Online Appendixes

B Estimates with a sample from F

In this appendix, we present the full results for all of the estimators that require a sample from the frame population. First, we describe the general requirements that our sampling design for F must satisfy (Section B.1). Then we describe how to estimate the total number of out-reports, $y_{F,H}$ (Section B.2). Next we turn to some background material on multisets (Section B.3), which is needed for the following section on the known population method for estimating network degree (Section B.4). Finally, we present an estimator for the frame ratio, ϕ_F, which makes use of the known population method results (Section B.5).

B.1 Requirements for sampling designs from F

We follow Sarndal et al. (1992)'s definition of a probability sampling design, which we repeat here for convenience. Suppose that we have a set of possible samples $\{s_1, \ldots, s_j, \ldots, s_{\max}\}$, with each $s_j \subset F$. Furthermore, suppose $p(s_j)$ gives the probability of selection for each possible sample s_j. If we select a sample s_F at random using a process that will produce each possible sample s_j with probability $p(s_j)$, and if every element $i \in F$ has a nonzero probability of inclusion $\pi_i > 0$, then we will say that we have selected a *probability sample* and we call $p(\cdot)$ the *sampling design*.

B.2 Estimating the total number of out-reports, $y_{F,H}$

If we have a probability sample from the frame then estimating the total number of out-reports is a straightforward application of a standard survey estimator.

Result B.1 *Suppose we have a sample s_F taken from the frame population using a probability sampling design with probabilities of inclusion given by π_i (Sec. B.1). Then* *the estimator given by*

$$\widehat{y}_{F,H} = \sum_{i \in s_F} y_{i,H} / \pi_i \tag{B.1}$$

is consistent and unbiased for $y_{F,H}$.

Proof: This follows from the fact that Equation B.1 is a Horvitz-Thompson estimator (Sarndal et al., 1992, Section 2.8). ■

B.3 Reporting about multisets

Appendix B.4 and Appendix C both describe strategies that involve asking respondents to answer questions about their network alters in specific groups. In this section, we develop the notation and some basic properties of responses generated this way; these properties will be then be used in the subsequent sections.

Suppose we have several groups $A_1, \ldots, A_J$ with $A_j \subset U$ for all j, and also a frame population F of potential interviewees. (Note that we do not require $A_j \subset F$.) Imagine concatenating all of the people in populations $A_1, \ldots, A_J$ together, repeating each individual once for each population she is in. The result, which we call the *probe alters*, $\mathcal{A}$, is a multiset. The size of $\mathcal{A}$ is $N_{\mathcal{A}} = \sum_j N_{A_j}$.

Let y_{i,A_j} be the number of members of group A_j that respondent i reports having among the members of her personal network. We also write $y_{i,\mathcal{A}} = \sum_j y_{i,A_j}$ for the sum of the responses for individual i across all of $A_1, \ldots, A_J$, and $y_{F,\mathcal{A}} = \sum_{i \in F} \sum_j y_{i,A_j}$ to denote the total number of reports from F to $\mathcal{A}$. Similarly, we write $d_{i,\mathcal{A}} = \sum_j d_{i,A_j}$ for the sum of the network connections from individual i to each $A_1, \ldots, A_J$, and $d_{F,\mathcal{A}} = \sum_{i \in F} \sum_j d_{i,A_j}$ for the total of the individual $d_{i,\mathcal{A}}$ taken over all i. As always, we will write averages with respect to the first subscript so that, for example, $\bar{d}_{\mathcal{A},F} =$

$d_{\mathcal{A},F}/N_{\mathcal{A}}$.

We now derive a property of estimation under multisets that will be useful later on. Roughly, this property says that we can estimate the total number of reports from the entire frame population to the entire multiset of probe alters using only a sample from the frame population with known probabilities of inclusion (Section B.1). While this property might seem intuitive, we state it formally for two reasons. First, by stating it explicitly, we show that this property is very general: it does not require any assumptions about the contact pattern between the frame population and probe alters, nor does it require any assumptions about the probe alters. Second, it will turn out to be useful in several later proofs, and so we state it for compactness.

Property B.2 *Suppose we have a sample s_F from F taken using a probability sampling design with probabilities of inclusion π_i (Section B.1). Then*

$$\widehat{y}_{F,\mathcal{A}} = \sum_{i \in s_F} y_{i,\mathcal{A}}/\pi_i \tag{B.2}$$

is a consistent and unbiased estimator for $y_{F,\mathcal{A}}$.

Proof: If we define $a_i = \sum_j y_{i,A_j}$, the sum of the responses to each A_j for individual i, then we can write our estimator as

$$\widehat{y}_{F,\mathcal{A}} = \sum_{i \in s_F} a_i/\pi_i. \tag{B.3}$$

This is a Horvitz-Thompson esimator (see, e.g., Sarndal et al., 1992, chap. 2); it is unbiased and consistent for the total $\sum_{i \in F} a_i = y_{F,\mathcal{A}}$. ■

B.4 Network degree and the known population method for estimating $\bar{d}_{F,F}$, $\bar{d}_{F,U}$, and $\bar{d}_{U,F}$

In order to conduct a scale-up study, we need a definition of the network that we will ask respondents to tell us about; that is, we need to define what it will mean for two members of the population to be connected by an edge. To date, most scale-up studies have used slight variations of the same definition: the respondent is told that she should consider someone a member of her network if she "knows" the person, where to know someone means (i) you know her and she knows you; (ii) you have been in contact in the past 2 years; and, (iii), if needed, you could get in touch with her (Bernard et al., 2010). Of course, many other definitions are possible, and an investigation of this issue is a matter for future study. The only restriction on the tie definition we impose here is that it be reciprocal; that is, the definition must imply that if the respondent is connected to someone, then that person is also connected to the respondent.

For a particular definition of a network tie an individual i's degree, $d_{i,U}$ may not be very easy to directly observe, even if the network is conceptually well-defined. For the basic scale-up estimator, the most commonly used technique for estimating respondents' network sizes is called the known population method (Killworth et al., 1998a; Bernard et al., 2010).[8] The known population method is based on the idea that we can estimate a respondent's network size by asking how many connections she has to a number of different groups whose sizes are known. The more connections a respondent reports to these groups, the larger we estimate her network to be. Current

[8] There are other techniques for estimating personal network size, including the summation method (McCarty et al., 2001; Bernard et al., 2010), which could be used in conjunction with many of our results. We focus on the known population method here because it is relatively easy to work with from a statistical perspective, and also because there is some evidence that it works better in practice (Salganik et al., 2011a; Rwanda Biomedical Center, 2012)

standard practice is to ask a respondent about her connections to approximately 20 groups of known size in order to estimate her degree (Bernard et al., 2010), although the exact number of groups used has no impact on the bias of the estimates as we show in Results B.3 and B.4.

The known population estimator was originally introduced to estimate the personal network size of each respondent individually (Killworth et al., 1998a), but in Sections 3 and 4.2 we showed that for the scale-up method the quantity of interest is actually the average number of connections from a member of the frame population F to the rest of the frame population F ($\bar{d}_{F,F}$), or the average number of connections from a member of the entire population U to the frame population F ($\bar{d}_{U,F}$).[9] This is fortunate, because it is easier to estimate an average degree over all respondents than it is to estimate the individual degree for each respondent.

B.4.1 Guidance for choosing the probe alters, $\mathcal{A}$

Result B.3, below, shows that the known population estimator will produce consistent and unbiased estimates of average network degree if (i) $y_{F,\mathcal{A}} = d_{F,\mathcal{A}}$ (*reporting condition*); and (ii) $\bar{d}_{\mathcal{A},F} = \bar{d}_{F,F}$ (*probe alter condition*). Stating these conditions precisely enables us to provide guidance about how the groups of known size $(A_1, A_2, \dots A_J)$ should be selected such that the probe alters $\mathcal{A}$ will enable consistent and unbiased estimates.

First, the reporting condition ($y_{F,\mathcal{A}} = d_{F,\mathcal{A}}$) in Result B.3 shows that researchers should select probe alters such that reporting will be accurate in aggregate. One way to make the reporting condition more likely to hold is to select groups that are unlikely to suffer from transmission error (Shelley et al., 1995, 2006; Killworth

[9] Although we have framed our discussion here in terms of $\bar{d}_{F,F}$, the same ideas apply to $\bar{d}_{U,F}$ and $\bar{d}_{F,U}$.

et al., 2006; Salganik et al., 2011b; Maltiel et al., 2015). Another way to make the reporting condition more likely to hold is to avoid selecting groups that may lead to recall error (Killworth et al., 2003; Zheng et al., 2006; McCormick and Zheng, 2007; McCormick et al., 2010; Maltiel et al., 2015). That is, previous work suggests that respondents seem to under-report the number of connections they have to large groups, although the precise mechanism behind this pattern is unclear (Killworth et al., 2003). Researchers who have data that may include recall error can consider some of the empirically-calibrated adjustments that have been used in earlier studies (Zheng et al., 2006; McCormick and Zheng, 2007; McCormick et al., 2010; Maltiel et al., 2015).

Second, the probe alter condition ($\bar{d}_{\mathcal{A},F} = \bar{d}_{F,F}$) in Result B.3 shows that researchers should select groups to be typical of F in terms of their connections to F. In most applied situations, we expect that F will consist of adults, so that researchers should choose groups of known size that are composed of adults, or that are typical of adults in terms of their connections to adults. Further, when trying to choose groups that satisfy the probe alter condition, it is useful to understand how connections from the individual known populations to the frame ($\bar{d}_{A_1,F}, \ldots, \bar{d}_{A_J,F}$) aggregate up into connections from the probe alters to the frame ($\bar{d}_{\mathcal{A},F}$). Basic algebraic manipulation shows that the probe alter condition can be written as:

$$\frac{\sum_j \bar{d}_{A_j,F} \; N_{A_j}}{\sum_j N_{A_j}} = \bar{d}_{F,F}. \tag{B.4}$$

Equation B.4 reveals that the probe alter condition requires that $\bar{d}_{F,F}$ is equal to a weighted average of the average number of connections between each individual known population A_j and the frame population F ($\bar{d}_{A_j,F}$). The weights are given by the size of each known population, N_{A_j}. The simplest way that this could be satisfied is if $\bar{d}_{A_j,F} = \bar{d}_{F,F}$ for every known population A_j. If this is not true, then the probe alter condition can still hold as long as groups for which $\bar{d}_{A_j,F}$ is too high are offset by other groups for which $\bar{d}_{A_{j'},F}$ is too low.

In practice it may be difficult to determine if the reporting condition and probe alter condition will be satisfied. Therefore, we recommend that researchers assess the sensitivity of their size estimates using the procedures described in Online Appendix D. Further, we note that in many realistic situations, N_{A_j} might not be known exactly. Fortunately, researchers only need to know $\sum_j N_{A_j}$, and they can assess the sensitivity of their estimates to errors in the size of known populations using the procedures described in Online Appendix D.

B.4.2 The known population estimators

Given that background about selecting the probe alters, we present the formal results for the known population estimators for $\bar{d}_{F,F}$, $\bar{d}_{U,F}$, and $\bar{d}_{F,U}$.

Result B.3 *Suppose we have a sample s_F taken from the frame population using a probability sampling design with probabilities of inclusion given by π_i (see Section B.1). Suppose also that we have a multiset of known populations, $\mathcal{A}$. Then the known population estimator given by*

$$\widehat{\bar{d}}_{F,F} = \frac{\sum_{i \in s_F} \sum_j y_{i,A_j}/\pi_i}{N_{\mathcal{A}}} \tag{B.5}$$

is consistent and unbiased for $\bar{d}_{F,F}$ if

$$y_{F,\mathcal{A}} = d_{F,\mathcal{A}}, \qquad \textit{(reporting condition)} \tag{B.6}$$

and if

$$\bar{d}_{\mathcal{A},F} = \bar{d}_{F,F}. \qquad \textit{(probe alter condition)} \qquad \text{(B.7)}$$

Proof: By Property B.2, we know that our estimator is unbiased and consistent for $y_{F,\mathcal{A}}/N_{\mathcal{A}}$. By the reporting condition in Equation B.6, this means it is unbiased and consistent for $d_{F,\mathcal{A}}/N_{\mathcal{A}}$. Then, by the probe alter condition in Equation B.7, it is also unbiased and consistent for $\bar{d}_{F,F}$. ■

Result B.4 *Suppose we have a sample s_F taken from the frame population using a probability sampling design with probabilities of inclusion given by π_i (see Section B.1). Suppose also that we have a multiset of known populations, $\mathcal{A}$. Then the known population estimator given by*

$$\widehat{\bar{d}}_{U,F} = \frac{\sum_{i \in s_F} \sum_j y_{i,A_j}/\pi_i}{N_{\mathcal{A}}} \qquad \text{(B.8)}$$

is consistent and unbiased for $\bar{d}_{U,F}$ if

$$y_{F,\mathcal{A}} = d_{F,\mathcal{A}}, \qquad \textit{(reporting condition)} \qquad \text{(B.9)}$$

and if

$$\bar{d}_{\mathcal{A},F} = \bar{d}_{U,F}. \qquad \textit{(probe alter condition)} \qquad \text{(B.10)}$$

Proof: By Property B.2, we know that our estimator is unbiased and consistent for $y_{F,\mathcal{A}}/N_{\mathcal{A}}$. By the reporting condition in Equation B.9, this means it is unbiased and consistent for $d_{F,\mathcal{A}}/N_{\mathcal{A}}$. Then, by the probe alter condition in Equation B.10, it is also unbiased and consistent for $\bar{d}_{U,F}$. ■

Since $\bar{d}_{F,U} = \frac{N}{N_F}\bar{d}_{U,F}$, as a direct consequence of Result B.4 we have the following corollary.

Corollary B.5 *If the conditions described in Result B.4 hold,*

$$\widehat{\bar{d}}_{F,U} = \widehat{\bar{d}}_{U,F}\;\frac{N}{N_F} \tag{B.11}$$

is consistent and unbiased for $\bar{d}_{F,U}$.

B.5 Estimating the frame ratio, ϕ_F

Given our estimator of $\bar{d}_{F,F}$ (Result B.3) and our estimator of $\bar{d}_{U,F}$ (Result B.4), we can estimate the frame ratio, ϕ_F.

Result B.6 *The estimator*

$$\widehat{\phi}_F = \frac{\widehat{\bar{d}}_{F,F}}{\widehat{\bar{d}}_{U,F}} \tag{B.12}$$

is consistent and essentially unbiased for ϕ_F *if* $\widehat{\bar{d}}_{F,F}$ *is consistent and essentially unbiased for* $\bar{d}_{F,F}$ *and* $\widehat{\bar{d}}_{U,F}$ *is consistent and essentially unbiased for* $\bar{d}_{U,F}$.

Proof: This follows from the properties of a ratio estimator (Sarndal et al., 1992, chap. 5). ■

More concretely, combining the estimator for $\bar{d}_{F,F}$ (Result B.3) and the estimator for $\bar{d}_{U,F}$ (Result B.4), and assuming that we have known populations $\mathcal{A}_{F_1}$ for $\bar{d}_{F,F}$, and $\mathcal{A}_{F_2}$ for $\bar{d}_{U,F}$, we obtain

$$\widehat{\phi}_F = \frac{N_{\mathcal{A}_{F_2}}}{N_{\mathcal{A}_{F_1}}}\;\frac{\sum_{i \in s_F} \sum_{A_j \in \mathcal{A}_{F_1}} y_{i,A_j}/\pi_i}{\sum_{i \in s_F} \sum_{A_k \in \mathcal{A}_{F_2}} y_{i,A_k}/\pi_i}. \tag{B.13}$$

In our discussion of $\widehat{\bar{d}}_{F,F}$ (Result B.3) and $\widehat{\bar{d}}_{U,F}$ (Result B.4), we concluded that we want the known populations $\mathcal{A}_{F_1}$ used for $\widehat{\bar{d}}_{F,F}$ to be typical of members of F in their connections to F. An analogous argument shows that we want the known populations $\mathcal{A}_{F_2}$ used for $\widehat{\bar{d}}_{U,F}$ to be typical of members of U in their connections to F. In general, we expect that it will not be appealing to assume that F and U are similar to each other in terms of their connections to F meaning that, unfortunately, it will not make sense to use the same set of known populations for $\widehat{\bar{d}}_{F,F}$ and $\widehat{\bar{d}}_{U,F}$. If researchers wish to estimate ϕ_F directly, one approach would be to choose $\mathcal{A}_{F_2}$ to be typical of U in such a way that some of the individual known populations are more typical of F, while others more typical of $U - F$. The multiset formed from only the ones that are more typical of F could then be our choice for $\mathcal{A}_{F_1}$. In this case, researchers would also want $\frac{N_{\mathcal{A}_{F_1}}}{N_{\mathcal{A}_{F_2}}} \approx \frac{N_F}{N}$. This complication is one of the reasons we recommend in Section 4 that future scale-up studies estimate $\bar{d}_{F,F}$ directly, thus avoiding the need to estimate ϕ_F entirely.

C Estimates with samples from F and H

In this appendix, we present the full results for all of the estimators that require a sample from the hidden population. Section C.1 defines the general requirements that our sampling design for H must satisfy. Section C.2 describes a flexible data collection procedure called the game of contacts. Section C.3 introduces some background material on estimation using questions about multisets and presents an estimator for $\bar{v}_{H,F}$, the average number of in-reports among the members of the hidden population. Section C.5 gives some guidance about how to choose the probe alters for the known population method. Section C.6 presents estimators for the two adjustment factors introduced in Section 3: the degree ratio, δ_F, and the true positive rate, τ_F. Finally,