BIOMETRICS 000, 000–000 DOI: 000
000 0000

Estimating the Size of Populations at High Risk for HIV using Respondent-Driven Sampling Data

Mark S. Handcock[1,*], Krista J. Gile[2,], and Corinne M. Mar[3,***]**

[1]Department of Statistics, University of California, Los Angeles, CA, U.S.A.

[2]Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA, U.S.A.

[3]CSDE, University of Washington, Seattle, WA U.S.A.

*email: handcock@ucla.edu

**email: gile@math.umass.edu

***email: cmmar@uw.edu

SUMMARY: The study of hard-to-reach populations presents significant challenges. Typically, a sampling frame is not available, and population members are difficult to identify or recruit from broader sampling frames. This is especially true of populations at high risk for HIV/AIDS. Respondent-Driven Sampling (RDS) is often used in such settings with the primary goal of estimating the prevalence of infection. In such populations, the number of people at risk for infection and the number of people infected are of fundamental importance. This paper presents a case-study of the estimation of the size of the hard-to-reach population based on data collected through RDS. We study two populations of female sex workers and men-who-have-sex-with-men in El Salvador. The approach is Bayesian and we consider different forms of prior information, including using the UNAIDS population size guidelines for this region. We show that the method is able to quantify the amount of information on population size available in RDS samples. As separate validation, we compare our results to those estimated by extrapolating from a capture-recapture study of El Salvadorian cities. The results of our case-study are largely comparable to those of the capture-recapture study when they differ from the UNAIDS guidelines. Our method is widely applicable to data from RDS studies and we provide a software package to facilitate this.

KEY WORDS: Hard-to-reach population sampling; Model-based survey sampling; Network sampling; Social networks; Successive sampling

This paper has been submitted for consideration for publication in *Biometrics*

1 Introduction

Respondent-Driven Sampling (RDS, introduced by Heckathorn 1997) is a widely-used link-tracing network-sampling method for hard-to-reach human populations. Beginning with a small convenience sample, each *respondent* is given a small number of uniquely identified coupons to distribute to other population members, making them eligible for participation. The coupon structure assuages confidentiality concerns in hidden populations, and restricting the number of coupons promotes many waves of sampling, decreasing the dependence on the initial sample. Additional details are given in Johnston (2007), Gile and Handcock (2010), and elsewhere.

Population size estimation is of critical importance in high-risk populations, especially among those most at risk for HIV. The most common use of RDS data is in estimating population disease prevalences as well as rates of risk behaviors, often in the service of fulfilling UNAIDS reporting requirements. Using the UNAIDS Estimation and Projection Package (EPP) (UNAIDS, 2009), population proportion estimates are combined with population size estimates derived by other methods to estimate total numbers of HIV infections in each population. This procedure is required of all countries with *concentrated* HIV epidemics, that is, epidemics in which HIV prevalence is low in the general population, but higher in certain high-risk populations, typically female sex workers, men who have sex with men, and injecting drug users. Johnston et al. (2008) summarizes 128 studies using RDS to estimate prevalence in these hard-to-reach populations around the world. Many more have since been completed. Results of the UNAIDS reporting are widely used in decisions regarding resource allocation, both within countries and among international funding agencies. Critically, to date, all such reports have relied on two sources of data: prevalence data (often collected using RDS), and population size data,

collected by other means. The method applied in the current paper is the first method allowing for population size estimation based on RDS data alone.

In addition to UNAIDS reporting, population size and population proportion are of joint interest in program evaluation. In recent decades the scale of HIV prevention and risk reduction programs has increased. As the resources devoted to HIV prevention have increased there has been an concomitant focus on the assessment of the effectiveness of the programs. In particular, international donors expect progress to be measured. Countries able to document progress are more likely to attract and retain funding. Longitudinal measures of the size of the populations at high risk are a fundamental part of this assessment. In particular, they are combined with measures of HIV prevalence to estimate the number of individuals with HIV over time, as well as combined with other estimated rates to estimate numbers of individuals in need of services. To date, many such assessments have relied on RDS data for prevalence estimates, but required additional data sources to measure population size.

Note that there is no direct or naive way to estimate population size from RDS data alone. These data are collected through a link-tracing design in a population of unknown size. Absolute sampling probabilities are not known, and are approximated only up to a constant of proportionality, which is, in fact, the population size. For this reason, RDS data is typically used to estimate population averages, but is not used to directly estimate population sums.

Because of the importance of the size of a hard-to-reach population there are several approaches to estimating it (Bao et al., 2010; Berchenko and Frost, 2011; Paz-Bailey et al., 2011; UNAIDS and World Health Organization, 2010). Most use a variant of capture-recapture, in which the overlap of two samples is used to infer population size (Fienberg et al., 1999; Paz-Bailey et al., 2011; Rocchetti et al., 2011). All other methods using RDS data are of this type, typically using a second capture based on an administrative list or

the distribution of an identifiable token (Salganik et al., 2011; UNAIDS and World Health Organization, 2010). The method we use in this paper is unique in requiring only the single RDS sample (Handcock et al., 2012). As RDS surveys are very widely used, this means that the approach is applicable in the typical situation where a secondary capture is not available. In addition it can be applied in combination with the secondary recapture when available.

Conceptually, our approach is to leverage the information in the sequence of sample *degrees*, or numbers of network contacts to infer the size of the hidden population. Link-tracing network samples are generally more likely to sample nodes with more network connections, or higher *degree*. Therefore, we would expect higher-degree nodes to be more likely to be sampled earlier in the sampling process. As the target population becomes depleted, we would expect higher-degree nodes to be sampled earlier, and lower-degree nodes to be sampled later. Therefore, the rate of decrease in the degrees of sampled nodes over the course of the sample provides information on the size of the hidden population. It is this information that we use to infer population size. A similar approach has been used by West (1996) to estimate the number of previously-unexplored oil fields. These ideas are formalised in Section 3.1.

In the next section (Section 2) we introduce the context of the study of HIV across major El Salvadorian cities (Paz-Bailey et al., 2011). Section 3 reviews the inferential framework and a flexible way to specify prior knowledge about the population size. In Section 4 we use the methodology to estimate the number of FSW in Sonsonate, El Salvador. We show how to elicit and incorporate different types of prior information and its effect on the interval estimates. Section 5 studies the population of MSM in San Miguel, El Salvador. In Section 6, we compare the results of the method to results from separate capture-recapture studies. Section 7 reviews limitations of the method, and Section 8 concludes the paper with a broader discussion.

2 Studies of Populations most at risk for HIV in El Salvadorian cities

El Salvador is a country with low HIV prevalence. As of 2010, the adult HIV prevalence was estimated at 0.8% (UNAIDS/WHO, 2010). However the virus remains a significant threat in groups who practice high-risk behaviors, such as female sex workers (FSW) and men who have sex with men (MSM) (Morales-Miranda et al., 2007; Soto et al., 2007).

From a global and public health perspective, it is crucial to assess the demographic characteristics of the population at risk. In particular, the number of people in the population is a primary measure used to allocate scarce resources. It is used to drive the scale and nature of HIV prevention interventions. Knowledge of the size of the population enables evidence-based approaches to be applied. The population size affects the choice of intervention, its scale and the delivery method (de Estadistica y Censos, 2007; Paz-Bailey et al., 2011; UNAIDS/WHO, 2003).

In this paper we analyze two datasets collected in 2010 as part of a series of RDS studies of populations most at risk for HIV across major El Salvadorian cities (Paz-Bailey et al., 2011). RDS was used as a data collection method because it is effective for hard-to-reach and/or stigmatised populations. The data was collected primarily to estimate the prevalence of infection in the populations and to better understand their demographics, behaviors and practices. The series of integrated behavioral and biological surveys, *Encuesta Centroamericana de Vigilancia de VIH y Comportamiento en Poblaciones Vulnerables* (ECVC) are described in detail in Paz-Bailey et al. (2011) and Guardado et al. (2010).

To provide a sense of the approach, we describe the study of FSW in the department of Sonsonate, which had a population of about 540,000 in 2008 (Guardado et al., 2010). This RDS study began with five initial FSW chosen as seeds. These were interviewed and their behavioral and biological information was collected. They were each given three coupons that they were asked to give to FSW whom they knew. Respondents were asked

how many other FSW they knew well enough to pass them a coupon. If the coupon recipients contacted the survey staff, and agreed to be interviewed, their recruiter received a financial incentive. The order in which the new recruits contacted the staff was carefully recorded. At the completion of their interview, each new recruit was given three coupons and the process of recruitment continued for eight more waves. Some coupons were unused or unreturned. The last two waves had 11 and 5 new recruits in them, respectively, and a total of 184 survey responses were recorded. The average wave number was 3.8. Figure 1 is a graph of the recruitment tree for the RDS. This gives a visual sense of the successive recruitment of the FSW and the chains of recruitment from the initial sample.

[Figure 1 about here.]

As is typical in these settings, the number of female sex workers in Sonsonate is unknown. The public health officials use the UNAIDS national HIV estimation working group recommendation to estimate the number of FSW based on a percent of the total adult female population (UNAIDS/WHO, 2003). In 2009, this group estimated that FSW constitute 0.4%-0.8% of the urban female population 15-49 years of age (139,804 in Sonsonate) (de Estadistica y Censos, 2007; Paz-Bailey et al., 2011). The range for Sonsonate is then 560 to 1120 FSW. It is important to note that the UNAIDS guidelines are not intended to be accurate estimates for a specific population, such as FSW in Sonsonate. Sometimes they are based on a study in another region or context.

The second group at high risk for HIV is MSM. These are significantly understudied in El Salvador. A similar survey was conducted for MSM, and here we consider the population resident in San Miguel, El Salvador. In 2009, the UNAIDS national HIV estimation work group estimated the number of MSM in El Salvador at 2%-5% of the urban male population 15-49 years of age (148,489 in San Miguel) (de Estadistica y Censos, 2007; Paz-

Bailey et al., 2011; UNAIDS/WHO, 2003). The range for San Miguel is then 2970 to 7425 MSM.

Our goal, then, is to use the RDS survey information to estimate the sizes of the two populations. We will assess the level of certainty that is possible from the RDS data and the available prior information. We can then compare the approach to the guidelines provided by UNAIDS. As a final assessment we compare the approach to that possible from a separate capture-recapture study applied to the same populations.

3 Bayesian Inference for the Population Size

In this section we describe an approach to infer the population size, N, using data from an RDS survey. The approach taken is Bayesian, treating N as an unknown parameter. This requires a probability model for the observed data given N, as well as a prior for N. This sampling model is non-amenable to the model (Handcock and Gile, 2010). In fact, most information about the population size is drawn from the pattern in the sampling process. For this reason, the probability model must represent the sampling structure.

The distribution of the sampling process is modeled as a function of the *sizes* of units. The sampling model, described in Section 3.1 below, follows Gile (2011) and is based on a successive sampling approximation to the RDS process. The superpopulation model for these unit sizes is given in Section 3.3. In Section 3.2 the likelihood function is formed from these two models and then combined with a prior to make inference for N. Section 3.4 presents the forms of the prior distributions for the population size and unit size distribution.

3.1 Pragmatic Modeling of the RDS process as Successive Sampling

Many estimators for RDS (Heckathorn, 2007; Salganik and Heckathorn, 2004; Volz and Heckathorn, 2008) begin with the assumption that the sampling distribution can be treated as independent draws from a distribution proportional to nodal *degrees*, or numbers of contacts in the target population. This approximation is based on treating the sampling

process as a random walk on the nodes along the graph of the underlying social network. The stationary distribution of this random walk is proportional to nodal degree. That is, if the probability distribution of the sample at step k of a random walk is proportional to degree, then the probability distribution of the sample at step $k+1$ of the random walk will also be proportional to degree.

Gile (2011) extends this approximation to account for without-replacement sampling. She argues that under certain conditions, the corresponding distribution without-replacement is equal in distribution to a successive sampling process. While our inferential goal, estimating population size, is different from Gile's goal of estimating the mean of a nodal covariate, we use the paradigm of the successive sampling approximation to RDS to inform our development of methodology for estimating population size. We now describe the basis for the successive sampling approximation, more fully described in Gile (2011).

Consider a so-called *configuration model* for networks, a popular null model for networks, especially in the physics literature (Molloy and Reed, 1995). This model places equal probability on all networks with a given set of fixed nodal degrees. Networks from such a distribution are sometimes generated by starting with an empty graph and randomly adding edges between nodes with insufficient edges until all degrees are attained. For maximum degree small enough, the resulting distribution on networks is close to the configuration model distribution (Boguna et al., 2004; Burda and Krzywicki, 2003; Catanzaro et al., 2005; Chung and Lu, 2002; Foster et al., 2007).

Suppose we execute a random walk on a graph with unknown edges, but drawn from this distribution. Then at any given step, when an edge is followed from the previous node, the next node sampled will be drawn with probability proportional to degree. Thus the transition probabilities at each step of the sample will be proportional to nodal degree.

Now consider a without-replacement random walk of the same structure. Here, at each step of the sample, subsequent samples are restricted to previously unsampled nodes. In

this case, each subsequent sample is drawn from a distribution proportional to degree *from the remaining unsampled nodes only*.

This sampling structure is equivalent to *successive sampling* or *probability proportional to size without replacement* sampling (Andreatta and Kaufman, 1986; Bickel et al., 1992; Murthy, 1957; Nair and Wang, 1989; Raj, 1956), a sampling design in which units are sampled without replacement with unequal probabilities, such that each successive sample is drawn with probability proportional to *unit size* from among the remaining unsampled units. In particular, under this design, the sampling probability of the observed sequence of units takes the form:

$$p(G = g|U = u) = \prod_{i=1}^{n} \frac{u_{g_i}}{\sum_{j=1}^{N} u_j - \sum_{j=1}^{i-1} u_{g_j}},$$

where n is the sample size, and N the population size, $G = (G_1, \ldots, G_n)$ is the ordered sequence of sampled unit sizes, $U = \{U_1, \ldots, U_N\}$ is the population of unit sizes, with realised values g and u, respectively. So u_{g_i} is the realised unit size of the i^{th} sample. Note that in RDS, the unit sizes are typically nodal degrees, according to the above argument, although other functions of nodal features are also possible.

Gile (2011) uses this distribution to approximate the sampling probabilities of RDS respondents in order to weight the resulting sample. In contrast, we use the successive sampling approximation to model the sampling structure directly in the interest of estimating N. Note that this model depends on both the observed and unobserved values of u, as well as the unknown population size N, making the sampling structure non-amenable to the model, and requiring the joint modeling of the sampling structure and superpopulation of unit sizes. Indeed, most of the information about the population size is contained in the sample order.

3.2 Jointly Modeling the Unit Size Distribution and the Sampling Process

The population unit sizes are treated as an i.i.d. sample of size N generated from a superpopulation model based on some (unknown) distribution. For simplicity of presentation, the unit sizes are presumed to have the natural numbers as their support (e.g., degrees). Specifically: $U_i \overset{\text{i.i.d.}}{\sim} f(\cdot|\eta)$ where $f(\cdot|\eta)$ is a probability mass function (PMF) with support $1, \ldots,$ and η a parameter.

If the ordered observed sample is denoted by the random vector $G = (G_1, \ldots, G_n)$, with realised values $g = (g_1, \ldots, g_n)$, let $\backslash g = \{1, \ldots, N\} \backslash \{g_1, \ldots, g_n\}$ represent the set of indices of the unobserved population units. Further, consider the observed and unobserved unit sizes. Let $U_{obs} = \{U_{g_1}, U_{g_2}, \ldots, U_{g_n}\}$, the random vector of observed unit sizes, with values $u_{obs} = (u_{g_1}, \ldots, u_{g_n})$. Similarly, let $U_{unobs} = \{U_i\}_{i \in \backslash g}$ and $u_{unobs} = \{u_i\}_{i \in \backslash g}$ represent the random and realised values of the unit sizes of the unobserved units. Thus the full observed data are $\{g, u_{obs}\}$. For simplicity of notation, denote the observed data by $D = (U_{obs} = u_{obs}, G = g)$.

The joint posterior is:

$$
\begin{aligned}
p(N, \eta|D) &\propto \pi(N, \eta) \cdot p(D|N, \eta) \qquad (1)\\
&= \pi(N, \eta) \cdot \sum_{u_{unobs} \in \mathcal{U}(u_{obs})} p(G = g, U = (u_{obs}, u_{unobs})|\eta)\\
&= \pi(N, \eta) \cdot \sum_{u_{unobs} \in \mathcal{U}(u_{obs})} p(G = g|U = (u_{obs}, u_{unobs})) \prod_{j=1}^{n} f(u_{g_j}|\eta) \cdot \prod_{j \in \backslash g} f(u_j|\eta),
\end{aligned}
$$

where $\pi(N, \eta)$ is a prior for the population size and the unit size distribution parameter, and $\mathcal{U}(u_{obs})$ is the set of possible u_{unobs} given u_{obs}, that is the unit sizes possible for the remaining $N - n$ units given that the first n sampled were u_{obs}. Typically, this will be the $N - n$ product support of $f(\cdot|\eta)$. Thus the correct model is related to the complete data model through the sampling design as well as the superpopulation model.

Note that this approach is an extension of the approach developed by West (1996),

who used a successive sampling approximation to estimate the number of un-discovered oil fields, and their volume of oil. The approach used here, and presented in Handcock et al. (2012) extends the work of West in three ways. First, the unit sizes are modeled as discrete rather than continuous. Second, the branching and network nature of the RDS sample may reduce or confound the information in the ordering of the sample. Third, the sample sizes of RDS samples are typically at least an order of magnitude larger, and with a different range of unit sizes than in the data available in ecological applications such as oil fields.

3.3 Models for the Unit Size Distribution

We now treat the parametric model for the distribution of the unit sizes in equation (1). The question of models for the degree distributions of social networks has been extensively studied in Handcock and Jones (2004) and broad classes are included in the accompanying software (Handcock, 2011). We will use the Conway-Maxwell-Poisson class of distributions as it allows both under-dispersion and over-dispersion relative to a Poisson distribution via a single additional parameter (Shmueli et al., 2005).

3.4 Specifying prior knowledge about the population size and unit size distribution

The model allows for an arbitrary prior distribution over the population size (N). However, this is an opportunity to choose priors that aid elicitation of expert prior information or easily incorporate previous or concomitant sources of information about the population size.

The most common parametric models for N (e.g., Negative Binomial) typically have too thin tails for large N. This issue has been treated by Fienberg et al. (1999). They suggest the prior:

$$\pi(N) = (N-l)!/N! \quad \text{for} \quad n < N < N_{max}, \tag{2}$$

where N_{max} covers the range where the likelihood is non-negligible. For their applications

they choose their Jeffrey's prior $l = 1, \pi(N) \propto 1/N, n < N < N_{max}$. In addition to these possibilities, we propose a new class of priors specifying knowledge about the sample proportion (i.e. n/N) as a Beta(α, β) distribution. The implied density function on N (considered as a continuous variable) is:

$$\pi(N) = \beta n (N-n)^{\beta-1}/N^{\alpha+\beta} \quad \text{for} \quad N > n. \tag{3}$$

The distribution has tail behavior $O(1/N^{\alpha+1})$. We have found this class of priors to be very useful: It is often relatively flat in regions where the likelihood is centered. The long right tail allows large population sizes but the rate of decline ameliorates this.

When $\alpha = l - 1 > 0$, this class is similar to that of Fienberg et al. (1999). The Beta prior class, however, is directly motivated as a proper prior on the sample proportion. Figure 2 presents three different versions of this prior, corresponding to a prior mean, median and mode of 1000.

[Figure 2 about here.]

For simplicity, in this paper we specify that N and η are *a priori* independent so that $\pi(N, \eta) = \pi(N) \cdot \pi(\eta)$. The Conway-Maxwell-Poisson distribution for unit sizes, can be parameterised in terms of its mean and standard deviation, and this can aid elicitation of prior information about them. In this study, the prior for the the mean given the standard deviation is normal and the variance is scaled inverse Chi-squared:

$$\mu|\sigma \sim N(\mu_0, \sigma/df_{\text{mean}}) \qquad \sigma \sim \text{Inv}\chi(\sigma_0; df_{\text{sigma}}).$$

The default prior on these parameters is diffuse with an equivalent sample size of $df_{\text{mean}} = 1$ for the mean of the unit size distribution and $df_{\text{sigma}} = 5$ for the variance of the unit size distribution.

3.5 Computation

The joint posterior $p(N, \eta, U_{unobs} = u_{unobs}|D)$ can be sampled from using a four component Gibbs sampler, the details of which are given in Handcock et al. (2012). This can then

be marginalised to produce samples from $p(N|D)$, $p(\eta|D)$, and the posterior predictive distribution of the unobserved unit sizes, $p(U_{unobs} = u_{unobs}|D)$. Hence it produces posterior predictive distributions of the full population of unit sizes $(u_i, i = 1, \ldots, N)$. These posteriors enable inference for such quantities as the population size, the mean unit size, the unit size distribution, etc.

4 Estimating the Number of Female Sex Workers in Sonsonate, El Salvador

In this section we infer the number of Female Sex Workers in Sonsonate, El Salvador based on the RDS survey described in Section 2.

A strength of our method is the ability of incorporate prior knowledge of different types and sources (3.4). We consider three prior specifications that reflect different frames of reference for the public-health officials.

4.1 Use of a reference prior

In this sub-section we consider the case where the prior for the population size is taken to be constant over the range of population sizes where the likelihood is non-negligible. This would not usually be used to produce a final estimate but could be used as a baseline for other specifications. In particular, the resulting posterior reflects the shape of the likelihood and divergences from it based on other prior information can be instructive.

The first panel of Figure 3 plots both the prior and posterior distributions in this case. The posterior mass ranges from the sample size (184) up to about 4000. The peakedness of the posterior shape indicates that there is information in the data about the population size, with a mode of around 1250 FSW. To help calibrate the information the plot of the posterior includes additional benchmarks. The lower purple line is the lower end of the UNAIDS guideline (560 = 0.4%). The upper purple line is at the upper UNAIDS guideline (1120 FSW). The UNAIDS guidelines fall in the mid to low part of the posterior distribution and are broadly consistent with it. The lower blue line is at the 2.5% quantile of the posterior and is close to the lower UNAIDS guideline. The upper blue line is at the

97.5% quantile, at about 2800. We note that the UNAIDS guidelines, although based on more general considerations, do fall within the 95% HPD interval (blue lines).

[Figure 3 about here.]

4.2 Use of a prior expression of central tendency

In many situations the field researchers are willing to express their prior belief about the population size but struggle to express it via a fully specified distribution. To aid in this process we ask them to express it either as (a) "a value it is as likely to be above as below"; (b) "the most likely value"; or (c) "the value averaged from all knowledgeable people". Based on this we find the prior in the class (3) that matches that value (i.e., median, mode, mean, respectively). This class has a long right tail, capturing the often expressed belief that there is significant probability mass on large population sizes. We have found the sub-class with $\alpha = 1$ to be the most useful, when a single measure of central tendency is elicited.

For the population of FSW, the expressed belief by the researchers was that the mean population size was at the mid-point of the UNAIDS suggested range $0.6\% \times 139804 = 838$. The middle panel of Figure 3 plots this prior and the resulting posterior. The mean, median and mode of the posterior fall within the UNAIDS guidelines (purple lines) and these guidelines fall within the the 95% HPD interval. As expected, this prior results in more mass in the posterior in the area of the UNAIDS estimates.

In addition to measures of central tendency, we may also ask field researchers to specify their prior beliefs via quantiles. The explicit questions were the population size values that there is: (d) "One chance in four of being less than"; (e) "One chance in four of being greater than"; (f) "Most reasonably lowest"; (g) "Most reasonably highest". The first two can be used to find the prior in the class (3) that matches that the two quantiles. This information specifies both α and β. The answers to (f) and (g) are often used to set the

extreme quantiles (e.g., 5% and 95%) and also are asked to allow the researcher to calibrate their answers to all the questions, so improving self-consistency.

4.3 Calibrating the prior information from the UNAIDS guidelines

The previous approaches are based on eliciting prior information from the researchers and reflects both their beliefs and the information in the RDS data. In this section we consider the additional approach based on a direct use of the UNAIDS guidelines. This effectively uses the UNAIDS guidelines as a specification of their prior belief about the population size. The approach then refines that belief using the survey data specific to the population.

As UNAIDS provides a range of values, it may be useful to specify a prior based on multiple points in that range. The parametric class of priors described by (3) allows the flexibility to choose a prior that reflects a range of values. In this case, two parameters (α, β) were chosen so that the prior mean is the midpoint of the range (0.6%) and the lower quartile of the prior is the lower UNAIDS estimate (0.4%). The third panel of Figure 3 plots this prior and resulting posterior distribution. Note that the prior reflects the guidelines, but is also quite right skewed. As this prior is intended to be a closer match to the UNAIDS guidelines, and better captures the wide range in the guidelines (missed by the prior in the previous sub-section), the posterior is slightly broader than the previous one as it better captures that uncertainty.

This posterior distribution has a mode at about 1000 FSW, and a 95% HPD interval from 481 to 1998 FSW. The posterior is slightly broader than in the previous case, reflecting the broader prior used. Note that both posteriors based on the use of the UNAIDS guidelines are narrower than that based on the reference prior.

5 Estimating the Number of Men-who-have-sex-with-Men in San Miguel, El Salvador

We turn now to a second high-risk population, that of MSM in San Miguel, El Salvador. We conducted the same analysis process as for the FSW. As noted in Section 2, an application of the UNAIDS guidelines suggested the population of MSM in San Miguel is between 2970 and 7425.

Figure 4 shows the same three prior specifications as the previous case. The first panel plots the posterior distribution and the prior when the prior is constant over the range of population sizes where the likelihood is non-negligible. The peakedness of the posterior shape again indicates that there is information in the data about the size, but it is diffuse and has a long upper tail compared to that for the FSW. The UNAIDS guidelines (purple lines) fall in the mid to upper part of the distribution, and are well within the 95% HPD interval (blue lines).

[Figure 4 about here.]

The second panel plots the posterior distribution based on the prior with mean the mid-point of the UNAIDS suggested value $3.5\% \times 148489 = 5197$. The majority of the posterior is below the UNAIDS estimates as the prior pulls in the larger values while the 95% interval mostly covers the UNAIDS estimates.

The last panel of Figure 4 plots the posterior distribution based on the prior fixing the mean at the midpoint of the range (3.5%) and the lower point (2%) at the lower quartile. This prior contains the most information from the UNAIDS work group and hence is perhaps the best choice. The resulting posterior distribution has a mode at about 2100 MSM, and a 95% HPD interval from 200 to 7048 MSM. Thus this method yields an estimate of the number of MSM in San Miguel with a wide interval.

6 Comparison to a Capture-Recapture Method

Because there is no other way to estimate the population size from RDS data alone, it is difficult to benchmark the performance of our method. We have already considered comparison to the region-wide guidelines provided by UNAIDS (UNAIDS/WHO, 2003). Further comparison requires additional data collection. Because of the importance of population size estimation, separate population size estimation studies are conducted in many areas. Indeed, we can compare our results to results of a separate capture-recapture based study used to estimate the number of MSM and FSW in San Salvador in 2008 (Paz-Bailey et al., 2011). Absent more local information, it is plausible that the population percentages of MSM and FSW in San Salvador may be similar to those in Sonsonate and San Miguel. This approach required the distribution of tokens (e.g., key chains) throughout the population followed by a recapture step with a follow-up survey. Paz-Bailey et al. (2011) estimate that the size of the FSW population in San Salvador is almost double the UNAIDS figures (1.4%). This population proportion in Sonsonate would translate to 2079 FSW, close to the posterior mean of the proposed method for the flat prior, but in the upper tail of the posteriors based on priors using the UNAIDS guidelines. Paz-Bailey et al. (2011) estimate that the size of the MSM population in San Salvador is close to the UNAIDS figures (3.4%). This is somewhat high but comparable to the MSM results in Figure 4. Thus the results of our case-study are largely comparable to those of Paz-Bailey et al. (2011) when they differ from the UNAIDS guidelines.

7 Assumptions and Limitations

The method relies on an approximation of the RDS sampling process which assumes that units (people) have observable sizes (here, network degree), and that sampling proceeds according to a successive sampling procedure in which each subsequent sample is selected from among the remaining units with probability proportional to size. In the RDS

context, this condition is satisfied if we consider network structures sampled from a so-called configuration model, in which network ties form completely at random among a population of people with fixed and observable degrees (Gile, 2011). In practice, we know that a configuration model is only an approximation to the underlying network structure, and from this, we intuit that the method should have degraded performance for networks with structure far from this distribution. In Handcock et al. (2012), we explore this phenomenon further through systematic simulation studies. There are several limitations of the method, of which users should be aware. First, as shown in our results, the amount of information in the data may be small enough that the method is sensitive to the prior chosen. Where possible, informative priors based on existing information should be used, and thereby incorporated into the estimator. Second, the performance of the method is degraded by substantial deviations from the assumed sampling structure. In particular, for highly structured data, such as very clustered populations, the method may not be valid, and the method may be sensitive to mis-reporting of network degrees. In highly clustered populations, we recommend conducting RDS separately within each cluster, and the same recommendation would apply to population size estimation. In ongoing work, we explore approaches to treating mis-reported network degrees.

8 Discussion

In studies of HIV/AIDS risk populations, the number of people at risk for infection is of fundamental importance. However public health officials often have little information about the specific population of interest and are forced to use generalised estimates from broad geographic regions or social contexts (UNAIDS and World Health Organization, 2010). In many such high-risk populations, surveys are conducted via RDS primarily as a means to estimate prevalence (Johnston et al., 2008).

In this paper we present a case-study of the use of a new methodology to estimate population size from RDS data that can incorporate expert prior information or data from

other sources. The study is of two different populations in El Salvador. The first is female sex workers in Sonsonate and the second is men-who-have-sex-with-men in San Miguel. Because so little was know about the populations, public health officials were using broad ranges of figures of unknown accuracy to estimate the specific population sizes (Paz-Bailey et al., 2011; UNAIDS/WHO, 2003). Our method provides interval estimates based on RDS surveys in the populations. In the case-study, we show how the method can be used to incorporate various forms of prior information including that from broad administrative guidelines. The method has the strength that it expresses a credible measure of the certainty in the population size based on the available information, especially in cases where the uncertainty is large.

Supplementary Materials

An R package implementing the methods used in this paper along with code using it on example data is available with this paper at the Biometrics website on Wiley Online Library.

ACKNOWLEDGEMENTS

The project described was supported by grant number 1R21HD063000 from NICHD and grant number MMS-0851555 from NSF, grant number N00014-08-1-1015 from ONR, and grant number SES-1230081 from NSF, including support from the National Agricultural Statistics Service. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Demographic & Behavioral Sciences (DBS) Branch, the National Science Foundation, the Office of Naval Research, or the National Agricultural Statistics Service. The authors would like to thank the members of the Hard-to-Reach Population Research Group (`hpmrg.org`), especially Lisa G. Johnston, for their helpful input. We would like to express our gratitude to the Ministry of Health of El Salvador for allowing us to use the results of the Encuesta Centroamericana de VIH y Comportamientos en Poblaciones Vulnerables, ECVC-EL Salvador. We would especially

like to thank the study principal investigators Dr. Ana Isabel Nieto and Gabriela Paz-Bailey, and the study coordinator Maria Elena Guardado for their assistance interpreting the results of this analysis. Funding for the ECVC-El Salvador study was provided by the United States Centers for Disease Control and Prevention, United States Agency for International Development, the Ministry of Health of El Salvador, and the World Bank.

REFERENCES

Andreatta, G. and Kaufman, G. M. (1986). Estimation of finite population properties when sampling is without replacement and proportional to magnitude. *Journal of the American Statistical Association* **81,** 657–666.

Bao, L., Raftery, A. E., and Reddy, A. (2010). Estimating the size of populations at high risk of HIV in Bangladesh using a Bayesian hierarchical model. Department of Statistics Technical Report no 573, University of Washington.

Berchenko, Y. and Frost, S. D. W. (2011). Capture-recapture methods and respondent-driven sampling: their potential and limitations. *Sexually Transmitted Infections* **87,** 267–268.

Bickel, P. J., Nair, V. N., and Wang, P. C. C. (1992). Nonparametric inference under biased sampling from a finite population. *The Annals of Statistics* **20,** 853–878.

Boguna, M., Pastor-Satorras, R., and Vespignani, A. (2004). Cut-offs and finite size effects in scale-free networks. *European Physical Journal B* **38,** 205–209.

Burda, Z. and Krzywicki, A. (2003). Uncorrelated random networks. *Physical Review E* **67,** 046118.

Catanzaro, M., Boguna, M., and Pastor-Satorras, R. (2005). Generation of uncorrelated random scale-free networks. *Physical Review E* **71,**.

Chung, F. and Lu, L. (2002). Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics* **6,** 125–145.

de Estadistica y Censos, D. G. (2007). Vi censo de poblacion y v de vivienda El Salvador.

Technical report, Ministerio de Economia de El Salvador.

Fienberg, S. E., Johnson, M. S., and Junker, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **162,** 383–405.

Foster, J. G., Foster, D. V., Grassberger, P., and Paczuski, M. (2007). Link and subgraph likelihoods in random undirected networks with fixed and partially fixed degree sequences. *Physical Review E* **76,**.

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

Guardado, M. E., Creswell, J., Monterroso, E., et al. (2010). Encuesta centroamericana de vigilancia de comportamiento sexual y prevalencia de VIH/ITS en poblaciones vulnerables, hombres que tienen sexo con hombres, trabajadoras sexuales y personas con VIH, ECVC El Salvador. Research report, Minisiterio de Salud de El Salvador.

Handcock, M. S. (2011). **size***: Estimating Population Size from Discovery Models using Successive Sampling Data*. Hard-to-Reach Population Methods Research Group, Los Angeles, CA. R package version 0.20.

Handcock, M. S. and Gile, K. J. (2010). Modeling networks from sampled data. *Annals of Applied Statistics* **272,** 383–426.

Handcock, M. S., Gile, K. J., and Mar, C. M. (2012). Estimating hidden population size using respondent-driven sampling data. Preprint at `http://arxiv.org/abs/1209.6241`.

Handcock, M. S. and Jones, J. H. (2004). Likelihood-based inference for stochastic models

of sexual network formation. *Theoretical Population Biology* **65,** 413–422.

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

Heckathorn, D. D. (2007). Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology* **37,** 151–207.

Johnston, L. G. (2007). Conducting respondent driven sampling (RDS) studies in diverse settings: A training manual for planning RDS studies. Centers for Disease Control and Prevention, Atlanta, GA and Family Health International, Arlington, VA.

Johnston, L. G., Malekinejad, M., Kendall, C., Iuppa, I. M., and Rutherford, G. W. (2008). Implementation challenges to using respondent-driven sampling methodology for HIV biological and behavioral surveillance: Field experiences in international settings. *AIDS and Behavior* **12,** 131–141.

Molloy, M. S. and Reed, B. A. (1995). A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* **6,** 161–179.

Morales-Miranda, S., Paz-Bailey, G., Alvarez, B., et al. (2007). Behavioral and HIV survey among female sex workers and men who have sex with men in Honduras using respondent driven sampling. Technical report, University of Valle of Guatemala.

Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya: The Indian Journal of Statistics* **18,** 379–390.

Nair, V. N. and Wang, P. C. C. (1989). Maximum likelihood estimation under a successive sampling discovery model. *Technometrics* **31,** 423–436.

Paz-Bailey, G., Jacobson, J. O., Guardado, M. E., Hernandez, F. M., Nieto, A. I., 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,** 279–282.

Raj, D. (1956). Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association* **51,** 269–284.

Rocchetti, I., Bunge, J., and Böhning, D. (2011). Population size estimation based upon ratios of recapture probabilities. *Annals of Applied Statistics* **5,** 1512–1533.

Salganik, M. J., Fazito, D., Bertoni, N., Abdo, A. H., Mello, M. B., and Bastos, F. I. (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.

Salganik, M. J. and Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology* **34,** 193–239.

Shmueli, G., Minka, T. P., Kadane, J. B., 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.

Soto, R. J., Ghee, A. E., Nunez, C. A., Mayorga, R., Tapia, K. A., Astete, S. G., Hughes, J. P., Buffardi, A. L., Holte, S. E., Holmes, K. K., and the Estudio Multicentrico Study Team (2007). Sentinel surveillance of sexually transmitted infections/HIV and risk behaviors in vulnerable populations in 5 Central American countries. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **46,** 101–111.

UNAIDS (2009). Estimating National Adult Prevalence of HIV-1 in Concentrated Epidemics. Technical report, UNAIDS - Joint United Nations Programme on HIV/AIDS.

UNAIDS and World Health Organization (2010). Guidelines on estimating the size of populations most at risk to HIV. Technical Report UNAIDS/00.03E, UNAIDS - Joint United Nations Programme on HIV/AIDS.

UNAIDS/WHO (2003). Estimating the size of populations at risk for HIV: Issues and

methods. Technical report, UNAIDS/WHO Working Group on HIV/AIDS.

UNAIDS/WHO (2010). HIV/AIDS health profile of El Salvador. Technical report, UNAIDS/WHO Working Group on HIV/AIDS.

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

West, M. (1996). Inference in successive sampling discovery models. *Journal of Econometrics* **75,** 217–238.

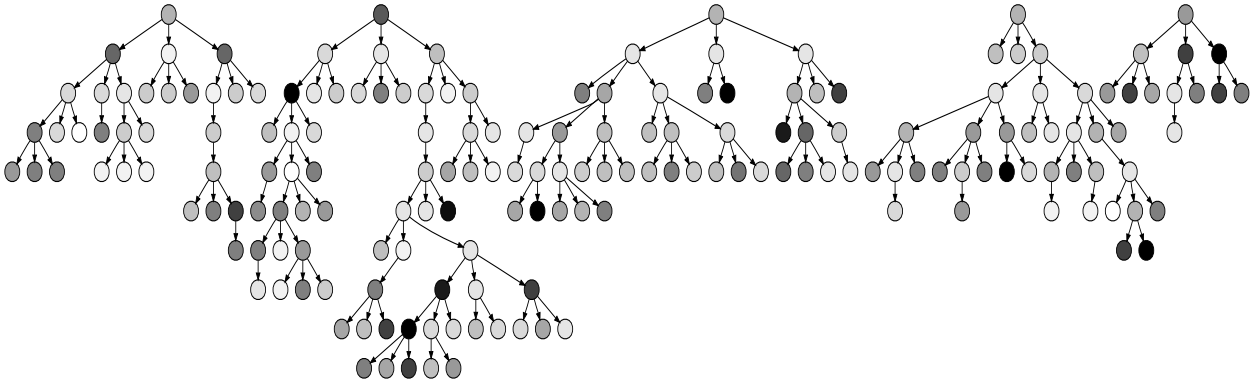

Figure 1: Graphical representation of the recruitment tree for the sampling of female sex workers in Sonsonate, El Salvador in 2010. The nodes are the respondents and the wave number increases as you go down the page. The node gray scale is proportional to the self-reported degree with white being degree one and black the maximum degree.

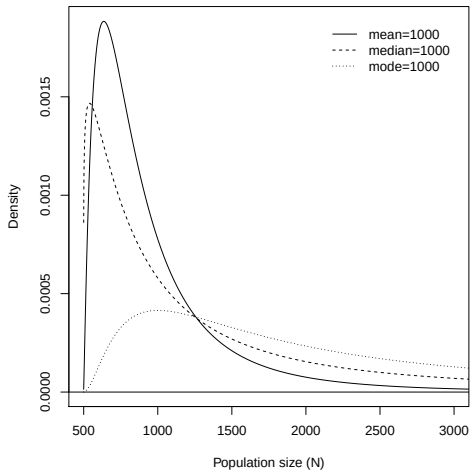

Figure 2: Three example prior distributions for the population size (N). They correspond to $\alpha = 1$ and $\beta = 1.55, 1.16$ and 3.

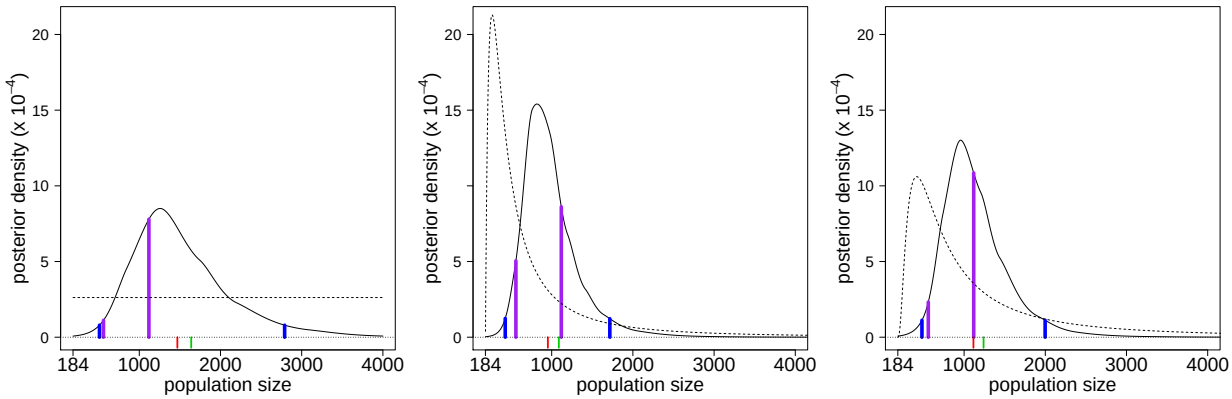

Figure 3: Posterior distribution for the number of female sex workers in Sonsonate based on three prior distributions for the population size: flat, matching the midpoint UNAIDS estimate, and interval-matching the UNAIDS estimate. The prior is dashed. The red mark is at the posterior median. The green mark is at the posterior mean. The blue lines are at the lower and upper bounds of the 95% highest-probability-density interval. The purple lines demark the lower and upper UNAIDS guidelines.

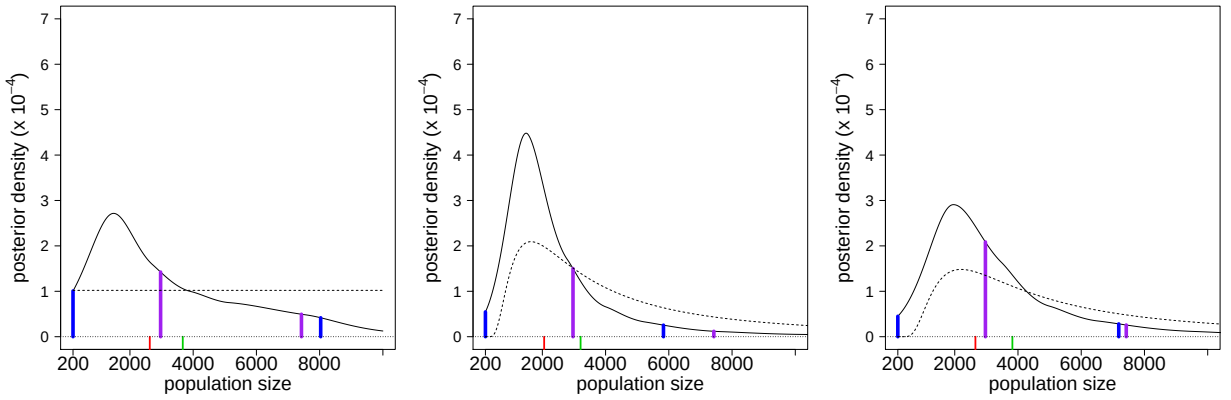

Figure 4: Posterior distribution for the number of MSM in San Miguel based on three prior distributions for the population size: flat, matching the midpoint UNAIDS estimate, and interval-matching the UNAIDS estimate. The prior is dashed. The red mark is at the posterior median. The green mark is at the posterior mean. The blue lines are at the lower and upper bounds of the 95% highest-probability-density interval. The purple lines demark the lower and upper UNAIDS guidelines.