In other words, we do not track unique matching between the two lists.

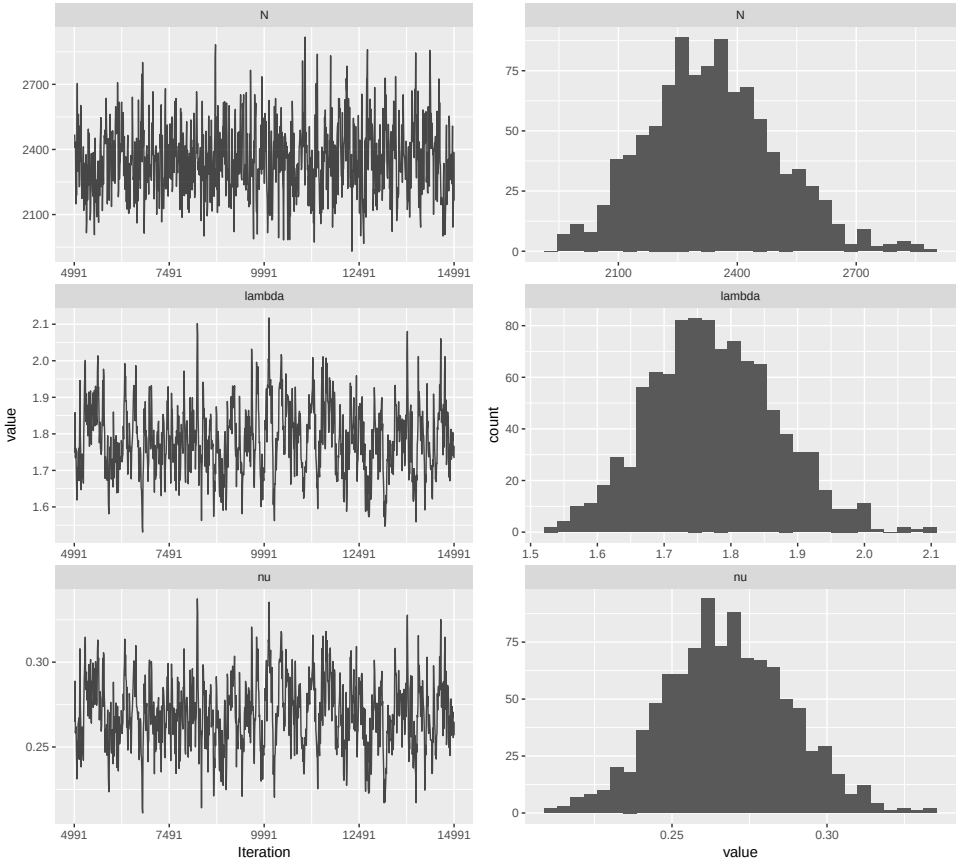

Figure 2: Trace plots and histogram of the posterior distribution for N, λ, and ν using the Add Health data set ($N = 2587$)

We started with a random sample of 10 initial seeds, which is consistent with a review of over 120 RDS studies that found an average of 10 seeds used (Malekinejad et al., 2008), and the number used in a pilot study run by the CDC (Abdul-Quader et al., 2006). We used two coupons for each respondent. In the simulation, starting with the seeds, we sampled two recruits randomly from the nodes connected to each respondent. Then, we used the newest wave of recruits to repeat the process. If a respondent had only one available link to an unsampled person, that person only recruited one person. We stopped the RDS recruitment when we reached our target sample size. The recruiting was assumed to have been done at random from each respondent's personal network.

4.2 Population Size Estimation Methods

We compare our method, CR-SS-PSE, to five other methods. We include two methods that used multiple list concepts: Simple capture-recapture using the Hypergeometric distribution (Cosenza et al., 2014) and the Non-Parametric Latent Class Model (NPLCM) (Manrique-Vallier, 2016). We used a Bayesian implementation of the Hypergeometric model so that we could use the same priors as in all of our other methods.

We also used three variants of the Successive Sampling - Population Size Estimation (SS-PSE) model: SS-PSE with just the first list (which we call SS-PSE); SS-PSE using the first list, then SS-PSE with the second list using the posterior from the first as the prior (which we call Independent SS-PSE); and SS-PSE with one combined list consisting of all unique units in the order they were sampled, removing any double-counting from the second list (which we call Combined SS-PSE).

4.3 Add Health Friendship Network ($N = 2587$)

We first used CR-SS-PSE to estimate the size of the Add Health Network. The Add Health data set was constructed from a series of questionnaires given to students in school. The students were asked to nominate friends either in the same school or in a "sister" school, and a friendship network was generated based on these responses (Harris et al. 2009). We used a network of $N = 2587$ for our simulations. Though the networks generated were directed, we treated them as undirected by turning each tie into an undirected one.

The trace plots and posterior distributions for the population size N and the unit size distribution parameters η are shown in Figure 2. The MCMC seems to be well-behaved, with good convergence. The posterior distribution looks reasonable based on our true population size of $N = 2587$.

We use 1000 simulated RDS trials, apply each of the six methods, and compare their posterior means and 95% highest posterior density as shown in Figure 3. We see that CR-SS-PSE performs the best, with every single other method besides the Combined SS-SPE underestimating the true population size. Every single point estimate in the SS-PSE and Independent SS-PSE methods coming under the true population size. Table 1 shows this difference even more clearly. CR-SS-PSE again has the lowest MSE, and the bias for the SS-PSE methods are all quite high. Combined SS-PSE again does well in coverage rate due to very wide interval estimates, but CR-SS-PSE performs better than every other method.

5. Conclusion

We have proposed a new method of estimating the size of a hidden population using multiple RDS surveys. This model, called Capture-Recapture Successive Sampling for Population Size Estimation, or CR-SS-PSE, uses information from RDS samples more efficiently than existing methods and is shown to have good properties when applied to networks with

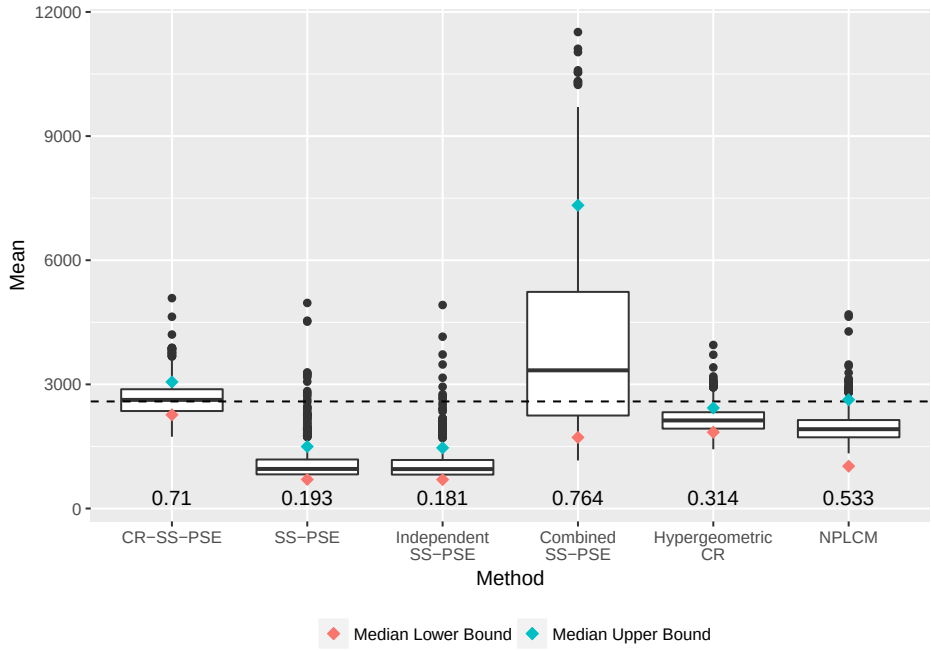

Figure 3: Boxplots of posterior means with six methods using the Add Health network. The value underneath each boxplot represents the proportion of 95% SCIs containing the true size. The median upper and lower bounds of the 95% SCI are also shown. The dashed line shows the true population size ($N = 2587$).

Method	MSE	Bias	Variance	Bias Proportion of MSE
CR-SS-PSE	1.7×10^5	80	1.7×10^5	0.037
SS-PSE	2.4×10^6	-1497	2.0×10^5	0.919
Independent SS-PSE	2.5×10^6	-1523	1.6×10^5	0.935
Combined SS-PSE	6.5×10^6	1373	4.7×10^6	0.288
Hypergeometric CR	2.9×10^5	-433	9.9×10^4	0.656
NPLCM	5.1×10^5	-617	1.3×10^5	0.749

Table 1: Mean Squared Error (MSE), Bias, Variance of the posterior means, and the Bias Proportion of MSE (Bias^2/MSE) for each of the six methods with the Add Health network.

known statistical properties, and has outperformed many existing methods. Due to the popularity of RDS because of the relative ease with which RDS surveys can be implemented, CR-SS-PSE can be useful in providing better estimates by using more of the available information rather than only the network information or a simple capture-recapture approximation.

References

Abdul-Quader, A., Heckathorn, D., Sabin, K., and Saidel, T. (2006), "Implementation and Analysis of Respondent Driven Sampling: Lessons Learned from the Field," *Journal of Urban Health*, 83.

Bengtsson, L., Lu, X., Nguyen, Q., Camitz, M., Hoang, N., Nguyen, T., Liljeros, F., and Thorson, A. (2012), "Implementation of Web-Based Respondent-Driven Sampling among Men Who Have Sex with Men in Vietnam," *PLoS ONE*, 7.

Berchenko, Y. and Frost, S. (2011), "Capture-Recapture Methods and Respondent-Driven Sampling: Their Potential and Limitations," *Sexually Transmitted Infections*, 87, 267–268.

Bernard, H., Hallett, T., Iovita, A., Johnsen, E., Lyerla, R., McCarty, C., Mahy, M., Salganik, M., Saliuk, T., Scutelniciuc, O., Shelley, G., Sirinirund, P., Weir, S., and Stroup, D. (2010), "Counting Hard-to-Count Populations: The Network Scale-Up Method for Public Health," *Sexually Transmitted Infections*, 86, ii11–ii15.

Centers for Disease Control and Prevention (2012), "HIV Infection and HIV-Associated Behaviors Among Injecting Drug Users — 20 Cities, United States, 2009," *Morbidity and Mortality Weekly Report*, 61, 133–138.

— (2015), "HIV Infection and HIV-Associated Behaviors Among Persons Who Inject Drugs — 20 Cities, United States, 2012," *Morbidity and Mortality Weekly Report*, 64, 270–275.

Cosenza, C. et al. (2014), "A Review of Methods for Point and Interval Estimation of Population Size in Capture-Recapture Studies," *American Review of Mathematics and Statistics*, 2.

Crawford, F. W., Wu, J., and Heimer, R. (2018), "Hidden Population Size Estimation From Respondent-Driven Sampling: A Network Approach," *Journal of the American Statistical Association*, 113, 755–766.

Fienberg, S., Johnson, M., and Junker, B. (1999), "Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Multiple Lists," *Journal of Royal Statistical Society*, 162, 383–405.

Gile, K. (2011), "Improved Inference for Respondent-Driven Sampling Data With Application to HIV Prevalence Estimation," *Journal of the American Statistical Association*, 106, 135–146.

Gile, K. and Handcock, M. (2010), "Respondent-Driven Sampling: An Assessment of Current Methodology," *Sociological Methodology*, 40, 285–327.

Handcock, M. and Gile, K. (2011), "Comment: On the Concept of Snowball Sampling," *Sociological Methodology*.

Handcock, M., Gile, K., and Mar, C. (2014), "Estimating Hidden Population Size using Respondent-Driven Sampling Data," *Electronic Journal of Statistics*, 8, 1491–1521.

Handcock, M. and Jones, J. (2006), "Interval Estimates for Epidemic Thresholds in Two-Sex Network Models," *Theoretical Population Biology*, 70, 125–134.

Handcock, M. S., Fellows, I. E., and Gile, K. J. (2016), *RDS: Respondent-Driven Sampling*, Los Angeles, CA, r package version 0.7-8.

Harris, K., Halpern, C., Whitsel, E., Hussey, J., Tabor, J., Entzel, P., and Udry, J. (2009), "The National Longitudinal Study of Adolescent to Adult Health: Research Design," .

Heckathorn, D. (1997), "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations," *Social Problems*, 44, 174–99.

Kendall, C., Kerr, L., Gondim, R., Werneck, G., Macena, R., Pontes, M., Johnston, L., Sabin, K., and McFarland, W. (2008), "An Empirical Comparison of Respondent-driven Sampling, Time Location Sampling, and Snowball Sampling for Behavioral Surveillance in Men Who Have Sex with Men, Fortaleza, Brazil," *AIDS and Behavior*, 12, S97–S104.

Magnani, R., Sabin, K., Saidel, T., and Heckathorn, D. (2005), "Review of Sampling Hard-to-Reach and Hidden Populations for HIV Surveillance," *AIDS*, 19, S67–S72.

Malekinejad, M., Johnston, L., Kendall, C., Kerr, L., Rifkin, M., and Rutherford, G. (2008), "Using Respondent-Driven Sampling Methodology for HIV Biological and Behavioral Surveillance in International Settings: A Systematic Review," *AIDS and Behavior*, 12, S105–S130.

Manrique-Vallier, D. (2016), "Bayesian Population Size Estimation Using Dirichlet Process Mixtures," *Biometrics*, 72, 1246–1252.

Paz-Bailey, G., Jacobson, J., Guardado, M., Hernandez, F., Nieto, A., Estrada, M., and Creswell, J. (2011), "How Many Men who have Sex with Men and Female Sex Workers Live in El Salvador? Using Respondent-Driven Sampling and Capture-Recapture to Estimate Population Sizes," *Sexually Transmitted Infections*, 87, 279–82.

Platt, L., Wall, M., Rhodes, T., Judd, A., Hickman, M., Johnston, L., Renton, A., Bobrova, N., and Sarang, A. (2006), "Methods to Recruit Hard-to-Reach Groups: Comparing Two Chain Referral Sampling Methods of Recruiting Injecting Drug Users Across Nine Studies in Russia and Estonia," *Journal of Urban Health*, 83, i39–i53.

Rivest, L. and Baillargeon, S. (2007), "Applications and Extensions of Chao's Moment Estimator for the Size of a Closed Population," *Biometrics*, 62, 999–1006.

Salganik, M., Fazito, D., Bertoni, N., Abdo, A., Mello, M., and Bastos, F. (2011), "Assessing Network Scale-up Estimates for Groups Most at Risk of HIV/AIDS: Evidence From a Multiple-Method Study of Heavy Drug Users in Curitiba, Brazil," *American Journal of Epidemiology*, 174, 1190–1196.

Semaan, S. (2010), "Time-Space Sampling and Respondent-Driven Sampling with Hard-To-Reach Populations," *Methodological Innovations Online*, 5, 60–75.

Shmueli, G., Minka, T., Kadane, J., Borle, S., and Boatwright, P. (2005), "A Useful Distribution for Fitting Discrete Data: Revival of the Conway–Maxwell–Poisson Distribution," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54, 127–142.

Snijders, T., Pattison, P., and Handcock, M. (2006), "New Specifications for Exponential Random Graph Models," *Sociological Methodology*, 36, 99–153.

Spiller, M., Gile, K., Handcock, M., Mar, C., and Wejnert, C. (2017), "Evaluating Variance Estimators for Respondent-Driven Sampling," *Journal of Survey Statistics and Methodology*.

Spreen, M. (1992), "Rare Populations, Hidden Populations, and Link-Tracing Designs: What and Why?" *Bulletin of Sociological Methodology*, 36, 34–58.

UNAIDS and Organization, W. H. (2010), "Guidelines on Estimating the Size of Populations Most at Risk to HIV," Tech. Rep. UNAIDS/00.03E, UNAIDS and World Health Organization.

Vincent, K. and Thompson, S. (2016), "Estimating Population Size With Link-Tracing Sampling," *Journal of the American Statistical Association*, 0, 1–10.

Volz, E. and Heckathorn, D. (2008), "Probability-Based Estimation Theory for Respondent Driven Sampling," *Journal of Official Statistics*, 24, 79–97.

Wattana, W., van Griensven, F., Rhucharoenpornpanich, O., Manopaiboon, C., Thienkrua, W., Bannatham, R., Fox, K., Mock, P., Tappero, J., and Levine, W. (2007), "Respondent-Driven Sampling to Assess Characteristics and Estimate the Number of Injection Drug Users in Bangkok, Thailand," *Drug and Alcohol Dependence*, 90, 228–233.

West, M. (1996), "Inference in Successive Sampling Discovery Models," *Journal of Econometrics*, 75, 217–238.