Online Appendixes

$\widehat{\text{cv}}(\widehat{\bar{d}})$	source
0.05	Rwanda
0.10	Curitiba
0.02	US

Table E.2: Estimated coefficients of variation for the average degree from 3 different scale-up surveys. These play a role in the approximate relative bias for the estimate of $\widehat{\delta}_F$. Our approximation tells us that the larger these values are, the worse the relative bias will be. The estimates were computed using the rescaled bootstrap procedure.

	estimated coef. of variation
$\sum_{i \in s_H} y_{i,\mathcal{A} \cap F} / c\pi_i$	0.08
$\sum_{i \in s_H} \tilde{v}_{i,\mathcal{A} \cap F} / c\pi_i$	0.08
$\sum_{i \in s_H} 1 / c\pi_i$	0.06

Table E.3: Estimated coefficients of variation for quantities derived from a sample from the hidden population. These quantities play a role in the approximate relative bias for the estimate of all of the nonlinear estimators we propose. The estimates were computed using the respondent-driven sampling bootstrap procedure (Salganik, 2006).

F Variance estimation and confidence intervals

In addition to producing point estimates, researchers must also produce confidence intervals around their estimates. The procedure currently used by scale-up researchers begins with the variance estimator proposed in Killworth et al. (1998b):

$$\widehat{se}(\widehat{N}_H) = \sqrt{\frac{N \cdot \widehat{N}_H}{\sum_{i \in s_F} \widehat{d}_{i,U}}}, \tag{F.1}$$

and then produces a confidence interval:

$$\widehat{N}_H \pm z_{1-\alpha/2} \widehat{se}(\widehat{N}_H), \tag{F.2}$$

where $1-\alpha$ is the desired confidence level (typically 0.95), and $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard Normal distribution.

	estimated correlation
$\widehat{\text{cor}}(\sum_{i \in s_H} y_{i,\mathcal{A} \cap F}/c\pi_i, \sum_{i \in s_H} \tilde{v}_{i,\mathcal{A} \cap F}/c\pi_i)$	0.92
$\widehat{\text{cor}}(\sum_{i \in s_H} y_{i,\mathcal{A} \cap F}/c\pi_i, \sum_{i \in s_H} 1/c\pi_i)$	0.71
$\widehat{\text{cor}}(\sum_{i \in s_H} \tilde{v}_{i,\mathcal{A} \cap F}/c\pi_i, \sum_{i \in s_H} 1/c\pi_i)$	0.68

Table E.4: Estimated pairwise correlations for quantities derived from a sample from the hidden population. These quantities play a role in the approximate relative bias for the estimate of all of the nonlinear estimators we propose.

	approx. rel. bias, B_d	estimate	estimated absolute bias
$\widehat{\tau}_F$	0.0005	0.77	0.0004
$\widehat{\delta}_F$	0.0105	0.69	0.0073
$\widehat{N}_H$	0.0026	114498.00	298.0000

Table E.5: Approximate relative bias in the estimates of the nonlinear quantities using data taken from the Curitiba study, the point estimates produced by the Curitiba study, and the estimated implied absolute bias. For each quantity, the bias is very small.

Unfortunately, the variance estimator (Equation F.1) was derived from the basic scale-up model (Equation 11), and so it suffers from the limitations of that model. In particular, it has three main problems, none of which seem to have been appreciated in the scale-up literature and all of which lead it to underestimate the variance in most situations. First, the variance estimator in Equation F.1 does not include any information about the procedure used to sample respondents, which can lead to problems when complex sampling designs, such as stratified, multi-stage designs, are used. Second, it implicitly assumes that the researchers have learned about $\sum_{i \in s_F} d_{i,U}$ independent alters, which is not true if there are barrier effects (i.e., non-random social mixing). Finally, like virtually all variance estimators, it only provides a measure of uncertainty introduced by sampling but not other possible sources of error.

To address the first two problems but not the third, we propose that researchers used the rescaled bootstrap variance estimation procedure (Rao and Wu, 1988; Rao et al., 1992; Rust and Rao, 1996) with the percentile method; a combination that, for

convenience, we will refer to as the rescaled bootstrap. This procedure, described in more detail below, has strong theoretical foundations; does not depend on the basic scale-up model; can handle both simple and complex sample designs; and can be used for both the basic scale-up estimator and the generalized scale-up estimator.

In addition to the theoretical reasons to prefer the rescaled bootstrap, empirically, we find that the rescaled bootstrap produces intervals with slightly better coverage properties in three real scale-up studies. In particular, using the internal consistency check procedure proposed in Killworth et al. (1998a) for all groups of known size in three real scale-up datasets—one collected via simple random sampling (McCarty et al., 2001) and two collected via complex sample designs (Salganik et al., 2011a; Rwanda Biomedical Center, 2012)—we produced a size estimate using the basic scale-up estimator (Equation 12), and we produced confidence intervals using (1) the current procedure (Equation F.1); (2) the simple bootstrap (which does not account for complex sample designs) with the percentile method; and (3) the rescaled bootstrap (which does account for complex sample designs) with the percentile method.

This empirical evaluation (Figure F.1) produced three main results. First, as expected, we found that the current confidence interval procedure produces intervals with bad coverage properties: purported 95% confidence intervals had empirical coverage rates of about 5%. This poor performance does not seem to have been widely appreciated in the scale-up literature. Second, also consistent with expectation, we found that the rescaled bootstrap produced wider intervals than both the current procedure and the simple bootstrap, especially in the case of complex sample designs. Third, and somewhat surprisingly, the rescaled bootstrap did not work well in an absolute sense: purported 95% confidence intervals had empirical coverage rates of about 10%, only slightly better than the current procedure.

We speculate that there are two possible reasons for the surprisingly poor cover-

age rates of the rescaled bootstrap. The first is bias in the basic scale-up estimator. As described in detail in Sarndal et al. (1992, Sec 5.2), bias in an estimator can degrade the coverage rates for confidence intervals. For example, if Native Americans (one of the groups in the study of McCarty et al. (2001)) have smaller personal networks than other Americans, then there will be a downward bias in the estimated number of Native Americans (Equation 20). This bias will necessarily degrade the coverage properties of any confidence interval procedure, especially if the bias ratio $\left(bias(\hat{N}_H)/se(\hat{N}_H)\right)$ is large (see Sarndal et al. (1992, Sec 5.2)). The second possible reason for the surprisingly poor coverage rates could also be some unknown problem with the rescaled bootstrap or the percentile method. Because (i) the rescaled bootstrap and percentile method have strong theoretical foundations (Rao and Wu, 1988; Rao et al., 1992; Rust and Rao, 1996; Efron and Tibshirani, 1993) and (ii) we expect that the basic scale-up estimates are biased in most situations (see Equation 20), we believe that the main reason for the poor coverage is the bias. However, we also believe that future research should explore the properties of the rescaled bootstrap and percentile method in greater detail.

An additional concern about these empirical results is that they only apply to the basic scale-up estimator and not the generalized scale-up estimator. Unfortunately, we cannot assess the performance of the rescaled bootstrap procedure when used with the generalized scale-up estimator because the generalized scale-up estimator has not yet been used for populations of known size.

These empirical results, and the theoretical arguments that follow, lead us to three conclusions. First, confidence intervals from the rescaled bootstrap are preferable to intervals from the current procedure. Second, researchers should expect that the confidence intervals from the rescaled bootstrap procedure will be anti-conservative (i.e., they will be too small). Third, creating confidence intervals around scale-up

estimates is an important area for further research.

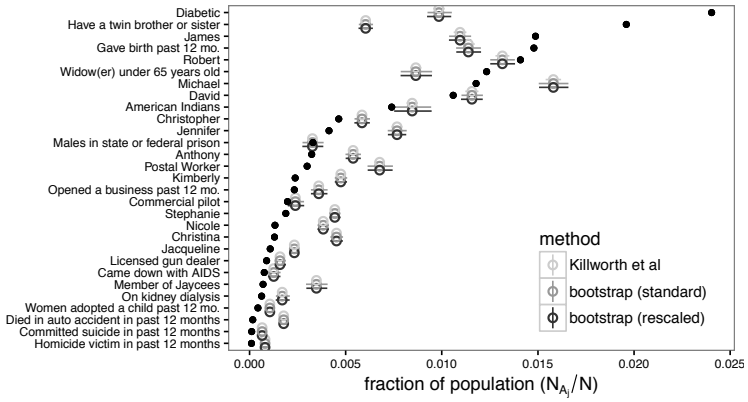
Next in Section F.1 we review the standard bootstrap and rescaled bootstrap; describe how we applied these methods to three real scale-up datasets; and describe the results in Figure F.1 in greater detail. Finally, in Section F.2 we describe how researchers can use the rescaled bootstrap with the generalized scale-up estimator.

F.1 Variance estimation with a sample from F

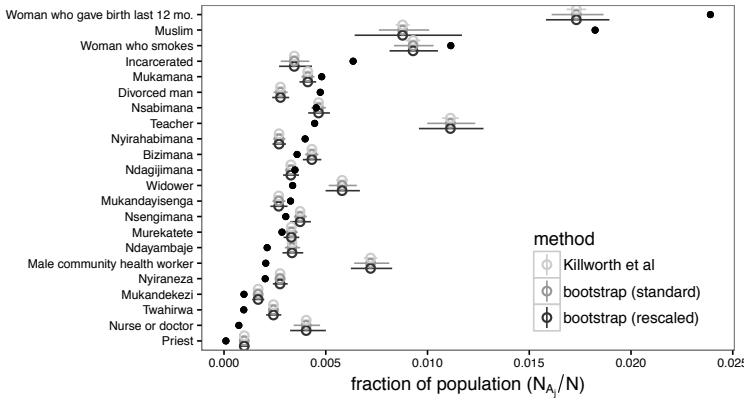
The goal of a bootstrap variance estimation procedure is to put a confidence interval around an estimate $\widehat{N}_H$ that is derived from a sample s_F. The most standard bootstrap procedure has three steps. First, researchers generate B replicate samples, $s_F^{(1)}, s_F^{(2)}, \ldots, s_F^{(B)}$ by randomly sampling with replacement from s_F. Second, these replicate samples are then used to produce a set of replicate estimates, $\widehat{N}_H^{(1)}, \widehat{N}_H^{(2)}, \ldots, \widehat{N}_H^{(B)}$. Finally, the replicate estimates are combined to produce a confidence interval; for example, by the percentile method which chooses the 2.5th and 97.5th percentiles of the B estimates (Fig. F.2) (Efron and Tibshirani, 1993).

When the original sample can be modeled as a simple random sample, this standard bootstrap procedure is appropriate. For example, consider the scale-up study of McCarty et al. (2001) that was based on telephone survey of 1,261 Americans selected via random digit dialing.[12] We can approximate the sampling design as simple random sampling, and draw $B = 10,000$ replicate samples of size 1,261. In this case the bootstrap confidence intervals are, as expected, larger than the confidence intervals from Equation F.1, since they account for the clustering of responses with respondent; on average, they are 2.05 times wider.

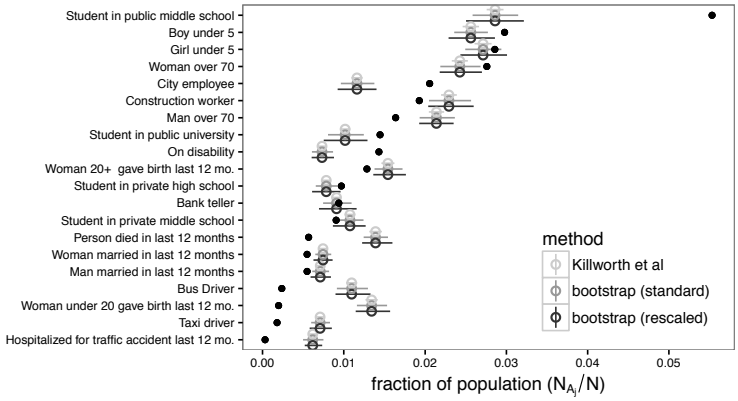
[12]The original data file includes 1,375 respondents. From these cases, 113 respondents who had missing data for some of the aggregated relational data questions and 1 respondent who answered 7 for all questions (see Zheng et al. (2006)). Further, consistent with common practice (e.g., Zheng et al. (2006)), we top coded all responses at 30, affecting 0.26% of responses.



(a) United States (simple random sample)



(b) Rwanda (stratified, multi-stage)



(c) Curitiba, Brazil (multi-stage)

Figure F.1: Assessing confidence interval procedures using scale-up studies in the United States (McCarty et al., 2001), Rwanda (Rwanda Biomedical Center, 2012), and Curitiba, Brazil (Salganik et al., 2011a). The true size of each group is shown with a black dot. Estimates made use the basic scale-up estimator are shown with circles. The rescaled bootstrap confidence intervals include the true group size for 3.4%, 9.1%, and 15.0% of the groups in the US, Rwanda, and Curitiba, respectively. The standard bootstrap confidence intervals include the true group size for 3.4%, 9.1%, and 10.0% of the groups. The currently used procedure (Equation F.1), contains the true group size for 3.4%, 9.1%, and 5.0% of the groups.

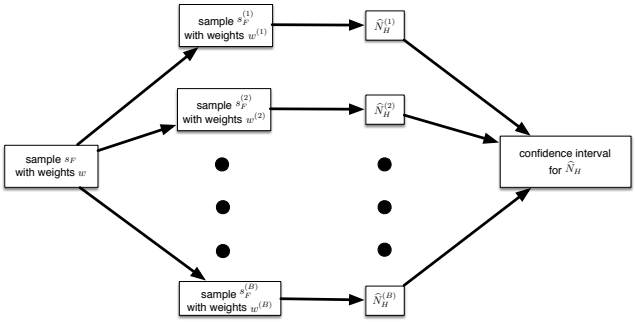


Figure F.2: Schematic of the bootstrap procedure to put a confidence interval around $\widehat{N}_H$ when there is a sample from the frame s_F.

This standard bootstrap procedure, however, can perform poorly when the original data are collected with a complex sample design (Shao, 2003). To deal with this problem, Rust and Rao (1996) proposed the rescaled bootstrap procedure that works well when the data are collected with a general multistage sampling design, a class of designs that includes most designs that would be used for face-to-face scale-up surveys. For example, it includes stratified two-stage cluster sampling with oversampling (as was used in a recent scale-up study in Rwanda (Rwanda Biomedical Center, 2012)) and three-stage element sampling (as was used in a recent scale-up study in Curitiba, Brazil (Salganik et al., 2011a)); a full description of the designs included in this class is presented in Rust and Rao (1996).

The rescaled bootstrap includes two conceptual changes from the standard bootstrap. First, it approximates the actual sampling design by a closely related one that is much easier to work with. In particular, if we assume that primary sampling units (PSUs) are selected with replacement and that all subsequent stages of sampling are conducted independently each time a given PSU is selected, then we can use the

with-replacement sampling framework in which variance estimation is much easier; see Sarndal et al. (1992) Result 4.5.1 for a more formal version of this claim. It is important to note that this approximation is generally conservative because with-replacement sampling usually results in higher variance than without-replacement sampling. Therefore, we will be estimating the variance for a design that has higher variance than the actual design. In practice, this difference is usually small because the sampling fraction in each stratum is usually small (Rao et al., 1992; Rust and Rao, 1996); see Sarndal et al. (1992) Section 4.6 for a more formal treatment. To estimate the variance in this idealized with-replacement design, resampling should be done independently in each stratum and the units that are resampled with replacement should be entire PSUs, not respondents.

This change—resampling PSUs, not respondents—introduces the need for a second change in the resampling procedure. It is known that the standard bootstrap procedure is off by a factor of $(n-1)/n$ where n is the sample size (Rao and Wu, 1988). Thus, when the sample size is very small, the bootstrap will tend to underestimate the variance. While this issue is typically ignored, it can become important when we resample PSUs rather than respondents. In particular, the number of sampled PSUs in stratum h, n_h, can be small in complex sample designs. At the extreme, in a design with two sampled PSUs per stratum, which is not uncommon, the standard bootstrap would be expected to produce a 50% underestimate of the variance. Therefore, Rao et al. (1992) developed the rescaled bootstrap, whereby the bootstrap sample size is slightly smaller than the original sample size and the sample weights are rescaled to account for this difference. Rust and Rao (1996) recommend that if the original sample includes n_h PSUs in strata h, then researchers should resample $n_h - 1$ PSUs and rescale the respondent weights by $n_h/(n_h - 1)$. That is, the weight for the j^{th}

person in PSU i in the b^{th} replicate sample is

$$w_{ij}^{(b)} = w_{ij} \times \frac{n_h}{(n_h - 1)} \times r_i^{(b)} \tag{F.3}$$

where w_{ij} is the original weight for the j^{th} unit in the i^{th} PSU, n_h is the number of PSUs in strata h, and $r_i^{(b)}$ is the number of times the i^{th} PSU was selected in replicate sample b.

In Figure 1, we compared the three different procedures for putting confidence intervals around the basic scale-up estimator: the current procedure (Killworth et al., 1998b), the standard bootstrap with the percentile method, and the rescaled bootstrap with the percentile method. We made this comparison using data from scale-up studies in the United States, Rwanda,[13] and Curitiba, Brazil.[14] As expected, the rescaled bootstrap produced confidence intervals that are larger than those from the standard bootstrap, which in turn are larger than those from the current scale-up

[13]The scale-up study in Rwanda used stratified two-stage cluster sampling with unequal probability of selection across strata in order to oversample urban areas. Briefly, the sample design divided Rwanda into five strata: Kigali City, North, East, South, and West. At the first stage, PSUs—in this case villages—were selected with probability proportional to size and without replacement within each stratum with oversampling in the Kigali City stratum. This approach resulted in a sample of 130 PSUs: 35 from Kigali City, 24 from East, 19 from North, 26 from South, and 26 from West. At the second stage, 20 households were selected via simple random sampling without replacement from each PSU in Kigali City and 15 households from each PSU in other strata. Finally, all members of the sampled household over the age of 15 were interviewed. The study included a survey experiment which randomized respondents to report about one of two different personal networks; to keep things simple, we use responses about only one personal network here. For full details see Rwanda Biomedical Center (2012). The original data file includes 2,406 respondents. From these cases, we removed 2 respondents who had missing data for some of the aggregated relational data questions. Further, consistent with common practice (e.g., Zheng et al. (2006)), we top coded all responses at 30, affecting 0.12% of responses.

[14]The scale-up study in Curitiba, Brazil used two-stage element sampling where 54 primary sampling units (PSUs)—in this case census tracks—were selected with probability proportional to their estimated number of housing units and without replacement. Then, within each cluster, eight secondary sampling units (SSUs)—in this case people—were selected with equal probability without replacement. For full details see Salganik et al. (2011a). The original data file includes 500 respondents. From these cases, we removed no respondents who had missing data for some of the aggregated relational data questions. Further, consistent with common practice (e.g., Zheng et al. (2006)), we top coded all responses at 30, affecting 0.58% of responses.

variance estimation procedure. In the study from Curitiba, the rescaled bootstrap procedure produced confidence intervals 1.17 times larger than the standard bootstrap and 2.84 times larger than the current procedure. In the Rwanda case, the rescaled bootstrap procedure produced confidence intervals 1.35 times larger than the standard bootstrap and 2.65 times larger than the current procedure.

Finally, Figure F.1 shows the estimated confidence intervals for the groups of known size in the three studies described above. The coverage rates for the bootstrap confidence intervals for the US, Rwanda, and Curitiba, are 3.4%, 9.1%, 15.0%. While this is far from ideal, we note that it is slightly better than the currently used procedure (Equation F.2), which produced coverage rates of 3.4%, 9.1%, 5.0%, and it is also slightly better than the standard bootstrap, which produced coverage rates of 3.4%, 9.1%, and 10.0%.

F.2 Variance estimation with sample from F and H

Producing confidence intervals around the generalized scale-up estimator is more difficult than the basic scale-up estimator because the generalized estimator has uncertainty from two different samples: the sample from the hidden population and the sample from the frame population. To capture all of this uncertainty, we propose combining replicate samples from the frame population with independent replicate samples from the hidden population in order to produce a set of replicate estimates. More formally, given s_F, a sample from the frame population, and an independent sample s_H from the hidden population, we seek to produce a set of B bootstrap replicate samples for s_F and s_H, $s_F^{(1)}, s_F^{(2)}, \ldots, s_F^{(B)}$ and $s_H^{(1)}, s_H^{(2)}, \ldots, s_H^{(B)}$, which are then combined to produce a set of B bootstrap estimates: $\widehat{N}_H^{(1)} = f(s_F^{(1)}, s_H^{(1)})$, $\widehat{N}_H^{(2)} = f(s_F^{(2)}, s_H^{(2)}), \ldots \widehat{N}_H^{(B)} = f(s_F^{(B)}, s_H^{(B)})$. Finally, these B replicate estimates are

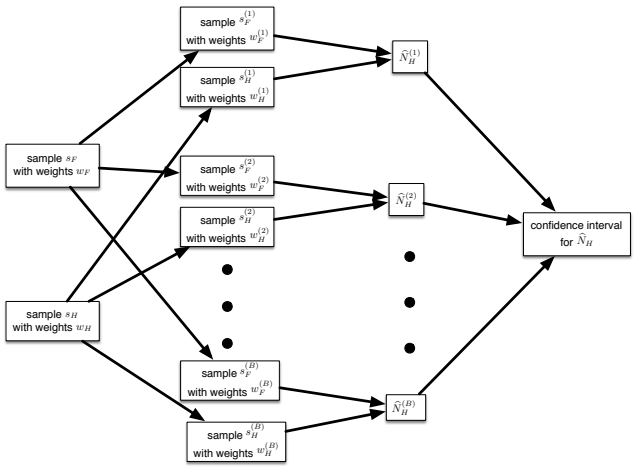


Figure F.3: Schematic of the bootstrap procedure to put a confidence interval around $\widehat{N}_H$ when there is a sample from the frame s_F and a sample from the hidden population s_H.

converted into a confidence interval using the percentile method (Fig. F.3).

Because of the challenges involved in sampling hard-to-reach populations, the two most likely sampling designs for s_H will be time-location sampling and respondent-driven sampling. If s_H was selected with time-location sampling, we recommend treating the design as a two-stage element sample (see Karon and Wejnert (2012)) and using the procedure of Rust and Rao (1996). If s_H was selected with respondent-driven sampling, as was done in a recent study of heavy drug users in Curitiba, Brazil (Salganik et al., 2011b), we recommend using the best available bootstrap method for respondent-driven sampling data, which at the present time is the procedure introduced in Salganik (2006). One implementation detail of that particular bootstrap procedure is that it requires researchers to divide the sample of the hidden population into two mutually exclusive groups. In this case, we recommend dividing the hidden population into those who are above and below the median of their estimated visibility $\widehat{v}_{i,F}$ in order to capture some of the extra uncertainty introduced if there are strong tendencies for more hidden members of the hidden population to

recruit each other.

Because the generalized scale-up estimator has never been used for groups of known size, we cannot explore the coverage rate of the proposed procedure. However, based on experience with respondent-driven sampling, we suspect that variance estimation procedures for hidden populations will underestimate the actual uncertainty in the estimates (Goel and Salganik, 2009, 2010; Yamanis et al., 2013; Verdery et al., 2013; Rohe, 2015). If this is the case, then the intervals around the generalized scale-up estimates will be anti-conservative.

In conclusion, Sec. F.1 presents a bootstrap procedure for simple and complex sample designs from the sampling frame, and Sec. F.2 extends these results to account for the sampling variability introduced by having a sample from the hidden population. We have shown that the performance of these procedures on three real scale-up datasets is consistent with theoretical expectations. Additional research in this area, which is beyond the scope of this paper, could adopt a total survey error approach and attempt to quantify all sources of uncertainty in the estimates, not just sampling uncertainty. Additional research could also explore the properties and sensitivity of these confidence interval procedures though simulation.

G Simulation study

In this appendix, we describe a simulation study comparing the performance of the generalized and basic network scale-up estimators. The results of these simulations confirm and illustrate several of the analytical results in Section 3 of the paper. Most importantly, the simulations show that the generalized network scale-up estimator is unbiased for all of the situations explored by the simulation, while the basic network scale-up estimator is biased for all but a few special cases. Moreover, our analytical