Inference for the Visibility Distribution for Respondent-Driven Sampling

Katherine R. McLaughlin[*] Mark S. Handcock[†] Lisa G. Johnston[‡]

Abstract
Respondent-Driven Sampling (RDS) is used throughout the world to estimate prevalences and population sizes for hard-to-reach populations. Although RDS is an effective method for enrolling people from key populations (KPs) in studies, it relies on an unknown sampling mechanism and thus each individual's inclusion probability is unknown. Current estimators rely on a participant's network size (degree) to compute their visibility and their inclusion probability in 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. We therefore propose a measurement error model to impute visibility in the context of the sample based on each participant's self-reported network size, number of recruits, and time to recruit. These imputed visibilities can also be thought of as a way to smooth the degree distribution and bring in outliers, as well as a mechanism to deal with missing and invalid network sizes. They can be used in place of degree in existing RDS estimators. Finally, we demonstrate the performance of inference for the visibility distribution on a population of men who have sex with men (MSM) from Prishtina, Kosovo in 2014.

Key Words: survey sampling, measurement error model, network sampling, Kosovo, heaping, bias, HIV/AIDS, visibility

1. Introduction

Respondent-driven sampling (RDS)[16] has been shown to be a cost-effective, culturally sensitive method to sample from hard-to-reach populations throughout the world. These are populations which cannot be reached through traditional probability samples and for which the sampling frames are unknown, so innovative methods are needed [8]. In particular, RDS is typically used for key populations (KPs) that are at high-risk for HIV/AIDS and related diseases. KPs identified by the World Health Organization (WHO) include people who inject drugs (PWID), female sex workers (FSW), and men who have sex with men (MSM). 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 KPs to the WHO and UNAIDS from samples conducted using RDS [9]. These estimates are used to inform policy decision, 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 [16]. An initial set of *seeds* is selected, usually via a convenience sample. After completing the survey instrument, each seed is given a small number of *coupons* (usually 3), which contain unique identifying information, 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 are given their own coupons to distribute after participating. Recruitment continues in this manner for many *waves* until the desired sample size is attained. Participants often receive a small *primary incentive* for completing the survey instrument, and a small *secondary incentive* for each recruit they successfully enroll in the study.

[*] Ph.D. Candidate, University of California, Los Angeles, Department of Statistics, USA
[†] Professor, University of California, Los Angeles, Department of Statistics, USA
[‡] Independent Consultant, University of California, San Francisco, Global Health Sciences, USA

Because study participants rather than researchers control recruitment into the study, RDS almost always has an *unknown sampling mechanism* [8]. Therefore the sample is a non-probability sample and key outcome measures cannot be computed by traditional 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. The three main estimators are described below. Assume we are trying to estimate the prevalence of characteristic A, denoted P_A, where A is, for example, being HIV positive. Let $i \in s$ denote that individual i is in the sample, 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 degree for individual i is $\tilde{d}_i$. The true selection probability for individual i is π_i.

The Salganik-Heckathorn (RDS-I) estimator [22] uses the estimated number of cross-group memberships to adjust for the RDS sampling process. The estimator is given by

$$\widehat{P_A^{SH}} = \frac{\widehat{C_{BA}}}{\widehat{C_{BA}} + \widehat{C_{AB}}\left(\frac{\widehat{\overline{D_B}}}{\widehat{\overline{D_A}}}\right)}, \tag{1}$$

where $\widehat{C_{AB}}$ is the proportion of all individuals recruited by members of group A who are members of group B. $\widehat{\overline{D_A}}$ is an estimate of the mean degree of individuals who are part of group A, given by $\widehat{\overline{D_A}} = n_A/(\sum_{i \in s_A} 1/\tilde{d}_i)$. If the mean degree of members of group A is the same as the mean degree of members of group B, then P_A is simply the number of people in group A recruited by those in group B divided by the total number of recruits.

The Volz-Heckathorn (RDS-II) estimator [25] is given by

$$\widehat{P_A^{VH}} = \frac{\sum_{i \in s_A} 1/\tilde{d}_i}{\sum_{i \in s} 1/\tilde{d}_i} \tag{2}$$

where $\tilde{d}_i$ is the self-reported degree of individual i. This is a generalized Hansen-Hurwitz estimator. It is asymptotically unbiased for P_A if $\pi_i \propto \tilde{d}_i$ for all i under the assumption of infinite population size [24].

The Successive Sampling (SS) estimator [7] 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

$$\widehat{P_A^{SS}} = \frac{\sum_{i \in s_A} 1/\tilde{\pi}_i}{\sum_{i \in s} 1/\tilde{\pi}_i} \tag{3}$$

where $\tilde{\pi}_i$ is the estimated sampling probability of individual i. This is determined using the successive sampling procedure [7]. This estimator also requires knowledge of N, the population size.

These estimators all rely on accurate measures of each person's degree d_i (personal network size). For RDS-I, the self-report $\tilde{d}_i$ must be accurate because it is used to calculate the mean degree for members belonging to the group of interest. For RDS-II, the self-report $\tilde{d}_i$ must be accurate because it is assumed that $\pi_i \propto \tilde{d}_i$ for all i. For RDS-SS, the self-report $\tilde{d}_i$ must be accurate because it is assumed that $\pi_i \propto \tilde{d}_i$ conditional on the people not yet sampled. However, in typical RDS studies, degree is self-reported (i.e., we observe only $\tilde{d}_i$, not d_i). This means that it is subject to misreporting and bias, especially because the KPs may practice stigmatized or illegal activities [9]. Beyond these problems with self-reports, degree itself may not correspond to inclusion probability. For all three estimators, the propensity for inclusion in the sample is a function of degree. More precisely, we think of this propensity as *visibility* u_i, where u_i is not necessarily equal to d_i. Visibility cannot be directly measured, but we can impute it from information already collected in RDS studies.

A further discussion of these biases and what we mean by visibility is provided in Section 2. Section 3 provides the model we propose to impute visibility. Results and discussion for MSM in Prishtina, Kosovo are given in Section 4, and Section 5 provides some concluding remarks.

2. Visibility in a Network Sample

The three main RDS estimators all 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 $d_i = u_i$, where u_i is the visibility of individual i.

The first assumption, $\tilde{d}_i = d_i$, is likely not true in practice for a variety of reasons, including: (1) heaping/rounding/coarsening or other approximation methods [8]; (2) intentional misreporting, perhaps in an attempt to minimize one's connection to a stigmatized population [2, 5, 6]; (3) unintentional misreporting, perhaps due to a lack of understanding of the question or memory recall problems [1, 3, 18, 20]. Consider a person with degree 23. An example of a heaped self-report is 20, as they are rounding to the nearest multiple of ten. Likewise, self-reporting 14 could be an example of either intentional or unintentional misreporting. Both of these values underestimate the individual's true degree. Obviously there are also cases where overestimation occurs.

To address these problems with self-reported network size, we could attempt to infer d_i. However, d_i is not the final quantity of interest because we believe the second assumption, $d_i = u_i$, might also be violated in many RDS studies. There are likely other factors in addition to degree that affect unit size. For example, a person may be well-connected in the network, but very unlikely to enroll in the study due to geographic barriers or scheduling concerns. What we want to impute, then, is not each participant's degree, but their *visibility* in the sample.

In a traditional simple random sample, each person has the same visibility and thus the prevalence estimate 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. However, in RDS studies, we often believe that the outcome measure (such as being HIV positive) is correlated with visibility. It is therefore important to use visibility in prevalence estimate calculations. Because we do not observe visibility directly, we need to impute it. A model to impute visibility is given in Section 3.

3. Inference for the Visibility Distribution

Visibility is not directly observable, so we impute a value for each individual based on information collected during the RDS study. 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 a recruits a person enrolls (r_i), and the time they had to recruit (t_i). Assuming $\tilde{d}_i$ and r_i are conditionally independent given the unit size u_i, then $p(\tilde{d}, r) = p(\tilde{d}|u)\, p(r|u)\, p(u)$. In this framework, degree depends only on unit size, but number of recruits depends on both unit size and the time a person had to recruit. Thus we can write

$$p(\tilde{d}, r) = p(\tilde{d}|u)\, p(r|u, t)\, p(u). \tag{4}$$

We now describe a model for each of these pieces.

3.1 Measurement error model for self-reported degree $\tilde{d}_i$

We would like the model for the $\tilde{d}_i|u_i$ 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 plus error (i.e., heaping), where the size of the error depends on the magnitude of the self-report. For $i = 1, \ldots, n$, let

$$\tilde{d}_i|u_i \sim \text{PoissonLogNormal}(\log u_i + \gamma, \sigma^2), \tag{5}$$

where PoissonLogNormal is the Poisson-Log Normal compound distribution [4]. Here, $\log \gamma$ allows for proportional inflation of the self-reported degree relative to the visibility, and σ allows for relative error around the inflated value. 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.

3.2 Model for number of recruits r_i

The number of recruits a person is able to enroll, r_i, is an integer between 0 and m_i, the maximum number of people respondent i was allowed to recruit (the number of coupons they have to distribute, usually 3). We assume r_i depends on person i's visibility, u_i, and their probability of being recruited, o_i. For $i = 1, \ldots, n$, let

$$r_i|u_i, t_i = \min(m_i, \text{Bin}(u_i, o_i)), \tag{6}$$

where we additionally model o_i based on the time to recruit via

$$\text{logit}(o_i|t_i) = \beta_0 + \beta_1 t_i. \tag{7}$$

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. If the number they actually recruited in less than this, their network size may overestimate their visibility.

3.3 Model for visibility u_i

We model the visibilities via a negative binomial distribution conditional on them being positive. This is motivated as a mixture of Poisson distributions for each individual with mean drawn from a Gamma distribution with shape $\mu > 0$ and scale $\alpha > 0$:

$$u_i \sim \text{NB}\left(\mu, \frac{\alpha}{1+\alpha}\right). \tag{8}$$

3.4 Joint distribution

We can then combine these three pieces to get the joint likelihood of the observed data:

$$\begin{aligned}\mathcal{L}[\gamma,\sigma,\beta_0,\beta_1,\mu,\alpha|\tilde{d},r] &\propto p(\tilde{d},r|\gamma,\sigma,\beta_0,\beta_1,\mu,\alpha) \\ &= p(\tilde{d}|u,\gamma,\sigma)\,p(r|u,\beta_0,\beta_1)\,p(u|\mu,\alpha)\end{aligned} \tag{9}$$

The joint likelihood can be maximized using standard statistical tools to obtain maximum likelihood estimates of γ, σ, β_0, β_1, μ, and α. We can then make a random draw from $p(u|\tilde{d},r;\gamma,\sigma,\beta_0,\beta_1,\mu,\alpha)$, and use this as the individual's imputed visibility. These imputed visibilities can then be used in place of self-reported degree in any of the prevalence estimators described in Section 1. In addition to prevalence estimates, imputed visibilities can also be used to improve performance for population size estimates [17], using the Successive Sampling-Population Size Estimation (SS-PSE) method [14, 15]. In this paper we demonstrate the application of imputed visibilities to prevalence estimation.

4. Analysis of HIV Prevalence Among MSM in Kosovo

We impute the participant visibilities for MSM in Prishtina, Kosovo from 2014 and examine performance for prevalence estimates. Prevalence estimates were obtained using the Salganik-Heckathorn (RDS-I), Volz-Heckathorn (RDS-II), and Successive Sample (SS) estimators.

The data are from the third round of the Integrated Behavioral and Biological Surveillance (IBBS) surveys among men who have sex with men (MSM) in Prishtina, Kosovo conducted in July, August, and September of 2014 [11]. The third round of the Kosovo HIV IBBS surveys were funded by by the Global Fund to fight HIV/AIDS, Tuberculosis and Malaria (GFATM) through the Community Development Fund (CDF), Prishtina, Kosovo, and implemented by the National Institute of Public Health (NIPH), Prishtina, Kosovo.

Figure 1 compares participants' self-reported network size to their imputed visibility. In the self-report data, note that many people are grouped around 0 with very low network size, but that there are also several large outliers over 100, including one person who reported 300. Note that the response of 300 is likely a rounded response. Although these values may represent the true personal network sizes of these individuals, they may not correspond to such a large difference in their relative visibilities. The histogram on the right shows imputed visibility. Here, the large outliers have been reduced to under 100, and the small values were also shifted away from 0. Overall, the imputation procedure smooths the distribution of network size and provides values that more plausibly correspond to individual's visibilities.

Although HIV status would generally be the outcome variable of interest in these populations, because few people in the sample were HIV positive (5 of $n = 215$ MSM) we instead demonstrate visibility imputation for other outcome measures. We consider two questions: (1) "Did you have any casual male partners[1] with whom you had anal sex in the past one year?" (2) "In the last six months, did you use a condom the last time you had anal intercourse with any male partner?" For the first question, possible answers were *Yes* and *No*. For the second question, possible answers were *Yes*, *No*, and *Not Applicable*. For both cases, we estimate the prevalence of a *Yes* response. Both of these questions are of interest because they are potentially correlated with HIV status.

[1] For this question, a *casual male partner* was defined as: a man with whom you occasionally have or once had sex without being in a relationship.

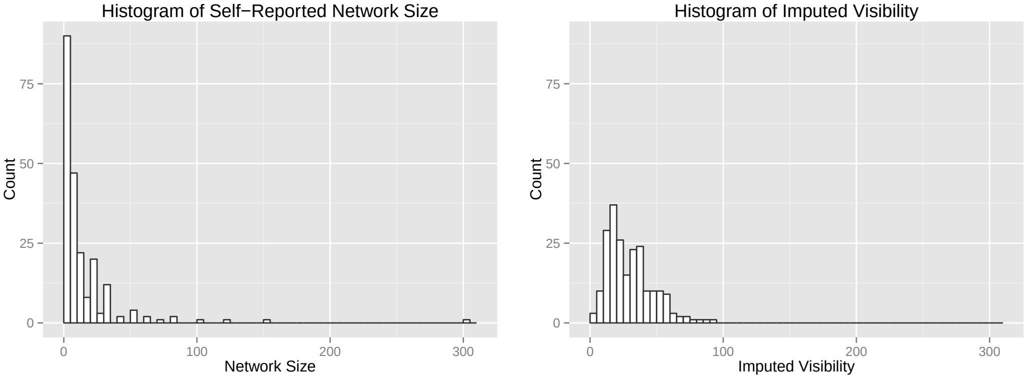

Figure 1: Comparison of self-reported network size (left) and imputed visibility (right) for MSM in Prishtina, Kosovo in 2014.

Estimator	Network size	Point estimate	Standard error	95% Confidence Interval
RDS-I	SR	0.8211	0.0289	(0.7645, 0.8777)
	IV	0.8345	0.0256	(0.7844, 0.8847)
RDS-II	SR	0.8154	0.0493	(0.7187, 0.9121)
	IV	0.8292	0.0450	(0.7410, 0.9174)
RDS-SS	SR	0.8156	0.0494	(0.7188, 0.9123)
	IV	0.8290	0.0543	(0.7402, 0.9177)

Table 1: Comparison of prevalence estimates for the question: “Did you have any casual male partners with whom you had anal sex in the past one year?” among MSM is Pristina, Kosovo. For network size, SR indicates that the self-report was used, and IV denotes imputed visibility.

Table 1 compares the prevalence estimates using the RDS-I, RDS-II, and RDS-SS estimators[2], using self-reported network size (SR) and imputed visibility (IV), for question (1). Using imputed visibility raises the prevalence estimates very slightly, although all point estimates remain within the 95% confidence interval for all other estimates. Note that when using self-reported network size the RDS-II and RDS-SS methods provide almost identical point estimates and confidence intervals. This property is maintained when using imputed visibility.

Table 2 compares the prevalence estimates using the RDS-I, RDS-II, and RDS-SS estimators, using self-reported network size and imputed visibility, for question (2). Although these estimates use the same self-reported network size and imputed visibility as those in Table 1, here all three estimates based on imputed visibility as smaller than those based on self-reported network size. The estimates based on imputed visibility have larger standard errors, but all point estimates are still contained within the confidence intervals for all other estimates.

In addition to providing improved estimates in cases where participants provided plausible but possibly biased network sizes, the visibility imputation method also provides a mechanism to deal with missing or erroneous cases. If a person's network size is missing, we can impute it based on the average degree in the sample and their own number of recruits and time to recruit. If a person reported an impossible value (for example 0 when

[2] The SS estimator requires knowledge of the population size, N. Here we use $N = 5214$, provided as the best guess for the population size in the HIV Integrated Behavioral and Biological Surveillance Surveys – Kosovo report[11]. This number is based on a combination of population size estimates using a variety of methods.

Estimator	Network size	Point estimate	Standard error	95% Confidence Interval
RDS-I	SR	0.7096	0.0335	(0.6440, 0.7753)
	IV	0.6518	0.0408	(0.5720, 0.7317)
RDS-II	SR	0.7200	0.0524	(0.6173, 0.8227)
	IV	0.6633	0.0818	(0.5030, 0.8235)
RDS-SS	SR	0.7192	0.0516	(0.6181, 0.8204)
	IV	0.6641	0.0798	(0.5077, 0.8205)

Table 2: Comparison of prevalence estimates for the question: "In the last six months, did you use a condom the last time you had anal intercourse with any male partner?" among MSM is Pristina, Kosovo. For network size, SR indicates that the self-report was used, and IV denotes imputed visibility.

they were recruited into the study by someone, or 1 when they recruited two people), we can impute using that person's smallest possible degree in place of the self-reported network size. For example, a person in wave 3 who recruited two people must have degree of at least 3. This situation occurred for the Pristina, Kosovo MSM data. Several participants self-reported their network size as 1, but recruited multiple people into the study. Inference for the visibility distribution provides a framework to handle missing and impossible values of network size.

5. Conclusion

Using self-reported personal network size 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 model visibility as a function of self-reported degree, number of recruits enrolled, and time to recruit to obtain imputed values that can be used in place of the self-reported degree in standard RDS estimators. Application of the method to MSM in Pristina, Kosovo provides promising results. Further study is warranted to assess the performance of visibility imputation in a broader setting. In particular, the authors wish to perform a variety of simulation studies to assess how visibility imputation adjusts for different types of bias and how it compares to standard estimators using self-reports in situations where the true prevalence is known.

Inference for the visibility distribution is implemented via the `impute.visibility` function in the `RDS` package [13] of the `R` programming language. It is also available as part of the RDS Analyst software package [12].

Acknowledgements

The authors would like to thank Xhevat Jakupi and Dafina Gexha-Bunjaku of the National Institute for Public Health of Kosova for their work designing, implementing, and lending their expertise to the RDS study used as an example in Section 4 of this paper.

References

[1] David C. Bell, Benedetta Belli-McQueen, and Ali Haider. Partner naming and forgetting: Recall of network members. *Social Networks*, 29(2):279–299, 2007.

[2] Linus Bengtsson and Anna Thorson. Global HIV surveillance among MSM: is risk behavior seriously underestimated? *AIDS (London, England)*, 24(June):2301–2303, 2010.

[3] Devon D. Brewer. Forgetting in the recall-based elicitation of personal and social networks. *Social Networks*, 22:29–43, 2000.

[4] M. G. Bulmer. On Fitting the Poisson Lognormal Distribution to Species-Abundance Data. *Biometrics*, 30(1):101–110, 1974.

[5] Kevin A. Fenton, Anne M. Johnson, Sally McManus, and Bob Erens. Measuring sexual behaviour: methodological challenges in survey research. *Sexually transmitted infections*, 77:84–92, 2001.

[6] Robert J. Fisher. Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research*, 20(2):303, 1993.

[7] Krista J. Gile. Improved Inference for Respondent-Driven Sampling Data With Application to HIV Prevalence Estimation. *Journal of the American Statistical Association*, 106(493):135–146, 2011.

[8] Krista J. Gile and Mark S. Handcock. Respondent-Driven Sampling: An Assessment of Current Methodology. *Sociological Methodology*, 40(1):285–327, 2010.

[9] Krista J. Gile, Lisa G. Johnston, and Matthew J. Salganik. Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society: Series A*, 178(1):241–269, 2015.

[10] Sharad Goel and Matthew J. Salganik. Respondent-driven sampling as Markov chain Monte Carlo. *Statistics in Medicine*, 28:2202–2229, 2009.

[11] HIV Integrated Behavioral Group and Biological Surveillance Survey Reference. HIV Integrated Behavioral and Biological Surveillance Surveys - Kosovo. Technical report, 2014.

[12] Mark S. Handcock, Ian E. Fellows, and Krista J. Gile. *RDS Analyst: Software for the Analysis of Respondent-Driven Sampling Data*. Los Angeles, CA, 2014. Version 0.52.

[13] Mark S. Handcock, Ian E. Fellows, and Krista J. Gile. *RDS: Respondent-Driven Sampling*. Los Angeles, CA, 2015. R package version 0.7-3.

[14] Mark S. Handcock, Krista J. Gile, and Corinne M. Mar. Estimating hidden population size using Respondent-Driven Sampling data. *Electronic Journal of Statistics*, 8(1):1491–1521, 2014.

[15] Mark S. Handcock, Krista J. Gile, and Corinne M. Mar. Estimating the size of populations at high risk for HIV using respondent-driven sampling data. *Biometrics*, 71(1):258–266, 2015.

[16] Douglas D. Heckathorn. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*, 44(2):174–199, 1997.

[17] Lisa G. Johnston, Katherine R. McLaughlin, Houssine El Rhilani, Amina Latifi, Abdalla Toufik, Aziza Bennani, Kamal Alami, Boutaina Elomari, and Mark S. Handcock. A novel method for estimating the size of hidden populations using respondent-driven sampling data: Case examples from Morocco. *Epidemiology*, (In press), 2015.

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

[19] Peter V. Marsden. Network Data and Measurement. *Annual Review of Sociology*, 16:435–463, 1990.

[20] Harriet L. Mills, Samuel Johnson, Matthew Hickman, Nick S. Jones, and Caroline Colijn. Errors in reported degrees and respondent driven sampling: Implications for bias. *Drug and Alcohol Dependence*, 142:120–126, 2014.

[21] Abby E Rudolph, Crystal M Fuller, and Carl Latkin. The importance of measuring and accounting for potential biases in respondent-driven samples. *AIDS and behavior*, 17(6):2244–52, July 2013.

[22] Matthew J. Salganik and Douglas D. Heckathorn. Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34(1):193–240, 2004.

[23] Kerstin E E Schroder, Michael P Carey, and Peter a Vanable. Methodological challenges in research on sexual risk behavior: II. Accuracy of self-reports. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine*, 26:104–123, 2003.

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

[25] E Volz and D D Heckathorn. Probability Based Estimation Theory for Respondent-Driven Sampling. *Journal of Official Statistics*, 24(1):79–97, 2008.

[26] Cyprian Wejnert. An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociological Methodology*, 39(1):73–116, 2009.