The Annals of Applied Statistics
2024, Vol. 18, No. 1, 683–703
https://doi.org/10.1214/23-AOAS1807

MODELING THE VISIBILITY DISTRIBUTION FOR RESPONDENT-DRIVEN SAMPLING WITH APPLICATION TO POPULATION SIZE ESTIMATION

BY KATHERINE R. MCLAUGHLIN[1,a], LISA G. JOHNSTON[2,b], XHEVAT JAKUPI[3,c], DAFINA GEXHA-BUNJAKU[3,d], EDONA DEVA[4,e] AND MARK S. HANDCOCK[4,f]

[1]*Department of Statistics, Oregon State University,* [a]*katherine.mclaughlin@oregonstate.edu*
[2]*LGJ Consultants, Inc.,* [b]*lsjohnston.global@gmail.com*
[3]*National Institute of Public Health of Kosovo,* [c]*xhevat.r.jakupi@rks-gov.net,* [d]*dafinagexha@gmail.com*
[4]*Community Development Fund,* [e]*edona.deva@kcdf.org*
[5]*Department of Statistics and Data Science, University of California, Los Angeles,* [f]*handcock@stat.ucla.edu*

Respondent-driven sampling (RDS) is used throughout the world to estimate prevalence and population size for hidden populations. Although RDS is an effective method for enrolling people from key populations in studies, it relies on a partially unknown sampling mechanism, and thus each individual's inclusion probability is unknown. Current estimators for population prevalence, population size, and other outcomes rely on a participant's network size (degree) to approximate their inclusion probability in the sample from the networked population. However, in most RDS studies, a participant's network size is attained via a self-report and is subject to many types of misreporting and bias. Because design-based inclusion probabilities cannot be exactly computed, we instead use the term visibility to describe how likely a person is to be selected to participate in the study. The commonly used successive sampling population size estimation (SS-PSE) framework to estimate population sizes from RDS data relies on self-reported network sizes in the model for the sampling mechanism. We propose an enhancement of the SS-PSE framework that adds a measurement error model for visibility used in place of the self-reported network size and a model for the number of recruits an individual can enroll. Inferred visibilities are a way to smooth the degree distribution and bring in outliers as well as a mechanism to deal with missing and invalid network sizes. We demonstrate the performance of visibility SS-PSE on three populations from Kosovo sampled in 2014 using RDS. We also discuss how the visibility modeling framework could be extended to prevalence estimation.

1. Introduction. Respondent-driven sampling (RDS) (Heckathorn (1997)) is a highly effective method to sample from hidden populations that cannot be reached through traditional probability samples and for which the sampling frames are unknown (Gile and Handcock (2010)). In particular, RDS is typically used for key populations that are at higher risk for the transmission of HIV/AIDS and related diseases, such as people who inject drugs (PWID), female sex workers (FSW), and men who have sex with men (MSM) (Johnston (2013)). These populations share much of the burden of the global HIV/AIDS epidemic. Countries report HIV/AIDS prevalence rates and population size estimates among these key populations to the World Health Organization (WHO) and UNAIDS from samples conducted using RDS (Gile, Johnston and Salganik (2015)). These estimates are used to inform policy decisions, budgetary considerations, and the allocation of resources for treatment and prevention efforts.

RDS utilizes the underlying social network of the population of interest and relies on participants in the study to recruit their peers (Heckathorn (1997)). An initial set of *seeds*

Received December 2021; revised July 2023.
Key words and phrases. Heaped data, hidden population, measurement error model, model-based survey sampling, network sampling.

is selected, usually via a convenience sample. After completing the survey instrument, each seed is given a small number of *coupons* (usually two to three) containing unique identifying information to track recruitment while maintaining anonymity and is told to distribute them to members of their social network who meet the study eligibility requirements. The recipients of these coupons form the first wave of the study and, upon completion of the survey, are given their own coupons to distribute. Recruitment continues in this manner for many *waves* until the desired sample size is attained or no additional participants are recruited. Participants often receive a small primary incentive for completing the survey and a small secondary incentive for each recruit they successfully enroll in the study.

Because study participants, rather than researchers, control recruitment into the study, RDS almost always has an *unknown sampling mechanism* (Gile and Handcock (2010)). Therefore, without accounting for the complex design, the sample is a nonprobability sample, and key outcome measures for the population cannot be computed using traditional survey sampling methods. Instead, several RDS estimators have been developed to attempt to account for the dependence among members of the sample and potential biases of their responses. Two commonly used RDS estimators, the Volz–Heckathorn estimator and successive sampling estimator, are described briefly below. Consider a population of N individuals, denoted by indices $1, \ldots, N$. Assume we are trying to estimate the prevalence of characteristic A in the population, denoted P_A, where A is, for example, being HIV positive. Let $i \in s$ denote that individual i is in the sample, let $n = |s|$ be the sample size, and further let $i \in s_A$ denote that individual i is in the sample and has characteristic A. Thus, $s_A \subseteq s$. There are $|s_A| = n_A$ people in the sample with characteristic A. The self-reported network size (degree) for individual i is $\tilde{d}_i$. The true selection probability for individual i is π_i, but π_i is unknown.

The Volz–Heckathorn (RDS-II) estimator (Volz and Heckathorn (2008)) is given by

$$\widehat{P_A^{\mathrm{VH}}} = \frac{\sum_{i \in s_A} 1/\tilde{d}_i}{\sum_{i \in s} 1/\tilde{d}_i}. \tag{1}$$

This is a generalized Hansen–Hurwitz estimator motivated by assuming an individual's inclusion probability is the drawwise selection probability from a simple random walk over the underlying social network that has reached equilibrium. For this to hold in RDS, we must assume that recruitment occurs via a single unbranching chain, that the chain is of a sufficient number of waves such that dependence on the original seed is removed, and that referral is random, along with other assumptions about the structure of the population social network (Gile and Handcock (2010)). It is asymptotically unbiased for P_A if $\pi_i \propto \tilde{d}_i$ for all i under the assumption of infinite population size (Tomas and Gile (2011)).

The successive sampling (SS) estimator (Gile (2011)) adjusts the Volz–Heckathorn estimator to account for the fact that sampling proceeds without replacement (i.e., a person cannot participate twice). It is given by

$$\widehat{P_A^{\mathrm{SS}}} = \frac{\sum_{i \in s_A} 1/\tilde{\pi}_i}{\sum_{i \in s} 1/\tilde{\pi}_i}, \tag{2}$$

where $\tilde{\pi}_i$ is the estimated sampling probability of individual i resulting from the assumption that selection is governed by a successive sampling (or probability proportional to size without replacement, PPSWOR) process (Gile (2011)). This estimator also requires knowledge of N, the population size, and performs particularly well in comparison to the RDS-II estimator for large sample fractions n/N.

These estimators both rely on accurate measures of each person's degree d_i (personal network size). Other common RDS estimators, such as the Salganik–Heckathorn (RDS-I) estimator (Salganik and Heckathorn (2004)) and the Homophily Configuration Graph (HCG) estimator (Fellows (2019)), as well as estimators for extensions of RDS sampling, such as

those for privatized network sampling (Fellows (2022a), Khan et al. (2018)), also rely on accurate measures of degree.

Degree is typically constructed using the elements in the study eligibility criteria. For example, a participant may first be asked how many people they know by name and face, who also know them by name and face. Then they would be asked how many of those people are between the ages of 18–65 and how many of those meet various other study eligibility criteria. Finally, it is common to ask a last targeted question such as "How many of those people have you shared a meal with in the past two weeks?" Ultimately, through these limiting questions, each individual will provide a self-report of their degree.

For the RDS-II estimator, an individual's inclusion probability is assumed to be proportional to their self-reported degree: $\pi_i \propto \tilde{d}_i$ for all i. Individuals with higher degrees are given lower weights, and so biases in self-reported degree could result in biased estimates if degree is related to the outcome of interest. Similarly, for the SS estimator, it is assumed that $\pi_i \propto \tilde{d}_i$ conditional on the people not yet sampled. The successive sampling-population size estimation (SS-PSE) method for population size estimation from RDS data also relies directly on degree because it incorporates the SS model for the sampling process (Handcock, Gile and Mar (2014), Handcock, Gile and Mar (2015)). Thus, biases in self-reported degree could also result in incorrect population size estimates. Further, because SS-PSE assesses the decrease in network size over the order of enrollment in the study, biases in self-reported degree could compound incorrect population size estimates if the bias is associated with time.

In typical RDS studies, degree is self-reported (i.e., we observe only $\tilde{d}_i$, not d_i). This means that degree is subject to misreporting and bias, especially because the members of key populations may practice stigmatized or illegal activities (Gile, Johnston and Salganik (2015)). Types of error that may occur due to self-reporting are discussed in more detail in Section 2.1. Beyond these problems, degree itself may not correspond to inclusion probability, or inclusion probability may depend on other factors besides degree. Prevalence and population size estimators for RDS model the propensity for inclusion in the sample as a function of degree. Because we cannot calculate an exact design-based inclusion probability for RDS, we refer to this propensity as *visibility*, u_i, where u_i is not necessarily equal to d_i. Visibility cannot be directly measured, but we can infer it from information already collected in the RDS study.

Given the unknown sampling mechanism of RDS, there have been many assessments of the sensitivity of inference to assumptions about the sampling mechanism; see Gile and Handcock (2010) and Gile et al. (2018) for a review. Lee et al. (2017) consider RDS from a total survey error perspective, emphasizing the importance of measurement error in the self-reported network size. They investigate measurement error in a total of four RDS samples from populations at high risk for HIV in Chicago, Los Angeles, and San Francisco. Fellows (2022b) studies two populations at high risk for HIV in the Dominican Republic. He also finds evidence of significant measurement error and analyzes its impact on prevalence estimates (e.g., our equations (1) and (2)). He finds that common RDS prevalence estimators remain consistent under measurement error, although the variance of the estimators is inflated.

A further discussion of these errors and the concept of visibility is provided in Section 2, motivated by data from three RDS studies in Kosovo. Section 3 details the model we propose to model the visibility distribution, and Section 4 discusses Bayesian inference within the SS-PSE framework as well as possible uses and extensions of the model. Results for the Kosovo data are given in Section 5. Section 6 provides a discussion of the results. Finally, concluding remarks are given in Section 7.

2. Visibility in a network sample. Common RDS estimators as well as SS-PSE rely on accurate measures of the self-reported network size $\tilde{d}_i$ and assume that $\tilde{d}_i = d_i$, meaning that

the degree is known without error. Additionally, the implicit assumption is made that $d_i \propto u_i$, where u_i is the visibility of individual i. If these assumptions are incorrect, our inferences may not be valid.

2.1. *Sources of error for network size.* The first assumption that the self-reported degree is equal to the true degree for each participant ($\tilde{d}_i = d_i$) is likely not true in practice for a variety of reasons, including:

(1) heaping/rounding/coarsening or other approximation methods (Gile and Handcock (2010)),
(2) intentional misreporting, perhaps in an attempt to minimize one's connection to a stigmatized population (Bengtsson and Thorson (2010), Fenton et al. (2001), Fisher (1993)),
(3) unintentional misreporting, perhaps due to a lack of understanding of the question or memory recall bias (Bell, Belli-McQueen and Haider (2007), Brewer (2000), Lu (2013), Mills et al. (2014)), and
(4) incorrect construction of the degree question, which can occasionally occur in practice (Johnston et al. (2010)).

Within the total survey error framework, all of these possible errors would be types of nonsampling error (Groves et al. (2009), Lee et al. (2017)). (1) and (2) are types of measurement error. (3) could be considered a measurement error as well if the unintentional misreporting is due to memory recall bias or a type of specification error if the unintentional misreporting is due to a poorly phrased question. Finally, (4) is a type of specification error that could be mitigated with formative research and planning. Nevertheless, incorrect construction of the degree question does occasionally occur, especially because of translation of the survey instrument and may be detected by participants giving impossibly large values of network size. In addition to these types of errors, it is also possible for the network size to be missing (nonresponse error) or to be incorrect due to an accidentally induced processing error. Because the researcher will never know which reason(s) led to $\tilde{d}_i \neq d_i$, we focus on a model framework that can account for different types of errors.

As an example, consider an individual with "true" degree $d_i = 23$. A heaped self-report could be $\tilde{d}_i = 20$, as the person is rounding to the nearest multiple of 10. Likewise, it is possible the individual could have reported $\tilde{d}_i = 18$ due to either intentional or unintentional misreporting. Both of these values underestimate the individual's "true" degree. Obviously, there are also cases where overestimation occurs, and it is common in RDS studies for at least one participant to report a network size deemed impossibly large. Commonly, these very large reported degrees are also rounded, for example, a value of 500. In cases where systematic differences exist between d_i and $\tilde{d}_i$ in a sample, it is possible that an overall prevalence estimate $\widehat{P_A}$ is biased if degree is related to the outcome of interest. Similarly, for population size estimation, it is possible that the estimated $\hat{N}$ is biased if degree is biased and/or the bias is related to the order of enrollment.

2.2. *Examples of possible error from three RDS studies in Kosovo.* To illustrate what observed distributions of self-reported degree $\tilde{d}$ look like, we introduce data from three RDS studies in Kosovo, which served as the motivation for the development of the visibility model. These studies were completed for the third round of the Integrated Behavioral and Biological Surveillance (IBBS) surveys in July, August, and September of 2014 among people who inject drugs (PWID) in Prishtina and in Prizren, Kosovo, and among men who have sex with men (MSM) in Prishtina, Kosovo (Kosovo HIV Integrated Behavioral Biological Surveillance Survey Reference Group (2014)). The third round of the Kosovo HIV IBBS surveys were funded by the Global Fund to Fight HIV/AIDS, Tuberculosis, and Malaria (GFATM)

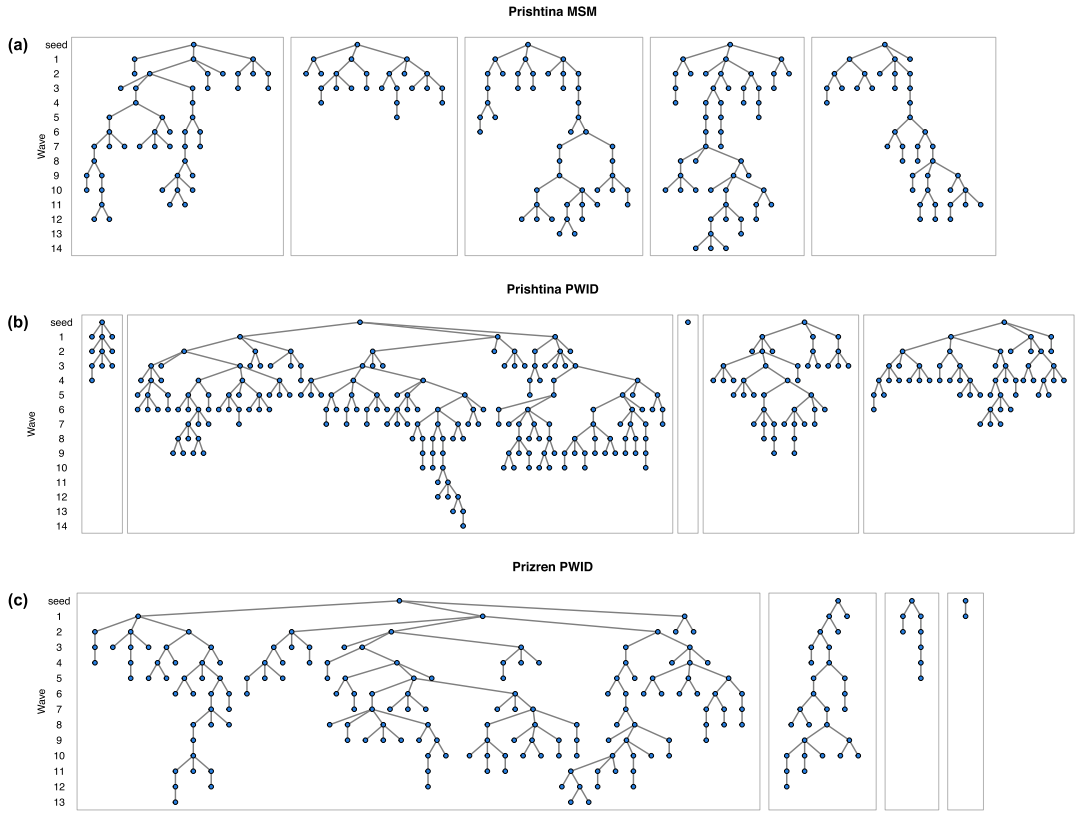

FIG. 1. *Recruitment chains for three populations in Kosovo sampled using RDS in* 2014: (*a*) *Prishtina MSM*, (*b*) *Prishtina PWID*, *and* (*c*) *Prizren PWID*.

through the Community Development Fund (CDF), Prishtina, Kosovo, and implemented by the National Institute of Public Health (NIPH), Prishtina, Kosovo. The recruitment chains for the three surveys are shown in Figure 1.

The Prishtina MSM survey began with five seeds and reached a maximum of 14 waves with a sample size of 217. The Prishtina PWID survey began with five seeds, one of whom did not recruit anyone, and reached a maximum of 14 waves with a sample size of 300. The Prizren PWID survey began with four seeds and reached a maximum of 13 waves with a sample size of 199. Histograms of the self-reported network size are shown in Figure 2, with values larger than 50 omitted but printed in text on each panel. Network size distributions are generally right-tailed, although it is common to observe heaped values, as can be seen in all three histograms. For Prishtina MSM, 64 participants (29.5%) reported a network size that was a multiple of 10. Similarly, that number is 103 (34.3%) for Prishtina PWID, and 70 (35.2%) for Prizren PWID. Although there were no missing values for network size in these data, that is also a possibility. Further, we can identify a minimum network size for each person based on the observed recruitment chains. For example, a seed who recruited two peers into the study should have a network size of at least 2. Similarly, someone in the third wave who recruited two peers into the study should have a network size of at least 3, since they know the person who recruited them as well as their two enrollees. In the Kosovo datasets, 15 participants (6.9%) among Prishtina MSM reported a network size smaller than their minimum, as did 20 (6.7%) among Prishtina PWID and 9 (4.5%) among Prizren PWID. In addition to these noted instances of possibly misreported network size, there may be other cases that are more difficult to detect.

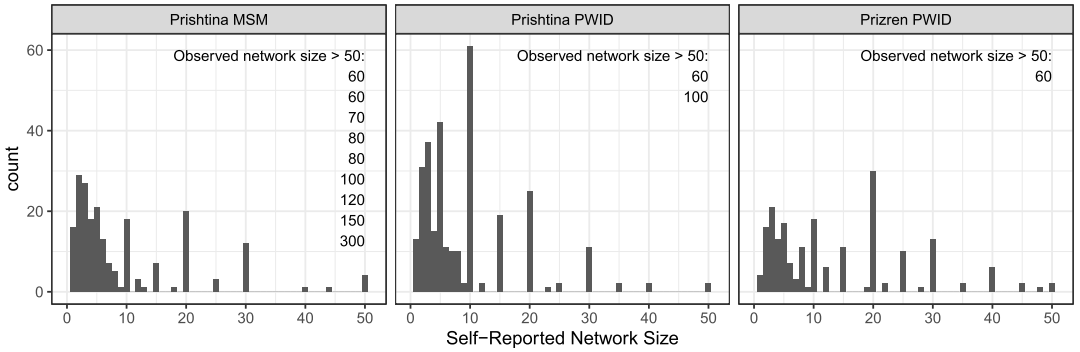

FIG. 2. *Histograms of self-reported network size for three populations in Kosovo. Network sizes larger than* 50 *are printed on each panel.*

2.3. *Visibility.* To address these issues with self-reported network size, we could attempt to infer d_i rather than assuming $d_i = \tilde{d}_i$. However, d_i is not the final quantity of interest because we believe the second assumption, $d_i \propto u_i$, might also be violated in many RDS studies. There are likely other factors in addition to degree that affect visibility. For example, a person may be well-connected in the network (high d_i) but very unlikely to enroll in the study due to constraints like geographic barriers or scheduling concerns. What we want to model, then, is not each participant's degree but their *visibility* in the sample. A person's visibility could be larger or smaller than their degree. Visibility is conceptually similar to inclusion probability, but we avoid using that term here since the design-based inclusion probabilities cannot be directly computed in RDS. A person's visibility does not refer to their own knowledge of their network.

In a traditional simple random sample, each person has the same visibility, and thus the prevalence estimate $\widehat{P}_A$ is just n_A/n because we would not need to weight the values at all. In the more complex situation of RDS, which has an unknown sampling mechanism, each individual's response needs to be weighted differently. Intuitively, people with large visibility (i.e., those who are included in the sample with relatively high probability) should have their responses down-weighted, while those with low visibility (i.e., people who are not very likely to be included in the sample) should have their responses up-weighted. If visibility is uncorrelated with the outcome of interest, this weighting will have no effect on the prevalence estimate for sufficiently large sample sizes. However, in RDS studies we may believe that the outcome measure (such as HIV status) is correlated with visibility.

Further, population size estimation is a key outcome of interest from many RDS studies and the focus of this paper. The commonly used SS-PSE method is susceptible to extreme network sizes and atypical observed network size distributions, which can cause poor estimates or lack of credible interval convergence (Johnston et al. (2015)). Intuitively, SS-PSE assumes that individuals with larger network size are more likely to be recruited earlier in the RDS process. Thus, if there are fewer large degrees in later waves, this reflects the depletion of the members of the population with larger network size (i.e., the sample represents a substantial portion of the population). SS-PSE has a Bayesian framework that allows for the incorporation of information about prior knowledge of the population size, which is often available via expert knowledge or population size estimates from previous studies or other sources, such as enumeration through mapping, network scale-up, multipliers, or capture-recapture (Johnston et al. (2015), Johnston et al. (2017)). However, SS-PSE also requires several tuning parameters, including a cap K on personal network size and the maximum population size $\max N$, which can be difficult to set in practice. The visibility extension of SS-PSE proposed here is designed to enhance fit in cases where SS-PSE alone may provide a poor estimate and obviate careful tuning for the choice of K.

In addition to providing improved estimates in cases where participants provided plausible, but possibly biased, network sizes, the visibility framework also provides a mechanism to deal with missing or erroneous cases. If a person's network size is missing, we can infer their visibility based on the observed values in the sample. Similarly, if a person reported an impossible network size value (for example 0 when they were recruited into the study by someone, or 1 when they recruited two people, or an impossibly large value), we can still model their visibility. Modeling and inference for the visibility distribution provide a framework to handle missing and impossible values of self-reported network size.

We do not observe visibility directly and thus infer it from information routinely collected during RDS studies. The model for the visibility distribution is presented in Section 3, and Bayesian inference within the SS-PSE framework is detailed in Section 4.

3. Modeling the visibility distribution.

3.1. *Motivation based on the error process.* Visibility is not directly observable, so we model it as a latent variable with a value for each individual. To do this, we use three pieces of information collected during the RDS study to make inferences about the visibility distribution for each individual: the self-reported degree ($\tilde{d}_i$), the number of recruits a person successfully enrolls (r_i), the time a person had to recruit (t_i), that is, the time interval from the time the person themselves were handed coupons (typically coincident with when they themselves were recruited) and the end of the study or the coupon expiration date (after which recruits are not accepted), and the number of coupons they had to distribute (m_i, typically constant). We assume that the self-reported degree $\tilde{d}_i$ and number of recruits r_i are conditionally independent for person i, given their visibility u_i. Then we can write the joint distribution of $\tilde{d}$, r, and u as

$$
\begin{aligned}
p(\tilde{d}, r, u|t, m) &= p(\tilde{d}, r|u, t, m)\, p(u) \\
&= p(\tilde{d}|u)\, p(r|u, t, m)\, p(u).
\end{aligned} \tag{3}
$$

In this framework degree depends only on visibility, but the number of recruits depends on both visibility and the time a person had to recruit from their enrollment date. We now describe a model for each of these pieces.

3.2. *Measurement error model for self-reported degree $\tilde{d}_i$.* We would like the model for $\tilde{d}_i|u_i$ to capture proportional inflation of the self-reported degree relative to the visibility and to allow for relative error of the self-reported degrees about this inflated value. This allows us to view the self-reported degrees as the visibility with some error (e.g., heaping, exaggeration), where the mean level of the error depends on the magnitude of the self-report.

For $i = 1, \ldots, n$, let

$$
\tilde{d}_i|u_i \sim \mathrm{EWP}(o_{\mathrm{mem}} u_i; \nu_{\mathrm{mem}}), \tag{4}
$$

where EWP is the exponentially weighted Poisson distribution of Ridout and Besbeas (2004) with location parameter $o_{\mathrm{mem}} u_i$ and concentration parameter ν_{mem}. The exponentially weighted Poisson distribution is based on the Poisson with an extra concentration parameter to model over- or underdispersion, which we expect in visibilities. Explicitly, let X be distributed as Poisson with mean μ with probability mass function (PMF), denoted $\mathrm{Poisson}(d; \mu)$. Suppose that when a person self-reports $X = d$ that we have probability of $w(d; \mu)$ of recording it. Then the PMF of the resultant random variable is proportional to $w(d; \mu)\,\mathrm{Poisson}(d; \mu)$. Each choice of $w(d; \mu)$ produces a different weighted Poisson distribution (Kokonendji (2014)). Here we choose $w(d; \mu) = \exp(-\nu_{\mathrm{mem}}|d - \mu|)$ to focus the

self-reported degree around the (scaled) visibility $o_{\text{mem}}u$. Explicitly, the PMF of a random variable with distribution $\text{EWP}(d; \nu_{\text{mem}})$ is

$$p(d; u, o_{\text{mem}}, \nu_{\text{mem}}) \propto \frac{u^d e^{-o_{\text{mem}}u - \nu_{\text{mem}}|d - o_{\text{mem}}u|}}{d!}, \quad d = 0, 1, \ldots; o_{\text{mem}} > 0, \nu_{\text{mem}} \in \mathbb{R}. \tag{5}$$

The location parameter is $o_{\text{mem}}u$ with the o_{mem} factor allowing for proportional inflation or deflation of the self-reported degree relative to the visibility. For ease of interpretation of the visibilities and o_{mem}, the expected value of the visibilities is set to 8. While this standardization value is arbitrary, this choice facilitates plotting and maps reasonably to many empirical distributions of degrees. The concentration parameter ν_{mem} allows for relative error around the visibility. For example, someone who reports a network size of 5 may have rounded that number from 4 or 6, but likely not from 27. However, a person who reports a network size of 200 may, in fact, have degree 227. A value of $\nu_{\text{mem}} = 0$ indicates that the self-reported degrees are Poisson distributed with mean equal to their inflated visibility, while $\nu_{\text{mem}} \to \infty$ indicates that there is no individual reporting error. In general, we expect the variation of the self-reported degrees about their visibilities to be underdispersed relative to the Poisson. However, this model allows flexibility in the specification including overdispersion (via $\nu_{\text{mem}} < 0$). Finally, we sometimes find it easier to refer to $1/\nu_{\text{mem}}$ as the scale of the measurement error model, as it is in the same units as the visibilities.

A value of $o_{\text{mem}} = 1$ indicates that the self-reported degrees are not inflated relative to the visibility, while $o_{\text{mem}} > 1$ indicates that the self-reported degrees are inflated relative to the visibility. Again, the parameterization allows both inflation and contraction relative to the visibilities.

3.3. *Model for number of recruits r_i.* The number of direct recruits a person is able to enroll, r_i, is an integer between 0 and m_i, where m_i is the maximum number of people respondent i was allowed to recruit (the number of coupons they had to distribute). We assume r_i is positively associated with person i's visibility u_i in the sense that if a person is highly visible they are more likely to recruit more people (up to the maximum). We assume that the recruitments are independent. Specifically, we model the number of recruits as a truncated Poisson regression. For $i = 1, \ldots, n$, we have

$$r_i | u_i, t_i, m_i = \min\bigl(m_i, \text{Poisson}(\lambda_i)\bigr), \tag{6}$$

where we additionally model λ_i based on the person's visibility and the time they had to recruit via

$$\log(\lambda_i | u_i, t_i) = \beta_0 + \beta_u \log u_i + \beta_t \log t_i. \tag{7}$$

Intuitively, we get a value for the number of recruits a person should have been able to enroll in the study, based on their visibility and time to recruit, and capped by the number of coupons. If the number they actually recruited is less than this, their self-reported network size may overestimate their visibility. We expect β_u to be nonnegative, as recruitment should be positively associated with a person's visibility. The model for the recruitment probability adjusts for the time the individual has to recruit. If t_i is close to the end of the study, then the individual will likely recruit less (i.e., the parameter β_t is likely to be nonnegative).

The assumption that each individual in the personal network is equally likely to be recruited is standard in the analysis of RDS data but could be relaxed here.

4. Bayesian inference for respondent driven sampling. The self-reported degrees are a result of the RDS process over a finite population of individuals. It is important to take into

account both the information in that sampling and the additional uncertainty about the sampling and underlying population being sampled in inference for the visibilities. To address this, we introduce a joint model for the sampling process and finite population. This model extends that in Handcock, Gile and Mar (2014) and Gile (2011). We use a Bayesian framework, treating the visibilities and population size N as unknown parameters. This requires a probability model for the observed data, given N, as well as a prior for N. The respondent driven sampling model is nonamenable 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.

4.1. *Jointly modeling the unit size distribution and the sampling process.* The population visibilities 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 visibilities 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 PMF with parameter η and support $1, \ldots,$. In this paper we consider a Conway–Maxwell–Poisson distribution for the visibilities, although other choices, such as negative binomial and Poisson-log-normal, are possible. The Conway–Maxwell–Poisson distribution (Conway and Maxwell (1962)) is motivated as a Poisson with an extra parameter to model underdispersion and overdispersion, which we expect in visibilities.

Let the ordered observed sample be denoted by the random vector $G = (G_1, \ldots, G_n)$, with realized values $g = (g_1, \ldots, g_n)$, and let $\backslash g = \{1, \ldots, N\} \backslash \{g_1, \ldots, g_n\}$ represent the set of indices of the unobserved population units. In addition, consider the (unobserved) visibilities. Let $U_{\text{obs}} = \{U_{g_1}, U_{g_2}, \ldots, U_{g_n}\}$ be the random vector of visibilities for the sampled individuals, with values $u_{\text{obs}} = (u_{g_1}, \ldots, u_{g_n})$. Similarly, let $U_{\text{unobs}} = \{U_i\}_{i \in \backslash g}$ and $u_{\text{unobs}} = \{u_i\}_{i \in \backslash g}$ represent the random and realized values of the visibilities of the unobserved units.

The respondent-driven sampling process is approximated by *successive sampling* or *probability proportional to size without replacement* sampling (Raj (1956), Murthy (1957), Andreatta and Kaufman (1986), Nair and Wang (1989), Bickel, Nair and Wang (1992)), a sampling design in which units are sampled without replacement with unequal probabilities, such that each successive sample is drawn with probability proportional to visibility 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 $G = (G_1, \ldots, G_n)$ is the ordered observed sample and $U = \{U_1, \ldots, U_N\}$ is the population of visibilities with realized values g and u, respectively. So u_{g_i} is the realized visibility of the ith unit in the sample.

The self-reported network sizes are represented by the random vector $D_{\text{obs}} = \{D_{g_1}, D_{g_2}, \ldots, D_{g_n}\}$ with realized values $d_{\text{obs}} = (d_{g_1}, \ldots, d_{g_n})$. Thus, the full observed data are $\{g, d_{\text{obs}}\}$. For simplicity of notation, denote the observed data by $D = (D_{\text{obs}} = d_{\text{obs}}, G = g, r, t)$, and let $\theta = (\eta, \beta_0, \beta_u, \beta_t, o_{\text{mem}}, \nu_{\text{mem}})$.

The primary focus of interest is the posterior distribution

$$p(N, U, \theta|D).$$

To draw on this distribution, we use a five component Gibbs sampler. First, consider the conditional

$$p(N, U_{\text{unobs}} = u_{\text{unobs}}, \theta|U_{\text{obs}} = u_{\text{obs}}, D), \tag{8}$$

which can be drawn from using the four component Gibbs sampler in Handcock, Gile and Mar (2014, Appendix B). Then

$$
\begin{aligned}
&p(U_{\text{obs}} = u_{\text{obs}} | N, U_{\text{unobs}} = u_{\text{unobs}}, \theta, D) \\
&\quad \propto p(D_{\text{obs}} | U_{\text{obs}} = u_{\text{obs}}, \theta, D) p(r | U_{\text{obs}} = u_{\text{obs}}, \theta) p(U_{\text{obs}} = u_{\text{obs}} | \eta) \\
&\quad = \prod_{j=1}^{n} \left[p(d_{g_j} | u_{g_j}, \theta) p(r_{g_j} | u_{g_j}, t_{g_j}, \theta) f(u_{g_j} | \eta) \right],
\end{aligned} \tag{9}
$$

where the terms are given by equations (4), (6) and (7). This can be drawn on directly as the right-hand side is computable.

The full conditional of $(\beta_0, \beta_u, \beta_t)$ is

$$
p(\beta_0, \beta_u, \beta_t | U = u, D) \propto p(\beta_0, \beta_u, \beta_t) \cdot \prod_{j=1}^{n} p(r_j | u_j, \beta_0, \beta_u, \beta_t).
$$

This can be sampled with a Metropolis–Hastings algorithm using equations (6) and (7). The full conditional of $(o_{\text{mem}}, \nu_{\text{mem}})$ is

$$
p(o_{\text{mem}}, \nu_{\text{mem}} | U = u, D) \propto p(o_{\text{mem}}, \nu_{\text{mem}}) \cdot \prod_{j=1}^{n} p(d_j | u_j, o_{\text{mem}}, \nu_{\text{mem}}).
$$

This can be sampled with a Metropolis–Hastings algorithm using equation (4). The full conditional of η is given by Step 2 of the algorithm in Handcock, Gile and Mar (2014, Appendix B).

4.2. *Specifying prior knowledge about the population size and visibility distribution.* The model allows for an arbitrary prior distribution over the population size (N). Following Handcock, Gile and Mar (2014), we use the class of priors specifying knowledge about the sample proportion (i.e., n/N) as a Beta(ϕ_1, ϕ_2) distribution. The implied density function on N (considered as a continuous variable) is

$$
\pi(N) = \phi_2 n (N - n)^{\phi_2 - 1} / N^{\phi_1 + \phi_2} \quad \text{for } N > n. \tag{10}
$$

The distribution has tail behavior $O(1/N^{\phi_1 + 1})$. 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.

For simplicity, in this paper we specify that N and θ are a priori independent so that $\pi(N, \theta) = \pi(N) \cdot \pi(\theta)$. The Conway–Maxwell–Poisson distribution for visibilities can be parameterized in terms of its mean and standard deviation, η, and this can aid elicitation of prior information about them. Further, we assume that $\pi(\theta) = \pi(\eta) \cdot \pi(\beta_0, \beta_u, \beta_t) \cdot \pi(o_{\text{mem}}, \nu_{\text{mem}})$.

In this paper the prior for the mean visibility, given the standard deviation is normal and the variance is scaled inverse Chi-squared, is

$$
\mu | \sigma \sim N\left(\mu_0, \sigma^2 / \text{df}_{\text{mean}}\right), \quad \sigma \sim \text{Inv}\chi^2(\sigma_0; \text{df}_{\text{sigma}}).
$$

The default prior on these parameters is diffuse with an equivalent sample size of $\text{df}_{\text{mean}} = 1$ for the mean of the visibility distribution and $\text{df}_{\text{sigma}} = 5$ for the variance of the visibility distribution. In this notation $\eta = (\mu, \sigma)$ and μ_0, df_{mean}, σ_0, df_{sigma} are hyperparameters.

In this paper we choose the the prior for $(\beta_0, \beta_u, \beta_t)$ to be multivariate Gaussian with mean β_{mem} and covariance matrix Σ_{mem}: $p(\beta_0, \beta_u, \beta_t) \sim N(\beta_{\text{mem}}, \Sigma_{\text{mem}})$. The prior for

$(o_{\text{mem}}, \nu_{\text{mem}})$ in the measurement error model for the self-reported degree, given in equation (4), has o_{mem} given ν_{mem} Gaussian and ν_{mem} is scaled inverse Chi-squared,

$$\log o_{\text{mem}} | \nu_{\text{mem}} \sim N(\log \mu_{\text{mem opt}}, \nu_{\text{mem}}/\text{df}_{\text{mem opt}}),$$
$$\nu_{\text{mem}} \sim \text{Inv}\chi^2(\sigma^2_{\text{mem scale}}; \text{df}_{\text{mem scale}}).$$

The default prior on these parameters is diffuse with an equivalent sample size of $\text{df}_{\text{mem opt}} = 1$ for the location of the optimism distribution and $\text{df}_{\text{mem scale}} = 5$ for the scale of the optimism distribution. In this notation the full set of hyperparameters is ϕ_1, ϕ_2, μ_0, df_{mean}, σ_0, df_{sigma}, $\mu_{\text{mem opt}}$, $\text{df}_{\text{mem opt}}$, $\sigma_{\text{mem scale}}$, $\text{df}_{\text{mem scale}}$, β_{mem}, and Σ_{mem}.

4.3. *Computation*. As noted in Section 4.1, the joint posterior $p(N, \theta, U|D)$ can be sampled from using a five component Gibbs sampler. While complex, these computations are manageable.

This distribution can then be marginalized to produce samples from $p(N|D)$, $p(\theta|D)$, and the posterior predictive distribution of the latent visibilities, $p(U|D)$. Hence, it produces posterior predictive distributions of the full population visibilities $(u_i, i = 1, \ldots, N)$. These posteriors enable inference for such quantities as the population size, the visibility distribution, etc.

The computational properties of the sampler can be investigated via standard Markov chain Monte Carlo (MCMC) diagnostic procedures to assess convergence and mixing (Plummer et al. (2006)).

The entire computational procedure has been implemented in `C` and `R` via the `sspse` package (Handcock et al. (2022)), which is available on CRAN (R Core Team (2020)). The function `posteriorsize()` allows estimation of population size with visibilities using the `visibility=TRUE` argument. Supplementary Material B (McLaughlin et al. (2024)) provides a vignette covering implementation and diagnostics for nonvisibility and visibility SS-PSE using the `sspse` package.

4.4. *Checking goodness-of-fit*. The model presented here has many components, and it is important to check the goodness-of-fit of the model to the data. A natural way to do this in our Bayesian framework is by considering the posterior predictive distributions of statistics. Specifically, we can calculate the tail-area probability corresponding to the observed values of those statistics as measures of goodness-of-fit, that is, extreme tail-area probabilities suggest the model is a poor fit to the data. These tail-area probabilities are analogous to p-values in a Frequentist framework, and they are referred to as posterior predictive p-values (Rubin (1984), Meng (1994)).

The core statistics here are the individual self-reported network sizes, $\tilde{d}_i$. We can compute the posterior predictive distribution of each individual's self-reported network size via

$$p(\tilde{D}_i|D) = \int p(\tilde{D}_i|N, \theta, U)\, p(N, \theta, U|D)\, dN\, d\theta\, dU, \quad i = 1, \ldots, n,$$

where $\tilde{D}_i$ is the self-reported degree size we could have observed under an identical RDS survey with the same data generating process (model, parameter values, population size, and visibilities) as the actual RDS survey. We write sums over the discrete elements as integrals for simplicity. The first term in the integral is given by equation (4), and the second is the posterior for the parameters and unknowns computed above. The posterior predictive p-values are then

$$p_i = p(\tilde{D}_i \geq \tilde{d}_i), \quad i = 1, \ldots, n.$$

These values can be considered as a measure of discrepancy between the observed data and the posited model. The distribution of the p_i need not be uniform on $[0, 1]$. If the model is correctly specified, the $\{p_i\}_{i=1}^n$ will be more concentrated near 0.5 than near 0 or 1. In particular, the over abundance of values close to 0 or 1 (relative to a uniform) is evidence for poor model fit and hence that the model's predictions are inconsistent with the observed data.

We will apply this goodness-of-fit check to the case study in Section 5.

4.5. *Comparison with other SS-PSE models*. This model extends the original SS-PSE work in Handcock, Gile and Mar (2014) and Handcock, Gile and Mar (2015) by adding the visibility modeling component. The prior work presumes the unit sizes are the self-reported network sizes (i.e., $u_i = d_i$). We treat the individual unit sizes as latent variables modeled from the self-reported degrees and other information through a novel measurement error model. This generalizes the model for the sampling mechanism, adding an additional layer to the model and requiring an additional component of the Gibbs sampler.

McLaughlin et al. (2019) considered a developmental version of SS-PSE that was similar conceptually to the model presented here but with some key differences. Here the measurement error model for self-reported degree $\tilde{d}_i$ (see Section 3.2) is an exponentially weighted Poisson (EWP) where McLaughlin et al. (2019) uses a Conway–Maxwell–Poisson (CMP). The authors found that although the CMP obviated the need to tune the tricky K parameter in the original SS-PSE model, it introduced a new overdispersion parameter that was also difficult to set. Thus, the EWP measurement model presented here represents an improvement based on the authors' experience working with a variety of real RDS data sets. In addition, McLaughlin et al. (2019) focuses on application and diagnostics; the statistical, mathematical, and computational aspects of the visibility SS-PSE model are presented here for the first time.

Finally, Kim and Handcock (2019) present a capture-recapture extension of SS-PSE that allows population size to be estimated from two RDS data sources. That approach utilizes the original (nonvisibility) version of SS-PSE. An extension of capture-recapture SS-PSE that utilizes the visibility model presented here has been developed and is implemented in the `sspse` package (Handcock et al. (2022)); however, it has not yet been extensively tested on real data. The capture-recapture application of the visibility framework is left for future work.

4.6. *Uses of the visibility model*. This paper focuses on the use of visibility for SS-PSE, where the visibilities are redrawn at each step of the MCMC algorithm. Conceptually, the visibility SS-PSE approach differs from the original SS-PSE model by redrawing these values for both the observed and unobserved units. In the original SS-PSE framework, the observed degrees are treated as known and not updated, while only the unobserved degrees are redrawn. In the visibility SS-PSE model, observed degrees are also redrawn using the measurement error model.

It may also be of interest to extend the work by inferring a single value for the visibility of individual i from the full posterior predictive distribution. Draws from this distribution are provided by the sampler Section 4.1.

There are several methods possible to obtain a single value for visibility: (1) make a random draw from the joint posterior predictive distribution of the latent visibilities, $p(U|D)$, or (2) use the median, mean, or mode of this distribution. These visibilities could be used in place of self-reported degree in a plug-in estimator for prevalence estimation, for example, using the Volz-Heckathorn or successive sampling estimators described in Section 1. This would be the most basic approach to extend use of the visibility model beyond population size estimation and into prevalence estimation, given the current predominantly design-based estimators. A related approach would be to consider the model presented in Section 4.1 but with

a fixed population size. If N is known, the posterior distribution of interest is $p(U, \theta | D, N)$, which can be sampled using a four component Gibbs sampler to get a posterior distribution for the visibilities U. A further extension could explore how to jointly model visibility with prevalence, for example, using Bayesian empirical likelihood.

While prevalence estimation with visibility is a very important contribution of the methodology, exploration is left for a future work because of the variety of possible approaches as well as the challenge of comparing prevalence estimates on real data without a gold standard.

5. Application to three populations in Kosovo. We assess the performance of incorporating modeled visibility for population size estimation for three key populations in Kosovo. The data were briefly introduced in Section 2.2, and the recruitment chains for the three surveys are shown in Figure 1. For each population we elicited information from two expert sources—the National Institute of Public Health (NIPH) of Kosovo and a nongovernmental organization (NGO)—about likely values of the population size, which serve as prior information for SS-PSE. We specify the median for the prior distribution for N as the average of the two medians provided by the experts. These are $N = 2700$ for Prishtina MSM, $N = 3375$ for Prishtina PWID, and $N = 700$ for Prizren PWID; see Supplementary Material A (McLaughlin et al. (2024)) for a sensitivity analysis performed using the individual values provided by experts as the prior median.

For each population we compared the nonvisibility (NV) SS-PSE estimates (Handcock, Gile and Mar (2014), Handcock, Gile and Mar (2015)) to the visibility (Vis) SS-PSE estimates developed in this paper using the same median prior value. Table 1 shows the posterior median, mean, and 90% credible interval for each of these six models. In each of the three populations, the visibility posterior median is closer to the prior median than the nonvisibility posterior median. However, both posterior medians for Prishtina MSM and Prishtina PWID are quite a bit smaller than the prior median. In the case of the nonvisibility SS-PSE fit for Prishtina MSM, the 90% credible interval does not contain the prior median. The visibility fit for Prishtina MSM does not have this issue.

The prior, nonvisibility posterior, and visibility posterior distributions are shown for each population in Figure 3. The 90% credible interval for each posterior distribution is shaded. Horizontal reference lines are given at the top of the panel for each of the two expert sources (NIPH and NGO), listing their best guess for the reasonably smallest, median, and reasonably largest values of the corresponding population size. Note that the average of the two expert medians was used as the prior median, but the reasonably smallest and reasonably largest values were not used in model fitting.

TABLE 1

Population size estimates using nonvisibility (NV) and visibility (Vis) SS-PSE based on a prior median for three populations in Kosovo

Survey	Method	Prior	Posterior			
		Median	Median	Mean	5%	95%
Prishtina MSM	NV	2700	529	497	363	1112
	Vis		1050	1598	550	4421
Prishtina PWID	NV	3375	1158	1528	612	3809
	Vis		1522	2261	742	6148
Prizren PWID	NV	700	828	1025	414	2426
	Vis		772	962	457	2202

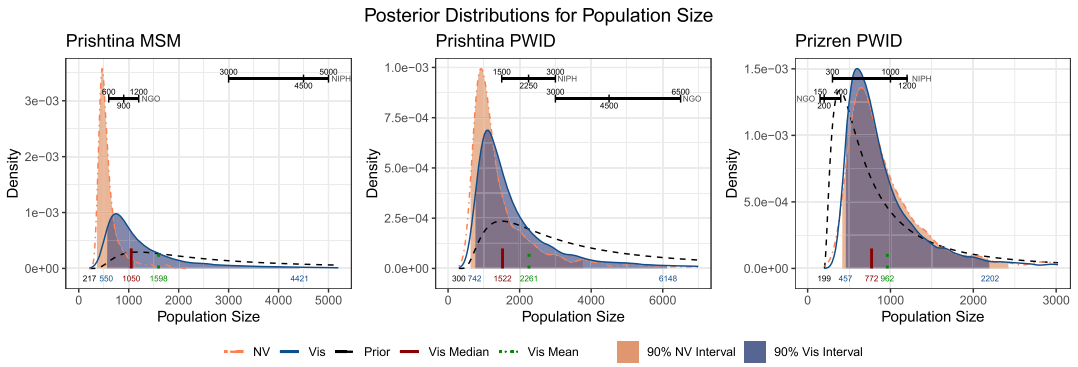

FIG. 3. *Posterior distributions for population size. The black dashed curve shows the common prior distribution, the orange dash-dot curve shows the posterior distribution from nonvisibility (NV) SS-PSE, and the blue solid curve shows the posterior distribution from visibility (Vis) SS-PSE. The* 90% *credible intervals for each posterior distribution are shaded. The two horizontal black line segments at the top of each panel provide the reasonably smallest, median, and reasonably largest expert guesses for population size from two sources, an NGO and the NIPH.*

For Prishtina MSM the nonvisibility fit places most of the mass of the posterior distribution for N at very small values, with a posterior median of 529 that is smaller than either of the reasonably smallest values provided by experts (600 and 3000). Fits like this are considered poor and were one of the motivating factors to develop the visibility SS-PSE method. Although the visibility fit also has a posterior median smaller than than the prior, the distribution is wider and overlaps more with the plausible population size values provided by the NGO. The plausible values, provided by the NIPH, are quite a bit bigger than those given by the NGO, demonstrating the overall uncertainty of population size estimation among hidden populations, even among experts.

The Prishtina PWID fits are more similar, although again the nonvisibility posterior distribution places more mass on slightly smaller values of population size. There is again a wide range of plausible values for population size provided by experts, although the posterior medians for both the nonvisibility (1158) and visibility (1522) fits are smaller than either of the reasonably smallest values provided by experts (1500 and 3000).

Finally, the nonvisibility and visibility SS-PSE fits for Prizren PWID are very similar and overlap well with the NIPH plausible values. The NGO reasonably largest value (400) is less than the lower bound of the credible interval for both fits, but in this case the expert values might be mistaken, since the sample size (199) was already about as big as the guess for the median. In this instance, unlike the two Prishtina populations, the population size estimate is larger than the prior used to fit the model.

Next, to illustrate the differences between nonvisibility and visibility SS-PSE, we examine the relationship between the augmented self-reported network size and the visibility for each person in each survey. The augmented self-reported network size for person i is defined as the maximum of their self-reported network size $\tilde{d}_i$ and the number of edges observed in the recruitment tree connected to person i. In other words, if a person reported a network size smaller than the number of peers they recruited plus 1 (unless they were a seed), their network size is augmented to this minimum. Augmented network size is used because nonvisibility SS-PSE recodes the network sizes this way. Figure 4 compares the augmented self-reported network size for each person with their visibility distribution. For each person the median of their visibility distribution is represented with a point, and the whiskers show the interquartile range of the visibility distribution. The posterior means of the optimism parameters, o_{mem}, for

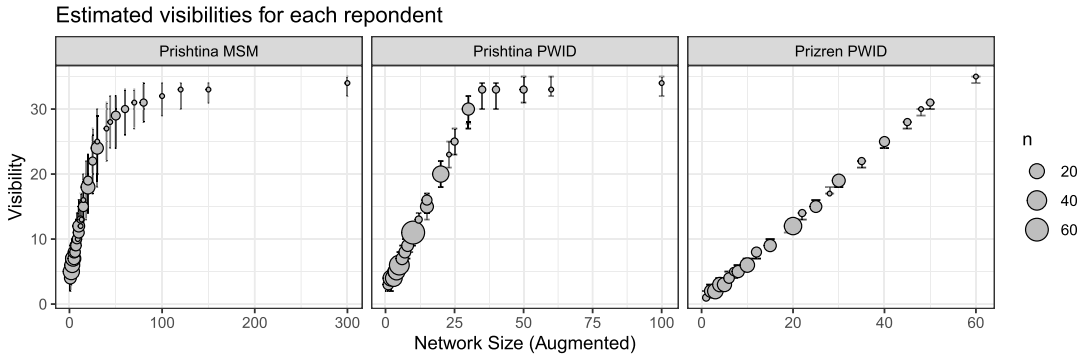

FIG. 4. *Estimated visibilities for each respondent* (*vs. augmented network size*) *for three populations in Kosovo.*

the three populations were 0.94, 0.95, 1.61, respectively, indicating typical network sizes of a little under 8 for Prishtina MSM and PWID and 13 for Prizen PWID. These are consistent with the scales of the figure.

Across all three populations, there is a strong association between network size and visibility: people with larger network sizes tend to have larger visibilities, as we would expect. The visibility distributions are capped at 35, based on the scaling chosen, so the effect of extremely large network sizes, for example, the value of 300 in the Prishtina MSM survey is lessened. The relationship between network size and visibility is more linear for Prizren PWID than for the two Prishtina populations, illustrating why the nonvisibility and visibility fits may have been more similar in this case.

Figure 5 compares the visibility and augmented network size distributions directly by using the median of the visibility distribution for each person as their value for visibility. Both distributions have longer right tails, but the visibility distribution demonstrates how the heaped self-reported network sizes have been smoothed out by the visibility SS-PSE method. There are no longer spikes at multiples of 10.

As a further diagnostic, we compute the posterior predictive p-values developed in Section 4.4. These can be used to assess the validity of the estimated visibility distribution and other model assumptions. We look at the percentile rank of the self-reported degree within the posterior predictive distribution of the self-reported degree. If the model is correct and computed correctly, the self-reported degrees should be in the center of the posterior, so the p-values are close to 0.5 (or at least not extreme). Figure 6 shows the posterior predictive

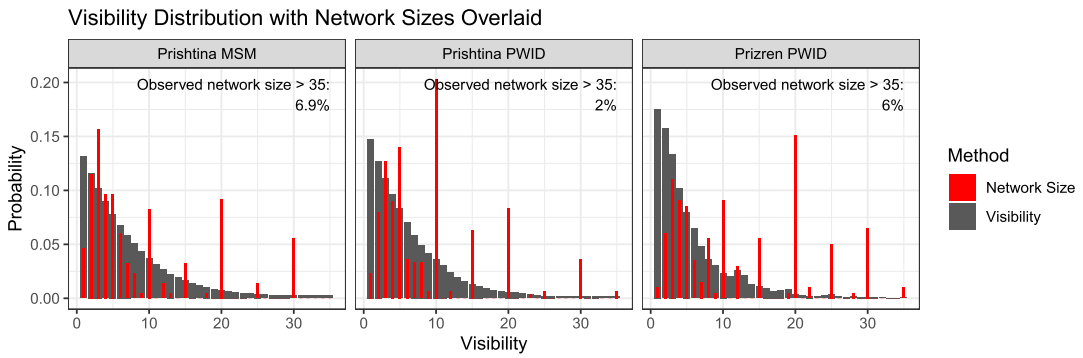

FIG. 5. *Visibility distribution with augmented network sizes overlaid for three populations in Kosovo. The percent of observed network sizes larger than* 35 (*and thus omitted from the plot*) *is displayed on each panel.*

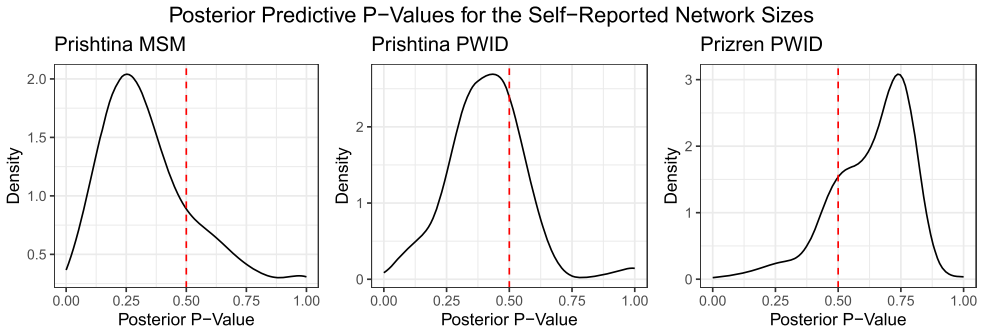

FIG. 6. *Posterior predictive p-values for the self-reported network sizes for three populations in Kosovo.*

p-values for the three populations in Kosovo, with reference lines at 0.5 overlaid. The plot shows the reference value of 0.5 occurring in the midrange of the posterior p-value distribution for all three populations, providing strong evidence supporting the model for the self-reported degrees and the modeling overall.

In addition to the posterior distribution for N and the visibilities, it is also useful to consider the other model parameters. Table 2 provides the maximum a posteriori (MAP) estimates of the other model parameters: μ and σ for nonvisibility SS-PSE and μ, σ, ν_{mem}, β_0, β_u, and β_t for visibility SS-PSE. In nonvisibility SS-PSE, μ and σ are the mean and standard deviation of the unit size distribution; in visibility SS-PSE, they are the mean and standard deviation of the visibility distribution. The posterior distributions for the model parameters for the visibility SS-PSE fits are shown in Figure 7. The estimates of the location and scale of the unit size distributions are largely comparable, especially given their uncertainties (Figure 7). All three estimates of β_t are positive, indicating that those with longer time-to-recruit tended to recruit more. All three estimates of β_u are close to 0, indicating that visibility did not have a big impact on the number of recruits. The measurement error model concentration parameter ν_{mem} is low for Prishtina MSM and higher for Prishtina PWID and Prizren PWID. This can be seen in Figure 4, where there is significant variation in the self-reported network sizes when the visibilities are above 25, while the visibilities and self-reported network sizes are highly correlated for Prishtina PWID and Prizren PWID. This suggests that the self-reported network sizes are much more reliable measures of visibility in Prishtina PWID and Prizren PWID, compared to Prishtina MSM. All MCMC diagnostics plots confirmed convergence and stability (not shown but available).

TABLE 2

Maximum a posteriori (MAP) estimates of model parameters for three populations in Kosovo. In nonvisibility (NV) SS-PSE, μ and σ are the mean and standard deviation of the unit size distribution; in visibility (Vis) SS-PSE they are the mean and standard deviation of the visibility distribution

Survey	Method	μ	σ	ν_{mem}	β_0	β_t	β_u	median days-to-recruit
Prishtina MSM	NV	7.19	6.57					
	Vis	7.22	6.25	0.048	−0.938	0.005	0.221	44
Prishtina PWID	NV	5.67	4.92					
	Vis	6.37	5.55	0.346	−0.846	0.019	0.056	27
Prizren PWID	NV	8.13	7.17					
	Vis	5.49	4.85	0.647	−0.554	0.004	0.068	27

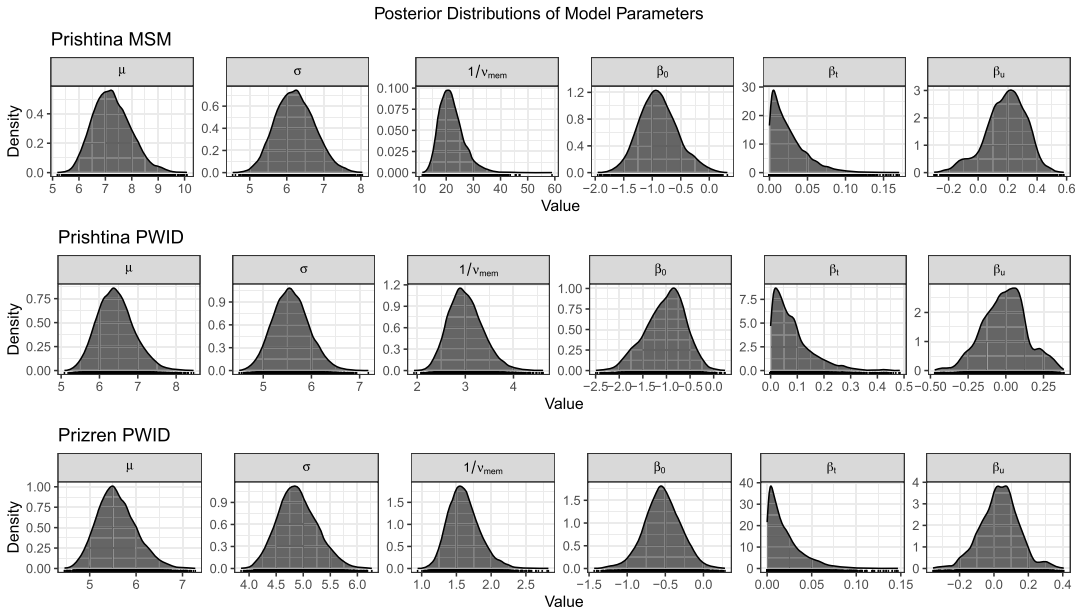

FIG. 7. *Posterior distributions for model parameters for each of three populations in Kosovo. Note that* $1/\nu_{\text{mem}}$ *is the scale of the measurement error model.*

6. Discussion. During the development of the visibility SS-PSE method, it was implemented in several contexts beyond the ones presented here and provided improved population size estimates. In Armenia, visibility SS-PSE was used to estimate population size for 15 datasets of FSW, MSM, and PWID populations and compared with nonvisibility SS-PSE as well as other common population size estimation methods (McLaughlin et al. (2019)). The visibility SS-PSE estimates performed favorably, compared to object and service multiplier methods and wisdom of the crowds, each of which commonly produced population size estimates that were either much too small or much too large. Further, the nonvisibility SS-PSE models would not converge in six of the 15 cases and produced poor estimates in others. A visibility SS-PSE estimate was possible in all cases, and estimates often compared favorably to other methods when judged against "true" population sizes provided by experts. A sensitivity analysis was also performed to assess the impact of the prior median for N, and it was found that changing the prior median to other expert values did not drastically change the population size estimate given the overall variability of the posterior distribution (McLaughlin et al. (2019)). Visibility SS-PSE was also used on four MSM populations in Europe to estimate the number of MSM living with HIV and produced reasonable estimates (Johnston et al. (2021)).

Despite those favorable results and the ones presented in this paper, it is important to note the limitations of visibility SS-PSE. The main strength of visibility SS-PSE is adjusting for the effects of extremely large and heaped network sizes, which can often result in poor fits using the nonvisibility SS-PSE method (Johnston et al. (2015)). Nevertheless, both types of SS-PSE rely on the successive sampling model for RDS, and if there were unanticipated issues with RDS data collection, SS-PSE will not address these. For example, bottlenecks in the underlying population network, disconnected parts of the population, or situations with few waves and high seed dependence may result in biased SS-PSE estimates. If RDS assumptions are violated or there are issues with convergence in the model, results from either type of SS-PSE should be interpreted with caution. Some guidance for goodness-of-fit is provided in McLaughlin et al. (2019).

In the case of the two Prishtina, Kosovo surveys, it is difficult to know if the smaller than expected population size estimates were due to an issue with RDS data collection or if the expert values used to assess the estimate may be inaccurate, especially since the NIPH and NGO prior medians are quite different; see Supplementary Material A (McLaughlin et al. (2024)) for further discussion of this case study comparing the nonvisibility and visibility SS-PSE estimates using the NIPH and NGO priors separately. In a separate, non-RDS study, a programmatic mapping approach was used to collect systematic information from key informants in 26 municipalities in Kosovo between February to April, 2016 (Gexha Bunjaku et al. (2019)). This study estimated the number of MSM in Prishtina to be 2613, the number of PWID in Prishtina to be 1426, and the number of PWID in Prizren to be 559. Although the time periods do not exactly overlap and there are likely differences between population members reachable via RDS and programmatic mapping as well as biases inherent in each method, these population size estimates provide additional benchmarks. The nonvisibility Prishtina MSM credible interval does not contain the programmatic mapping estimate, but the other five credible intervals do contain their respective estimates (Table 1). The Prishtina PWID programmatic mapping estimate (1426) is smaller than either of the medians provided by experts (2250 and 2500) and is closer to the visibility SS-PSE estimate (1522). Validation of population size estimates relies on expert input from many stakeholders, including governmental and nongovernmental organizations working with the population as well as people directly involved with sampling. Other population size estimates, such as those presented here using programmatic mapping, can also be useful to validate results.

7. Conclusion. Common RDS prevalence and population size estimators incorporate self-reported degree to account for the unknown sampling mechanism. However, using self-reported degree as a proxy for an individual's inclusion probability is problematic, both because of misreporting and bias on the network size variable and because other factors besides degree influence a person's likelihood to be sampled. We introduced a measurement error model for visibility as a function of self-reported degree, number of recruits enrolled, and time to recruit as well as a method for Bayesian inference within the existing SS-PSE framework. An advantage of the Bayesian framework is the prior beliefs about the ranges or signs of parameters can be incorporated into the analysis. Visibility SS-PSE serves to smooth the degree distribution, which commonly has heaped values, and bring in extremely large values as well as providing a framework to handle missing and impossible values of self-reported network size. We demonstrated the visibility SS-PSE methodology on RDS studies of MSM and PWID in Prishtina, Kosovo and PWID in Prizren, Kosovo, and compared the results with nonvisibility SS-PSE. When observed degrees exhibit possible characteristics of misreporting, visibility SS-PSE can provide reasonable population size estimates in instances where nonvisibility SS-PSE estimates are unreasonably small or large or cannot be obtained at all. The visibility framework also has promising extensions for prevalence estimation, and the methodology is broadly applicable in any situation where RDS is used.

Acknowledgments. The authors would like to thank the National Institute for Public Health of Kosovo for their work designing, implementing, and lending their expertise to the RDS study used as examples in Section 5. The authors would also like to thank members of the Hard-to-Reach Populations Methods Research Group (HPMRG) and the RDS Analyst Users Group for helpful comments on the implementation of the method.

Funding. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144087.

SUPPLEMENTARY MATERIAL

Supplement A: Kosovo population size estimates with different prior medians (DOI: 10.1214/23-AOAS1807SUPPA; .pdf). Extended case study considering nonvisibility and visibility SS-PSE applied to three populations in Kosovo using prior medians provided by the National Institute of Public Health (NIPH) of Kosovo, a nongovernmental organization (NGO), and the average of the two. Sensitivity of the methods to the choice of prior is examined.

Supplement B: Visibility SS-PSE vignette (DOI: 10.1214/23-AOAS1807SUPPB; .pdf). A vignette providing an introduction to successive sampling population size estimation (SS-PSE) in R using the `sspse` package, including model fitting and diagnostics. The package details the visibility extension and assesses how the addition of measurement error impacts the population size estimates. Examples are provided using the simulated RDS data set `fauxmadrona` from the `RDS` package.

REFERENCES

ANDREATTA, G. and KAUFMAN, G. M. (1986). Estimation of finite population properties when sampling is without replacement and proportional to magnitude. *J. Amer. Statist. Assoc.* **81** 657–666. MR0860497

BELL, D. C., BELLI-MCQUEEN, B. and HAIDER, A. (2007). Partner naming and forgetting: Recall of network members. *Soc. Netw.* **29** 279–299. https://doi.org/10.1016/j.socnet.2006.12.004

BENGTSSON, L. and THORSON, A. (2010). Global HIV surveillance among MSM: Is risk behavior seriously underestimated? *AIDS* **24** 2301–2303. https://doi.org/10.1097/QAD.0b013e32833d207d

BICKEL, P. J., NAIR, V. N. and WANG, P. C. C. (1992). Nonparametric inference under biased sampling from a finite population. *Ann. Statist.* **20** 853–878. MR1165596 https://doi.org/10.1214/aos/1176348660

BREWER, D. D. (2000). Forgetting in the recall-based elicitation of personal and social networks. *Soc. Netw.* **22** 29–43. https://doi.org/10.1016/S0378-8733(99)00017-9

CONWAY, R. W. and MAXWELL, W. L. (1962). A queuing model with state dependent service rates. *Int. J. Ind. Eng.* **12** 132–136.

FELLOWS, I. E. (2019). Respondent-driven sampling and the homophily configuration graph. *Stat. Med.* **38** 131–150. MR3887272 https://doi.org/10.1002/sim.7973

FELLOWS, I. E. (2022a). Estimating population size from a privatized network sample. *J. Surv. Stat. Methodol.* **10** 1346–1369. https://doi.org/10.1093/jssam/smac010

FELLOWS, I. E. (2022b). On the robustness of respondent-driven sampling estimators to measurement error. *J. Surv. Stat. Methodol.* **10** 377–396. https://doi.org/10.1093/jssam/smab056

FENTON, K. A., JOHNSON, A. M., MCMANUS, S. and ERENS, B. (2001). Measuring sexual behaviour: Methodological challenges in survey research. *Sex. Transm. Infect.* **77** 84–92. https://doi.org/10.1136/sti.77.2.84

FISHER, R. J. (1993). Social desirability bias and the validity of indirect questioning. *J. Consum. Res.* **20** 303. https://doi.org/10.1086/209351

GEXHA BUNJAKU, D., DEVA, E., GASHI, L., KAÇANIKU-GUNGA, P., COMINS, C. A. and EMMANUEL, F. (2019). Programmatic mapping to estimate size, distribution, and dynamics of key populations in Kosovo. *JMIR Public Health Surveill.* **5** e11194. https://doi.org/10.2196/11194

GILE, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J. Amer. Statist. Assoc.* **106** 135–146. MR2816708 https://doi.org/10.1198/jasa.2011.ap09475

GILE, K. J., BEAUDRY, I. S., HANDCOCK, M. S. and OTT, M. Q. (2018). Methods for inference from respondent-driven sampling data. *Annu. Rev. Stat. Appl.* **5** 65–96. MR3774740 https://doi.org/10.1146/annurev-statistics-031017-100704

GILE, K. J. and HANDCOCK, M. S. (2010). Respondent-driven sampling: An assessment of current methodology. *Sociol. Method.* **40** 285–327. https://doi.org/10.1111/j.1467-9531.2010.01223.x

GILE, K. J., JOHNSTON, L. G. and SALGANIK, M. J. (2015). Diagnostics for respondent-driven sampling. *J. Roy. Statist. Soc. Ser. A* **178** 241–269. MR3291770 https://doi.org/10.1111/rssa.12059

GROVES, R. M., FOWLER, F. J. JR., COUPER, M. P., LEPKOWSKI, J. M., SINGER, E. and TOURANGEAU, R. (2009). *Survey Methodology*, 2nd ed. *Wiley Series in Survey Methodology*. Wiley, Hoboken, NJ. MR3241261

HANDCOCK, M. S. and GILE, K. J. (2010). Modeling networks from sampled data. *Ann. Appl. Stat.* **272** 383–426.

HANDCOCK, M. S., GILE, K. J., KIM, B. J. and MCLAUGHLIN, K. R. (2022). sspse: Estimating hidden population size using respondent driven sampling data, Los Angeles, CA. R package version 1.1.0.

HANDCOCK, M. S., GILE, K. J. and MAR, C. M. (2014). Estimating hidden population size using respondent-driven sampling data. *Electron. J. Stat.* **8** 1491–1521. MR3263129 https://doi.org/10.1214/14-EJS923

HANDCOCK, M. S., GILE, K. J. and MAR, C. M. (2015). Estimating the size of populations at high risk for HIV using respondent-driven sampling data. *Biometrics* **71** 258–266. MR3335370 https://doi.org/10.1111/biom.12255

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

JOHNSTON, L. G. (2013). *Introduction to Respondent-Driven Sampling*. World Health Organization, Geneva, Switzerland.

JOHNSTON, L. G., MCLAUGHLIN, K. R., EL RHILANI, H., LATIFI, A., TOUFIK, A., BENNANI, A., ALAMI, K., ELOMARI, B. and HANDCOCK, M. S. (2015). Estimating the size of hidden populations using respondent-driven sampling data: Case examples from Morocco. *Epidemiology* **26** 846–852.

JOHNSTON, L. G., MCLAUGHLIN, K. R., GIOS, L., CORDIOLI, M., STANEKOVÁ, D. V., BLONDEEL, K., TOSKIN, I., MIRANDOLA, M. and FOR THE SIALON II NETWORK* (2021). Populations size estimations using SS-PSE among MSM in four European cities: How many MSM are living with HIV? *Eur. J. Public Health* **31** 1129–1136. https://doi.org/10.1093/eurpub/ckab148

JOHNSTON, L. G., MCLAUGHLIN, K. R., ROUHANI, S. A. and BARTELS, S. A. (2017). Measuring a hidden population: A novel technique to estimate the population size of women with sexual violence-related pregnancies in South Kivu Province, Democratic Republic of Congo. *J. Epidemiol. Glob. Health* **7** 45–53. https://doi.org/10.1016/j.jegh.2016.08.003

JOHNSTON, L. G., WHITEHEAD, S., SIMIC-LAWSON, M. and KENDALL, C. (2010). Formative research to optimize respondent-driven sampling surveys among hard-to-reach populations in HIV behavioral and biological surveillance: Lessons learned from four case studies. *AIDS Care* **22** 784–792. https://doi.org/10.1080/09540120903373557

KHAN, B., LEE, H.-W., FELLOWS, I. and DOMBROWSKI, K. (2018). One-step estimation of networked population size: Respondent-driven capture-recapture with anonymity. *PLoS ONE* **13** 1–39. https://doi.org/10.1371/journal.pone.0195959

KIM, B. J. and HANDCOCK, M. S. (2019). Population size estimation using multiple respondent-driven sampling surveys. *J. Surv. Stat. Methodol.* **9** 94–120. https://doi.org/10.1093/jssam/smz055

KOKONENDJI, C. C. (2014). Over- and underdispersion models. In *Methods and Applications of Statistics in Clinical Trials. Vol. 2. Wiley Ser. Methods Appl. Statist.* 506–526. Wiley, Hoboken, NJ. MR3287469 https://doi.org/10.1002/9781118596333.ch30

KOSOVO HIV INTEGRATED BEHAVIORAL BIOLOGICAL SURVEILLANCE SURVEY REFERENCE GROUP (2014). HIV integrated behavioral and biological surveillance surveys—Kosovo. Technical report.

LEE, S., SUZER-GURTEKIN, T., WAGNER, J. and VALLIANT, R. (2017). Total survey error and respondent driven sampling: Focus on nonresponse and measurement errors in the recruitment process and the network size reports and implications for inferences. *J. Off. Stat.* **33** 335–366. https://doi.org/doi:10.1515/jos-2017-0017

LU, X. (2013). Linked ego networks: Improving estimate reliability and validity with respondent-driven sampling. *Soc. Netw.* **35** 669–685. https://doi.org/10.1016/j.socnet.2013.10.001

MCLAUGHLIN, K. R., JOHNSTON, L. G., GAMBLE, L. J., GRIGORYAN, T., PAPOYAN, A. and GRIGORYAN, S. (2019). Population size estimations among hidden populations using respondent-driven sampling surveys: Case studies from Armenia. *JMIR Public Health Surveill.* **5** e12034. https://doi.org/10.2196/12034

MCLAUGHLIN, K. R., JOHNSTON, L. G., JAKUPI, X., GEXHA-BUNJAKU, D., DEVA, E. and HANDCOCK, M. S. (2024). Supplement to "Modeling the visibility distribution for respondent-driven sampling with application to population size estimation." https://doi.org/10.1214/23-AOAS1807SUPPA, https://doi.org/10.1214/23-AOAS1807SUPPB

MENG, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.* **22** 1142–1160. MR1311969 https://doi.org/10.1214/aos/1176325622

MILLS, H. L., JOHNSON, S., HICKMAN, M., JONES, N. S. and COLIJN, C. (2014). Errors in reported degrees and respondent driven sampling: Implications for bias. *Drug Alcohol Depend.* **142** 120–126. https://doi.org/10.1016/j.drugalcdep.2014.06.015

MURTHY, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā* **18** 379–390. MR0094869

NAIR, V. N. and WANG, P. C. C. (1989). Maximum likelihood estimation under a successive sampling discovery model. *Technometrics* **31** 423–436. MR1041563 https://doi.org/10.2307/1269993

PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6** 7–11.

R CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAJ, D. (1956). Some estimators in sampling with varying probabilities without replacement. *J. Amer. Statist. Assoc.* **51** 269–284. MR0082236

RIDOUT, M. S. and BESBEAS, P. (2004). An empirical model for underdispersed count data. *Stat. Model.* **4** 77–89. MR2037815 https://doi.org/10.1191/1471082X04st064oa

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681 https://doi.org/10.1214/aos/1176346785

SALGANIK, M. J. and HECKATHORN, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Method.* **34** 193–240. https://doi.org/10.1111/j.0081-1750.2004.00152.x

TOMAS, A. and GILE, K. J. (2011). The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electron. J. Stat.* **5** 899–934. MR2831520 https://doi.org/10.1214/11-EJS630

VOLZ, E. and HECKATHORN, D. D. (2008). Probability based estimation theory for respondent-driven sampling. *J. Off. Stat.* **24** 79–97.