Comparing the Robustness of Simple Network Scale-Up Method (NSUM) Estimators

Jessica P. Kunke[1]*, Ian Laga[2], Xiaoyue Niu[3], Tyler H. McCormick[1]

[1] *University of Washington,* [2] *Montana State University,*
[3] *Pennsylvania State University*

March 15, 2023

Abstract: The network scale-up method (NSUM) is a cost-effective approach to estimating the size or prevalence of a group of people that is hard to reach through a standard survey. The basic NSUM involves two steps: estimating respondents' degrees by one of various methods (in this paper we focus on the probe group method which uses the number of people a respondent knows in various groups of known size), and estimating the prevalence of the hard-to-reach population of interest using respondents' estimated degrees and the number of people they report knowing in the hard-to-reach group. Each of these two steps involves taking either an average of ratios or a ratio of averages. Using the ratio of averages for each step has so far been the most common approach. However, we present theoretical arguments that using the average of ratios at the second, prevalence-estimation step often has lower mean squared error when a main model assumption is violated, which happens frequently in practice; this estimator which uses the ratio of averages for degree estimates and the average of ratios for prevalence was proposed early in NSUM development but has largely been unexplored and unused. Simulation results using an example network data

*Correspondence: jkunke@uw.edu.

set also support these findings. Based on this theoretical and empirical evidence, we suggest that future surveys that use a simple estimator may want to use this mixed estimator, and estimation methods based on this estimator may produce new improvements.

Keywords: Aggregated relational data (ARD), hard-to-reach populations, hidden populations, network scale-up method (NSUM), prevalence estimation, size estimation, surveys.

1 Introduction

Surveys are a standard approach to estimating the size of subpopulations, or groups of people with a particular trait. Many key subpopulations of interest, however, are hard to reach with standard surveys (Bernard et al., 1991; Killworth et al., 1998). For example, (1) it may be hard to get a list of the members of this subpopulation or they may be hard to contact, as in the case of the homeless subpopulation; (2) it may be hard to accurately determine their membership in the subpopulation because their group status is stigmatized, as in the case of heavy drug users; and/or (3) the subpopulation is often fairly rare.

There are generally two strategies to this estimation problem that use social networks. One is to adapt the survey methods to find more members of the subpopulation, then provide guarantees that the results are probabilistic. This category includes methods such as respondent driven sampling (RDS) and snowball sampling (Salganik and Heckathorn, 2004; Handcock et al., 2014; Crawford et al., 2018). These methods involve surveying members of the population of interest, and therefore they have the advantage that researchers may ask additional questions to study other aspects of the population in addition to estimating prevalence. For example, one could not only estimate the number of people who have been trafficked in a given region but also study how they entered trafficking, what enabled them to leave if they left, and what factors increased or decreased their vulnerability to trafficking.

However, this approach is not always feasible, and when it is, it can be expensive, particularly if the survey aims to estimate sizes for multiple subpopulations.

The other strategy, the network scale-up method (NSUM), is to conduct a traditional survey with a representative sample from the general or frame population and ask respondents how many people they know in the hidden populations of interest, then use information about their personal network sizes (degrees) to scale up their data into an estimate for the general prevalence of those populations (Bernard et al., 1991, 2010). Respondents' degrees are also estimated by asking similar questions about groups of known sizes (McCormick et al., 2010). The responses to questions of the form "How many people do you know with X trait?" are known as aggregated relational data or ARD (McCormick and Zheng, 2015). This approach does not require knowing whether the respondents themselves are in the hidden populations.

In this paper we focus on the NSUM strategy. We conducted an initial literature search and found that the majority of studies use one of the simplest estimators. However, we find that when a key model assumption is violated, as it often is in practice, the mean squared error (MSE) may actually tend to be much smaller for another simple estimator that is equally easy to implement but much less commonly used. We demonstrate this through theoretical derivations as well as simulations with a real network data set.

The paper is structured as follows: Section 2 introduces the key estimators of interest in this study. Section 3 presents our framework for studying the behavior of these estimators in the presence of barrier effects. Section 4 details the analytical results comparing estimator bias, variance, and RMSE in this setting. Section 5 examines the results of simulated surveys using real network data, the Facebook 100 data set. Code for the analytical and simulation results is available at `https://github.com/jpierkunke/simple-nsum-robust`. Section 6 concludes with a discussion.

2 NSUM estimators

In this section, we introduce the estimators of interest in this study as compositions of degree and prevalence estimators. We discuss the assumptions underlying the models that led to these estimators, which estimators are currently used in practice, and why we reevaluate which estimator if any should be the standard.

The target of estimation is the prevalence R of the hidden group H in the general population, which equals the ratio of the hidden group size to the size N of the general or frame population. For each respondent in a sample of size n, let y_i represent respondent i's response to the ARD question, "How many people do you know in H?" Let d_i represent the true degree of respondent i. In practice, these degrees are unknown and must be estimated. Therefore, we can decompose the estimation problem into two steps: first estimating respondents' degrees, then using these degree estimates $\hat{d}_i$ with the responses y_i to estimate the prevalence R.

The NSUM approach is based on the idea that given the response y_i and degree d_i for one person, a rough estimate of the hidden group prevalence is y_i/d_i (Bernard et al., 1991). To obtain a better estimate, we can pool the responses and degrees from a larger sample of people. This information can be pooled in two basic ways, either the ratio of average response to average degree (ratio of averages, hereafter RoA) or the average ratio of response to degree (average of ratios, hereafter AoR):

$$\hat{R}_{\text{RoA}} = \frac{\sum_i y_i}{\sum_i d_i}, \quad \hat{R}_{\text{AoR}} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{d_i}. \tag{1}$$

We could call the RoA estimator a ratio of sums since we have dropped the factor of $1/n$ from the numerator and denominator, but we still refer to it as the ratio of averages to emphasize that the difference between the two estimators is simply the order of two operations, averaging and dividing.

Similarly, the degrees can be estimated by asking respondents how many people they

Table 1: Each of the three estimators can be viewed as a composition of a hidden group prevalence estimator and a degree estimator. Each component estimator may be the ratio of averages (RoA) or the average of ratios (AoR). The fourth combination, using the AoR degree estimator and the RoA prevalence estimator, has not been proposed to our knowledge and, based on our analysis of the other estimators, seems unlikely to perform well. We list MLE in quotes to indicate that the dRpR is not itself an MLE. The ROA prevalence estimator is the MLE for R conditional on degrees, and the dRpR uses this prevalence estimator with estimated degrees.

	Degree estimator		Prevalence estimator		Other names
dRpR	RoA	$\hat{d}_i = N \cdot \frac{\sum_j y_{ij}}{\sum_j N_j}$	RoA	$\hat{R} = \frac{\sum_i y_i}{\sum_i \hat{d}_i}$	"MLE"
dRpA	RoA	$\hat{d}_i = N \cdot \frac{\sum_j y_{ij}}{\sum_j N_j}$	AoR	$\hat{R} = \frac{1}{n}\sum_i \frac{y_i}{\hat{d}_i}$	PIMLE
dApA	AoR	$\hat{d}_i = N \cdot \frac{1}{K}\sum_j \frac{y_{ij}}{N_j}$	AoR	$\hat{R} = \frac{1}{n}\sum_i \frac{y_i}{\hat{d}_i}$	MoS
dApR	AoR	$\hat{d}_i = N \cdot \frac{1}{K}\sum_j \frac{y_{ij}}{N_j}$	RoA	$\hat{R} = \frac{\sum_i y_i}{\sum_i \hat{d}_i}$	(Unmentioned)

know in each of K subpopulations of known size, often called probe groups; here, known size means that the number or prevalence of each probe group in the general population can be obtained from census information or other data sources. For example, respondents might be asked how many people named Jamal they know and how many firefighters they know; this would provide responses y_{ij} for each person i about probe group $j \in \{1, 2\}$, where the two probe groups are the subsets of the general population (1) named Jamal and (2) serving as firefighters, respectively. Note that probe groups can and often do overlap; people named Jamal who serve as firefighters would be counted in both of these group sizes. Then analogously to the prevalence estimates based on respondents' degrees, the degrees can be estimated based on these K probe groups by either the RoA or the AoR:

$$\hat{d}_{i,\text{RoA}} = N \cdot \frac{\sum_j y_{ij}}{\sum_j N_j}, \quad \hat{d}_{i,\text{AoR}} = N \cdot \frac{1}{K}\sum_{j=1}^{K} \frac{y_{ij}}{N_j}.$$

These separate degree and prevalence estimators can be combined in four ways to obtain a two-step prevalence estimator that incorporates the degree estimation step (see Table 1). To be explicit and concise, we will refer to these estimators by the choice of degree estimator

followed by the choice of prevalence estimator; for example, we will use the name dRpA to refer to the estimator which plugs the ratio of averages for the degree estimates into the average of ratios for the prevalence estimates.

The NSUM is cost effective but also depends on fairly strong assumptions that are known to be non-negligibly violated in many settings. The assumption that $y_i/d_i = R$ either for each person i or on average in one of the senses presented in (1) above is known as the constant proportion assumption. Violations of this assumption are called barrier effects. The binomial model also assumes perfect visibility (each respondent knows whether each person in their network is in H), perfect recall (each respondent can enumerate everyone they know, or report the correct total), truthful answers, and the absence of other survey and response error. In this paper, we focus specifically on barrier effects.

We often expect there to be barrier effects in practice. Typically, we expect that not everyone in the population is equally likely to know people in the hidden population of interest (see for example McCormick and Zheng (2012) for more discussion). For instance, homophily often drives connections, and people who are more similar to people in the hidden population may be more likely to know them (McPherson et al., 2001). Additionally, there is evidence that people in the hidden subpopulation often have smaller degrees than people in the general population, which we will see necessarily violates the constant proportion assumption (Shelley et al., 1995).

In the years since the early NSUM papers, a body of research has extended this model to handle barrier effects and relax the constant proportion assumption; McCormick (2021) and Laga et al. (2021) provide detailed reviews of these methods. However, these approaches are more complex and require additional data. For example, the generalized NSUM developed by (Feehan and Salganik, 2016) can be used in the presence of barrier effects and imperfect visibility, but this approach requires sampling from the hidden population in addition to the original probability sample from the general population. For this reason, they also describe how correction factors can be applied to the dRpR estimator if sufficient data and expert

knowledge exist to estimate those factors. We conducted an initial literature search of papers published in 2021 using the network scale-up method to investigate what methods researchers currently seem to use in practice, and we found no papers that use the dRpA and only two reviews or guidance references that even mentioned it (see online supplement for details). Most of the studies used either the dRpR, sometimes with scaling factors as recommended by Feehan and Salganik (2016), or a more complicated model developed by Maltiel et al. (2015).

The most commonly used of these four estimators seems to be the dRpR estimator, first proposed to our knowledge by Killworth et al. (1998). An earlier paper by Killworth et al. (1998) compares the dRpA and dApA estimators, and a more recent paper by Habecker et al. (2015) promotes the use of the dApA estimator, which they call the mean-of-sums estimator or dApA. We are not aware of any literature considering or proposing the use of the dApR estimator, and our results in this study suggest it is unlikely to outperform the other estimators, but it could be studied in future work. In the present study, we focus on evaluating the other three estimators.

When the binomial model is true, both the AoR and RoA prevalence estimators with fixed or known degrees are unbiased, but the latter has a smaller variance. Killworth et al. (1998) compute and empirically evaluate the dRpR and dRpA on a data set, but their theoretical analysis assumes the degrees are known; thus their theoretical analysis concerns only the (one-step) AoR and RoA prevalence estimators, not the (two-step) dRpR and dRpA estimators. The degree estimation step in the dRpR and dRpR estimators not only involves additional uses of the conditional proportion assumption to estimate the degree, but also accounts for the distribution of degrees rather than conditioning on them. Perhaps this is the reason that surveys tend to use the dRpR if they use one of the simple estimators, and that researchers tend to start from the dRpR when they develop methods to extend the NSUM approach and relax modeling assumptions.

3 A framework for studying estimator behavior under barrier effects

Our question is, how do these estimators compare under violations of the constant proportion assumption, since such violations are common in practice? To investigate this, we assume a model for link formation in the general population, then consider the distribution of the estimators over simple random samples from that population without replacement. The binomial model can be viewed as an approximation to the Erdös-Renyi network model in which the presence or absence of a link between each pair of nodes is independently drawn from the same Bernoulli(p) distribution (see online supplement and Cheng et al. (2020) for details). In this study, we generalize this model to incorporate barrier effects by using a stochastic block model (SBM) with two groups, the hidden group of interest (H) and everyone else (L).

Under the two-group SBM, the probability of a link between any two nodes takes one of three values depending on the membership of the two nodes involved; let us denote the within-group probabilities by p_{HH} and p_{LL} and the between-group probability by p_{HL}. This model will have barrier effects as long as $p_{HH} > p_{HL}$, since in that case members of H will be more connected than members of L to people in H. The dissortative condition $p_{HH} < p_{HL}$ would also create barrier effects but is typically less realistic in practice. The Erdös-Renyi model is a special case of this SBM with $p_{HH} = p_{HL} = p_{LL}$.

Many network models can be approximated by a stochastic block model (Olhede and Wolfe, 2014), so the two-group SBM is a motivating choice for studying the impact of barrier effects. The two-group SBM may provide insights that have more general relevance, such as behavior based on network assortativity, even if a given problem is not believed to follow a two-group SBM.

For the sake of interpretability, we start with a simple case using estimated degrees: we suppose that respondents' degrees are estimated using one probe group K. Furthermore

we suppose that $K \subset L$ to avoid having to specify the prevalence of H within K, which would be necessary to compute expectations and variances. This renders the probe group unrepresentative of the general population, since it contains no one in the hidden group H, but this is sometimes a concern with the choice of probe groups and is relevant to evaluate here. Additionally, the real data example in Section 5 assumes more than one probe group and does not require the probe group to be disjoint from H.

For the case of a single probe group, the AoR and RoA degree estimators are identical; therefore the dApA and dRpA estimators are equivalent when degrees are estimated using a single probe group, and our analytical results will only compare the choice of prevalence estimator. However, this analysis still accounts for the additional use of the constant proportion assumption in estimating the degrees, and it also accounts for the distribution of degrees instead of conditioning on them. The degree estimators are no longer identical when more than one probe group is used; therefore we save comparison with the dApA for Section 5.

Killworth et al. (1998) assume a simple random sample without replacement, and Habecker et al. (2015) propose a way to extend this to general probability survey designs. We provide a note about this in the online supplement and suggest modifications to this extension, but for simplicity and to stay consistent with the original estimators, in this study we assume simple random sampling without replacement.

4 Analytical results on estimator bias and variance

We begin by deriving approximations to the bias and variance of each estimator under a two-group stochastic block model assuming degrees are estimated using a single probe group K contained in L. Then to facilitate interpretation, we restrict further to the case in which the within-group link probabilities are equal to the same scaling factor a times the between-group link probability p_{HL}. We present closed-form approximations for the bias and variance, and

we numerically compute the bias, variance, and RMSE for the two estimators over a range of parameter values to characterize the regions in which one estimator outperforms the other. As mentioned previously, the dApA and dRpA estimators are equivalent when degrees are estimated using a single probe group; hence this section discusses only the dRpR and the dRpA. However, the dApA and dRpA are distinct once there is more than one probe group, so we compare all three estimators in Section 5.

Under the two-group SBM, the number of people person i knows in the hidden group H and probe group K, respectively, is given by

$$Y_{iH} = \sum_{j=1}^{N_H^*} A_i j \sim \text{Binom}(N_H^*, p_{H\,g_i}), \quad Y_{iK} = \sum_{j=1}^{N_K^*} A_i j \sim \text{Binom}(N_K^*, p_{g_i\,L}),$$

where $g_i \in \{H, L\}$ denotes the group membership of person i; Y_{iH} and Y_{jH} are independent for any i, j ($i = j$ or $i \neq j$); $N_H^* = N_H - 1$ if $i \in H$ and $N_H^* = N_H$ otherwise; and N_K^* is defined analogously to N_H^*. Henceforth we assume N_H and N_K are sufficiently large such that $N_H^* \approx N_H$ and $N_K^* \approx N_K$.

Using first-order Taylor approximations for the expectation and variance of ratios, we can estimate the expectation and variance of the estimators over the SBM superpopulation for a given sample, then take the limit $n_H/n \to R$ as in simple random sampling without replacement. In each case, the expectation is a function only of R and the three link probabilities, while the variance is also a function of n, N, and N_K (see Appendix A for further details):

$$\begin{aligned}
E\left(\hat{R}_{\text{dRpR}}\right) &\to R\frac{Rp_{HH} + (1-R)p_{HL}}{Rp_{HL} + (1-R)p_{LL}} \\
E\left(\hat{R}_{\text{dRpA}}\right) &\to R\left[R\frac{p_{HH}}{p_{HL}} + (1-R)\frac{p_{HL}}{p_{LL}}\right] \\
\text{Var}\left(\hat{R}_{\text{dRpR}}\right) &\to \frac{R}{nN}\frac{\left(Rp_{HL} + (1-R)p_{LL}\right)^2\left[Rp_{HH}(1-p_{HH}) + (1-R)p_{HL}(1-p_{HL})\right]}{\left(Rp_{HL} + (1-R)p_{LL}\right)^4} +
\end{aligned}$$

$$\frac{R^2}{nN_K}\frac{\left(Rp_{HH}+(1-R)p_{HL}\right)^2\left[Rp_{HL}(1-p_{HL})+(1-R)p_{LL}(1-p_{LL})\right]}{\left(Rp_{HL}+(1-R)p_{LL}\right)^4}$$

$$\mathrm{Var}\left(\hat{R}_{\mathrm{dRpA}}\right) \to \frac{R}{nNp_{HL}^2}\left[Rp_{HH}(1-p_{HH})+(1-R)p_{LL}(1-p_{HL})\right] + \frac{R^2}{nN_Kp_{HL}^3}\left[Rp_{HH}^2(1-p_{HL})+(1-R)p_{HL}^2(1-p_{LL})\right]$$

Note that when the three link probabilities are equal, corresponding to the binomial model, the dRpR does not have smaller variance: both estimators have the same first-order variance. We believe this is the first time this result has been shown for these estimators. In this case, the first-order approximations of both estimators' expectations equal the true prevalence; in the language of Feehan and Salganik (2016) and other literature, both the dRpR and dRpA estimators are essentially unbiased.

Thus we have expressions for the bias and variance of each estimator under a general two-group stochastic block model when degrees are estimated using a single probe group $K \subset L$. However, we would like to be able to characterize the regions of parameter space in which each estimator performs better than the other, and interpretation is difficult with this many parameters. Therefore we now analyze a slightly simpler case: Fix $p_{HH} = p_{LL} = ap_{HL}$ for some $0 < a < \infty$. Notice that $a = 1$ corresponds to the Erdös-Renyi case, $a > 1$ corresponds to assortativity, and $a < 1$ corresponds to dissortativity. This reduces the number of degrees of freedom in the parameters by one, since we can characterize the three link probabilities with just the two parameters a and p_{HL}, the latter of which we will now denote simply by p.

The biases are now a function only of a and R:

$$\mathrm{Bias}\left(\hat{R}_{\mathrm{dRpR}}\right)(a,R) \to R\left[\frac{(a-1)(2R-1)}{(1-R)a+R}\right] = \begin{cases} > 0 & \{a>1\}\cap\{R>0.5\} \text{ or } \{a<1\}\cap\{R<0.5\}, \\ = 0 & \{a=1\}\cup\{R=0.5\}, \\ < 0 & \text{else}, \end{cases}$$

$$\text{Bias}\left(\hat{R}_{\text{dRpA}}\right)(a,R) \to R\left[\frac{(a-1)\big[(a+1)R-1\big]}{a}\right]$$
$$= \begin{cases} > 0 & \{a>1\}\cap\{a>\frac{1-R}{R}\} \text{ or } \{a<1\}\cap\{a<\frac{1-R}{R}\}, \\ = 0 & \{a=1\}\cup\{a=\frac{1-R}{R}\}, \\ < 0 & \text{else.} \end{cases}$$

The dRpR is unbiased if and only if the Erdös-Renyi case holds ($a = 1$) or the hidden population prevalence is exactly 50%. The dRpA is unbiased if and only if the Erdös-Renyi case holds ($a = 1$) or $a = (1 - R)/R$. It is unlikely that the parameters take these exact values such that the estimators are exactly unbiased, but these conditions serve as boundary cases to define regions in which one estimator or the other has smaller bias.

Now we approximate the variances. Defining $r_K = N_K/N$, the dependence simplifies to effectively five parameters: R, a, p, r_K, and nN, since the dependence on sample size n and population size N is only through their product.

$$\text{Var}\left(\hat{R}_{\text{dRpR}}\right) \to \frac{R}{nNp}\left[\frac{(R+(1-R)a)^2\left(Ra+(1-R)-\left[Ra^2+(1-R)\right]p\right)}{\left[R+(1-R)a\right]^4} + \frac{R}{r_K}\frac{(Ra+(1-R))^2\left(R+(1-R)a-\left[R+(1-R)a^2\right]p\right)}{\left[R+(1-R)a\right]^4}\right]$$
$$\text{Var}\left(\hat{R}_{\text{dRpA}}\right) \to \frac{R}{nNp}\left(pa(1-a)R+(1-p)a+\frac{1}{r_K}\left[\big([1-p]a^2+pa-1\big)R^2+(1-pa)R\right]\right)$$

To characterize the conditions under which each estimator outperforms the other, we evaluate the approximate bias and variance expressions above over a range of parameter values and visualize the regions in which each estimator has lower bias, lower variance, and lower RMSE. In the top row of Figure 1 we show the results for $\log(a)$ ranging from -4 to 4 in increments of 0.1, and R ranging from 0.01 to 0.99 in increments of 0.02. The bottom row shows the results restricted to assortative cases with small R, with $\log(a)$ ranging from 0 to 4 in increments of 0.05, and R ranging from 0.001 to 0.1 in increments of 0.001. In both

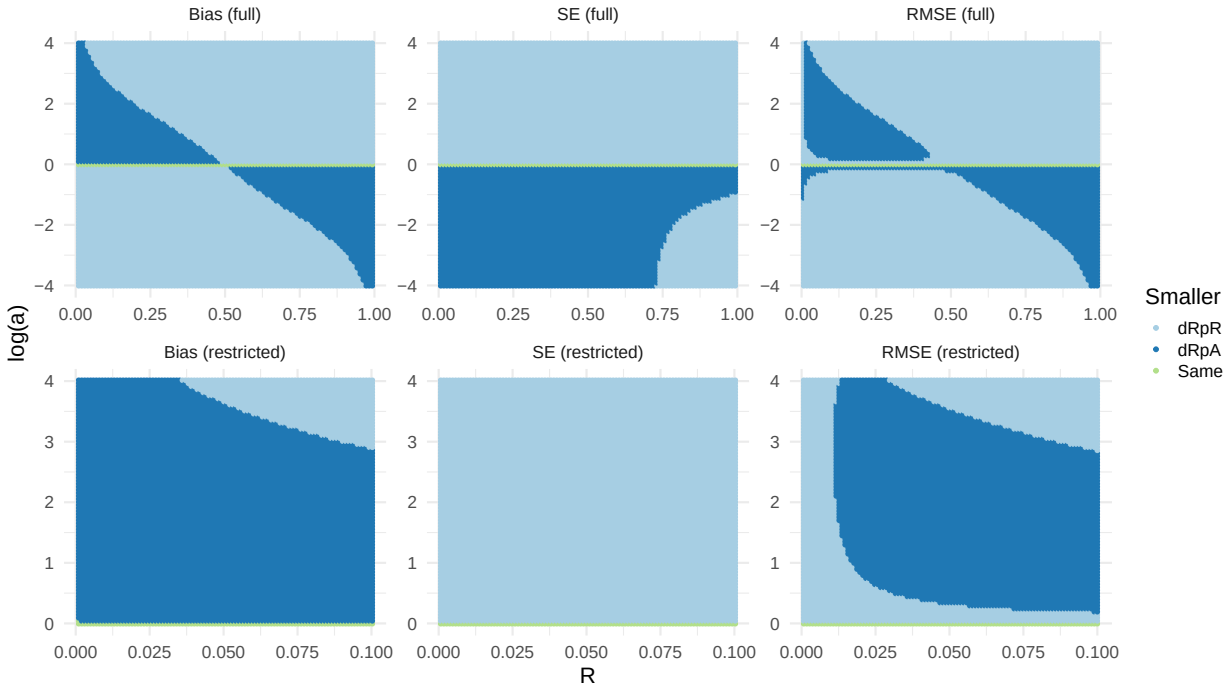

Figure 1: Comparing estimator bias (left panel), variance (center panel), and RMSE (right panel) as a function of a and R. The darkest regions indicate the combinations of a and R for which the dRpA has the lower value of the quantity examined in that subplot. Note the log scale on a, such that the x-axis corresponds to the Erdös-Renyi case $a = 1$ in which the two estimators have the same bias and variance. Under assortativity and for prevalences less than 10 or 20%, the dRpA generally has smaller bias and RMSE than the dRpR. Here, $p = 0.01$, $nN = 500,000$, and $r_K = 0.1$. The top row shows results over a wider range of parameter values while the bottom row shows results for a smaller range of parameter values thought to reflect most practical settings, namely $a > 1$ (assortative) and prevalence smaller than 10%. Under these conditions, the dRpA generally has smaller bias and RMSE than the dRpR.

these cases we fix $p = 0.01$, $nN = 500,000$, and $r_K = 0.1$ since the variance changes with these parameters.

The leftmost panels of Figure 1 compare the bias of the two estimators as a function of a and R; the darkest shaded region is the region in which the dRpA has smaller bias in magnitude. We have already shown the two estimators have the same bias and variance and therefore the same RMSE for $a = 0$, which is the x-axis in these figures. In practice, we typically expect assortativity ($a > 1$) and a fairly small value of R, probably less than 10 or 20%; in this region of the parameter space (for any values of the other parameters, since the estimator biases depend only on a and R), the dRpA has smaller bias.

The middle panels of Figure 1 illustrate that the variance is smaller for the dRpR under assortativity and generally smaller for the dRpA under dissortativity. The rightmost panels comparing the RMSE of the two estimators resemble the leftmost panels comparing the bias. Therefore, in the settings most likely to be practically relevant, and for sample sizes and population sizes likely to be realistic, the dRpA tends to have smaller RMSE. For assortative settings with small prevalence R, the dRpA has lower RMSE. When the assortativity is weaker, the dRpA has lower RMSE for a wider range of R, and under stronger assortativity the upper bound on R for this region shrinks.

Therefore, although the dRpR is the most commonly used estimator in the literature, the dRpA often has lower bias and RMSE in the presence of barrier effects.

The relative importance of the bias and variance in the RMSE are determined by nN, r_K, and p. The region of parameter space in which the dRpA has lower RMSE increases with nN and with p and decreases with r_K; see the online supplement for figures and additional details. The size of this region depends more strongly on nN and p than on r_K.

5 Facebook 100 data example

We also conduct simulations using an example data set, and the results are consistent with the analytical results in Section 4 even though we use multiple probe groups to estimate the degrees.

For these simulations we use the Facebook 100 data set, which consists of the intra-school links in the September 2005 Facebook networks of 100 colleges and universities (Traud et al., 2011, 2012). The networks range in size from 769 to 41,554 nodes, typically fewer than 25,000. Similar to Feehan et al. (2022), we create candidate probe groups from the following five variables: status (such as faculty or student), gender, year, dorm, and major. We treat them as categorial variables, with an indicator for each level of each variable, and each indicator variable whose prevalence in that school network is 0.1-10% of the population is a candidate probe group for that school network. We compare the bias, variance, and RMSE of the dRpA, dRpR, and dApA.

For each school network, we select the 16 largest candidate probe groups to constitute one hidden group and 15 probe groups used in estimating respondents' degrees; we will refer to a choice of school network and hidden group as a case. For each of the 100 schools there are 16 ways to choose one of the 16 largest groups to be the hidden group, resulting in 1,600 cases total. For each case, we draw 500 "survey" samples of 500 people each using simple random sampling without replacement to represent our survey respondents, and for each survey sample we compute the dRpR, dRpA, and dApA estimates for that case. We approximate the mean and variance of each estimator for each case as the sample mean and sample standard deviation across the 500 surveys for that case, then use this to compute the estimated bias and RMSE for each estimator.

We categorize the cases based on whether the degree ratio is "low" (< 0.8, 205 cases), "high" (> 1.2, 378 cases), or near 1 (between 0.8 and 1.2, 1017 cases). The assortativity coefficient of each case ranges from -0.09 to 0.88, with first, second, and third quartiles of 0.05, 0.19, and 0.35. All the cases with low degree ratios are assortative.

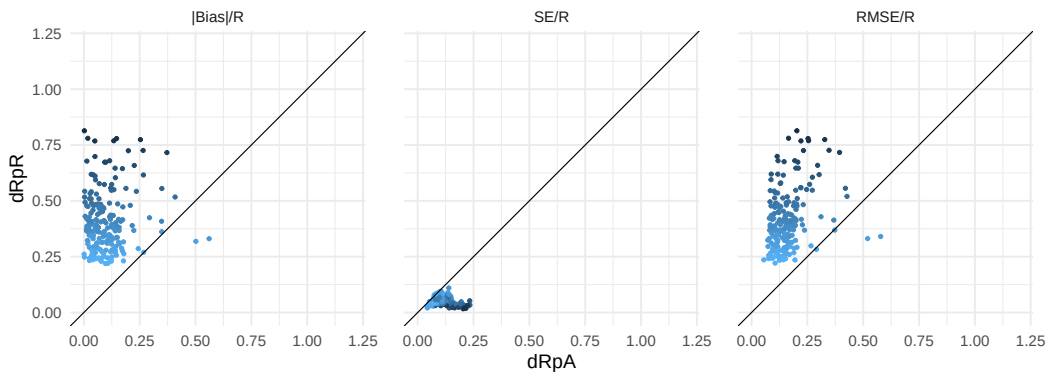

Figure 2: Comparing the absolute-value bias (left panel), standard error (center panel), and RMSE (right panel), all standardized by the true prevalence, of the dRpR and the dRpA in the 205 "low" degree ratio cases (< 0.8, with darker points representing lower degree ratios) from the Facebook 100 simulations. The diagonal line is the one-to-one line; points above the line have lower values for the dRpA than the dRpR. Each point represents the average across 500 surveys of size 500 for one combination of school network and hidden group. Among these cases, the dRpA has almost strictly smaller bias, larger variance, and smaller RMSE.

These simulations result in some degree estimates equal to zero. As a result, the dRpA and dApA sample estimates for a given survey are NaN when both the numerator (response) and denominator (estimated degree) are zero for a given person and Inf when only the estimated degree is zero. For each case (combination of school network and choice of hidden group from among the 16 largest groups), we exclude the NaN and Inf survey estimates in computing the mean and standard deviation of the survey estimates, reducing the effective sample size.

Figure 2 compares the absolute value of the bias, standard error, and RMSE of the dRpR and dRpA estimators for the cases with low degree ratios, thought to be more relevant to hidden population settings. We standardize these metrics by the true prevalence of each case since the true prevalence varies widely across cases. The dRpA has lower bias and RMSE than the dRpR in all except two of these 205 cases, which have degree ratios 0.7 and 0.72. Across the low degree ratio cases, the dRpA has 0.1-56% error while the dRpR has 22-81% error.

Figures comparing the dRpR and dRpA estimators for the cases with near-one and high

degree ratios are provided in the online supplement, but we provide some comments here. We find that when the degree ratio is near one, the dRpA and dRpR have comparable bias and RMSE and the dRpR tends to have smaller variance. For cases with high degree ratios, the dRpA tends to have lower bias and RMSE than the dRpR and their variance is comparable. Therefore, researchers that are confident the degree ratio is near 1 for the hidden and general population of interest to them may want to use the dRpR, but otherwise the dRpA may generally have lower RMSE.

Figure 3 illustrates that in these Facebook 100 simulations, the dRpA tends to have smaller bias, variance, and RMSE than the dApA estimator regardless of degree ratio. Recall that the dApA estimator uses the same size estimator as the dRpA but uses the average of ratios for the degree estimator as well (Table 1). This choice increases not only the variance but the bias. We suspect the increased bias follows from the fact that the probe groups are intended to be collectively but not necessarily individually representative of the general population (McCormick et al., 2010). If the set of chosen probe groups satisfy this property, then taking the ratio of averages keeps the numerator and denominator representative while taking the average of ratios changes the relative weighting of the probe groups.

6 Discussion

We have presented theoretical and empirical evidence that over what seems likely to be a realistic range of true prevalence, sample size, and population size, the dRpA estimator often has lower bias and RMSE than the dRpR estimator under a two-block stochastic block model with assortativity. In other words, using the average-of-ratios size estimator and the ratio-of-averages degree estimator often has lower bias and RMSE than using the ratio of averages for both the degree and size estimators. We have also shown empirical evidence that the dRpA estimator has lower bias, RMSE, and variance than the dApA estimator, suggesting that using the average-of-ratios size estimator with the ratio-of-averages degree estimator is

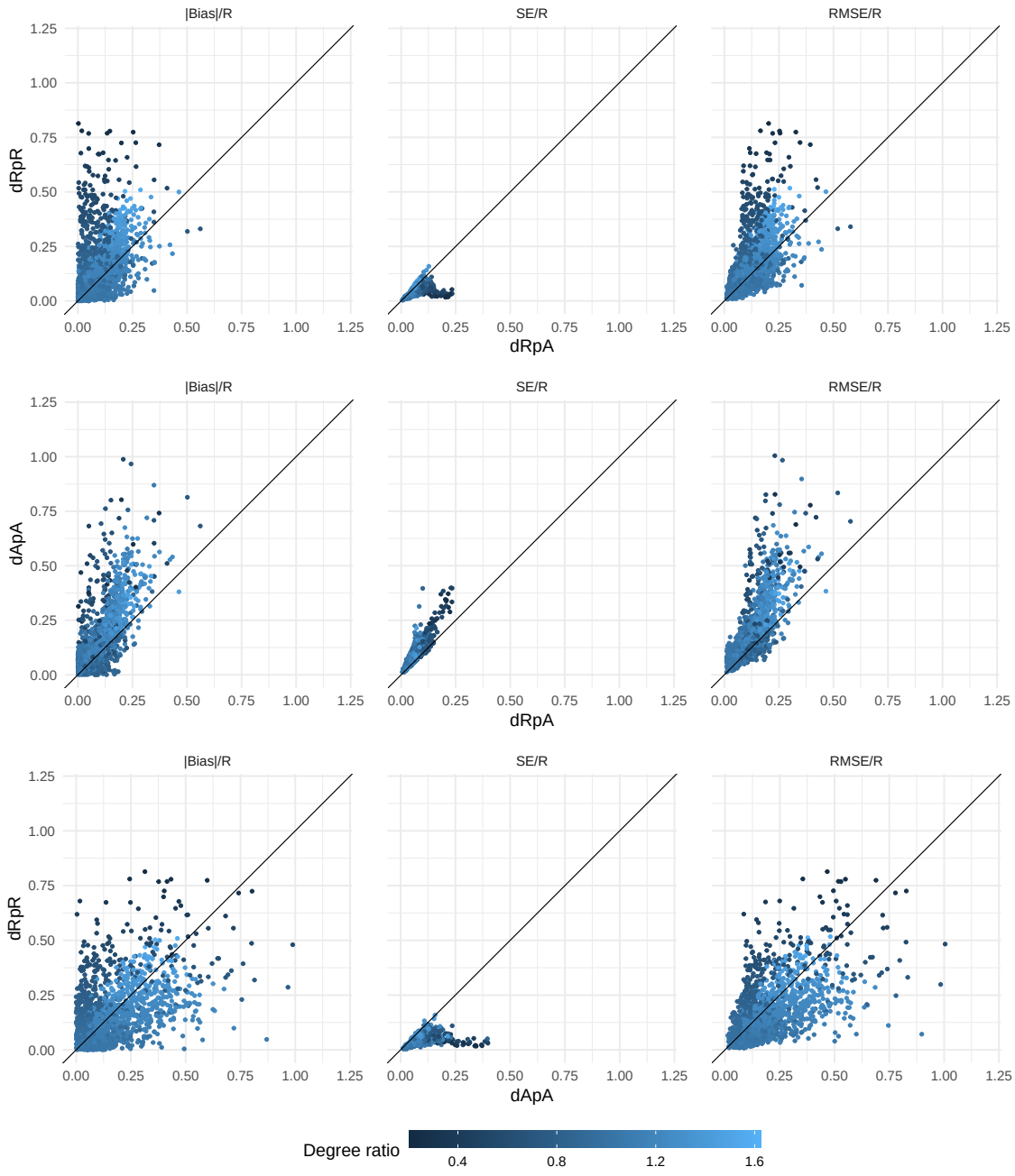

Figure 3: As in Figure 2 except comparing all three estimators for all 1,600 cases. The dRpA tends to have lower bias and RMSE than both the other estimators and lower variance than the dApA.

better than using the average-of-ratios size estimator for both the size and degree estimators. We have provided some reasoning for these findings.

The evidence in favor of one estimator over another still depends on the range of parameter values expected to be practically relevant. The more we know about the typical network and sample size, assortativity, approximate prevalence, typical degree sizes, and the fraction of the population that is in one or more of the probe groups in hard-to-reach population studies, the better we can constrain the relative performance of the estimators. Additionally, the Facebook 100 network data do not represent subpopulations we expect to be hidden. Comparing the robustness of these estimators on additional data sets beyond our initial example, particularly using network data sets involving hidden populations, would provide more direct evidence for evaluating these estimators.

Generalized NSUM and other complex methods are likely to be preferable to these simple estimators when feasible. However, in studies that use one of these simple estimators without any corrections, the dRpA may be the most robust choice. When the necessary data is available, it may be easier to correct bias in the dRpR estimator than the dRpA since the dRpR bias can be expressed directly in terms of the degree ratio; however, this requires having reliable estimates of the degree ratio. When ad hoc corrections to the simple estimators are applied to the dRpR, it may be helpful to also compute the dRpA to help bound the result; it is easily done and does not require additional data or computational power.

Additionally, these results raise the question whether it may be more robust to develop methods based on the dRpA instead of the dRpR. This may depend on how successfully we can characterize the bias of the dRpA and correct for it. It may turn out that we can more readily understand and correct for the bias of the dRpR.

Hidden populations are often studied in the context of estimating the impact of a social or epidemiological concern; for instance, estimating the number of people who have experienced labor trafficking is part of an effort to understand, intercept and prevent trafficking.

Studies may benefit from using RDS and related methods to learn about possible causes and interventions, while using NSUM to do more frequent monitoring. For example, researchers and monitoring agencies could use NSUM regularly to estimate how many people are being trafficked over time, and implement less frequent RDS studies to both validate the NSUM estimates and learn from people who have been trafficked how they entered their situation, how they were able to leave, or what prevented them.

Supplementary Materials

Section S1 contains proofs that the average-of-ratios degree estimators and prevalence estimators have greater variance than their ratio-of-averages counterparts under the binomial model. Section S2 presents the results of our initial literature search of current NSUM practice. Section S3 demonstrates that the binomial model approximates the Erdös-Renyi model; this is not a new result but is included here for completeness and convenience. Section S4 contains additional figures referenced in the body of the paper that pertain to either (a) the analytical results using a single probe group and estimated degrees or (b) the Facebook 100 data example.

Acknowledgements

The authors would like to thank Shane Lubold for thoughtful comments and feedback on this work.

A Derivations

$$E(Y_{iH}) = \begin{cases} N_H p_{HH} & i \in H \\ N_H p_{HL} & i \in L \end{cases} \qquad E(Y_{iK}) = \begin{cases} N_K p_{HL} & i \in H \\ N_K p_{LL} & i \in L \end{cases}$$

$$\mathrm{Var}(Y_{iH}) = \begin{cases} N_H p_{HH}(1-p_{HH}) & i \in H \\ N_H p_{HL}(1-p_{HL}) & i \in L \end{cases} \qquad \mathrm{Var}(Y_{iK}) = \begin{cases} N_K p_{HL}(1-p_{HL}) & i \in H \\ N_K p_{LL}(1-p_{LL}) & i \in L \end{cases}$$

$$\sum_{i=1}^{n} E(Y_{iH}) = N_H \left(n_H p_{HH} + n_L p_{HL}\right) \qquad \sum_{i=1}^{n} E(Y_{iK}) = N_K \left(n_H p_{HL} + n_L p_{LL}\right)$$

$$\sum_{i=1}^{n} \mathrm{Var}(Y_{iH}) = N_H \left(n_H p_{HH}(1-p_{HH}) + n_L p_{HL}(1-p_{HL})\right)$$

$$\sum_{i=1}^{n} \mathrm{Var}(Y_{iK}) = N_K \left(n_H p_{HL}(1-p_{HL}) + n_L p_{LL}(1-p_{LL})\right)$$

First-order Taylor approximations for the mean and variance of a ratio of random variables are given by $E(A/B) \approx E(A)/E(B)$ and, if A and B are independent, $\mathrm{Var}(A/B) \approx [E(B)^2\mathrm{Var}(A) + E(A)^2\mathrm{Var}(B)]/[E(B)^4]$. For handling more than one probe group and allowing probe groups and the hidden group to overlap with each other, one can either assume approximate independence or include the covariance term in the variance approximation; for now we consider the case of a single probe group K that is disjoint from H.

Taking the expectation over the SBM superpopulation for a given sample,

$$\begin{aligned} E\left(\hat{R}_{\mathrm{dRpR}}\right) &= E\left(\frac{N_K}{N}\frac{\sum_{i=1}^{n} Y_{iH}}{\sum_{i=1}^{n} Y_{iK}}\right) \\ &\approx \frac{N_K}{N}\frac{E\left(\sum_{i=1}^{n} Y_{iH}\right)}{E\left(\sum_{i=1}^{n} Y_{iK}\right)} && \text{1st order Taylor approximation} \\ &= \frac{N_K}{N}\frac{\sum_{i=1}^{n} E\left(Y_{iH}\right)}{\sum_{i=1}^{n} E\left(Y_{iK}\right)} \\ &= \frac{N_K}{N}\frac{n_H N_H p_{HH} + n_L N_H p_{HL}}{n_H N_K p_{HL} + n_L N_K p_{LL}} \\ &= \frac{N_H}{N}\frac{n_H p_{HH} + n_L p_{HL}}{n_H p_{HL} + n_L p_{LL}} \end{aligned}$$

$$= R\frac{n_H p_{HH} + n_L p_{HL}}{n_H p_{HL} + n_L p_{LL}}.$$

Similarly, for the dRpA estimator,

$$\begin{aligned}
E\left(\hat{R}_{\text{dRpA}}\right) &= E\left(\frac{N_K}{N}\frac{1}{n}\sum_{i=1}^{n}\frac{Y_{iH}}{Y_{iK}}\right) \\
&= \frac{N_K}{N}\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{Y_{iH}}{Y_{iK}}\right) \\
&\approx \frac{N_K}{N}\frac{1}{n}\sum_{i=1}^{n}\frac{E\left(Y_{iH}\right)}{E\left(Y_{iK}\right)} \\
&= \frac{N_K}{N}\left[\frac{n_H}{n}\frac{N_H p_{HH}}{N_K p_{HL}} + \frac{n_L}{n}\frac{N_H p_{HL}}{N_K p_{LL}}\right] \\
&= R\left[\frac{n_H}{n}\frac{p_{HH}}{p_{HL}} + \frac{n_L}{n}\frac{p_{HL}}{p_{LL}}\right].
\end{aligned}$$

If $n_H/n \to R$, as in simple random sampling without replacement, then

$$\begin{aligned}
E\left(\hat{R}_{\text{dRpR}}\right) &\to R\frac{Rp_{HH} + (1-R)p_{HL}}{Rp_{HL} + (1-R)p_{LL}}, \\
E\left(\hat{R}_{\text{dRpA}}\right) &\to R\left[R\frac{p_{HH}}{p_{HL}} + (1-R)\frac{p_{HL}}{p_{LL}}\right].
\end{aligned}$$

$$\begin{aligned}
\text{Var}\left(\hat{R}_{\text{dRpR}}\right) &= \text{Var}\left(\frac{N_K}{N}\frac{\sum_{i=1}^{n} Y_{iH}}{\sum_{i=1}^{n} Y_{iK}}\right) \\
&= \frac{N_K^2}{N^2}\text{Var}\left(\frac{\sum_{i=1}^{n} Y_{iH}}{\sum_{i=1}^{n} Y_{iK}}\right) \\
&\approx \frac{N_K^2}{N^2}\cdot\frac{E(\sum_{i=1}^{n} Y_{iK})^2\text{Var}(\sum_{i=1}^{n} Y_{iH}) + E(\sum_{i=1}^{n} Y_{iH})^2\text{Var}(\sum_{i=1}^{n} Y_{iK})}{E(\sum_{i=1}^{n} Y_{iK})^4} \qquad (2) \\
&= \frac{N_K^2}{N^2}\cdot\frac{\left(\sum_{i=1}^{n} E[Y_{iK}]\right)^2\sum_{i=1}^{n}\text{Var}(Y_{iH}) + \left(\sum_{i=1}^{n} E[Y_{iH}]\right)^2\sum_{i=1}^{n}\text{Var}(Y_{iK})}{\left(\sum_{i=1}^{n} E[Y_{iK}]\right)^4} \qquad (3) \\
&= \frac{N_K^2}{N^2}\Bigg(\frac{N_K^2\left(n_H p_{HL} + n_L p_{LL}\right)^2 N_H\left[n_H p_{HH}(1-p_{HH}) + n_L p_{HL}(1-p_{HL})\right]}{N_K^4\left(n_H p_{HL} + n_L p_{LL}\right)^4} +
\end{aligned}$$

$$\left.\frac{N_H^2\left(n_H p_{HH}+n_L p_{HL}\right)^2 N_K\left[n_H p_{HL}(1-p_{HL})+n_L p_{LL}(1-p_{LL})\right]}{N_K^4\left(n_H p_{HL}+n_L p_{LL}\right)^4}\right)$$

$$= R\Bigg(\frac{1}{N}\frac{\left(n_H p_{HL}+n_L p_{LL}\right)^2\left[n_H p_{HH}(1-p_{HH})+n_L p_{HL}(1-p_{HL})\right]}{\left(n_H p_{HL}+n_L p_{LL}\right)^4}+$$

$$\frac{R}{N_K}\frac{\left(n_H p_{HH}+n_L p_{HL}\right)^2\left[n_H p_{HL}(1-p_{HL})+n_L p_{LL}(1-p_{LL})\right]}{\left(n_H p_{HL}+n_L p_{LL}\right)^4}\Bigg).$$

Step (2) above uses the Taylor approximation and Step (3) holds if Y_{iG}, Y_{jG} are independent for $i \neq j$, $G = H, K$. A similar derivation for the dRpA yields

$$\mathrm{Var}\left(\hat{R}_{\mathrm{dRpA}}\right) \approx \frac{R}{n^2 N p_{HL}^2}\left[n_H p_{HH}(1-p_{HH})+n_L p_{LL}(1-p_{HL})\right]+$$
$$\frac{R^2}{n^2 N_K p_{HL}^3}\left[n_H p_{HH}^2(1-p_{HL})+n_L p_{HL}^2(1-p_{LL})\right].$$

If $n_H/n \to R$, then

$$\mathrm{Var}\left(\hat{R}_{\mathrm{dRpR}}\right) \to \frac{R}{nN}\frac{\left(Rp_{HL}+(1-R)p_{LL}\right)^2\left[Rp_{HH}(1-p_{HH})+(1-R)p_{HL}(1-p_{HL})\right]}{\left(Rp_{HL}+(1-R)p_{LL}\right)^4}+$$
$$\frac{R^2}{nN_K}\frac{\left(Rp_{HH}+(1-R)p_{HL}\right)^2\left[Rp_{HL}(1-p_{HL})+(1-R)p_{LL}(1-p_{LL})\right]}{\left(Rp_{HL}+(1-R)p_{LL}\right)^4}$$

$$\mathrm{Var}\left(\hat{R}_{\mathrm{dRpA}}\right) \to \frac{R}{nNp_{HL}^2}\left[Rp_{HH}(1-p_{HH})+(1-R)p_{LL}(1-p_{HL})\right]+$$
$$\frac{R^2}{nN_K p_{HL}^3}\left[Rp_{HH}^2(1-p_{HL})+(1-R)p_{HL}^2(1-p_{LL})\right].$$

References

Bernard, H. R., T. Hallett, A. Iovita, E. C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M. J. Salganik, T. Saliuk, O. Scutelniciuc, G. A. Shelley, P. Sirinirund, S. Weir, and D. F. Stroup (2010). Counting hard-to-count populations: the network scale-up method for public health. *Sexually Transmitted Infections 86*(Suppl. 2), 1368–4973.

Bernard, H. R., E. C. Johnsen, P. D. Killworth, and S. Robinson (1991). Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Social Science Research 20*(2), 109–121.

Cheng, S., D. J. Eck, and F. W. Crawford (2020). Estimating the size of a hidden finite set: Large-sample behavior of estimators. *Statistics Surveys 14*(none), 1–31.

Crawford, F. W., J. Wu, and R. Heimer (2018). Hidden population size estimation from respondent-driven sampling: A network approach. *Journal of the American Statistical Association 113*(522), 755–766. PMID: 30828120.

Feehan, D. M. and M. J. Salganik (2016). Generalizing the network scale-up method: A new estimator for the size of hidden populations. *Sociological Methodology 46*(1), 153–186.

Feehan, D. M., V. H. Son, and A. Abdul-Quader (2022). Survey methods for estimating the size of weak-tie personal networks. *Sociological Methodology 52*(2), 193–219.

Habecker, P., K. Dombrowski, and B. Khan (2015, 12). Improving the network scale-up estimator: Incorporating means of sums, recursive back estimation, and sampling weights. *PLOS ONE 10*(12), 1–16.

Handcock, M. S., K. J. Gile, and C. M. Mar (2014). Estimating hidden population size using respondent-driven sampling data. *Electronic Journal of Statistics 8*(1), 1491–1521.

Killworth, P. D., E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard (1998). A social network approach to estimating seroprevalence in the United States. *Social Networks 20*, 23–50.

Killworth, P. D., C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen (1998). Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach. *Evaluation Review 22*(2), 289–308.

Laga, I., L. Bao, and X. Niu (2021). Thirty years of the network scale-up method. *Journal of the American Statistical Association 116*(535), 1548–1559.

Maltiel, R., A. E. Raftery, T. H. McCormick, and A. J. Baraff (2015). Estimating population size using the network scale up method. *The Annals of Applied Statistics 9*(3), 1247–1277.

McCormick, T. H. (2021, 01). The Network Scale-Up Method. In *The Oxford Handbook of Social Networks*. Oxford University Press.

McCormick, T. H., M. J. Salganik, and T. Zheng (2010). How many people do you know?: Efficiently estimating personal network size. *Journal of the American Statistical Association 105*(489), 59–70. PMID: 23729943.

McCormick, T. H. and T. Zheng (2012). Latent demographic profile estimation in hard-to-reach groups. *The Annals of Applied Statistics 6*(4), 1795.

McCormick, T. H. and T. Zheng (2015). Latent surface models for networks using aggregated relational data. *Journal of the American Statistical Association 110*(512), 1684–1695.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology 27*(1), 415–444.

Olhede, S. C. and P. J. Wolfe (2014). Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences 111*(41), 14722–14727.

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

Shelley, G. A., H. R. Bernard, P. Killworth, E. Johnsen, and C. McCarty (1995). Who knows your HIV status? What HIV+ patients and their network members know about each other. *Social Networks 17*(3), 189–217.

Traud, A. L., E. D. Kelsic, P. J. Mucha, and M. A. Porter (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM Review 53*(3), 526–543.

Traud, A. L., P. J. Mucha, and M. A. Porter (2012). Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications 391*(16), 4165–4180.

Supplement: Comparing the Robustness of Simple Network Scale-Up Method (NSUM) Estimators

Jessica P. Kunke[1]*, Ian Laga[2], Xiaoyue Niu[3], Tyler H. McCormick[1]

[1] *University of Washington,* [2] *Montana State University,*
[3] *Pennsylvania State University*

March 15, 2023

Section S1 contains proofs that the average-of-ratios degree estimators and prevalence estimators have greater variance than their ratio-of-averages counterparts under the binomial model. Section S2 presents the details of our initial literature search to understand what NSUM approaches are currently being used in practice. Section S3 demonstrates that the binomial model approximates the Erdös-Renyi model; this is not a new result but is included here for completeness and convenience. Section S4 contains additional figures referenced in the body of the paper that pertain to either (a) the analytical results using a single probe group and estimated degrees or (b) the Facebook 100 data example.

S1 Variance of AoR and RoA estimators

The two degree estimators under consideration are as follows:

$$\hat{d}_{i,\text{RoA}} = N \cdot \frac{\sum_j y_{ij}}{\sum_j N_j}, \quad \hat{d}_{i,\text{AoR}} = N \cdot \frac{1}{K} \sum_{j=1}^{K} \frac{y_{ij}}{N_j}.$$

In the unlikely case that the probe groups are all the same size, then the two estimators are identical and therefore also have the same variance. We explore how the variances compare

*Correspondence: jkunke@uw.edu.

outside of this special case. Under the binomial model, $y_{ij} \sim \text{Binom}(d_i, N_j/N)$. Therefore,

$$\text{Var}\left(\hat{d}_{i,\text{RoA}}\right) = d_i \left[N \cdot \frac{1}{L} \frac{1}{\sum_j N_j/L} - \frac{\sum_j N_j^2}{\left(\sum_j N_j\right)^2} \right].$$

The second term within the brackets is less than $1/L$ because $N_j > 1$ for each probe group j. Therefore, if the average probe group size is smaller than $O(N)$, the first term dominates:

$$\text{Var}\left(\hat{d}_{i,\text{RoA}}\right) \approx d_i \left[N \cdot \frac{1}{L} \frac{1}{\sum_j N_j/L} \right] = \frac{d_i N}{L} \frac{1}{\text{mean}(N_j)}.$$

Following similar reasoning,

$$\text{Var}\left(\hat{d}_{i,\text{AoR}}\right) = d_i \left[\frac{N}{L} \frac{1}{L} \frac{1}{\sum_j N_j} - \frac{1}{L} \right] \approx \frac{d_i N}{L} \frac{1}{L} \frac{1}{\sum_j N_j} = \frac{d_i N}{L} \frac{1}{\text{hmean}(N_j)},$$

where $\text{hmean}(N_j)$ denotes the harmonic mean of the probe group sizes. Since $N_j > 0$ for all j, the harmonic mean is strictly smaller than the arithmetic mean, and therefore $\text{Var}\left(\hat{d}_{i,\text{RoA}}\right) < \text{Var}\left(\hat{d}_{i,\text{AoR}}\right)$ if the average probe group size is sufficiently smaller than N.

A similar but simpler argument demonstrates the same result for the two prevalence estimators (with fixed or known degrees) by comparing the arithmetic and harmonic means of the degrees:

$$\hat{R}_{\text{RoA}} = \frac{\sum_i y_i}{\sum_i d_i}, \quad \hat{R}_{\text{AoR}} = \frac{1}{n} \sum_i \frac{y_i}{d_i}.$$

Again, we ignore the trivial and impractical case in which all degrees are identical. Under the binomial model, $y_i \overset{\text{indep}}{\sim} \text{Binom}(d_i, R)$. Therefore,

$$\text{Var}\left(\hat{R}_{\text{RoA}}\right) = \frac{R(1-R)}{n} \cdot \frac{1}{\sum_i d_i/n} = \frac{R(1-R)}{n} \cdot \frac{1}{\text{mean}(d_i)},$$

$$\text{Var}\left(\hat{R}_{\text{AoR}}\right) = \frac{R(1-R)}{n} \cdot \frac{1}{n} \sum_i \frac{1}{d_i} = \frac{R(1-R)}{n} \cdot \frac{1}{\text{hmean}(d_i)}.$$

For $d_i > 0$, outside of the case that all degrees are equal, the harmonic mean is strictly

smaller than the arithmetic mean of the degrees. Therefore, $\text{Var}\left(\hat{R}_{\text{RoA}}\right) < \text{Var}\left(\hat{R}_{\text{AoR}}\right)$.

S2 Literature search of current NSUM practice

We conducted an initial literature search to examine which NSUM models applied researchers currently tend to use. We restricted our search to articles published in 2021 because we are interested in current trends. A Google Scholar search for English-text articles published in 2021 with the term "network scale-up" initially yielded 109 references. We further refined the results to articles that used or reviewed a specific model when discussing size or prevalence estimation for hidden populations. We excluded papers that talked only briefly about the general NSUM approach and only cited papers without referencing specific equations/models, as well as preprints, social network size estimates, and references for which we were unable to track down the paper itself. This procedure resulted in a final set of eight papers. None of the resulting references use the dRpA.

Table 1: Articles published in 2021 in journals or conference proceedings that either analyzed data using a NSUM model or discussed a specific NSUM model. Related papers are included in the same row, where the title refers to the first citation.

Citation	**Title**	**NSUM Models**
Jami et al. (2021a) Jami et al. (2021b)	Population Size Estimation of Drug Users in Isfahan City (Iran) Using Network Scale-up Method in 2018	dRpR; Scaling Factors

Balvardi et al. (2021)	Investigating the Prevalence of Substance Use Among Students of Medical Science Universities in the Eighth Macro-region of Iran	dRpR
Baneshi et al. (2021)	Estimating the Size of Hidden Groups	Basic; Scaling Factors
Baquero et al. (2021) Garcia-Agundez et al. (2021)	The CoronaSurveys System for COVID-19 Incidence Data Collection and Processing	dRpR
Ocagli et al. (2021)	Using Social Networks to Estimate the Number of COVID-19 Cases: The Incident (Hidden COVID-19 Cases Network Estimation) Study Protocol	Maltiel et al. (2015); Modified Version of Maltiel et al. (2015); References dRpR, but does not apply

Sakhno et al. (2021)	Estimating the size of key populations, bridge populations and other categories in Ukraine, 2020: the network scale up method	dRpR; Scaling Factors

S3 The binomial and Erdös-Renyi models

Given a set of nodes, a graph can be simulated from an Erdös-Renyi model by conducting an iid Bernoulli trial ℓ_{ij} for each pair of nodes to determine whether there is a link between them:

$$\ell_{ij} \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$$

Here we will suppose that the nodes have group memberships (either hidden or not hidden), but that these memberships do not impact link formation.

We can then derive random variables for the number of people each person i knows who are in the hidden group:

$$y_i = \sum_{j \in H, j \neq i} \ell_{ij} \sim \begin{cases} \text{Binom}\left(N_H,\ p\right) & i \notin H, \\ \text{Binom}\left(N_H - 1,\ p\right) & i \in H, \end{cases}$$

and the number of people each person i knows who are not in the hidden group:

$$z_i = \sum_{j \notin H, j \neq i} \ell_{ij} \sim \begin{cases} \text{Binom}\left(N - N_H - 1,\ p\right) & i \notin H, \\ \text{Binom}\left(N - N_H,\ p\right) & i \in H. \end{cases}$$

Let $N_H^* = N_H$ if $i \notin H$ and $N_H - 1$ if $i \in H$. Let $N_L^* = N - N_H - 1$ if $i \notin H$ and $N - N_H$ if $i \in H$. Note that $N_H^* + N_L^* = N - 1$ in either case, so $N_L^* = N - N_H^* - 1$. The distributions also simplify notationally to

$$y_i \sim \text{Binom}\left(N_H^*,\ p\right), \quad z_i \sim \text{Binom}\left(N_L^*,\ p\right).$$

These two variables are independent of one another for a given person, and their sum is that person's degree:

$$d_i = y_i + z_i.$$

Consider two people i and j. Their degrees are not completely independent because there is one potential link between them, so their degrees (sums of potential links) each include ℓ_{ij}. Additionally, their responses are not independent if they are both in the hidden group, because in that case y_i and y_j both correspond to sums that include ℓ_{ij}. Therefore, their conditional responses $y_i|d_i$ are not strictly independent of one another. However, for sufficiently large N and N_H this departure from independence is negligible.

The following derivation demonstrates that $y_i|d_i$ follows a hypergeometric distribution:

$$\begin{aligned}
p(y_i = y \mid d_i = d) &= \frac{p(y_i = y, d_i = d)}{p(d_i = d)} \\
&= \frac{p(y_i = y, z_i = d - y)}{p(d_i = d)} \\
&= \frac{p(y_i = y)p(z_i = d - y)}{p(d_i = d)} && y_i, z_i \text{ indep} \\
&= \frac{p(y_i = y)p(z_i = d - y)}{\sum_{k=0}^{d_i} p(d_i = d \mid y_i = k)p(y_i = k)} \\
&= \frac{p(y_i = y)p(z_i = d - y)}{\sum_{k=0}^{d} p(y_i = k)p(z_i = d - k)}
\end{aligned}$$

$$p(y_i = y)p(z_i = d - y) = \binom{N_H^*}{y}\binom{N_L^*}{d-y} p^y (1-p)^{N_H^* - y} p^{d-y} (1-p)^{N_L^* - d + y}$$

$$= \binom{N_H^*}{y}\binom{N_L^*}{d-y}p^d(1-p)^{N-1-d}$$

The only dependence on y appears in the binomial coefficients:

$$p(y_i = y \mid d_i = d) \propto \binom{N_H^*}{y}\binom{N_L^*}{d-y}.$$

Therefore, $y_i|d_i$ follows a hypergeometric distribution for the number y of successes out of d draws without replacement from a population of size $N-1$ containing N_H^* many successes.

For sufficiently large N and N_H, the hypergeometric distribution of $y_i|d_i$ under the Erdös-Renyi model converges to the binomial distribution of $y_i|d_i$ under the binomial model, and the approximate independence of $y_i|d_i$ and $y_j|d_j$ for different people i and j converges to independence. In both models, the response y_i can be interpreted as building person i's personal network by drawing one person simply at random from the population and counting how many people were drawn that were in the hidden population; however, these draws are modeled without replacement under the Erdös-Renyi model and with replacement under the binomial model. Modeling without replacement seems more natural since people should not be double-counted in a given person's personal network.

The binomial model is an approximation for the conditional distribution of ARD responses in networks generated from an Erdös-Renyi model, and the approximation improves as N and N_H increase.

S4 Additional figures

Figure 1 examines the dependence of the region in which dRpA has lower RMSE on the value of nN, the product of the sample size and the population size, when $p = 0.01$ and the other parameters are allowed to vary. For nN as small as five or ten thousand, the dRpA no longer has lower RMSE. The size of the region increases with nN, expanding to encompass smaller values of R and a wider range of a.

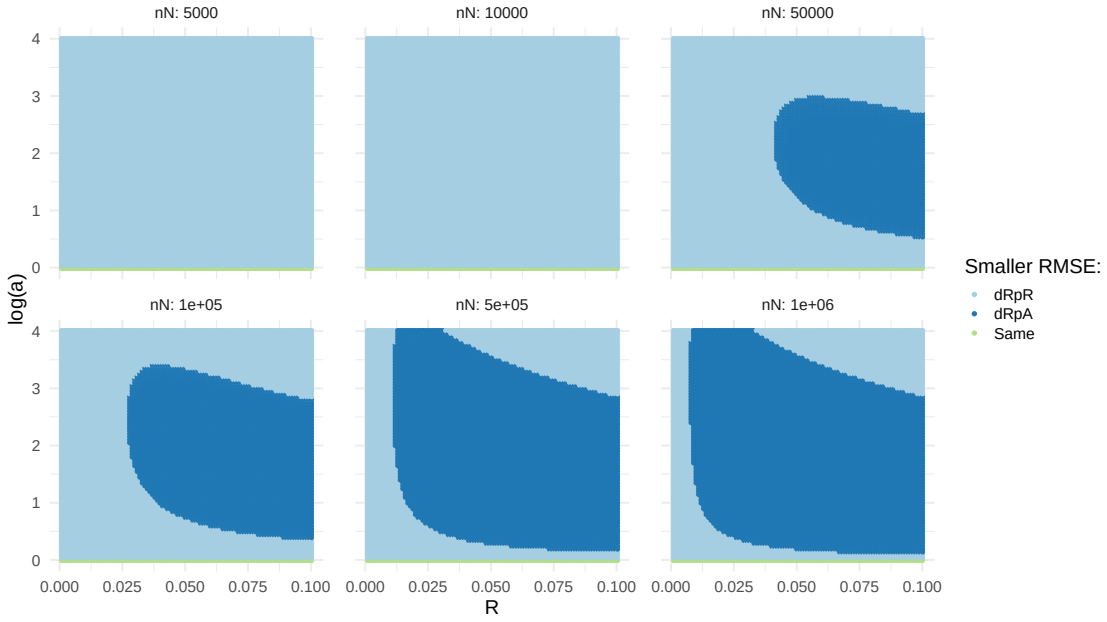

Figure 1: The six subpanels here correspond to six different values of nN, the product of the sample size and the population size. The size of the darkest region, the region in which the dRpA has smaller RMSE than the dRpR, grows with nN (larger sample sizes and larger populations). All assortative and Erdös-Renyi simulations for each value of nN are shown; R ranges from 0.01 to 0.99, $\log(a)$ ranges from 0 to 4, $p_{HL} = 0.01$, and r_K ranges from 0.01 to 0.8.

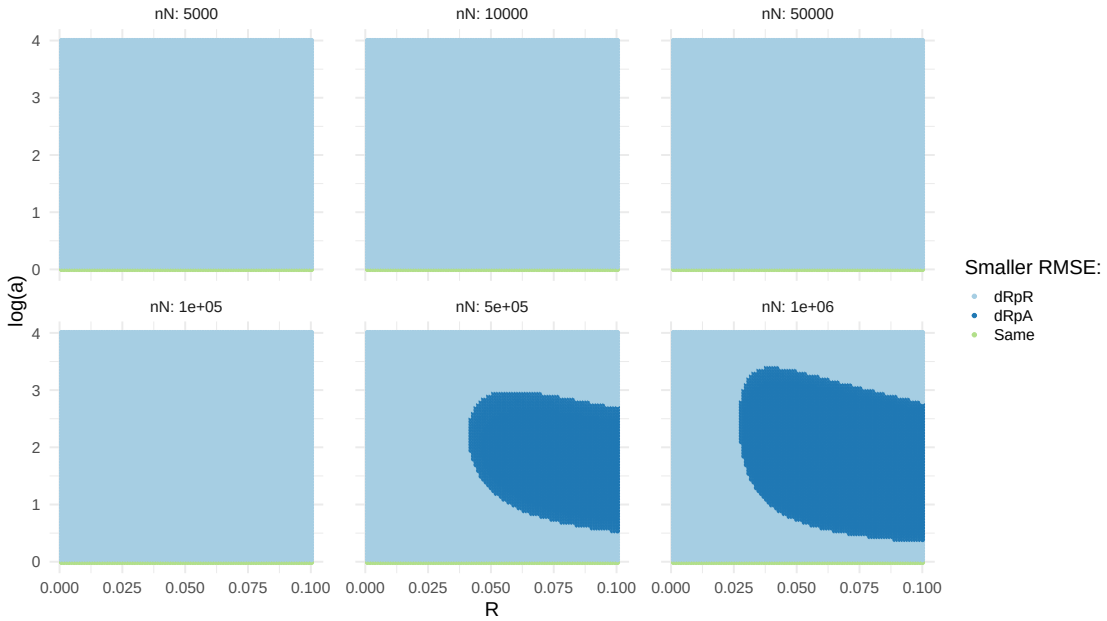

Figure 2: As in Figure 1 but with $p_{HL} = 0.001$ instead of 0.01. The size of the region in which the dRpA has smaller RMSE than the dRpR is smaller for smaller p.

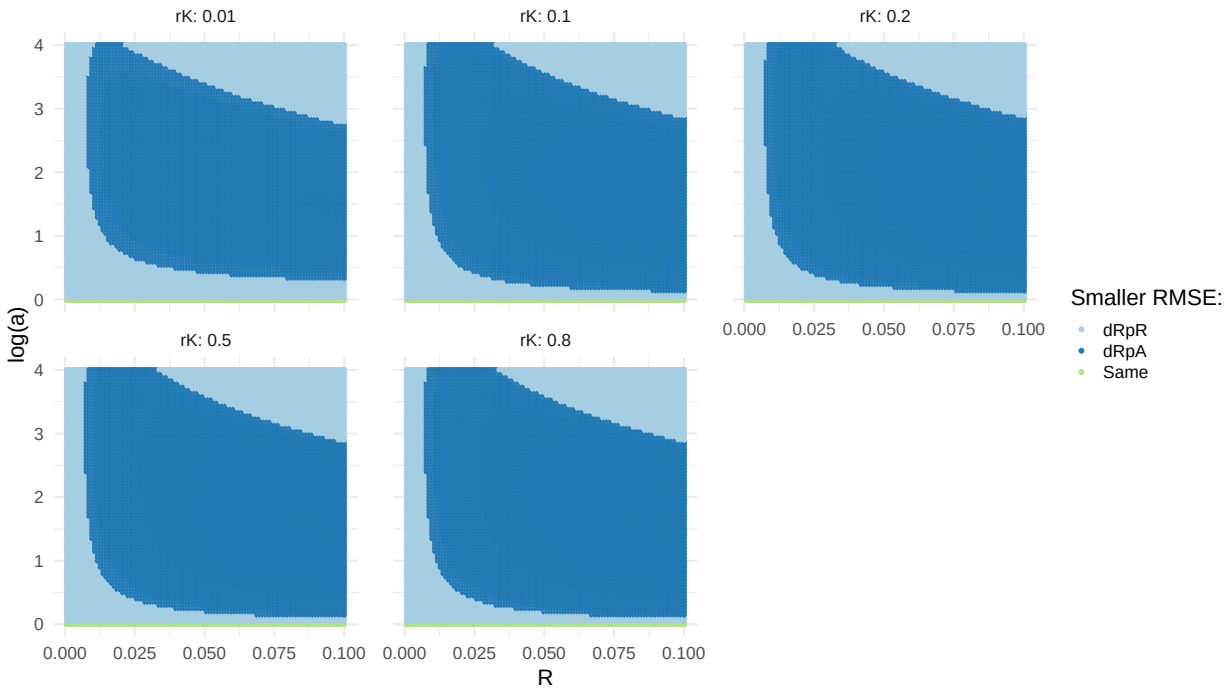

Figure 3: The size of the region in which the dRpA has smaller RMSE than the dRpR depends on r_K, the ratio of the probe group size to the population size, to a smaller extent than on nN. The five subpanels here correspond to five different values of r_K. All assortative and Erdös-Renyi simulations for each value of nN are shown; R ranges from 0.01 to 0.99, $\log(a)$ ranges from 0 to 4, $p_{HL} = 0.01$, and nN ranges from five thousand to one million.

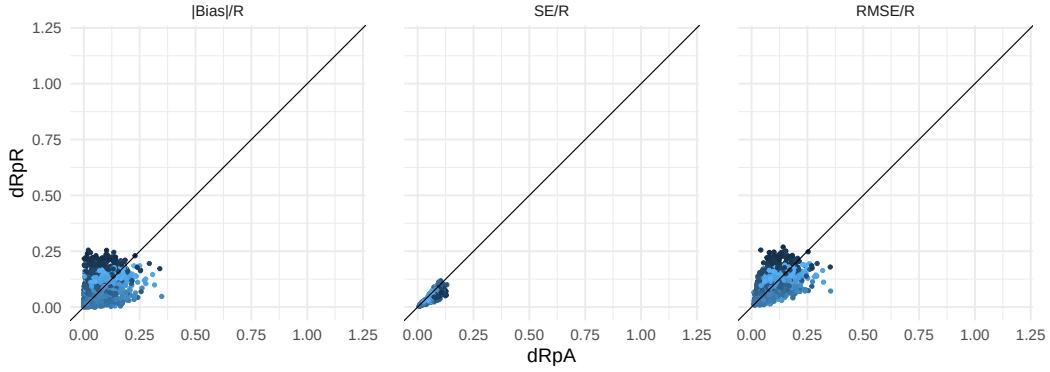

Figure 4: Comparing the bias (left panel), standard error (center panel), and RMSE (right panel) of the dRpR and the dRpA in the 1,017 cases with degree ratios close to 1 ($0.8 - 1.2$, with darker points representing lower degree ratios) from the Facebook 100 simulations. The diagonal line is the one-to-one line; points above the line have lower values for the dRpA than the dRpR. Each point represents the average across 500 surveys of size 500 for one combination of school network and hidden group. In these cases, the dRpA and dRpR have comparable bias and RMSE and the dRpR tends to have smaller variance.

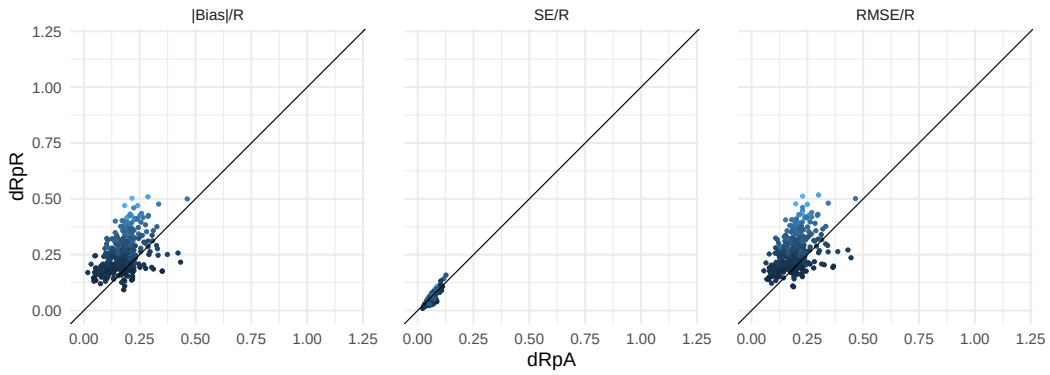

Figure 5: As in Figure 4 except for the 378 cases with "high" degree ratios (> 1.2). In this setting, the dRpA tends to have lower bias and RMSE than the dRpR.

Figure 2 shows the same results as Figure 1 except for a smaller value of p, 0.001 instead of 0.01. With smaller p, the region in which dRpA has smaller RMSE shrinks. Figure 3 shows that to a smaller extent, the size of this region also depends on the value of r_K, the ratio of probe group size to population size.

Figures 4 and 5 compare the dRpR and dRpA estimators for the cases with near-one and high degree ratios, respectively. When the degree ratio is near one, the dRpA and dRpR have comparable bias and RMSE and the dRpR tends to have smaller variance than the dRpA. For cases with high degree ratios, the dRpA tends to have lower bias and RMSE than the dRpR and their variance is comparable.

References

Balvardi, M., N. Dehdashti, Z. Imani-Goghary, M. Ghaljeh, H. Bashiri, K. Babaee, S. Daneshi, and M. Raei (2021). Investigating the prevalence of substance use among students of medical science universities in the eighth macro-region of Iran. *International Journal of High Risk Behaviors and Addiction 10*(4), e113237.

Baneshi, M. R., F. Zolala, S. Haji-Maghsoudi, M. Zamanian, A. A. Haghdoost, and A. Mirzazadeh (2021). Estimating the size of hidden groups. In *Methods in Epidemiology*, pp. 39–59. Springer.

Baquero, C., P. Casari, A. Fernandez Anta, A. García-García, D. Frey, A. Garcia-Agundez, C. Georgiou, B. Girault, A. Ortega, M. Goessens, et al. (2021). The CoronaSurveys system for COVID-19 incidence data collection and processing. *Frontiers in Computer Science 3*, 641237.

Garcia-Agundez, A., O. Ojo, H. A. Hernández-Roig, C. Baquero, D. Frey, C. Georgiou, M. Goessens, R. E. Lillo, R. Menezes, N. Nicolaou, et al. (2021). Estimating the COVID-19 prevalence in Spain with indirect reporting via open surveys. *Frontiers in Public Health 9*, 658544.

Jami, M. A., M. Baneshi, and M. Nasirian (2021a). Population size estimation of drug users in Isfahan City (Iran) using network scale-up method in 2018. *Addiction & Health 13*(4), 249.

Jami, M. A., M. Baneshi, and M. Nasirian (2021b). Population size estimation of high-risk behavior in Isfahan, Iran: Using the network scale-up method in 2018. *Journal of Biostatistics and Epidemiology 7*(2), 120–130.

Maltiel, R., A. E. Raftery, T. H. McCormick, and A. J. Baraff (2015). Estimating population size using the network scale up method. *The Annals of Applied Statistics 9*(3), 1247.

Ocagli, H., D. Azzolina, G. Lorenzoni, S. Gallipoli, M. Martinato, A. S. Acar, P. Berchialla, D. Gregori, I. S. Group, et al. (2021). Using social networks to estimate the number of COVID-19 cases: The INCIDENT (Hidden COVID-19 Cases Network Estimation) study protocol. *International Journal of Environmental Research and Public Health 18*(11), 5713.

Sakhno, Y., V. Paniotto, N. Kharchenko, O. Lyshtva, O. Kovtun, and T. Saliuk (2021). Estimating the size of key populations, bridge populations and other categories in Ukraine, 2020: The network scale up method. Technical report, ICF Alliance for Public Health.