**Technical Supplement**

**Estimator Details**

Because of the complexity of the true RDS process, estimators for the resulting data rely on simplified approximations to that sampling process. Differences between estimators are determined by the approximations used, and how those approximations are applied to produce estimators. We refer readers to the papers introducing the respective estimators for their full description and justification, but present a brief overview here.

**Volz-Heckathorn (V-H)**

The Volz-Heckathorn estimator is based on approximating the sampling process as a non-branching with-replacement random walk on the space of the population members, treated as nodes in the network connected by social ties ([Volz and Heckathorn 2008](#_ENREF_6)). The social network is assumed to be undirected, so all ties are bi-directional. In this approximating random walk, each successive step is chosen completely at random among the ties of the current node. This process is treated as a Markov chain on the space of the nodes (population members). In the stationary distribution of this Markov chain, each node has sampling probability proportional to its number of ties in the network, known as its network “degree.” Assuming the sample can be considered drawn from this stationary distribution, the Volz-Heckathorn estimator is of the Hájek form ([Hájek 1971](#_ENREF_2)), substituting the degree  of each sampled node for the sampling probability:

where is the size of the population,  indexes the population,  is the vector of nodal attributes whose population average is to be determined, and is the sampling indicator. We use the estimator as implemented in the open-source R package RDS ([Handcock et al. 2015](#_ENREF_3)).

**Salganik-Heckathorn (S-H)**

The Salganik-Heckathorn estimator assumes the same Markov chain approximation for sampling as the Volz-Heckathorn estimator ([Salganik and Heckathorn 2004](#_ENREF_5)). Its form, however, is slightly different, as it uses a method-of-moments-style approach, equating the estimated number of ties between the two nodal classes whose proportions are being estimated.

To compute this estimator, one first estimates the average degree of each nodal class, and, using an estimator of the following form:

where is the nodal class whose average degree is being estimated, indexes the population, is the set of population members in nodal class , is the sampling indicator, and is the degree of the sampled node. This estimator is equivalent to the harmonic mean of the nodal degrees for sample members of nodal class .

Adding further regularity assumptions about the recruitment process, the observed proportion of sample referrals between each directed pair of classes is used as an estimate of the proportion of ties from the referring class that go to the referred class. We denote these estimates , Then, because the network is assumed to be undirected, the number of ties from class to class must be equal to the number of ties from class  to If the population proportion of 1’s is μ, this means:

 (3)

Solving for μ gives the form of the Salganik-Heckathorn estimator:

(4)

We use the estimator as implemented in the open-source R package RDS ([Handcock et al. 2015](#_ENREF_3)).

**Successive Sampling (SS)**

The Successive Sampling estimator takes a Hájek form similar to that shown in equation (1), but using a different approximation to the sampling process, and uses different estimated sampling probabilities, denoted ([Hájek 1971](#_ENREF_2), [Gile 2011](#_ENREF_1)). In particular, this estimator uses a without-replacement random-walk approximation to the sampling process. Under certain approximations to the network structure, this results in a sampling process analogous to successive sampling, also known as probability proportional to size without replacement (PPSWOR) sampling, where nodal ‘sizes’ are given by nodal degrees. The ’s are therefore determined using a successive sampling approximation, and substituted for the ’s in equation (1), giving the form of the successive sampling estimator. We use the algorithm given in Gile 2011, as implemented in the open-source R package RDS ([Gile 2011](#_ENREF_1), [Handcock et al. 2015](#_ENREF_3)).

**Bootstrap Details**

**Salganik Bootstrap (Sal-BS)**

The Sal-BS first classifies each sample member by the infection status of his recruiter to create two pools of recruits, those recruited by an infected sample member and those recruited by an uninfected sample member ([Salganik 2006](#_ENREF_4)). The resampling procedure creates each replicate sample using an iterative process. First, an initial participant is sampled uniformly at random from the entire sample; this participant functions as the replicate sample’s seed. Based on the infection status of the seed, the procedure samples a second participant from the appropriate pool of recruits uniformly at random with replacement (for example, if the seed is infected the procedure chooses a person from the pool of participants who were recruited by an infected participant). The bootstrap then chooses a third participant from the appropriate pool of recruits based on the second participant’s infection status, and so on. This process continues until the replicate sample is the same size as the original sample. This bootstrap re-sampling procedure is repeated to create a large number of replicate samples, the estimator of interest is computed for each replicate sample, and the resulting bootstrap distribution of estimates is used to estimate the uncertainty of the estimator. The Sal-BS is typically applied in conjunction with the SH or VH RDS point estimators. We use the estimator as implemented in the open-source R package RDS ([Handcock et al. 2015](#_ENREF_3)).

**Successive Sampling Bootstrap (SS-BS)**

Unlike the Sal-BS, the SS-BS conducts resampling without replacement. SS-BS first estimates the probability of population network ties across infection status groups (cross-group probabilities) based on the recruitment patterns in the original sample ([Gile 2011](#_ENREF_1)). It begins by choosing *j* members of the original sample with probability proportional to degree, where *j* is the number of seeds in the original sample. For each of the *j* seeds, SS-BS then chooses *k* sample members, where *k* is the number of recruitments made by the seed. For each of the *k* recruits, it chooses the next replicate sample member’s infection status based on the cross-group probabilities and chooses a sample member from sample members with that infection status with probability proportional to degree. It continues this process for each recruit until all sample members have been re-sampled. Because the SS-BS samples without replacement, the cross-group probabilities are recalculated after each re-sample member is selected. This bootstrap re-sampling procedure is repeated a large number of times, the estimator of interest is computed for each, and the resulting bootstrap distribution of estimates is used to estimate the uncertainty of the estimator. The SS-BS is applied in conjunction with the Successive Sampling (SS) RDS point estimator. Details of the specific form of the SS bootstrap estimator can be found in the supplementary materials for Gile 2011, and the implementation used for our analysis is implemented in the RDS R package ([Gile 2011](#_ENREF_1), [Handcock et al. 2015](#_ENREF_3)).

**Bootstrap Types**

**Percentile**

For a two-sided confidence interval (CI), the percentile method calculates the 100\*α/2 and 100\*(1 – α/2) percentiles of the bootstrap sampling distribution of estimates (e.g., the 2.5 and 97.5 percentiles for a 95% CI). The values of these percentiles are the lower and upper bounds, respectively, for the CI. This method allows for asymmetrical CIs by assuming the bootstrap sampling distribution is a true approximation of the distribution of possible estimates. By definition, a percentile estimated 95% CI will contain 95% of replicate sample estimates produced by the bootstrap procedure. The SH/Sal-BS and VH-Sal-BS RDS estimator pairs have traditionally calculated CIs using the percentile method.

**Studentized**

Rather than relying on the bootstrap sampling distribution directly, the studentized bootstrap CI method employs an approximation based on the standard studentized distribution. It calculates the standard deviation of the bootstrap sampling distribution, then multiplies the standard deviation by the critical t-value associated with the CI’s α and the degrees of freedom (n-1, where n is the sample size), and adds and subtracts that product to the point estimate to calculate the upper and lower CI bounds, respectively. This produces CIs that are symmetric around the point estimate.

**Supplement Figures and Tables**

**Supplement Figure 1: Flowchart**

This diagram shows the steps that went into creating the population networks, simulating RDS sampling on these networks, and then calculating the point and variance estimates of the trait prevalence from the samples.

**Data Tables**

NOTE: The descriptions of the data table structures below abbreviate average as “Avg.” and standard deviation as “Std. Dev.”

**Supplement Table 1: Population Parameters**

Population parameters were estimated from the NHBS samples

**Column name Description**

Sample Unique identifier for a single NHBS sample

Trait Prevalence Population prevalence of trait

Mean Degree Population mean degree of trait

Differential Activity Population differential activity of trait

Homophily Population homophily for trait

**Supplement Table 2: Sample Parameters**

Although sample members were given varying numbers of coupons to hand to other population members, that number was not used in the simulation. What was relevant to the simulation was the number of coupons that successfully resulted in recruitment of other sample members. Sample members who were given one or more coupons were able to recruit from 0 to 5 (the maximum number of coupons given out) other persons. The distribution of this behavior is reflected in "Number of sample members who recruited" 0,...,5 persons. Note that some sample members were not given any coupons (as happened when the sample size had been reached) and are not included in these numbers.

**Column name Description**

Sample Unique identifier for a single NHBS sample

Seeds with Trait Number of seeds with trait, (i.e., number of infected seeds)

Seeds without Trait Number of seeds without trait, (i.e., number of uninfected seeds)

Seeds Missing Trait Number of seeds where trait was unknown; for the purposes of the simulation, these seeds were considered to be without the trait

Number of sample members who recruited

0 persons Number who recruited 0 other persons into sample

1 persons Number who recruited 1 other person into sample

2 persons Number who recruited 2 other persons into sample

3 persons Number who recruited 3 other persons into sample

4 persons Number who recruited 4 other persons into sample

5 persons Number who recruited 5 other persons into sample

**Supplement Table 3: Coverage Results**

**Condition: Without replacement sampling, studentized bootstrap**

"Coverage" Percentage of times 95% Confidence Interval included the Population Prevalence

**Column name Description**

Sample Unique identifier for a single NHBS sample

Coverage: sample mean Coverage of sample mean and SRS confidence interval

Coverage: SH Coverage of the Salganik-Heckathorn estimate/Salganik bootstrap

Coverage: VH Coverage of the Volz-Heckathorn estimate/Salganik bootstrap

Coverage: SS Coverage of the Successive-Sampling estimate/Successive-Sampling bootstrap

**Supplement Table 4: Confidence Interval Width Results**

**Condition: Without replacement sampling, studentized bootstrap**

For the repeated (1000) samples per each site condition, the following was calculated:

"Avg CI Width" Mean of the lengths of each sample's 95% Confidence Interval

"Std. Dev. Width" Standard deviation of the lengths of each sample's 95% Confidence Interval

**Column name Description**

Sample Unique identifier for a single NHBS sample

Avg. CI Width: sample mean Avg. CI Width of sample mean

Avg. CI Width: SH Avg. CI Width of the Salganik-Heckathorn estimate/Salganik bootstrap

Avg. CI Width: VH Avg. CI Width of the Volz-Heckathorn estimate/Salganik bootstrap

Avg. CI Width: SS Avg. CI Width of the Successive-Sampling estimate/Successive-Sampling bootstrap

Std. Dev. CI Width: sample mean Std. Dev. CI Width of sample mean

Std. Dev. CI Width: SH Std. Dev. CI Width of the Salganik-Heckathorn estimate/Salganik bootstrap

Std. Dev. CI Width: VH Std. Dev. CI Width of the Volz-Heckathorn estimate/Salganik bootstrap

Std. Dev. CI Width: SS Std. Dev. CI Width of the Successive-Sampling estimate/Successive-Sampling bootstrap

**Supplement Table 5: Actual Design Effect Results**

**Condition: Without replacement sampling**

"Actual DE" The design effect (DE) is a relative measure of the variability of a complex sampling method applied to a given population. It is calculated as the ratio of the variance of an estimate using the complex sampling method to that of an estimate from a hypothetical simple random sample of the same population. The most direct way to calculate the variance of such an estimate is to collect repeated samples and observe the variation in estimates based on those samples. That approach provides the actual, as opposed to the estimated, variance of the estimate. In the real world, we cannot repeat a given sampling method on the same population, so we cannot calculate the actual variance or DE of a given method and population. However, in a simulation setting we can repeat the method on a simulated population. In our study, for the simulated population corresponding to a given NHBS sample we simulate 1,000 RDS resamples and produce the point estimate for each resample. We calculate the variance of RDS on that simulated population as the variance of the 1,000 point estimates. We calculate the actual (true) design effect (“actual DE”) as the ratio of that variance to the SRS variance for a binary variable, where the SRS variance includes a finite population adjustment based on the proportion of the population included in the sample:

var(RDS resample point estimates)/[[p\*(1 - p)/n]\*fpc], where p is the population prevalence, n is the sample size, N is the population size, and fpc = (N-n)/(N-1)

**Column name Description**

Sample Unique identifier for a single NHBS sample

Actual DE: sample mean True design effect for sample mean

Actual DE: SH True design effect for Salganik-Heckathorn estimates

Actual DE: VH True design effect for Volz-Heckathorn estimates

Actual DE: SS True design effect for Successive-Sampling estimates

**Supplement Table 6: Estimated Design Effect Results**

**Condition: Without replacement sampling, studentized bootstrap**

**Column name Description**

Mean Estimated DE Mean of estimated design effects

Std. Dev. Estimated DE Standard deviation of estimated design effects

Median Estimated DE Median of estimated design effects

Minimum Estimated DE Minimum of estimated design effects

Maximum Estimated DE Maximum of estimated design effects

SHSalganik-Heckathorn estimates

VH Volz-Heckathorn estimates

SS Successive-Sampling estimates

The values marked by “>1000“ represent situations where the estimated design effect was extremely large in a number of samples, and hence the specific value of the mean is not meaningful.

**Supplement Table 7: CI Coverage, Percentile Bootstrap**

**Condition: Without replacement sampling, percentile bootstrap**

"Coverage" Percentage of times 95% Confidence Interval included the Population Prevalence

**Column name Description**

Sample Unique identifier for a single NHBS sample

Coverage: SH Coverage of the Salganik-Heckathorn estimate/Salganik bootstrap

Coverage: VH Coverage of the Volz-Heckathorn estimate/Salganik bootstrap

Coverage: SS Coverage of the Successive-Sampling estimate/Successive-Sampling bootstrap

**Supplement Table 8: CI Coverage, VH/Sal-BS With Replacement**

**Condition: With replacement sampling, Volz-Heckathorn estimate/Salganik bootstrap**

Coverage Percentage of times 95% Confidence Interval included the Population Prevalence

Estimated DE Design Effect estimated from a single sample using RDS variance estimates

**Column name Description**

Sample Unique identifier for a single NHBS sample

Coverage: Studentized Bootstrap Coverage studentized bootstrap

Coverage: Percentile Bootstrap Coverage percentile bootstrap

Actual DE Actual (true) Design Effect:

var(RDS resample point estimates)/[[p\*(1 - p)/n]\*fpc], where p is the population prevalence, n is the sample size, N is the population size, and fpc = (N-n)/(N-1)

**Studentized bootstrap:**

Mean Estimated DE Mean of estimated design effects

Std. Dev. Estimated DE Standard deviation of estimated design effects

Median Estimated DE Median of estimated design effects

Minimum Estimated DE Minimum of estimated design effects

Maximum Estimated DE Maximum of estimated design effects

**Design Effect Figures**

**Supplement Figure 2: Volz-Heckathorn Design Effect Comparison: With versus Without Replacement Sampling**

This figure shows that with replacement sampling results in consistently higher actual (true) design effects. On the x-axis are the samples ordered by design effect size for with replacement sampling with that design effect shown in red where the design effect magnitudes are shown on the left y-axis. The blue points are the corresponding design effect for without replacement sampling. The purple points are the ratio (right y-axis) of the with replacement/without replacement design effects. Note that ratio is always above 1, that is, with replacement sampling design effects are always higher than without replacement sampling design effects and that this ratio is almost 2.5 for site A-14.

**Supplement Figure 3: Actual Design Effect Cumulative Distribution**

This figure shows the cumulative distribution of the actual (true) design effect for each estimator for *without* replacement sampling (SS=Successive Sampling, SH=Salganik-Heckathorn, VH=Volz-Heckathorn) and the VH estimator for *with* replacement sampling. This figure enables the viewer to estimate the proportion of DE's below a certain cutoff for a particular estimator. For example, 70% of the SS design effects are below 2, 95% below 3.

**Supplement Figure 4: Estimated and Actual DE Ratio Boxplots**

This figure shows the distributions of the ratio of the estimated DE to the actual DE for the 1,000 RDS simulations for each population network, by estimator pair used to estimate the DE and sampling condition (with or without replacement). For each estimator pair, 40 box plots are shown (one per population network). The x-axis is the population network, and the y-axis is the estimated DE to actual DE ratio. The dark line in the middle of the box is the median ratio, the bottom and top of the box are the 25th and 75th percentiles of the ratio distribution (respectively), and the bottom and top whiskers are the minimum/maximum data point or 1.5 times the inter-quartile range [1.5 \* (75th percentile – 25th percentile)], whichever is lower/higher. Points are individual estimated DEs outside the whisker range. The horizontal red line at 1 on the y-axis represents an accurate estimated DE. Ratios below the lower blue horizontal line are those for which the estimated DEs were smaller than 50% of the actual DE, and ratios above the upper blue horizontal line are those for which the estimated DEs were greater than 200% the actual DE. The figure shows that there is meaningful variance in the accuracy of estimated DEs across population networks. It also shows that the SS/SS-BS estimated DEs exhibit greater accuracy (are closer to the red line) and less variance (more of the distribution is between the blue lines) than those for the other estimators. Finally, the bottom panel shows that when sampling is **with** replacement the VH/Sal-BS exhibits much worse performance compared to **without** replacement sampling, and its estimated DEs are usually too small (the anti-conservative direction).

**References**

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

Hájek, J. (1971), "Comment on an Essay on the Logical Foundations of Survey Sampling by Basu, D," *Foundations of Statistical Inference*, 236.

Handcock, M. S., Fellows, I. E., and Gile, K. J. (2015), "Rds: Respondent-Driven Sampling."

Salganik, M. J. (2006), "Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling," *J Urban Health*, 83, i98-112.

Salganik, M. J., and Heckathorn, D. D. (2004), "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling," *Sociological Methodology, Vol 34*, 34, 193-239.

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