Online Appendixes

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 results correctly predict the bias of the basic network scale-up estimator in each case.

Our simulation study is intentionally simple in order to clearly illustrate our analytical results; it is not designed to be a realistic model of any scale-up study. Concretely, our simulations compare the performance of generalized and basic scale-up estimators as three important quantities vary: (1) the size of the frame population F, relative to the size of the entire population, U; (2) the extent to which people's network connections are not formed completely at random, also called the amount of inhomogenous mixing; and (3) the accuracy of reporting, as captured by the true positive rate τ_F (see Equation 18).

We simulate populations consisting of $N = 5,000$ people, using a stochastic blockmodel (White et al., 1976; Wasserman and Faust, 1994) to randomly generate networks with different amounts of inhomogenous mixing. Stochastic block models assume population members can be grouped into different *blocks*. For any pair of people, i and j, the probability that there is an edge between i and j is completely determined by the block memberships of i and j.

In our simulation model, each person can be either in or out of the frame population F and each person can also be either in or out of the hidden population H, producing four possible blocks: FH, $F\neg H$, $\neg F\neg H$, and $\neg FH$. (Here, we use the logical negation symbol, $\neg$, to denote not being in a group.) The probability of an edge between any two people i and j is then governed by a Bernoulli distribution whose mean is a function of the two block memberships:

$$\Pr(i \leftrightarrow j) \sim \text{Bernoulli}(\mu_{g(i),g(j)}), \tag{G.1}$$

where $g(i)$ is the block containing i, $g(j)$ is the block containing j, $i \leftrightarrow j$ denotes an undirected edge between i and j, and $\mu_{g(i),g(j)}$ is the probability of an edge between a member of group $g(i)$ and a member of group $g(j)$. In a network with a no inhomogenous mixing (equivalent to an Erdos-Renyi random graph), $\mu_{g(i),g(j)}$ will be the same for all i and j. On the other hand, in a network with a high level of inhomogenous mixing, $\mu_{g(i),g(j)}$ will be relatively small when $g(i) \neq g(j)$ and $\mu_{g(i),g(j)}$ will be relatively large when $g(i) = g(j)$[15].

Each random network drawn under our simulation model depends on seven parameters. The first four parameters describe population size and group memberships; they are:

- N, the size of the population
- p_F, the fraction of people in the frame population
- p_H, the fraction of people in the hidden population
- $p_{F|H}$, the fraction of hidden population members also in the frame population

The next three parameters govern the amount of inhomogenous mixing in the network that connects people to each other:

- ζ, the probability of an edge between two people who are both in the same block.
- ξ, the relative probability of an edge between two vertices that differ in frame population membership. For example, a value of 0.6 would mean that the chances of having a connection between a particular person in F and a particular person not in F is 60% of the chance of a connection between two members of F or two members of $\neg F$.

[15] Computer code to perform the simulations was written in R (R Core Team, 2014) and used the following packages: devtools (Wickham and Chang, 2013); functional (Danenberg, 2013); ggplot2 (Wickham, 2009); igraph (Csardi and Nepusz, 2006); networkreporting (Feehan and Salganik, 2014); plyr (Wickham, 2011); sampling (Tillé and Matei, 2015); and stringr (Wickham, 2012).

$$\mathbf{M} = \begin{array}{c|cccc} & F\ H & F\neg H & \neg F\ H & \neg F\neg H \\ \hline F\ H & \zeta & \rho\cdot\zeta & \xi\cdot\zeta & \xi\cdot\rho\cdot\zeta \\ F\neg H & \rho\cdot\zeta & \zeta & \xi\cdot\rho\cdot\zeta & \xi\cdot\zeta \\ \neg F\ H & \xi\cdot\zeta & \xi\cdot\rho\cdot\zeta & \zeta & \rho\cdot\zeta \\ \neg F\neg H & \xi\cdot\rho\cdot\zeta & \xi\cdot\zeta & \rho\cdot\zeta & \zeta \end{array} \tag{G.2}$$

Figure G.1: The mixing matrix used to generate a random network using the stochastic block model. Entry (i, j) in the matrix describes the probability of an edge between two people, one of whom is in group i and one in group j. The probabilities are governed by ζ, ξ, and ρ. In our simulations, we generate networks with different amounts of inhomogenous mixing between hidden population members and non-hidden population members by fixing $\zeta = 0.05$ and $\xi = 0.4$, and then varying ρ from 0.1 (extreme inhomogenous mixing between hidden and non-hidden population members) to 1 (perfectly random mixing between hidden and non-hidden population members).

- ρ, the relative probability of an edge between two vertices that differ in hidden population membership. For example, a value of 0.8 would mean that the chances of having a connection between a particular person in H and a particular person not in H is 80% of the chance of a connection between two members of H or two members of $\neg H$.

Together, the parameters ζ, ξ, and ρ are used to construct the mixing matrix M (Figure G). Note that varying the parameter ρ will change several structural features of the network in addition to the amount of inhomogenous mixing; for example, changing ρ will alter the degree distribution. Our analytical results show that the generalized network scale-up estimator is robust to changes in these structural features.

The final parameter, τ_F, is used to control the amount of imperfect reporting. After randomly drawing a network using the stochastic block model, we generate a reporting network as follows:

1. convert all undirected edges $i \leftrightarrow j$ in the social network into two directed reporting edges in the reporting network: one $i \rightarrow j$ and one $j \rightarrow i$

2. select a fraction, $1 - \tau_F$, of the edges that lead from members of the frame pop-

ulation to members of the hidden population uniformly at random and remove them from the reporting graph.

Given a simple random sample of 500 members of the frame population and a relative probability sample of 30 members of the hidden population, the reporting graph is then used to compute the basic and generalized scale-up estimates for the size of the hidden population.

Across our simulations, we fix five of the parameters at constant values ($N = 5,000$; $p_F = 0.03$; $p_{F|H} = 1$; $\zeta = 0.05$; $\xi = 0.4$). We systematically explore varying the remaining parameters: we investigate ρ for values from 0.1 to 1 in increments of 0.1; we investigate p_F for values 0.1, 0.5, and 1; and we investigate τ_F for 0.1, 0.5, and 1. For each combination of the parameter values, we generate 10 random networks. Within each random network, we simulate 500 surveys. Each survey consists of two samples: a probability sample from the frame population, with sample size of 500; and a relative probability from the hidden population of size 30, with inclusion proportional to each hidden population member's personal network size. For each unique combination of parameters, we averaged the results across the surveys and across the randomly generated networks.