



Published in final edited form as:

Sociol Methodol. 2016 August ; 46(1): 153–186. doi:10.1177/0081175016665425.

Generalizing the Network Scale-Up Method: A New Estimator for the Size of Hidden Populations*

Dennis M. Feehan[†] and Matthew J. Salganik^{‡,§}

[†]Department of Demography, University of California, Berkeley, CA, USA

[‡]Office of Population Research, Princeton University, Princeton, NJ, USA

[§]Department of Sociology, Princeton University, Princeton, NJ, USA

Abstract

The network scale-up method enables researchers to estimate the size of hidden populations, such as drug injectors and sex workers, using sampled social network data. The basic scale-up estimator offers advantages over other size estimation techniques, but it depends on problematic modeling assumptions. We propose a new generalized scale-up estimator that can be used in settings with non-random social mixing and imperfect awareness about membership in the hidden population. Further, the new estimator can be used when data are collected via complex sample designs and from incomplete sampling frames. However, the generalized scale-up estimator also requires data from two samples: one from the frame population and one from the hidden population. In some situations these data from the hidden population can be collected by adding a small number of questions to already planned studies. For other situations, we develop interpretable adjustment factors that can be applied to the basic scale-up estimator. We conclude with practical recommendations for the design and analysis of future studies.

1 Introduction

Many important problems in social science, public health, and public policy require estimates of the size of hidden populations. For example, in HIV/AIDS research, estimates of the size of the most at-risk populations—drug injectors, female sex workers, and men who have sex with men—are critical for understanding and controlling the spread of the epidemic. However, researchers and policy makers are unsatisfied with the ability of current statistical methods to provide these estimates (Joint United Nations Programme on HIV/AIDS, 2010). We address this problem by improving the network scale-up method, a promising approach to size estimation. Our results are immediately applicable in many substantive domains in which size estimation is challenging, and the framework we develop advances the understanding of sampling in networks more generally.

*The authors thank Alexandre Abdo, Francisco Bastos, Russ Bernard, Neilane Bertoni, Dimitri Fazito, Sharad Goel, Wolfgang Hladik, Jake Hofman, Mike Hout, Karen Levy, Rob Lyerla, Mary Mahy, Chris McCarty, Maeve Mello, Tyler McCormick, Damon Phillips, Justin Rao, Adam Slez, and Tian Zheng for helpful discussions. This research was supported by The Joint United Nations Programme on HIV/AIDS (UNAIDS), NSF (CNS-0905086), and NIH/NICHD (R01-HD062366, R01-HD075666, & R24-HD047879). Some of this research was conducted while MJS was an employee Microsoft Research. The opinions expressed here represent the views of the authors and not the funding agencies.

The core insight behind the network scale-up method is that ordinary people have embedded within their personal networks information that can be used to estimate the size of hidden populations, if that information can be properly collected, aggregated, and adjusted (Bernard et al., 1989, 2010). In a typical scale-up survey, randomly sampled adults are asked about the number of connections they have to people in a hidden population (e.g., “How many people do you know who inject drugs?”) and a series of similar questions about groups of known size (e.g., “How many widowers do you know?”; “How many doctors do you know?”). Responses to these questions are called *aggregate relational data* (McCormick et al., 2012).

To produce size estimates from aggregate relational data, previous researchers have begun with the *basic scale-up model*, which makes three important assumptions: (i) social ties are formed completely at random (i.e., random mixing), (ii) respondents are perfectly aware of the characteristics of their alters, and (iii) respondents are able to provide accurate answers to survey questions about their personal networks. From the basic scale-up model Killworth et al. (1998b) derived the *basic scale-up estimator*. This estimator, which is widely used in practice, has two main components. For the first component, the aggregate relational data about the hidden population are used to estimate the number of connections that respondents have to the hidden population. For the second component, the aggregate relational data about the groups of known size are used to estimate the number of connections that respondents have in total. For example, a researcher might estimate that members of her sample have 5,000 connections to people who inject drugs and 100,000 connections in total. The basic scale-up estimator combines these pieces of information to estimate that 5% (5,000/100,000) of the population injects drugs. This estimate is a sample proportion, but rather than being taken over the respondents, as would be typical in survey research, the proportion is taken over the respondents’ alters. Researchers who desire absolute size estimates multiply the alter sample proportion by the size of the entire population, which is assumed to be known (or estimated using some other method).

Unfortunately, the three assumptions underlying the basic scale-up model have all been shown to be problematic. Scale-up researchers call violations of the random mixing assumption *barrier effects* (Killworth et al., 2006; Zheng et al., 2006; Maltiel et al., 2015); they call violations of the perfect awareness assumption *transmission error* (Shelley et al., 1995, 2006; Killworth et al., 2006; Salganik et al., 2011b; Maltiel et al., 2015); and they call violations of the respondent accuracy assumption *recall error* (Killworth et al., 2003, 2006; McCormick and Zheng, 2007; Maltiel et al., 2015).

In this paper, we develop a new approach to producing size estimates from aggregate relational data. Rather than depending on the basic scale-up model or its variants (e.g., Maltiel et al. (2015)), we use a simple identity to derive a series of new estimators. Our new approach reveals that one of the two main components of the basic scale-up estimator is problematic. Therefore, we propose a new estimator—the *generalized scale-up estimator*—that combines the aggregate relational data traditionally used in scale-up studies with similar data collected from the hidden population. Collecting data from the hidden population is a major departure from current scaleup practice, but we believe that it enables a more principled approach to estimation. For researchers who are not able to collect data from the hidden population, we propose a series of adjustment factors that highlight the possible

biases of the basic scale-up estimator. Ultimately, researchers must balance the trade-offs between the basic scale-up estimator, generalized scale-up estimator, and other size estimation techniques based on the specific features of their research setting.

In the next section, we derive the generalized scale-up estimator, and we describe the data collection procedures needed to use it. In Section 3, we compare the generalized and basic scale-up approaches analytically and with simulations; our comparison leads us to propose a decomposition that separates the difference between the two approaches into three measurable and substantively meaningful factors (Equation 15). In Section 4 we make practical recommendations for the design and analysis of future scale-up studies, and in Section 5, we conclude with an discussion of next steps. Online Appendices A – G provide technical details and supporting arguments.

2 The generalized scale-up estimator

The generalized scale-up estimator can be derived from a simple accounting identity that requires no assumptions about the underlying social network structure in the population. Figure 1 helps illustrate the derivation, which was inspired by earlier research on multiplicity estimation (Sirken, 1970) and indirect sampling (Lavallée, 2007). Consider a population of 7 people, 2 of whom are drug injectors (Figure 1(a)). In this population, two people are connected by a directed edge $i \rightarrow j$ if person i would count person j as a drug injector when answering the question “How many drug injectors do you know?” Whenever $i \rightarrow j$, we say that i makes an *out-report* about j and that j receives an *in-report* from i .¹

Each person can be viewed as both a source of out-reports and a recipient of in-reports, and in order to emphasize this point, Figure 1(b) shows the population with each person represented twice: on the left as a sender of out-reports and on the right as a receiver of in-reports. This visual representation highlights the following identity:

$$\text{total out- reports} = \text{total in- reports}. \quad (1)$$

Despite its simplicity, the identity in Equation 1 turns out to be very useful because it leads directly to the new estimator that we propose.

In order to derive an estimator from Equation 1, we must define some notation. Let U be the entire population, and let $H \subset U$ be the hidden population. Further, let $y_{i,H}$ be the total number of out-reports from person i (i.e., person i 's answer to the question “How many drug injectors do you know?”). For example, Figure 1(b) shows that person 5 would report knowing 1 drug injector, so $y_{5,H} = 1$. Let $v_{i,U}$ be the total number of in-reports to i if everyone in U is interviewed; that is, $v_{i,U}$ is the *visibility* of person i to people in U . For example, Figure 1(b) shows person 5 would be reported as a drug injector by 3 people so $v_{5,U} = 3$. Since total out-reports must equal total in-reports, it must be the case that

¹Throughout the paper, we only consider the case where i never reports j more than once.

$$y_{U,H} = v_{U,U}, \quad (2)$$

where $y_{U,H} = \sum_{i \in U} y_{i,H}$ and $v_{U,U} = \sum_{i \in U} v_{i,U}$. Multiplying both sides of Equation 2 by N_H , the number of people in the hidden population, and then rearranging terms, we get

$$N_H = \frac{y_{U,H}}{v_{U,U}/N_H}. \quad (3)$$

Equation 3 is an expression for the size of the hidden population that does not depend on any assumptions about network structure or reporting accuracy; it is just a different way of expressing the identity that the total number of out-reports must equal the total number of in-reports. If we could estimate the two terms on the right side of Equation 3—one term related to out-reports ($y_{U,H}$) and one term related to in-reports ($v_{U,U}/N_H$)—then we could estimate N_H .

However, in order to make the identity in Equation 3 useful in practice we need to modify it to account for an important logistical requirement of survey research. In real scale-up studies, researchers do not sample from the entire population U , but instead they sample from a subset of U called the frame population, F . For example, in almost all scale-up studies the frame population has been adults (but note that our mathematical results hold for any frame population). In standard survey research, restricting interviews to a frame population does not cause problems because inference is being made about the frame population. In other words, when respondents report about themselves it is clear to which group inferences apply. However, with the scaleup method, respondents report about others, so the group that inferences are being made about is not necessarily the same as the group that is being interviewed. As we show in Section 4.2, failure to consider this fact requires the introduction of an awkward adjustment factor that had previously gone unnoticed. Here, we avoid this awkward adjustment factor by deriving an identity explicitly in terms of the frame population. Restricting our attention to out-reports coming from people in the frame population, it must be the case that

$$N_H = \frac{y_{F,H}}{v_{U,F}/N_H}, \quad (4)$$

where $y_{F,H} = \sum_{i \in F} y_{i,H}$ and $v_{U,F} = \sum_{i \in U} v_{i,F}$. The only difference between Equation 3 and Equation 4 is that Equation 4 restricts out-reports and in-reports to come from people in the frame population (Figure 1(c)). The identity in Equation 4 is extremely general: it does not depend on any assumptions about the relationship between the entire population U , the frame population F , and the hidden population H . For example, it holds if no members of the hidden population are in the frame population, if there are barrier effects, and if there are transmission errors. Thus, if we could estimate the two terms on the right side of Equation 4

—one term related to out-reports ($y_{F,H}$) and one term related to in-reports ($v_{U,F}N_H$)—then we could estimate N_H under very general conditions.

Unfortunately, despite repeated attempts, we were unable to develop a practical method for estimating the term related to in-reports ($v_{U,F}N_H$). However, if we make an assumption about respondents’ reporting behavior, then we can re-express Equation 4 as an identity made up of quantities that we can actually estimate. Specifically, if we assume that the out-reports from people in the frame population only include people in the hidden population, then it must be the case that the visibility of everyone not in the hidden population is 0: $v_{i,F} = 0$ for all $i \notin H$. In this case, we can re-write Equation 4 as

$$N_H = \frac{y_{F,H}}{v_{H,F}/N_H} = \frac{y_{F,H}}{\bar{v}_{H,F}} \quad \text{if } v_{i,F} = 0 \text{ for all } i \notin H, \quad (5)$$

where $\bar{v}_{H,F} = v_{H,F}/N_H$

To understand the reporting assumption substantively, consider the two possible types of reporting errors: false positives and false negatives. Previous scale-up research on transmission error focused on the problem of false negatives, where a respondent is connected to a member of the hidden population but does not report this, possibly because she is not aware that the person she is connected to is in the hidden population (Bernard et al., 2010). Since hidden populations like drug injectors are often stigmatized, it is reasonable to suspect that false negatives will be a serious problem for the scale-up method. Fortunately, Equation 5 holds even if there are false negative reporting errors. However, false positives—which do not seem to have been considered previously in the scale-up literature—are also possible. For example, a respondent who is not connected to any drug injectors might report that one of her acquaintances is a drug injector. These false positive reports are not accounted for in the identity in Equation 5 and the estimators that we derive subsequently. If false positive reports exist, they will introduce a positive bias into estimates from the generalized scale-up estimator. Therefore, in Online Appendix A we (i) formally define an interpretable measure of false positive reports, the *precision of out-reports*; (ii) analytically show the bias in size estimates as a function of the precisions of out-reports; and (iii) discuss two research designs that could enable researchers to estimate the precision of out-reports.

2.1 Estimating N_H from sampled data

Equation 5 relates our quantity of interest, the size of the hidden population (N_H), to two other quantities: the total number of out-reports from the frame population ($y_{F,H}$) and the average number of in-reports in the hidden population ($\bar{v}_{H,F}$). We now show how to estimate $y_{F,H}$ with a probability sample from the frame population and $\bar{v}_{H,F}$ with a relative probability sample from the hidden population.

The total number of out-reports ($y_{F,H}$) can be estimated from respondents’ reported number of connections to the hidden population,

$$\hat{y}_{F,H} = \sum_{i \in s_F} \frac{y_{i,H}}{\pi_i}, \quad (6)$$

where s_F denotes the sample, $y_{i,H}$ denotes the reported number of connections between i and H , and π_i is i 's probability of inclusion from a conventional probability sampling design from the frame population. Because $\hat{y}_{F,H}$ is a standard Horvitz-Thompson estimator, it is consistent and unbiased as long as all members of F have a positive probability of inclusion under the sampling design (Sarndal et al., 1992); for a more formal statement, see Result B.1. This estimator depends only on an assumption about the sampling design for the frame population, and in Table D.2 we show the sensitivity of our estimator to violations of this assumption.

Estimating the average number of in-reports for the hidden population ($\bar{v}_{H,F}$) is more complicated. First, it will usually be impossible to obtain a conventional probability sample from the hidden population. As we show below, however, estimating $\bar{v}_{H,F}$ only requires a relative probability sampling design in which hidden population members have a nonzero probability of inclusion and respondents' probabilities of inclusion are known up to a constant of proportionality, $c\pi_i$ (see Online Appendix C.1 for a more precise definition). Of course, even selecting a relative probability sample from a hidden population can be difficult.

A second problem arises because we do not expect respondents to be able to easily and accurately answer direct questions about their visibility ($v_{i,F}$). That is, we do not expect respondents to be able to answer questions such as "How many people on the sampling frame would include you when reporting a count of the number of drug injectors that they know?" Instead, we propose asking hidden population members a series of questions about their connections to certain groups and their visibility to those groups. For example, each sampled hidden population respondent could be asked "How many widowers do you know?" and then "How many of these widowers are aware that you inject drugs?" This question pattern can be repeated for many groups (e.g., widowers, doctors, etc.). We call data with this structure *enriched aggregate relational data* to emphasize its similarity to the aggregate relational data that is familiar to scale-up researchers. An interviewing procedure called the *game of contacts* enables researchers to collect enriched aggregated relational data, even in realistic field settings (Salganik et al., 2011b; Maghsoudi et al., 2014).

Given a relative probability sampling design and enriched aggregate relational data, we can now formalize our proposed estimator for $\bar{v}_{H,F}$. Let A_1, A_2, \dots, A_j be the set of groups about which we collect enriched aggregate relational data (e.g., widowers, doctors, etc). Here, to keep the notation simple, we assume that these groups are all contained in the frame population, so that $A_j \subset F$ for all j ; in Online Appendix C.4 we extend the results to groups that do not meet this criterion. Let \mathcal{A} be the concatenation of these groups, which we call the *probe alters*. For example, if A_1 is widowers and A_2 is doctors, then the probe alters \mathcal{A} is the collection of all widowers and all doctors, with doctors who are widowers included twice. Also, let \tilde{v}_{i,A_j} be respondent i 's report about her visibility to people in A_j and let v_{i,A_j}

be respondents i 's actual visibility to people in A_j (i.e., the number of times that this respondent would be reported about if everyone in A_j was asked about their connections to the hidden population).

The estimator for $\bar{v}_{H,F}$ is:

$$\hat{v}_{H,F} = \frac{N_F}{N_{\mathcal{A}}} \frac{\sum_{i \in s_H} \sum_j \tilde{v}_{i,A_j} / (c\pi_i)}{\sum_{i \in s_H} 1 / (c\pi_i)}, \quad (7)$$

where $N_{\mathcal{A}}$ is the number of probe alters, c is the constant of proportionality from the relative probability sample, and s_H is a relative probability sample of the hidden population. Equation 7 is a standard weighted sample mean (Sarndal et al., 1992, Sec. 5.7) multiplied by a constant, $N_F/N_{\mathcal{A}}$. Result C.2 shows that, this estimator is consistent and *essentially unbiased*², when three conditions are satisfied: one about the design of the survey, one about reporting behavior and one about sampling from the hidden population.

The first condition underlying the estimator in Equation 7 is related to the design of the survey, and we call it the *probe alter condition*. This condition describes the required relationship between the visibility of the hidden population to the probe alters and the visibility of the hidden population to the frame population:

$$\frac{v_{H,\mathcal{A}}}{N_{\mathcal{A}}} = \frac{v_{H,F}}{N_F}, \quad (8)$$

where $v_{H,\mathcal{A}}$ is the total visibility of the hidden population to the probe alters, $v_{H,F}$ is the total visibility of the hidden population to the frame population, $N_{\mathcal{A}}$ is the number of probe alters, and N_F is the number of people in the frame population. In words, Equation 8 says that the rate at which the hidden population is visible to the probe alters must be the same as the rate at which the hidden population is visible to the frame population. For example, in a study to estimate the number of drug injectors in a city, drug treatment counselors would be a poor choice for membership in the probe alters because drug injectors are probably more visible to drug treatment counselors than to typical members of the frame population. On the other hand, postal workers would probably be a reasonable choice for membership in the probe alters because drug injectors are probably about as visible to postal workers as they are to typical members of the frame population. Additional results about the probe alter condition are presented in the Online Appendixes: (i) Result C.3 presents three other algebraically equivalent formulations of probe alter condition, some of which offer additional intuition; (ii) Result C.4 provides a method to empirically test the probe alter

²We use the term “essentially unbiased” because Equation 7 is not, strictly speaking, unbiased; the ratio of two unbiased estimators is not itself unbiased. However, a large literature confirms that the biases caused by the nonlinear form of ratio estimators are typically insignificant relative to other sources of error in estimate (e.g. Sarndal et al., 1992, chap. 5). Unfortunately, many of the estimators we propose are actually ratios of ratios, sometimes called “compound ratio estimators” or “double ratio estimators.” In Online Appendix E we demonstrate that the bias caused by the nonlinear form of our estimators is not a practical cause for concern.

condition; and (iii) Table D.1 quantifies the bias introduced when the probe alter condition is not satisfied.

The second condition underlying the estimator $\hat{v}_{H,F}$ (Equation 7) is related to reporting behavior, and we call it *accurate aggregate reports about visibility*.

$$\tilde{v}_{H,\mathcal{A}} = v_{H,\mathcal{A}}, \quad (9)$$

where $\tilde{v}_{H,\mathcal{A}}$ is the total reported visibility of members of the hidden population to the probe alters ($\sum_{i \in H} \sum_{j \in J} \tilde{v}_{i,A_j}$) and $v_{H,\mathcal{A}}$ is the total actual visibility of members of the hidden population to the probe alters ($\sum_{i \in H} \sum_{j \in J} v_{i,A_j}$). In words, Equation 9 says that hidden population members must be correct in their reports about their visibility to probe alters in aggregate, but Equation 9 does not require the stronger condition that each individual report be accurate. In practice, we expect that there are two main ways that there might not be accurate aggregate reports about visibility. First, hidden population members might not be accurate in their assessments of what others know about them. For example, research on the “illusion of transparency” suggests that people tend to over-estimate how much others know about them (Gilovich et al., 1998). Second, although we propose asking hidden population members what other people know about them (e.g., “How many of these widowers know that you are a drug injector?”) what actually matters for the estimator is what other people would report about them (e.g., “How many of these widowers would include you when reporting a count of the number of drug injectors that they know?”). In cases where the hidden population is extremely stigmatized, some respondents to the scale-up survey might conceal the fact that they are connected to people whom they know to be in the hidden population, and if this were to occur, it would lead to a difference between the information that we collect ($\tilde{v}_{i,\mathcal{A}}$) and the information that we want ($v_{i,\mathcal{A}}$). Unfortunately, there is currently no empirical evidence about the possible magnitude of these two problems in the context of scale-up studies. However, Table D.1 quantifies the bias introduced into estimates if the accurate aggregate reports about visibility condition is not satisfied.

Finally, the third condition underlying the estimator $\hat{v}_{H,F}$ (Equation 7) is that researchers have a relative probability sample from the hidden population. Currently the most widely used method for drawing relatively probability samples from hidden populations is respondent-driven sampling (Heckathorn, 1997); see Volz and Heckathorn (2008) for a set of conditions under which respondent-driven sampling leads to a relative probability sample. Although respondent-driven sampling has been used in hundreds of studies around the world (White et al., 2015), there is active debate about the characteristics of samples that it yields (Heimer, 2005; Scott, 2008; Bengtsson and Thorson, 2010; Goel and Salganik, 2010; Gile and Handcock, 2010; McCreesh et al., 2012; Salganik, 2012; Mills et al., 2012; Rudolph et al., 2013; Yamanis et al., 2013; Li and Rohe, 2015; Gile and Handcock, 2015; Gile et al., 2015; Rohe, 2015). If other methods for sampling from hidden populations are demonstrated to be better than respondent-driven sampling (see e.g., Kurant et al. (2011); Mouw and Verdery (2012); Karon and Wejnert (2012)), then researchers should consider using these methods when using the generalized scale-up estimator. Further, researchers can use Table

D.2 to quantify the bias that results if the condition requiring a relative probability sample is not satisfied.

To recap, using two different data collection procedures—one with the frame population and one with the hidden population—we can estimate the two components of the expression for N_H given in Equation 5. The estimator for the numerator ($\hat{y}_{F,H}$) depends on an assumption about the ability to select a probability sample from the frame population (see Result B.1), and the estimator for the denominator ($\hat{v}_{H,F}$) depends on assumptions about survey construction, reporting behavior, and the ability to select a relative probability sample from the hidden population (see Result C.2).

We can combine these component estimators to form the *generalized scale-up estimator*:

$$\hat{N}_H = \frac{\hat{y}_{F,H}}{\hat{v}_{H,F}}. \tag{10}$$

Result C.8 proves that the generalized scale-up estimator will be consistent and essentially unbiased if (i) the estimator for the numerator ($\hat{y}_{F,H}$) is consistent and essentially unbiased; (ii) the estimator for the denominator ($\hat{v}_{H,F}$) is consistent and essentially unbiased; and (iii) there are no false positive reports.

One attractive feature of the generalized scale-up estimator (Equation 10) is that it is a combination of standard survey estimators. This structure enabled us to derive very general sensitivity results about the impact of violations of assumptions, either individually or jointly. We return to the issue of assumptions and sensitivity analysis when discussing recommendations for practice (Section 4).

3 Comparison between the generalized and basic scale-up approaches

In Section 2, we derived the generalized network scale-up estimator by using an identity relating in-reports and out-reports as the basis for a design-based estimator. The approach we followed differs from previous scale-up studies, which have posited the basic scale-up model and derived estimators conditional on that model. In this section, we compare these two different approaches from a design-based perspective.

We begin our comparison by reviewing the basic scale-up model, which was used in most of the studies listed in Table 1. In order to review this model, we need to define another quantity: we call $d_{i,U}$ person i 's degree, the number of undirected network connections she has to everyone in U .

The basic scale-up model assumes that each person's connections are formed independently, that reporting is perfect, and that visibility is perfect (Killworth et al., 1998b). Together, these three assumptions lead to the probabilistic model:

$$y_{i,A_j} = d_{i,A_j} \sim \text{Binomial} \left(d_{i,U}, \frac{N_{A_j}}{N} \right), \quad (11)$$

for all i in U and for any group A_j . In words, this model suggests that the number of connections from a person i to members of a group A_j is the result of a series of $d_{i,U}$ independent random draws, where the probability of each edge being connected to A_j is $\frac{N_{A_j}}{N}$.

The basic scale-up model leads to what we call the basic scale-up estimator:

$$\hat{N}_H = \frac{\sum_{i \in s_F} y_{i,H}}{\sum_{i \in s_F} \hat{d}_{i,U}} \times N, \quad (12)$$

where $\hat{d}_{i,U}$ is the estimated degree of respondent i from the known population method (Killworth et al., 1998a). Killworth et al. (1998b) showed that Equation 12 is the maximum-likelihood estimator for N_H under the basic scale-up model, conditional on the additional assumption that $d_{i,U}$ is known for each $i \in s_F$.

Given this background, we can now compare the basic and generalized scale-up approaches by comparing their estimands; that is, we compare the quantities that they produce in the case of a census with perfectly observed degrees. The basic scale-up estimand can be written

$$\hat{N}_H = \frac{y_{F,H}}{d_{F,U}} \times N = \frac{y_{F,H}}{\bar{d}_{U,F}}, \quad (13)$$

where $d_{F,U} = \sum_{i \in s_F} d_{i,U}$ and $\bar{d}_{U,F} = d_{U,F}/N = d_{F,U}/N$. Further, as shown in Section 2, the generalized scale-up estimand is

$$\hat{N}_H = \frac{y_{F,H}}{\bar{v}_{H,F}}. \quad (14)$$

Comparing Equations 13 and 14 reveals that both estimands have the same numerator but they have different denominators. The network reporting identity from Section 2 (total out-reports = total in-reports) shows that the appropriate way to adjust the out-reports is based on in-reports, as in the generalized scale-up approach. However, the basic scale-up approach instead adjusts out-reports with the degree of respondents. While using the degree of respondents cleverly avoids any data collection from the hidden population, our results reveal that it will only be correct under a very specific special case ($\bar{d}_{U,F} = \bar{v}_{H,F}$).

In order to further clarify the relationship between the basic and generalized scale-up approaches, we propose a decomposition that separates the difference between the two estimands into three measurable and substantively meaningful *adjustment factors*:

$$N_H = \underbrace{\left(\frac{y_{F,H}}{\bar{d}_{U,F}}\right)}_{\text{basic scale-up}} \times \underbrace{\frac{1}{\bar{d}_{F,F}/\bar{d}_{U,F}}}_{\text{frame ratio } \phi_F} \times \underbrace{\frac{1}{\bar{d}_{H,F}/\bar{d}_{F,F}}}_{\text{degree ratio } \delta_F} \times \underbrace{\frac{1}{\bar{v}_{H,F}/\bar{d}_{H,F}}}_{\text{true positive rate } \tau_F} = \underbrace{\left(\frac{y_{F,H}}{\bar{v}_{H,F}}\right)}_{\text{generalized scale-up}}.$$

(15)

The decomposition shows that when the product of the adjustment factors is 1, the two estimands are both correct. However, when the product of the adjustment factors is not 1, then the generalized scale-up estimand is correct but the basic scale-up estimand is incorrect. We now describe each of the three adjustment factors in turn.

First, we define the frame ratio, ϕ_F to be

$$\phi_F = \frac{\text{avg \# connections from a member of } F \text{ to the rest of } F}{\text{avg \# connections from a member of } U \text{ to } F} = \frac{\bar{d}_{F,F}}{\bar{d}_{U,F}}. \tag{16}$$

ϕ_F can range from zero to infinity, and in most practical situations we expect ϕ_F will be greater than one. Result B.6 shows that we can make consistent and essentially unbiased estimates of ϕ_F from a sample of F .³

Next, we define the degree ratio δ_F to be

$$\delta_F = \frac{\text{avg \# connections from a member of } H \text{ to } F}{\text{avg \# connections from a member of } F \text{ to the rest of } F} = \frac{\bar{d}_{H,F}}{\bar{d}_{F,F}}. \tag{17}$$

δ_F ranges from zero to infinity, and it is less than one when the hidden population members have, on average, fewer connections to the frame population than frame population

³Note that, since $\bar{d}_{U,F} = (NF/N) \bar{d}_{F,U}$, an equivalent expression for the frame ratio is

$$\phi_F = \frac{\bar{d}_{F,F}}{\bar{d}_{F,U} (N_F/N)} = \frac{\bar{d}_{F,F}}{\bar{d}_{F,U}} \frac{N}{N_F}.$$

members. Result C.6 shows that we can to make consistent and essentially unbiased estimates of δ_F from samples of F and H .

Finally, we define the true positive rate, τ_F , to be

$$\tau_F = \frac{\# \text{ in- reports to } H \text{ from } F}{\# \text{ edges connecting } H \text{ and } F} = \frac{v_{H,F}}{d_{H,F}} = \frac{\bar{v}_{H,F}}{\bar{d}_{H,F}}. \quad (18)$$

τ_F relates network degree to network reports.⁴ τ_F ranges from 0, if none of the edges are correctly reported, to 1 if all of the edges are reported. Substantively, the more stigmatized the hidden population, the closer we would expect τ_F to be to 0. Result C.7 shows that we can to make consistent and essentially unbiased estimates of τ_F from a sample of H .

Further, the decomposition in Equation 15 can be used to derive an expression for the bias in the basic scale-up estimator when we have a census and degrees are known:

$$\text{bias}(\hat{N}_H^{\text{basic}}) \equiv \hat{N}_H^{\text{basic}} - N_H \quad (19)$$

$$= \hat{N}_H^{\text{basic}} \left[1 - \frac{1}{\phi_F \delta_F \tau_F} \right]. \quad (20)$$

The comparison between the basic and generalized scale-up approaches leads to two main conclusions. First, the estimand of the basic scale-up approach is correct only in one particular situation: when the product of the three adjustment factors is 1. The estimand of generalized scale-up approach, in contrast, is correct more generally. Second, as Equation 15 shows, if the adjustment factors are known (or have been estimated), then they can be used to improve basic scale-up estimates.

3.1 Illustrative simulation

In order to illustrate our comparison between the basic and generalized scale-up approaches, we conducted a series of simulation studies. The simulations were not meant to be a realistic model of a scale-up study, but rather, they were designed to clearly illustrate our analytic results. More specifically, the simulation investigated the performance of the estimators as three important quantities vary: (1) the size of the frame population F , relative to the size of

⁴Note that the fact that in-reports must equal out-reports means that τ_F can also be defined

$$\tau_F = \frac{\# \text{ reported edges from } F \text{ actually connected to } H}{\# \text{ edges connecting } F \text{ and } H} = \frac{y_{F,H}^+}{d_{F,H}}.$$

Here we have written $y_{F,H}^+$ to mean the true positive reports among the $y_{F,H}$ see Online Appendix A for a detailed explanation.

the entire population U ; (2) the extent to which people's network connections are not formed completely at random; and (3) the accuracy of reporting, as captured by the true positive rate τ_F (see Equation 18).⁵

As described in detail in Online Appendix G, we created populations of 5,000 people with different proportions of the population on the sampling frame (p_F). Next, we connected the people with a social network created by a stochastic block-model (White et al., 1976; Wasserman and Faust, 1994) in which the randomness of the mixing was controlled by a parameter ρ such that $\rho = 1$ is equivalent to random mixing (i.e., an Erdos-Reyni random graph) and the mixing becomes more non-random as $\rho \rightarrow 0$. Then, for each combination of parameters, we drew 10 populations, and within each of these populations, we simulated 500 surveys. For each survey, we drew a probability sample of 500 people from the frame population, a relative probability sample of 30 people from the hidden population, and simulated responses with a specific level of reporting accuracy (τ_F). Finally, we used these reports and the appropriate sampling weights to calculate the basic and generalized scale-up estimates.

Figure 2 shows that the simulations support our analytic results. First, the simulations show that the generalized scale-up estimator is unbiased even in the presence of incomplete sampling frames, non-random mixing, and imperfect reporting. Second, they show that the basic scale-up estimator is unbiased in a much smaller set of situations. More concretely, the basic scale-up estimator is unbiased in situations where the basic scale-up model holds—when everyone is in the frame population ($p_F = 1$), there is random mixing ($\rho = 1$), and respondents' reports are perfect ($\tau_F = 1$).⁶ Further, Figure 3 illustrates that our analytic approach (Equation 3) can correctly predict the bias of the basic scale-up estimator.

4 Recommendations for practice

The results in Sections 2 and 3 lead to us to recommend a major departure from current scale-up practice. In addition to collecting a sample from the frame population, we recommend that researchers consider collecting a sample from the hidden population so that they can use the generalized scale-up estimator. As our results clarify, researchers using the scale-up method face a decision: they can collect data from the hidden population or they can make assumptions about the adjustment factors described in Section 3. The appropriate decision depends on a number of factors, but we think that two are most important: (i) the difficulty of sampling from the hidden population and (ii) the availability of high-quality estimates of the adjustment factors in Section 3. For example, if it is particularly difficult to sample from a specific hidden population and high-quality estimates of the adjustment factors are already available, then a basic scale-up estimator may be appropriate. If however, it is possible to sample from the hidden population and there are no high-quality estimates of adjustment factors, then the generalized scale-up estimator may be appropriate. Many

⁵Computer code to perform the simulations was written in R (R Core Team, 2014) and used the following packages: devtools (Wickham and Chang, 2013); functional (Daneberg, 2013); gg-plot2 (Wickham, 2009); igraph (Csardi and Nepusz, 2006); networkreporting (Feehan and Salganik, 2014); plyr (Wickham, 2011); sampling (Tillé and Matei, 2015); and stringr (Wickham, 2012).

⁶In addition to the settings where the basic scale-up model holds, the basic scale-up estimator can also be unbiased when its different biases cancel (e.g., when the product of the adjustment factors is 1).

realistic situations will be somewhere between these two extremes, and the trade-offs must be weighed on a case-by-case basis.

In order to aid researchers deciding between basic and generalized scale-up approaches, we collected the conditions needed for consistent and essentially unbiased estimates into Table 2; formal proofs of these results are presented in Online Appendices B and C. We find it helpful to group these conditions into four broad categories: sampling, survey construction, network structure, and reporting behavior.

A review of the conditions in Table 2 necessarily raises practical concerns. In situations where researchers are trying to make estimates about real hidden populations, they probably won't know how close they are to meeting these conditions. Therefore, researchers may wonder how their estimates will be impacted by violations of these assumptions, both individually (e.g., "How would my estimates be impacted if there was a problem with the survey construction?") and jointly (e.g., "How would my estimate be impacted if there was a problem with my survey construction and reporting behavior?"). To address this concern, in Online Appendix D, we develop a framework for sensitivity analysis that shows researchers exactly how estimates will be impacted by violations of all assumptions, either individually or jointly. Table 3 summarizes the results of our sensitivity framework.

Another problem that researchers face in practice is putting appropriate confidence intervals around estimates. The procedure currently used in scale-up studies was proposed in Killworth et al. (1998b), but it has a number of conceptual problems, and in practice, it produces intervals that are anti-conservative (e.g., the actual coverage rate is lower than the desired coverage rate). Both of these problems—theoretical and empirical—do not seem to be widely appreciated in the scale-up literature. Therefore, instead of the current procedure, we recommend that researchers use the rescaled bootstrap procedure (Rao and Wu, 1988; Rao et al., 1992; Rust and Rao, 1996), which 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 Online Appendix F we review the current scale-up confidence interval procedure and the rescaled bootstrap, highlighting the conceptual advantages of the rescaled bootstrap. Further, we show that the rescaled bootstrap produces slightly better confidence intervals 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). Finally, and somewhat disappointingly, our results show that none of the confidence interval procedures work very well in an absolute sense, a finding that highlights an important problem for future research.

We now provide more specific guidance for researchers based on the data they decide to collect. In Section 4.1 we present recommendations for researchers who collect a sample from both the frame population, F , and the hidden population, H ; and, in Section 4.2, we present recommendations for researchers who only select a sample from the frame population.

4.1 Estimation with samples from F and H

We recommend that researchers who have samples from F and H use a generalized scale-up estimator to produce estimates of N_H (see Section 2):

$$\hat{N}_H = \frac{\hat{y}_{F,H}}{\hat{v}_{H,F}}. \quad (21)$$

For researchers using the generalized scale-up estimator we have three specific recommendations. Of all the conditions needed for consistent and essentially unbiased estimation, the ones most under the control of the researcher are those related to survey construction, and so we recommend that researchers focus on these during the study design phase. In particular, we recommend that the probe alters be designed so that the rate at which the hidden population is visible to the probe alters is the same as the rate at which the hidden population is visible to the frame population (see Result C.2 for a more formal statement, and see Section C.5 for more advice about choosing probe alters). Second, when presenting estimates, we recommend that researchers use the results in Table 3 to also present sensitivity analyses highlighting how the estimates may be impacted by assumptions that are particularly problematic in their setting. Finally, we recommend that researchers produce confidence intervals around their estimate using the rescaled bootstrap procedure, keeping in mind that this will likely produce intervals that are anti-conservative.

We also have three additional recommendations that will facilitate the cumulation of knowledge about the scale-up method. First, although the generalized scale-up estimator does not require aggregate relational data from the frame population about groups of known size, we recommend that researchers collect this data so that the basic and generalized estimators can be compared. Second, we recommend that researchers publish estimates of δ_F and τ_F , although these quantities play no role in the generalized scale-up estimator (Fig. 4). As a body of evidence about these adjustment factors accumulates (e.g., Salganik et al. (2011a); Maghsoudi et al. (2014)), studies that are not able to collect a sample from the hidden population will have an empirical foundation for adjusting basic scale-up estimates, either by borrowing values directly from the literature, or by using published values as the basis for priors in a Bayesian model. Finally, we recommend that researchers design their data collections—both from the frame population and the hidden population—so that size estimates from the generalized scale-up method can be compared to estimates from other methods (see e.g., Salganik et al. (2011a)). For example, if respondent-driven sampling is used to sample from the hidden population, then researchers could use methods that estimate the size of a hidden population from recruitment patterns in the respondent-driven sampling data (Berchenko et al., 2013; Handcock et al., 2014, 2015; Crawford et al., 2015; Wesson et al., 2015; Johnston et al., 2015).

4.2 Estimation with only a sample from F

If researchers cannot collect a sample from the hidden population, we have three recommendations. First, we recommend two simple changes to the basic scale-up estimator

that remove the need to adjust for the frame ratio, ϕ_F . Recall, that the basic scale-up estimator that has been used in previous studies is (see Section 3):

$$\hat{N}_H = \frac{\hat{y}_{F,H}}{\hat{d}_{F,U}} \times N = \frac{\hat{y}_{F,H}}{\hat{d}_{F,U}/N}. \quad (22)$$

Instead of Equation 22, we suggest a new estimator, called the modified basic scale-up estimator, that more directly deals with the fact that researchers sample from the frame population F (typically adults), and not from the entire population U (adults and children):

$$\hat{N}_H = \frac{\hat{y}_{F,H}}{\hat{d}_{F,F}} \times N_F = \frac{\hat{y}_{F,H}}{\hat{d}_{F,F}/N_F} \quad (23)$$

There are two differences between the modified basic scale-up estimator (Equation 23) and the basic scale-up estimator (Equation 22). First, we recommend that researchers estimate $\hat{d}_{F,F}$ (i.e., the total number of connections between adults and adults) rather than $\hat{d}_{F,U}$ (i.e., the total number of connections between adults and everyone). In order to do so, researchers should design the probe alters for the frame population so that they have similar personal networks to the frame population; in Online Appendix B.4 we define this requirement formally, and in Section B.4.1 we provide guidance for choosing the probe alters. Second, we recommend that researchers use N_F rather than N .⁷ These two simple changes remove the need to adjust for the frame ratio ϕ_F and thereby eliminate an assumption about an unmeasured quantity. An improved version of the basic scale-up estimator would then be:

$$\hat{N}_H = \underbrace{\frac{\hat{y}_{F,H}}{(\hat{d}_{F,F}/N_F)}}_{\text{modified basic scale-up}} \times \underbrace{\frac{1}{\hat{\delta}_F} \times \frac{1}{\hat{\tau}_F}}_{\text{adjustment factors}} \quad (24)$$

Our second recommendation is that researchers using the modified basic scale-up estimator (Equation 23) perform a sensitivity analysis using the results in Table 3. In particular, we think that researchers should be explicit about the values that they assume for the adjustment factors δ_F and τ_F . Our third recommendation is that researchers construct confidence intervals using the rescaled bootstrap procedure, while explicitly accounting for the fact that there is uncertainty around the assumed adjustment factors and bearing in mind that this procedure will likely produce intervals that are anti-conservative.

⁷In some cases this difference between N_F and N can be substantial. For example, if F is adults, then in many developing countries, $N \approx 2N_F$.

5 Conclusion and next steps

In this paper, we developed the generalized network scale-up estimator. This new estimator improves upon earlier scale-up estimators in several ways: it enables researchers to use the scale-up method in populations with non-random social mixing and imperfect awareness about membership in the hidden population, and it accommodates data collection with complex sample designs and incomplete sampling frames. We also compared the generalized and basic scale-up estimators, leading us to introduce a framework that makes the design-based assumptions of the basic scale-up estimator precise. Finally, researchers who use either the basic or generalized scale-up estimator can use our results to assess the sensitivity of their size estimates to assumptions.

The approach that we followed to derive the generalized scale-up estimator has three elements, and these elements may prove useful in other problems related to sampling in networks. First, we distinguished between the network of reports and the network of relationships. Second, using the network of reports, we derived a simple identity that permitted us to develop a design-based estimator free of any assumptions about the structure of the network of relationships. Third, we combined data from different types of samples. Together, these three elements may help other researchers in other situations derive relatively simple, design-based estimators that are an important complement to complex, model-based techniques.

Although the generalized scale-up estimator has many attractive features, it also requires that researchers obtain two different samples, one from the frame population and one from the hidden population. In cases where studies of the hidden population are already planned (e.g., the behavioral surveillance studies of the groups most at-risk for HIV/AIDS), the necessary information for the generalized scale-up estimator could be collected at little additional cost by appending a modest number of questions to existing questionnaires. In cases where these studies are not already planned, researchers can either collect their own data from the hidden population, or they can use the modified basic scale-up estimator and borrow estimated adjustment factors from other published studies.

The generalized scale-up estimator, like all estimators, depends on a number of assumptions and we think three of them will be most problematic in practice. First, the estimator depends on the assumption that there are no false positive reports, which is unlikely to be true in all situations. Although we have derived an estimator that works even in the presence of false positive reports (Online Appendix A), we were not able to design a practical data collection procedure that would allow us to estimate one of the terms it requires. Second, the generalized scale-up estimator depends on the assumption that hidden population members have accurate aggregate awareness about visibility (Equation 9). That is, researchers have to assume that hidden population respondents can accurately report whether or not their alters would report them, and we expect this assumption will be difficult to check in most situations. Third, the generalized scale-up estimator depends on having a relative probability sample from the hidden population. Unfortunately, we cannot eliminate any of these assumptions, but we have stated them clearly and we have derived the sensitivity of the estimates to violations of these assumptions, individually and jointly.

Our results and their limitations highlight several directions for further work, in terms of both of improved modeling and improved data collection. We think the most important direction for future modeling is developing estimators in a Bayesian framework, and a recent paper by Maltiel et al. (2015) offers some promising steps in this direction. We see two main advantages of the Bayesian approach in this setting. First, a Bayesian approach would allow researchers to propagate the uncertainty they have about the many assumptions involved in scale-up estimates, whereas our current approach only captures uncertainty introduced by sampling. Further, as more empirical studies produce estimates of the adjustment factors (τ_F and δ_F), a Bayesian framework would permit researchers to borrow values from other studies in a principled way. In terms of future directions for data collection, researchers need practical techniques for estimating the rate of false positive reporting. These estimates, combined with the estimator in Online Appendix A, would permit the relaxation of one of the most important remaining assumptions made by all scale-up studies to date. We hope that the framework introduced in this paper will provide a basis for these and other developments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Aramrattan, A., Kanato, M. Network scale-up method: Application in Thailand. Presented at Consultation on estimating population sizes through household surveys: Successes and challenges; New York, NY. 2012 Mar 28–30.
- Bengtsson L, Thorson A. Global HIV surveillance among MSM: is risk behavior seriously underestimated? *AIDS*. 2010; 24(15):2301–2303. [PubMed: 20634663]
- Berchenko Y, Rosenblatt J, Frost SDW. Modeling and analysing respondent driven sampling as a counting process. 2013 arXiv:1304.3505 [stat].
- Bernard HR, Hallett T, Iovita A, Johnsen EC, Lyerla R, McCarty C, Mahy M, Salganik MJ, Saliuk T, Scutelnicuic O, Shelley GA, Sirinirund P, Weir S, Stroup DF. Counting hard-to-count populations: the network scale-up method for public health. *Sexually Transmitted Infections*. 2010; 86(Suppl 2):ii11–ii15. [PubMed: 21106509]
- Bernard HR, Johnsen EC, Killworth P, Robinson S. Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Social Science Research*. 1991; 20(2):109–121.
- Bernard, HR., Johnsen, EC., Killworth, PD., Robinson, S. Estimating the size of an average personal network and of an event subpopulation. In: Kochen, M., editor. *The Small World*. Ablex Publishing; Norwood, NJ: 1989. p. 159-175.
- Crawford FW, Wu J, Heimer R. Hidden population size estimation from respondent-driven sampling: a network approach. 2015 arXiv:1504.08349 [stat].
- Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal, Complex Systems*. 2006:1695.
- Danenberg, P. R package version 0.4. 2013. functional: Curry, compose, and other higher-order functions.
- Efron, B., Tibshirani, RJ. *An introduction to the bootstrap*. Chapman and Hall/CRC; 1993.
- Ezoe S, Morooka T, Noda T, Sabin ML, Koike S. Population size estimation of men who have sex with men through the network scale-up method in Japan. *PLoS ONE*. 2012; 7(1):e31184. [PubMed: 22563366]
- Feehan DM, Salganik MJ. *The networkreporting package*. 2014

- Gile KJ, Handcock MS. Respondent-driven sampling: An assessment of current methodology. *Sociological methodology*. 2010; 40(1):285–327. [PubMed: 22969167]
- Gile KJ, Handcock MS. Network model-assisted inference from respondent-driven sampling data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2015; 178(3):619–639.
- Gile KJ, Johnston LG, Salganik MJ. Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2015; 178(1):241–269.
- Gilovich T, Savitsky K, Medvec VH. The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology*. 1998; 75(2): 332–346. [PubMed: 9731312]
- Goel S, Mason W, Watts DJ. Real and perceived attitude agreement in social networks. *Journal of Personality and Social Psychology*. 2010; 99(4):611–621. [PubMed: 20731500]
- Goel S, Salganik MJ. Respondent-driven sampling as Markov chain Monte Carlo. *Statistics in Medicine*. 2009; 28(17):2202–2229. [PubMed: 19572381]
- Goel S, Salganik MJ. Assessing respondent-driven sampling. *Proceedings of the National Academy of Science, USA*. 2010; 107(15):6743–6747.
- Guo W, Bao S, Lin W, Wu G, Zhang W, Hladik W, Abdul-Quader A, Bulterys M, Fuller S, Wang L. Estimating the size of HIV key affected populations in Chongqing, China, using the network scale-up method. *PLoS ONE*. 2013; 8(8):e71796. [PubMed: 23967246]
- Handcock MS, Gile KJ, Mar CM. Estimating hidden population size using respondent-driven sampling data. *Electronic Journal of Statistics*. 2014; 8(1):1491–1521. [PubMed: 26180577]
- Handcock MS, Gile KJ, Mar CM. Estimating the size of populations at high risk for HIV using respondent-driven sampling data. *Biometrics*. 2015; 71(1):258–266. [PubMed: 25585794]
- Hartley HO, Ross A. Unbiased ratio estimators. *Nature*. 1954; 174(4423):270–271.
- Heckathorn DD. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*. 1997; 44(2):174–199.
- Heimer R. Critical issues and further questions about respondent-driven sampling: comment on Ramirez-Valles, et al.(2005). *AIDS and Behavior*. 2005; 9(4):403–408. [PubMed: 16344920]
- Jing L, Qu C, Yu H, Wang T, Cui Y. Estimating the sizes of populations at high risk for HIV: A comparison study. *PLoS ONE*. 2014; 9(4):e95601. [PubMed: 24755549]
- Johnston LG, McLaughlin KR, El Rhilani H, Latifi A, Toufik A, Bennani A, Alami K, Elomari B, Handcock MS. Estimating the Size of Hidden Populations Using Respondent-driven Sampling Data: Case Examples from Morocco. *Epidemiology*. 2015; 26(6):846–852. [PubMed: 26258908]
- Joint United Nations Programme on HIV/AIDS. Guidelines on estimating the size of populations most at risk to HIV. UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance; Geneva, Switzerland: 2010.
- Kadushin C, Killworth PD, Bernard HR, Beveridge AA. Scale-up methods as applied to estimates of heroin use. *Journal of Drug Issues*. 2006; 36(2):417–440.
- Karon J, Wejnert C. Statistical methods for the analysis of time–location sampling data. *Journal of Urban Health*. 2012; 89(3):565–586. [PubMed: 22421885]
- Khounigh AJ, Haghdoost AA, SalariLak S, Zeinalzadeh AH, Yousefi-Farkhad R, Mohammadzadeh M, Holakouie-Naieni K. Size estimation of most-at-risk groups of HIV/AIDS using network scale-up in Tabriz, Iran. *Journal of Clinical Research & Governance*. 2014; 3(1):21–26.
- Killworth PD, Johnsen EC, McCarty C, Shelley GA, Bernard H. A social network approach to estimating seroprevalence in the United States. *Social Networks*. 1998a; 20(1):23–50.
- Killworth PD, McCarty C, Bernard HR, Johnsen EC, Domini J, Shelly GA. Two interpretations of reports of knowledge of subpopulation sizes. *Social Networks*. 2003; 25(2):141–160.
- Killworth PD, McCarty C, Bernard HR, Shelley GA, Johnsen EC. Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach. *Evaluation Review*. 1998b; 22(2):289–308. [PubMed: 10183307]
- Killworth PD, McCarty C, Johnsen EC, Bernard HR, Shelley GA. Investigating the variation of personal network size under unknown error conditions. *Sociological Methods & Research*. 2006; 35(1):84–112.

- Kurant M, Markopoulou A, Thiran P. Towards unbiased BFS sampling. Selected Areas in Communications, IEEE Journal on. 2011; 29(9):1799–1809.
- Laumann EO. Friends of urban men: An assessment of accuracy in reporting their socioeconomic attributes, mutual choice, and attitude agreement. Sociometry. 1969; 32(1):54–69.
- Lavallée, P. Indirect sampling. Springer; 2007.
- Li X, Rohe K. Central limit theorems for network driven sampling. 2015 arXiv:1509.04704 [math, stat].
- Maghsoudi A, Baneshi MR, Neydavoodi M, Haghdoost A. Network scale-up correction factors for population size estimation of people who inject drugs and female sex workers in Iran. PLoS ONE. 2014; 9(11):e110917. [PubMed: 25365341]
- Maltiel R, Raftery AE, McCormick TH. Estimating population size using the network scale up method. Annals of Applied Statistics. 2015 (Forthcoming).
- McCarty C, Killworth PD, Bernard HR, Johnsen E, Shelley GA. Comparing two methods for estimating network size. Human Organization. 2001; 60:28–39.
- McCormick T, He R, Kolaczyk E, Zheng T. Surveying hard-to-reach groups through sampled respondents in a social network. Statistics in Biosciences. 2012:1–19.
- McCormick T, Salganik MJ, Zheng T. How many people do you know?: Efficiently estimating personal network size. Journal of the American Statistical Association. 2010; 105(489):59–70. [PubMed: 23729943]
- McCormick, TH., Zheng, T. Adjusting for recall bias in “How many X’s do you know?” surveys. Proceedings of the joint statistical meetings; Salt Lake City, UT. 2007.
- McCreech N, Frost S, Seeley J, Katongole J, Tarsh MN, Ndunguse R, Jichi F, Lunel NL, Maher D, Johnston LG, et al. Evaluation of respondent-driven sampling. Epidemiology (Cambridge, Mass). 2012; 23(1):138.
- Mills HL, Colijn C, Vickerman P, Leslie D, Hope V, Hickman M. Respondent driven sampling and community structure in a population of injecting drug users, Bristol, UK. Drug and alcohol dependence. 2012; 126(3):324–332. [PubMed: 22728045]
- Mouw T, Verdery AM. Network sampling with memory: A proposal for more efficient sampling from social networks. Sociological methodology. 2012; 42(1):206–256. [PubMed: 24159246]
- Paniotto, V., Petrenko, T., Kupriyanov, V., Pakhok, O. Technical report. Kiev International Institute of Sociology; 2009. Estimating the size of populations with high risk for HIV using the network scale-up method.
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2014.
- Rao J, Wu C, Yue K. Some recent work on resampling methods for complex surveys. Survey Methodology. 1992; 18(2):209–217.
- Rao JN, Wu CFJ. Resampling inference with complex survey data. Journal of the American Statistical Association. 1988; 83(401):231–241.
- Rao JNK, Pereira NP. On double ratio estimators. Sankhy : The Indian Journal of Statistics, Series A (1961–2002). 1968; 30(1):83–90.
- Rohe K. Network driven sampling; a critical threshold for design effects. 2015 arXiv:1505.05461 [math, stat].
- Rudolph AE, Fuller CM, Latkin C. The importance of measuring and accounting for potential biases in respondent-driven samples. AIDS and Behavior. 2013; 17(6):2244–2252. [PubMed: 23515641]
- Rust K, Rao J. Variance estimation for complex surveys using replication techniques. Statistical Methods in Medical Research. 1996; 5(3):283–310. [PubMed: 8931197]
- Rwanda Biomedical Center. Technical report. Calverton, Maryland, USA: RBC/IHDPC, SPF, UNAIDS and ICF International; 2012. Estimating the size of key populations at higher risk of HIV through a household survey (ESPHS) Rwanda 2011.
- Salganik MJ. Variance estimation, design effects, and sample size calculations for respondent-driven sampling. Journal of Urban Health. 2006; 83(7):98–112.
- Salganik MJ. Commentary: respondent-driven sampling in the real world. Epidemiology. 2012; 23(1): 148–150. [PubMed: 22157310]

- Salganik MJ, Fazito D, Bertoni N, Abdo AH, Mello MB, Bastos FI. Assessing network scale-up estimates for groups most at risk of HIV/AIDS: Evidence from a multiple-method study of heavy drug users in Curitiba, Brazil. *American Journal of Epidemiology*. 2011a; 174(10):1190–1196. [PubMed: 22003188]
- Salganik MJ, Mello MB, Abdo AH, Bertoni N, Fazito D, Bastos FI. The game of contacts: Estimating the social visibility of groups. *Social Networks*. 2011b; 33(1):70–78. [PubMed: 21318126]
- Sarndal, C-E., Swensson, B., Wretman, J. Model assisted survey sampling. Springer; New York: 1992.
- Scott G. “They got their program, and I got mine”: A cautionary tale concerning the ethical implications of using respondent-driven sampling to study injection drug users. *International Journal of Drug Policy*. 2008; 19(1):42–51. [PubMed: 18226516]
- Scutelnicu, O. Network scale-up method experiences: Republic of Kazakhstan. Presented at Consultation on estimating population sizes through household surveys: Successes and challenges; New York, NY. Mar 28–30. 2012a
- Scutelnicu, O. Network scale-up method experiences: Republic of Moldova. Presented at Consultation on estimating population sizes through household surveys: Successes and challenges; New York, NY. Mar 28–30. 2012b
- Shao J. Impact of the Bootstrap on Sample Surveys. *Statistical Science*. 2003; 18(2):191–198.
- Sheikhzadeh K, Baneshi MR, Afshari M, Haghdoost AA. Comparing direct, network scale-up, and proxy respondent methods in estimating risky behaviors among collegians. *Journal of Substance Use*. 2014:1–5.
- Shelley GA, Bernard HR, Killworth P, Johnsen E, McCarty C. Who knows your HIV status? What HIV+ patients and their network members know about each other. *Social Networks*. 1995; 17(3–4):189–217.
- Shelley GA, Killworth PD, Bernard HR, McCarty C, Johnsen EC, Rice RE. Who knows your HIV status II?: Information propagation within social networks of seropositive people. *Human Organization*. 2006; 65(4):430–444.
- Shokoohi M, Baneshi MR, Haghdoost AA. Size estimation of groups at high risk of HIV/AIDS using network scale up in Kerman, Iran. *International Journal of Preventive Medicine*. 2012; 3(7):471–476. [PubMed: 22891148]
- Sirken MG. Household surveys with multiplicity. *Journal of the American Statistical Association*. 1970; 65(329):257–266.
- Snidero S, Morra B, Corradetti R, Gregori D. Use of the scale-up methods in injury prevention research: An empirical assessment to the case of choking in children. *Social Networks*. 2007; 29(4):527–538.
- Snidero S, Soriani N, Baldi I, Zobec F, Berchiolla P, Gregori D. Scale-up approach in cati surveys for estimating the number of foreign body injuries in the aero-digestive tract in children. *International Journal of Environmental Research and Public Health*. 2012; 9(11):4056–4067. [PubMed: 23202831]
- Snidero S, Zobec F, Berchiolla P, Corradetti R, Gregori D. Question order and interviewer effects in CATI scale-up surveys. *Sociological Methods & Research*. 2009; 38(2):287–305.
- Tillé, Y., Matei, A. R package version 2.7. 2015. *sampling: Survey sampling*.
- Verdery AM, Mouw T, Bauldry S, Mucha PJ. Network structure and biased variance estimation in respondent driven sampling. 2013 arXiv preprint arXiv:1309.5109.
- Volz E, Heckathorn DD. Probability-based estimation theory for respondent-driven sampling. *Journal of Official Statistics*. 2008; 24(1):79–97.
- Wang J, Yang Y, Zhao W, Su H, Zhao Y, Chen Y, Zhang T, Zhang T. Application of Network Scale Up Method in the Estimation of Population Size for Men Who Have Sex with Men in Shanghai, China. *PLoS ONE*. 2015; 10(11)
- Wasserman, S., Faust, K. *Social network analysis*. Cambridge University Press; New York, NY: 1994.
- Wesson P, Handcock MS, McFarland W, Raymond HF. If You Are Not Counted, You Don’t Count: Estimating the Number of African-American Men Who Have Sex with Men in San Francisco Using a Novel Bayesian Approach. *Journal of Urban Health*. 2015; 92(6):1052–1064. [PubMed: 26392276]

- White HC, Boorman SA, Breiger RL. Social structure from multiple networks. I. Blockmodels of roles and positions. *American journal of sociology*. 1976:730–780.
- White RG, Hakim AJ, Salganik MJ, Spiller MW, Johnston LG, Kerr LR, Kendall C, Drake A, Wilson D, Orroth K, et al. Strengthening the reporting of observational studies in epidemiology for respondent-driven sampling studies: ‘STROBE-RDS’ statement. *Journal of Clinical Epidemiology*. 2015
- Wickham, H. *ggplot2: Elegant graphics for data analysis*. Springer; New York: 2009.
- Wickham H. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*. 2011; 40(1):1–29.
- Wickham, H. R package version 0.6.2. 2012. *stringr: Make it easier to work with strings*.
- Wickham, H., Chang, W. R package version 1.4.1. 2013. *devtools: Tools to make developing R code easier*.
- Yamanis TJ, Merli MG, Neely WW, Tian FF, Moody J, Tu X, Gao E. An empirical analysis of the impact of recruitment patterns on RDS estimates among a socially ordered population of female sex workers in China. *Sociological methods & research*. 2013; 42(3):392–425.
- Zheng T, Salganik MJ, Gelman A. How many people do you know in prison?: Using overdispersion in count data to estimate social structure in networks. *Journal of the American Statistical Association*. 2006; 101(474):409–423.

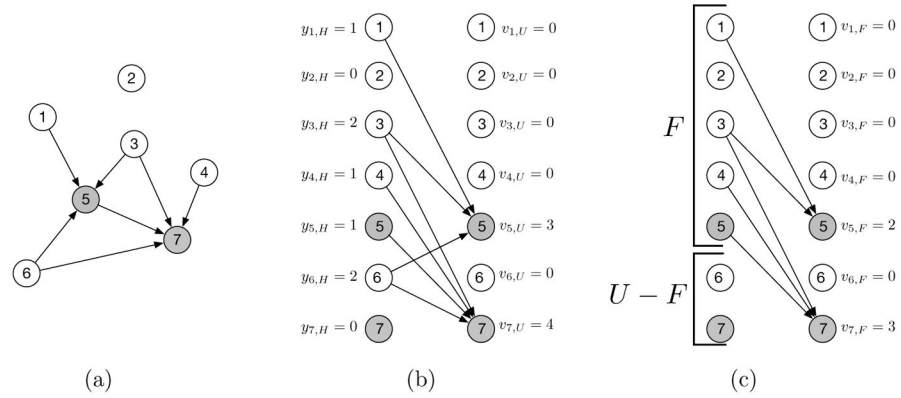


Figure 1. Illustration of the derivation of the generalized scale-up estimator. Panel (a) shows a population of 7 people, 2 of whom are drug injectors (shown in grey). A directed edge $i \rightarrow j$ indicates that i counts j as a drug injector when answering the question “How many drug injectors do you know?” Panel (b) shows the same population, but redrawn so that each person now appears twice: as a source of out-reports, on the left, and as a recipient of in-reports, on the right. This arrangement shows that total out-reports and total in-reports must be equal. Panel (c) shows the same population again, but now some of the people are in the frame population F and some are not. In real scale-up studies, we can only learn about out-reports from the frame population.

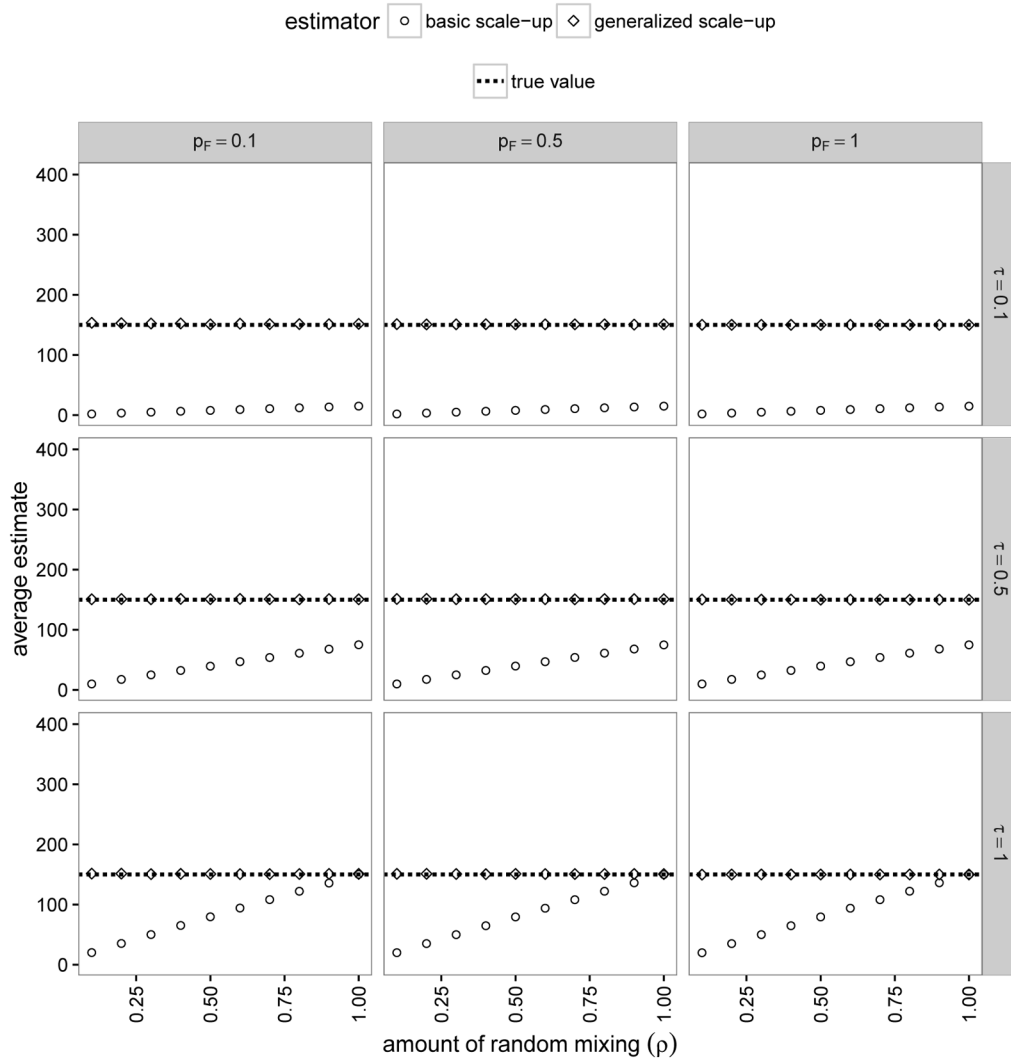


Figure 2. Estimated size of the hidden population for the generalized and basic scale-up estimators. Each panel shows how the two estimators change as the amount of random mixing is varied from low ($\rho = 0.1$; members of the hidden population are relatively unlikely to form contacts with nonmembers) to high ($\rho = 1$; members of the hidden population form contacts independent of other people’s hidden population membership). The columns show results for different sizes of the frame population, from small (left column, $p_F = 0.1$), to large (right column, $p_F = 1$). The rows show results for different levels of reporting accuracy, from a small amount of true positives (top row, $\tau_F = 0.1$), to perfect reporting (bottom row, $\tau_F = 1$). For example, looking at the middle of the center panel, when $p_F = 0.5$, $\tau_F = 0.5$, and $\rho = 0.5$, we see that the average basic scale-up estimate is about 50, while the average generalized scale-up estimate is 150 (the true value). The generalized scale-up estimator is unbiased for all parameter combinations, while the basic scale-up estimator is only unbiased for certain special cases (e.g., when $\rho = 1$, $\tau_F = 1$, and $p_F = 1$). Full details of the simulation are presented in Online Appendix G.

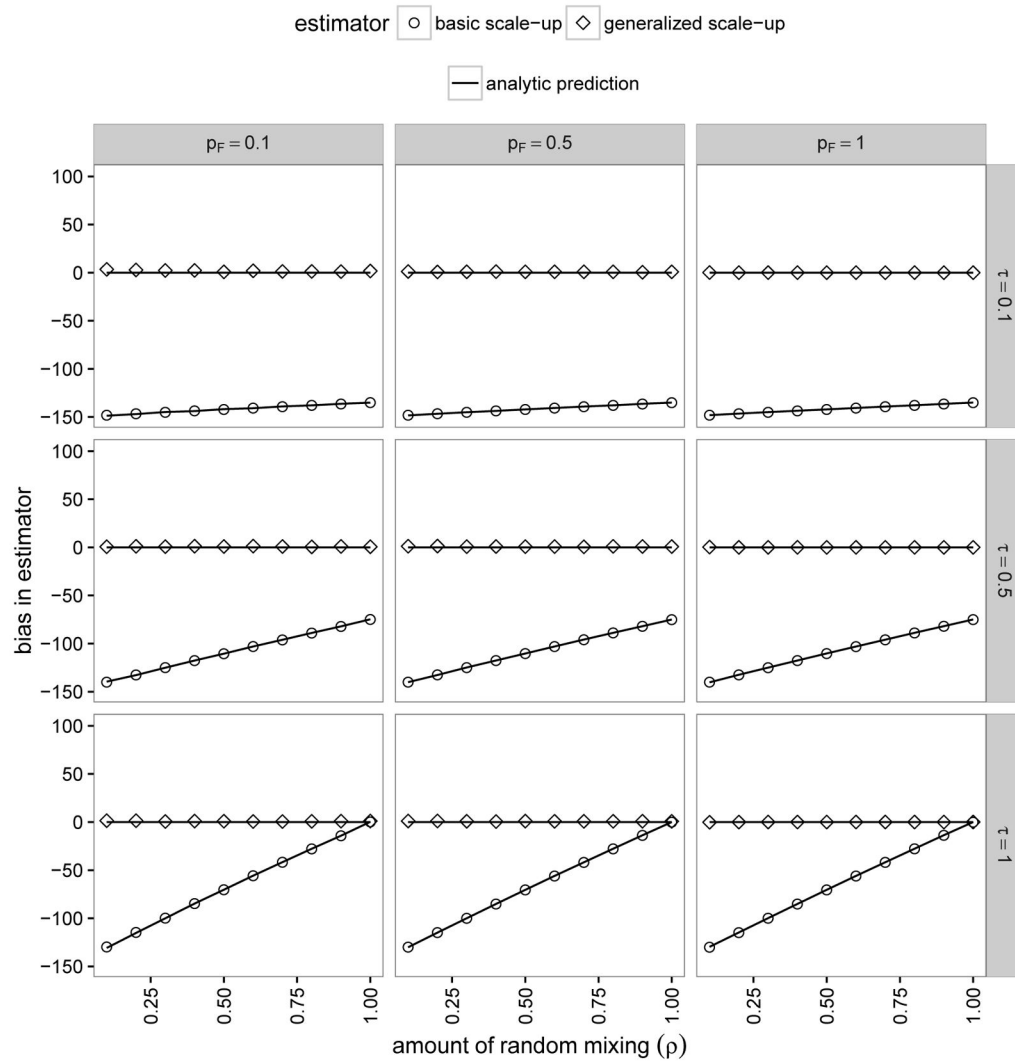


Figure 3. Bias (open circles and diamonds) and predicted bias (solid lines) in the basic scale-up estimates and generalized scale-up estimates for the same parameter configurations depicted in Figure 2. Our analytical results (Equation 20) accurately predict the bias observed in our simulation study.

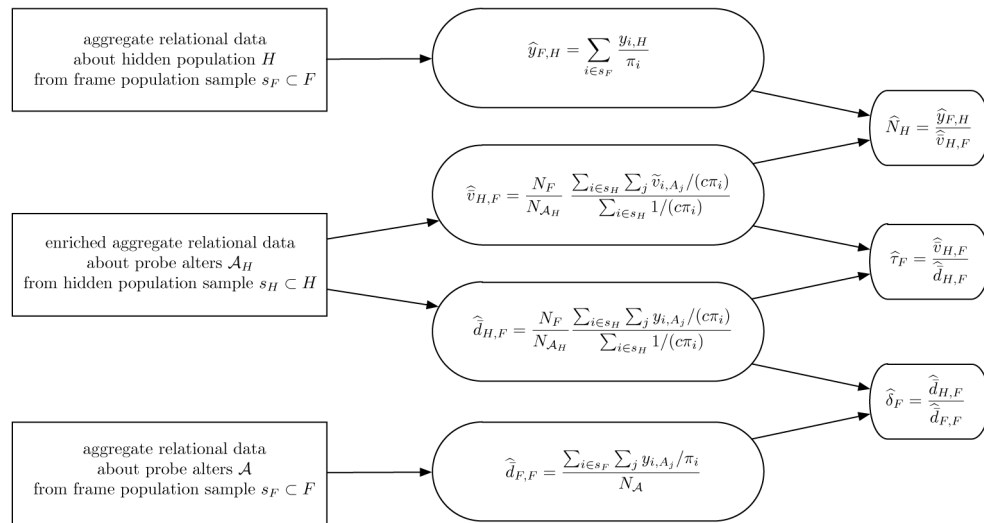


Figure 4. Recommended schematic of inputs and outputs for a study using the generalized scale-up estimator. We recommend that researchers produce size estimates using the generalized scale-up estimator, and that researchers produce estimates of the adjustment factors δ_F and τ_F in order to aid other researchers.

Table 1

Network scale-up studies that have been completed.

Hidden population(s)	Location	Citation
Mortality in earthquake	Mexico City, Mexico	(Bernard et al., 1989)
Rape victims	Mexico City, Mexico	(Bernard et al., 1991)
HIV prevalence, rape, and homelessness	U.S.	(Killworth et al., 1998b)
Heroin use	14 U.S. cities	(Kadushin et al., 2006)
Choking incidents in children	Italy	(Snidero et al., 2007, 2009, 2012)
Groups most at-risk for HIV/AIDS	Ukraine	(Paniotto et al., 2009)
Heavy drug users	Curitiba, Brazil	(Salganik et al., 2011a)
Groups most at-risk for HIV/AIDS	Kerman, Iran	(Shokoohi et al., 2012)
Men who have sex with men	Japan	(Ezoe et al., 2012)
Groups most at-risk for HIV/AIDS	Almaty, Kazakhstan	(Scutelnicu, 2012a)
Groups most at-risk for HIV/AIDS	Moldova	(Scutelnicu, 2012b)
Groups most at-risk for HIV/AIDS	Thailand	(Aramrattan and Kanato, 8 30)
Groups most at-risk for HIV/AIDS	Rwanda	(Rwanda Biomedical Center, 2012)
Groups most at-risk for HIV/AIDS	Chongqing, China	(Guo et al., 2013)
Groups most at-risk for HIV/AIDS	Tabriz, Iran	(Khounigh et al., 2014)
Men who have sex with men	Taiyuan, China	(Jing et al., 2014)
Drug and alcohol users	Kerman, Iran	(Sheikhzadeh et al., 2014)
Men who have sex with men	Shanghai, China	(Wang et al., 2015)

Table 2

Summary of the conditions needed for the generalized and modified basic network scale-up estimators, and their components, to produce estimates that are consistent and essentially unbiased. This table uses the version of the basic scale-up estimator we recommend in Section 4.2.

Quantity	Conditions required	Condition type	Result
reported connections to $H(\hat{y}_{F,H})$	1. probability sample from F	sampling	B.1
average personal network size of F ($\hat{d}_{F,F}$)	1. probability sample from F 2. groups of known size total is accurate $N_{\mathcal{A}}$ 3. probe alter condition ($\bar{d}_{\mathcal{A},F} = \bar{d}_{F,F}$) 4. accurate reporting condition ($y_{F,\mathcal{A}} = d_{F,\mathcal{A}}$)	sampling survey construction survey construction reporting behavior	B.3
average visibility of $H(\hat{v}_{H,F})$	1. relative probability sample from H 2. groups of known size total is accurate $N_{\mathcal{A} \cap H \cap F}$ 3. probe alter condition ($\frac{v_{H,\mathcal{A} \cap F}}{N_{\mathcal{A} \cap F}} = \frac{v_{H,F}}{N_F}$) 4. accurate aggregate reports about visibility ($\bar{v}_{H,\mathcal{A} \cap F} = v_{H,\mathcal{A} \cap F}$)	sampling survey construction survey construction reporting behavior	C.2
generalized scale-up ($\hat{N}_H = \frac{\hat{y}_{F,H}}{\hat{v}_{H,F}}$)	1. conditions needed for $\hat{y}_{F,H}$ 2. conditions needed for $\hat{v}_{H,F}$ 3. no false positive reports about connections to $H(\eta_F = 1)$	sampling sampling, survey construction, reporting behavior reporting behavior	C.8
modified basic scale-up ($\hat{N}_H = \frac{\hat{y}_{F,H}}{\hat{d}_{F,F}}$)	1. conditions needed for $\hat{y}_{F,H}$ 2. condition needed for $\hat{d}_{F,F}$ 3. no false positive reports about connections to $H(\eta_F = 1)$ 4. members of H and members of F have same average personal network size ($\delta_F = 1$) 5. no false negative reports about connections to H ($\tau_F = 1$)	sampling sampling, survey construction, reporting behavior reporting behavior network structure reporting behavior	Sections 2–3

Table 3

Analytical expressions that researchers can use to perform sensitivity analysis for estimates made using scale-up estimators (see Online Appendix D for more detail). K_{F_1} , K_{F_2} , and K_H are indices that reflect how imperfect the sampling weights researchers use to make estimates are; when these K values are 0, the weights are exactly correct; the farther they are from 0, the more imperfect the weights are. (NB: we use the symbol \rightsquigarrow as a shorthand for ‘is consistent and essentially unbiased for’.)

Quantity	Conditions required	Adjusted estimand for sensitivity analysis
generalized scale-up $\hat{N}_H = \frac{y_{F,H}}{\bar{v}_{H,F}}$	1	probability sample from F with accurate weights ($K_{F_2} = 0$ and $\bar{e}_F = 1$)
	2	relative probability sample from H with accurate weights ($K_H = 0$)
	3	conditions needed for $\hat{v}_{H,F}$ ($\hat{v}_{F,H} = \kappa \bar{v}_{F,H}$)
	4	no false positive reports about connections to H ($\eta_F = 1$)
modified basic scale-up $\hat{N}_H = \frac{y_{F,H}}{\bar{d}_{F,F}}$	1	probability sample from F with accurate weights for $y_{F,H}$ ($K_{F_2} = 0$)
	2	probability sample from F with accurate weights for $y_{F,\mathcal{A}}$ ($K_{F_1} = 0$)
	3	condition needed for $\hat{d}_{F,F}$ ($\hat{d}_{F,F} = \kappa \bar{d}_{F,F}$)
	4	no false positive reports about connections to H ($\eta_F = 1$)
	5	members of H and members of F have same average personal network size ($\delta_F = 1$)
	6	no false negative reports about connections to H ($\tau_F = 1$)

$$\hat{N}_H \cdot \frac{(1+K_H)}{\bar{e}_F(1+K_{F_2})} \cdot \kappa \cdot \eta_F \rightsquigarrow N_H$$

$$\hat{N}_H \cdot \frac{(1+K_{F_1})}{(1+K_{F_2})} \cdot \kappa \cdot \frac{\eta_F}{\delta_F \tau_F} \rightsquigarrow N_H$$