Online Appendixes

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, Section C.7 presents formal results for four different estimators for N_H.

C.1 Requirements for sampling designs from H

For the results that involve a sample from the hidden population s_H, we do not need a probability sample (Appendix B); instead, we need a weaker type of design. We require that every element $i \in H$ have a nonzero probability of selection $\pi_i > 0$, and that we can determine the probability of selection up to a constant factor c; that is, we only need to know $c\pi_i$. We are not aware of any existing name for this situation, so we will call it a *relative probability sample*. Because of the challenges involved in sampling hard-to-reach populations, the two most likely sampling designs for s_H will probably be time-location sampling (Karon and Wejnert, 2012) and respondent-driven sampling (Heckathorn, 1997). A relative probability sample allows us to use weighted sample means to estimate averages, but not totals. See Sarndal et al. (1992, Section 5.7) for more details on weighted sample means, also sometimes called Hájek estimators, which is what we use to estimate averages from a sample of hidden population members.

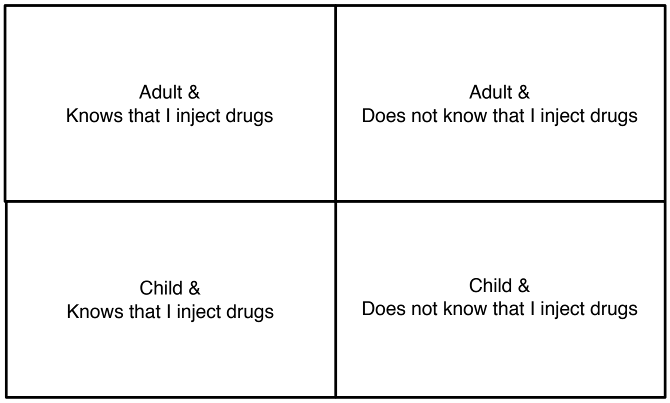
C.2 Data collection

In order to make estimates about the hidden population's visibility to the frame population, researchers will need to collect what we call *enriched aggregate relational data* from each respondent, and a procedure called the *game of contacts* has produced promising results from a study of heavy drug users in Brazil (Salganik et al., 2011b). In the main text, we assumed that the groups in the probe alters $A_1, \ldots, A_J$ were all contained in the frame population ($A_j \subset F$ for all j). However, the estimators in this Online Appendix are more general because they allow for the possibility that some of the groups $A_1, \ldots A_J$ may not be contained entirely in F. For example, if the frame population is adults, then this flexibility enables researchers to use groups based on names, such as Michael, even though not all people named Michael are adults.

In order to allow for this flexibility, we need to introduce some new notation: let $A_1 \cap F, A_2 \cap F, \ldots, A_J \cap F$ be the intersection of these groups and the frame population, and let $\mathcal{A} \cap F$ be the concatenation of these intersected groups. For example, if the frame population is adults, A_1 is people named Michael, and A_2 is doctors, then $A_1 \cap F$ is adults named Michael, $A_2 \cap F$ is adult doctors, and $\mathcal{A} \cap F$ is the collection of all adult Michaels and all adult doctors, with adult doctors named Michael included twice. (In the special case discussed in the main text, $A_1 \cap F, \ldots A_J \cap F = A_1, \ldots, A_J$.)

The data collection begins with a relative probability sample (Section C.1) from the hidden population. For a set of groups, $A_1, A_2, \ldots A_J$, each respondent in the hidden population is asked, "How many people do you know in group A_j?" We call the response y_{i,A_j}. Next for each of the y_{i,A_j} alters, the respondent picks up a token and places it on a game board like the one in Figure C.1. From the location of the tokens on the board, the researcher can record whether each alter is in the frame population (or not) and whether the alter is aware that the respondent is in the hidden population (or not) (Table C.2). This process is then repeated until the respondent has been asked about all groups.

If all of the probe alters are in the frame population, then the process is much easier for respondents and the game board can be modified to collect alternative information. If all of the probe alters are not in the frame population, then it is important for the researcher to define the frame population as clearly as possible. If the respondents are not able to correctly indicate whether the alters are in the frame population or not, it could lead to biased estimates of $\bar{v}_{H,F}$. For more on the operational implementation of this procedure, see Salganik et al. (2011b).



Adult & Knows that I inject drugs	Adult & Does not know that I inject drugs
Child & Knows that I inject drugs	Child & Does not know that I inject drugs

Figure C.1: Example of a game board that could be used in the game of contacts interviewing procedure if the hidden population was people who inject drugs and the frame was made up of adults. This board is a variation of the board used in Salganik et al. (2011b).

	aware	not aware	total
frame population	$\widetilde{v}_{i,A_j \cap F}$	$\widetilde{h}_{i,A_j \cap F}$	$y_{i,A_j \cap F}$
not frame population	$\widetilde{v}_{i,A_j \cap (U-F)}$	$\widetilde{h}_{i,A_j \cap (U-F)}$	$y_{i,A_j \cap (U-F)}$
total	$\widetilde{v}_{i,A_j}$	$\widetilde{h}_{i,A_j}$	y_{i,A_j}

Table C.1: Responses collected during the game of contacts for each respondent i and each group A_j. We use $\widetilde{\ }$ to indicate reported values. For example, $\widetilde{v}_{i,A_j}$ is the respondent's reported visibility to people in A_j and v_{i,A_j} is respondent's actual visiblility to people in A_j. Also, using this notational convention, it is the case that $y_{i,A_j} = \widetilde{d}_{i,A_j}$, but we have written y_{i,A_j} in order to be consistent with the rest of the paper.

C.3 Estimation using aggregated relational data from the hidden population

In this section, we follow Section B.3 and present another useful property about estimates made using aggregate relational data from the hidden population. Roughly, this property says that we can estimate the average number of reports from the entire hidden population to the probe alters using only a relative probability sample from the hidden population (Section C.1). Similar to Property B.2, the result we present below does not require any assumptions about the contact pattern between the hidden population and the probe alters, nor about the probe alters themselves.

Property C.1 *Suppose we have a sample s_H from H taken using a relative probability design, allowing us to compute the relative probabilities of inclusion $c\pi_i$ for all sampled elements (Sec. C.1). Then*

$$\widehat{\bar{y}}_{H,\mathcal{A}} = \frac{\sum_{i \in s_H} y_{i,\mathcal{A}}/(c\pi_i)}{\sum_{i \in s_H} 1/(c\pi_i)} \tag{C.1}$$

is a consistent and essentially unbiased estimator for $\bar{y}_{H,\mathcal{A}} = y_{H,\mathcal{A}}/N_H$.

Proof: Note that the c in the relative probabilities of inclusion $c\pi_i$ cancel, so that

$$\widehat{\bar{y}}_{H,\mathcal{A}} = \frac{\sum_{i \in s_H} y_{i,\mathcal{A}}/(\pi_i)}{\sum_{i \in s_H} 1/(\pi_i)}. \tag{C.2}$$

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

$$\widehat{\bar{y}}_{H,\mathcal{A}} = \frac{\sum_{i \in s_H} a_i/\pi_i}{\sum_{i \in s_H} 1/\pi_i}. \tag{C.3}$$

Now we have a standard weighted mean estimator (e.g. Sarndal et al., 1992, chap. 5); it is consistent and essentially unbiased for the average $\frac{1}{N_H}\sum_{i \in H} a_i = y_{H,\mathcal{A}}/N_H$. ■

C.4 Estimating the average visibility, $\bar{v}_{H,F}$

Given the data collection procedure described in Sec. C.2, we can estimate the average visibility ($\bar{v}_{H,F}$) as long as three conditions are satisfied: one about reporting, one about the visibility of the hidden population to the probe alters, and one about sampling.

Result C.2 *Assume that we have a sample s_H taken from the hidden population using a relative probability design with relative probabilities of inclusion $c\pi_i$ for all sampled elements (Sec. C.1). Then*

$$\widehat{\bar{v}}_{H,F} = \frac{N_F}{N_{\mathcal{A} \cap F}} \frac{\sum_{i \in s_H} \sum_j \widetilde{v}_{i,A_j \cap F}/(c\pi_i)}{\sum_{i \in s_H} 1/(c\pi_i)} \tag{C.4}$$

is consistent and essentially unbiased for $\bar{v}_{H,F}$ if

$$\widetilde{v}_{H,\mathcal{A} \cap F} = v_{H,\mathcal{A} \cap F}, \qquad \textit{(reporting condition)} \tag{C.5}$$

and

$$\frac{v_{H,\mathcal{A} \cap F}}{N_{\mathcal{A} \cap F}} = \frac{v_{H,F}}{N_F}. \qquad \textit{(probe alter condition)} \tag{C.6}$$

Proof: Property C.1 holds for estimating $\overline{\widetilde{v}}_{F,\mathcal{A} \cap F}$ from $\widetilde{v}_{i,\mathcal{A} \cap F}$, just as it holds for estimating $\bar{y}_{H,\mathcal{A} \cap F}$ from $y_{i,\mathcal{A} \cap F}$. Applying Property C.1 here, we conclude that the estimator is consistent and essentially unbiased for

$$\frac{N_F}{N_{\mathcal{A}\cap F}} \bar{\widetilde{v}}_{H,\mathcal{A}\cap F} = \frac{N_F}{N_{\mathcal{A}\cap F}} \frac{\widetilde{v}_{H,\mathcal{A}\cap F}}{N_H}. \tag{C.7}$$

Next, by applying the reporting condition in Equation C.5 we can conclude that

$$\frac{N_F}{N_{\mathcal{A}\cap F}} \frac{\widetilde{v}_{H,\mathcal{A}\cap F}}{N_H} = \frac{N_F}{N_{\mathcal{A}\cap F}} \frac{v_{H,\mathcal{A}\cap F}}{N_H}. \tag{C.8}$$

Finally, by applying the probe alter condition in Equation C.6 and rearranging terms, we conclude that

$$\frac{N_F}{N_{\mathcal{A}\cap F}} \frac{v_{H,\mathcal{A}\cap F}}{N_H} = \frac{N_F}{N_H} \frac{v_{H,F}}{N_F} \tag{C.9}$$

$$= \bar{v}_{H,F} \tag{C.10}$$

■

Note that Result C.2 requires us to know the size of the probe alters in the frame population, $N_{\mathcal{A}\cap F}$. In some cases, this may not be readily available, but it may be reasonable to assume that

$$N_{\mathcal{A}\cap F} = \frac{N_F}{N}\ N_{\mathcal{A}}. \tag{C.11}$$

Furthermore, if $\mathcal{A}$ is chosen so that all of its members are in F, then $N_{\mathcal{A}\cap F} = N_{\mathcal{A}}$ and $v_{i,A_j\cap F} = v_{i,A_j}$. In this situation, we do not need to specifically ask respondents about connections to $\mathcal{A} \cap F$; we can just ask about connections to $\mathcal{A}$.

The reporting condition required for Result C.5 states that the hidden population's total reported visibility from the probe alters on the frame must be correct. This might not be the case, if for example, respondents systematically over-estimate how much others know about them (see e.g., Gilovich et al. (1998)).

The required condition for the probe alters is slightly more complex. It needs to be the case 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. There are several equivalent ways of stating this condition, as we show in a moment. First, we need to define two new quantities: the individual-level true positive rate and the average of the individual-level true positive rates.

Definition 1 *We define the individual-level true positive rate for respondent $i \in F$ to be*

$$\tau_i = \frac{v_{H,i}}{d_{i,H}}, \tag{C.12}$$

where $v_{H,i} = \sum_{j \in H} v_{j,i}$.

Definition 2 *We define the average of the individual true positive rates over a set F of respondents as*

$$\overline{\tau}_F = \frac{1}{N_F} \sum_{i \in F} \tau_i. \tag{C.13}$$

In general, $\overline{\tau}_F \neq \tau_F$. To see this, note that while $\overline{\tau}_F$ is the average of the individual-level true positive rates with each individual weighted equally, τ_F can be written as the weighted average of the individual true positive rates, with the weights given by each individual's degree. We can see the exact relationship between the two by expressing τ_F in terms of the τ_i:

$$\tau_F = \frac{\sum_{i \in F} \tau_i \ d_{i,H}}{\sum_{i \in F} d_{i,H}}, \tag{C.14}$$

since multiplying each τ_i by $d_{i,H}$ and summing is the same as summing the $v_{H,i}$.

Result C.3 *The following conditions are all equivalent.*

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

(ii) $\tau_{\mathcal{A}\cap F}\ \bar{d}_{\mathcal{A}\cap F,H} = \tau_F\ \bar{d}_{F,H}$

(iii) $\overline{\tau}_{\mathcal{A}\cap F}\ \bar{d}_{\mathcal{A}\cap F,H} + cov_{\mathcal{A}\cap F}(\tau_i, d_{i,H}) = \overline{\tau}_F\ \bar{d}_{F,H} + cov_F(\tau_i, d_{i,H})$

(iv) $\bar{y}^{+}_{F,H} = \frac{\sum_j \bar{y}^{+}_{A_j\cap F,H}\ N_{A_j\cap F}}{\sum_j N_{A_j\cap F}}$,

where cov_F is the finite-population covariance taken over the set F.[10]

Proof: First, we show that

$$\tau_{\mathcal{A}\cap F}\ \bar{d}_{\mathcal{A}\cap F,H} = \tau_F\ \bar{d}_{F,H} \Longleftrightarrow \frac{v_{H,\mathcal{A}\cap F}}{N_{\mathcal{A}\cap F}} = \frac{v_{H,F}}{N_F}. \tag{C.15}$$

By definition, $\tau_F\ \bar{d}_{F,H} = (v_{H,F}/d_{F,H}) \times (d_{F,H}/N_F) = v_{H,F}/N_F$. The same argument demonstrates that $\tau_{\mathcal{A}\cap F}\ \bar{d}_{\mathcal{A}\cap F,H} = v_{H,\mathcal{A}\cap F}/N_{\mathcal{A}}$. We conclude that $(i) \Longleftrightarrow (ii)$.

Next, we show that (ii) is equivalent to (iii). We can use the relationship between τ_F and the τ_i, Equation C.14, to deduce that

$$\tau_F\ d_{F,H} = \sum_{i\in F} \tau_i\ d_{i,H} = N_F\ [\overline{\tau}_F\ \bar{d}_{F,H} + \text{cov}_F(\tau_i, d_{i,H})]. \tag{C.16}$$

Dividing the left-most and right-most sides by N_F, we conclude that

$$\tau_F\ \bar{d}_{F,H} = \overline{\tau}_F\ \bar{d}_{F,H} + \text{cov}_F(\tau_i, d_{i,H}). \tag{C.17}$$

[10] We define the finite-population covariance to have a denominator of N_F; this differs from some other authors, who define the finite-population covariance to have $N_F - 1$ in the denominator.

The same argument shows that

$$\bar{d}_{\mathcal{A}\cap F,H}\ \tau_{\mathcal{A}\cap F} = \overline{\tau}_{\mathcal{A}\cap F}\ \bar{d}_{\mathcal{A}\cap F,H} + \mathrm{cov}_{\mathcal{A}\cap F}(\tau_i, d_{i,H}). \tag{C.18}$$

So we conclude that $(ii) \iff (iii)$.

Finally, we show that (iv) is equivalent to (i). In Appendix A, showed that $y^{+}_{F,H} = v_{H,F}$ (Equation A.3). Dividing both sides by N_F, we have $\bar{y}^{+}_{F,H} = v_{H,F}/N_H$, which is the right-hand side of the identity in (i). Similarly, starting with the left-hand side of the identity in (i), we have

$$\frac{v_{H,\mathcal{A}\cap F}}{N_{\mathcal{A}\cap F}} = \frac{\sum_j v_{H,A_j\cap F}}{\sum_j N_{A_j\cap F}} = \frac{\sum_j y^{+}_{A_j\cap F,H}}{\sum_j N_{A_j\cap F}} = \frac{\sum_j \bar{y}^{+}_{A_j\cap F,H}\ N_{A_j\cap F}}{\sum_j N_{A_j\cap F}}. \tag{C.19}$$

So we conclude that $(i) \iff (iv)$.

Since $(i) \iff (ii)$ and $(ii) \iff (iii)$, it follows that $(i) \iff (iii)$. Furthermore, since $(i) \iff (iv)$, it follows that (iv) is equivalent to (ii) and (iii). ■

Result C.3 shows that the probe alter condition can be expressed in many equivalent ways. One of these alternate expressions is especially useful because it leads to an empirical check of the probe alter condition that future scale-up studies can implement. This empirical check is a direct consequence of Result C.4, below. Intuitively, Result C.4 and the empirical check are a consequence of the identity in Equation 1, which says that in-reports from the perspective of H are also out-reports from the perspective of F.

Result C.4 *Suppose that the precision of out-reports from the frame population is the same as the precision of the out-reports from $\mathcal{A}\cap F$:*

$$\frac{y^{+}_{F,H}}{y_{F,H}} = \frac{y^{+}_{\mathcal{A}\cap F,H}}{y_{\mathcal{A}\cap F,H}} \tag{C.20}$$

Then the probe alter condition (C.6) is satisfied if and only if

$$\bar{y}_{F,H} = \bar{y}_{\mathcal{A} \cap F,H}. \tag{C.21}$$

Proof: First, note that, by Result C.3, the probe alter condition is equivalent to

$$\bar{y}^{+}_{F,H} = \frac{\sum_j \bar{y}^{+}_{A_j \cap F,H} \; N_{A_j \cap F}}{\sum_j N_{A_j \cap F}}. \tag{C.22}$$

Since $\bar{y}^{+}_{A_j \cap F,H} = y^{+}_{A_j \cap F,H}/N_{A_j \cap F}$ for all j, the right-hand side of Equation C.22 is equal to $\bar{y}^{+}_{\mathcal{A} \cap F,H}$, meaning that the probe alter condition is also equivalent to

$$\bar{y}^{+}_{F,H} = \bar{y}^{+}_{\mathcal{A} \cap F,H}. \tag{C.23}$$

Second, note that the assumption in Equation C.20 can be re-written as

$$\frac{\bar{y}^{+}_{F,H}}{\bar{y}_{F,H}} = \frac{\bar{y}^{+}_{\mathcal{A} \cap F,H}}{\bar{y}_{\mathcal{A} \cap F,H}}, \tag{C.24}$$

by multiplying the left-hand side by $\frac{N_F}{N_F}$ and the right-hand side by $\frac{N_{\mathcal{A} \cap F}}{N_{\mathcal{A} \cap F}}$. So we are left with the task of showing that if Equation C.24 is true, then Equation C.23 is satisfied if and only if Equation C.21 is satisfied. But this is the case, since Equation C.23 equates the numerators of the two fractions in Equation C.24 and Equation C.21 equates the denominators of the two fractions in Equation C.24. Two fractions that are equal will have equal numerators if and only if they have equal denominators. (Formally, if $a/b = c/d$ then $a = c$ if and only if $b = d$.) ■

The implication of Result C.4 is that if (i) researchers design the probe alters so that the frame population sample s_F can be used to estimate $\bar{y}_{\mathcal{A} \cap F,H}$; and (ii) researchers assume that the precision of out-reports from the frame population is the same as the precision of out-reports from $\mathcal{A} \cap F$, then they can evaluate how well the probe alter condition is satisfied empirically by comparing $\widehat{\widehat{y}}_{F,H}$ and $\widehat{\widehat{y}}_{\mathcal{A} \cap F,H}$.

Finally, we can foresee four practical problems that might arise when researchers try to estimate $\bar{v}_{H,F}$. First, researchers might not be able to choose the probe alters to satisfy the probe alter condition (Equation C.6) because of limited information about the true visibility of the hidden population with respect to different social groups. A second problem might arise if researchers are not able to choose the probe alters to satisfy the reporting condition (Equation C.5) because of limited information about the hidden population's awareness about visibility. A third problem might arise due to errors in administrative records that would cause researchers to have incorrect information about the size of the multiset of probe alters on the frame ($N_{\mathcal{A} \cap F}$). Finally, a fourth problem might arise due to errors in the sampling method researchers use. Fortunately, as we show in Online Appendix D (Result D.6), it is possible to quantify the effect of these problems on the resulting estimates. In some cases they can cancel out, but in other cases they magnify each other.

C.5 Guidance for choosing the probe alters for the game of contacts, $\mathcal{A}$

Turning the results in Online Appendix C into easy to follow steps for selecting the probe alters for the game of contacts is an open and important research problem. Here, we briefly offer three recommendations for selecting the probe alters for the game of contacts. We realize that these recommendations may be difficult to follow exactly in practice. Therefore, we also discuss the sensitivity of the estimators to errors in the construction of the probe alters. Finally, we discuss one type of data that should be collected from the frame population in order to help the researchers evaluate their choice of probe alters for the game of contacts.

First, we recommend that probe alters for the game of contacts be in the frame population. For example, if the frame population is adults, we recommend that all members of the probe alters be adults. This choice will simplify the data collection task in the game of contacts, and for all the advice listed below, we assume that it has been followed. If it is not possible, researchers can still use the more general procedures developed in this Online Appendix.

Second, we recommend that the probe alters be selected such that the probe alter condition in Result C.2 is satisfied. That is, the probe alters as a whole should be typical of the frame population in the following way: it should be the case 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 ($\frac{v_{H,\mathcal{A}}}{N_{\mathcal{A}}} = \frac{v_{H,F}}{N_F}$). 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.

Third, we recommend that the probe alters be selected so that the reporting condition in Result C.2 is satisfied ($\widetilde{v}_{H,\mathcal{A}} = v_{H,\mathcal{A}}$). One way to help ensure that this condition holds is to avoid selecting large groups that may cause recall error (Killworth et al., 2003; Zheng et al., 2006; McCormick and Zheng, 2007; McCormick et al., 2010; Maltiel et al., 2015). In practice it might be difficult to meet each of these three conditions exactly, therefore we recommend a sensitivity analysis using the results in Online Appendix D.

Finally, the choice of probe alters for the game of contacts also has two implications for the design of the survey of the frame population. First, if researchers wish to estimate the degree ratio, δ_F, then they should design the probe alters $\mathcal{A}$ so that they can be asked of both members of the hidden population sample and members of the frame population sample (see Result C.6). Second, if researchers wish to test the probe alter condition using the approach in Result C.4, then additional information needs to be collected from each member of the frame population sample. For example, if one group in the probe alters for the game of contacts is postal workers, then members of the frame population sample should be asked if they are postal workers.

C.6 Term-by-term: δ_F and τ_F

In this section we describe how to estimate two adjustment factors: the degree ratio,

$$\delta_F = \frac{\bar{d}_{H,F}}{\bar{d}_{F,F}} \tag{C.25}$$

and the true positive rate,

$$\tau_F = \frac{\bar{v}_{H,F}}{\bar{d}_{H,F}}. \tag{C.26}$$

Estimating the degree ratio requires information from the survey of the hidden population and the survey of the frame population, while estimating the true positive rate only requires information from the survey of the hidden population (Fig. C.2). As Equations C.25 and C.26 make clear, both adjustment factors involve $\bar{d}_{H,F}$ so we first present an estimator for that quantity.

Result C.5 *Suppose we have a sample s_H taken from the hidden population using a relative probability sampling design with relative probabilities of inclusion denoted $c\pi_i$*

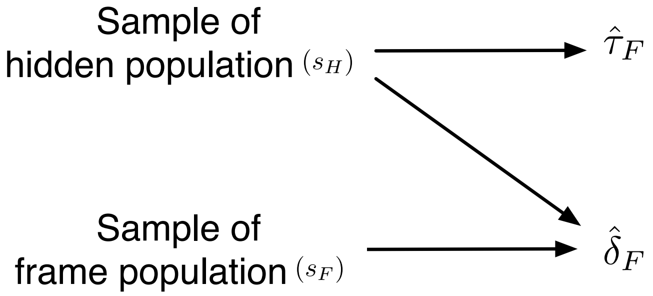

Sample of hidden population (s_H)
Sample of frame population (s_F)
$\hat{\tau}_F$
$\hat{\delta}_F$

Figure C.2: We estimate the true positive rate $\widehat{\tau}_F$ using data from the survey of the hidden population, and we estimate the degree ratio $\widehat{\delta}_F$ using the sample of the hidden population and the sample of the frame population.

(Sec C.1). Then the estimator given by

$$\widehat{\bar{d}}_{H,F} = \frac{N_F}{N_{\mathcal{A}\cap F}} \frac{\sum_{i \in s_H} \sum_j y_{i,(A_j \cap F)}/(c\pi_i)}{\sum_{i \in s_H} 1/(c\pi_i)} \tag{C.27}$$

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

$$y_{H,\mathcal{A}\cap F} = d_{H,\mathcal{A}\cap F}, \qquad \textit{(reporting condition)} \tag{C.28}$$

and

$$\bar{d}_{\mathcal{A}\cap F,H} = \bar{d}_{F,H}. \qquad \textit{(probe alter condition)} \tag{C.29}$$

Proof: From Property C.1, we can see that our estimator is consistent and essentially unbiased for

$$\frac{N_F}{N_{\mathcal{A}\cap F}} \frac{y_{H,\mathcal{A}\cap F}}{N_H} = \frac{N_F}{N_H} \frac{y_{H,\mathcal{A}\cap F}}{N_{\mathcal{A}\cap F}}. \tag{C.30}$$

Under the reporting condition (Equation C.28) this becomes

$$\frac{N_F}{N_H}\frac{y_{H,\mathcal{A}\cap F}}{N_{\mathcal{A}\cap F}} = \frac{N_F}{N_H}\frac{d_{H,\mathcal{A}\cap F}}{N_{\mathcal{A}\cap F}} \tag{C.31}$$

Finally, applying the probe alter condition in Equation C.29, we have

$$\frac{N_F}{N_H}\frac{d_{H,\mathcal{A}\cap F}}{N_{\mathcal{A}\cap F}} = \frac{N_F}{N_H}\frac{d_{F,H}}{N_F} \tag{C.32}$$

$$= \bar{d}_{H,F}. \tag{C.33}$$

■

Result C.5 requires that reports are, in total, correct (Equation C.28). Like Result C.2, Result C.5 also requires us to know the size of the probe alters on the frame, $N_{\mathcal{A}\cap F}$. In some cases, this may not be readily available, but it may be reasonable to assume that

$$N_{\mathcal{A}\cap F} = \frac{N_F}{N}\, N_{\mathcal{A}}. \tag{C.34}$$

Furthermore, if $\mathcal{A}$ is chosen so that all of its members are in F, then $N_{\mathcal{A}\cap F} = N_{\mathcal{A}}$ and $y_{i,A_j\cap F} = y_{i,A_j}$. In this situation, we do not need to specifically ask respondents about connections to $\mathcal{A}\cap F$; we can just ask about connections to $\mathcal{A}$. Result C.5 also requires a specific rate of connectivity between the probe alters and the hidden population (Equation C.29). We discussed some of the consequences of these assumption in the main text, where we made recommendations for practice (Section 4).

C.6.1 Estimating the degree ratio, δ_F

We can combine our estimator for $\bar{d}_{H,F}$ (Result C.5) and our estimator for $\bar{d}_{F,F}$ (Result B.3), to estimate the degree ratio, δ_F.

Result C.6 *The estimator*

$$\widehat{\delta}_F = \frac{\widehat{\bar{d}}_{H,F}}{\widehat{\bar{d}}_{F,F}} \tag{C.35}$$

is consistent and essentially unbiased for δ_F if $\widehat{\bar{d}}_{H,F}$ is consistent and essentially unbiased for $\bar{d}_{H,F}$ and $\widehat{\bar{d}}_{F,F}$ is consistent and essentially unbiased for $\bar{d}_{F,F}$.

Proof: This follows from the properties of a compound ratio estimator (Online Appendix E). ■

More concretely, combing the estimators in Result C.5 and Result B.3, results in an estimator for $\widehat{\delta}_F$ with the following form:

$$\widehat{\delta}_F = \frac{\frac{N_F}{N_{\mathcal{A}_H \cap F}} \frac{\sum_{i \in s_H} \sum_{A_j \in \mathcal{A}_H} y_{i,(A_j \cap F)}/(c\pi_i^H)}{\sum_{i \in s_H} 1/(c\pi_i^H)}}{\frac{1}{N_{\mathcal{A}_F}} \sum_{i \in s_F} \sum_{A_k \in \mathcal{A}_F} y_{i,A_k}/\pi_i^F}. \tag{C.36}$$

If the probe alters for the frame population and the hidden population are the same, so that $\mathcal{A}_H = \mathcal{A}_F = \mathcal{A}$, and if the probe alters are randomly distributed in the frame population in the sense that

$$N_{\mathcal{A} \cap F} = N_{\mathcal{A}} \, \frac{N_F}{N}, \tag{C.37}$$

then we can reduce the constants in front of Equation C.36 to

$$\frac{\frac{N_F}{N_{\mathcal{A} \cap F}}}{\frac{1}{N_{\mathcal{A}}}} = \frac{\frac{N}{N_{\mathcal{A}}}}{\frac{1}{N_{\mathcal{A}}}} = N. \tag{C.38}$$

In other words, when the probe alters for the frame and hidden population are the same, and when the probe alters are randomly distributed in the frame population, all of the factors involving the size of $\mathcal{A}$ drop out. This fact allows researchers to use groups defined by first names (e.g., people named Michael) in the probe alters $\mathcal{A}$, even if the size of these groups is not known, as long as it is reasonable to assume that $\mathcal{A}$ satisfies Equation C.37 (c.f., Salganik et al. (2011a)).

C.6.2 Estimating the true positive rate, τ_F

We can combine our estimator for $\bar{v}_{H,F}$ (Result C.2) and our estimator for $\bar{d}_{H,F}$ (Result C.5) to estimate the true positive rate τ_F.

Result C.7 *The estimator*

$$\widehat{\tau}_F = \frac{\widehat{\bar{v}}_{H,F}}{\widehat{\bar{d}}_{H,F}} \tag{C.39}$$

is consistent and essentially unbiased for τ_F if $\widehat{\bar{v}}_{H,F}$ is a consistent and essentially unbiased estimator of $\bar{v}_{H,F}$ and if $\widehat{\bar{d}}_{H,F}$ is a consistent and essentially unbiased estimator of $\bar{d}_{H,F}$.

Proof: This follows directly from the properties of a compound ratio estimator (Online Appendix E). ■

More concretely, combing the estimator in Result C.2 and Result C.5 yields an estimator for $\widehat{\tau}_F$ with the following form:

$$\widehat{\tau}_F = \frac{\sum_{i \in s_H} \widetilde{v}_{i,\mathcal{A}_H}/(c\pi_i)}{\sum_{i \in s_H} y_{i,\mathcal{A}_H}/(c\pi_i)}. \tag{C.40}$$

All of the factors involving the size of $\mathcal{A}$ drop out of Equation C.40. This fact allows researchers to use groups defined by first names (e.g., people named Michael) in the probe alters $\mathcal{A}$, even if the size of these groups is not known (c.f., Salganik et al. (2011b)).

C.7 Estimating the size of the hidden population, N_H

We now make use of all of the results for the individual terms we derived above to present four different estimators for the size of the hidden population, N_H.

Result C.8 *The generalized scale-up estimator given by*

$$\widehat{N}_H = \frac{\widehat{y}_{F,H}}{\widehat{\bar{v}}_{H,F}} \tag{C.41}$$

is consistent and essentially unbiased for N_H if there are no false positive reports, if $\widehat{y}_{F,H}$ is consistent and unbiased for $y_{F,H}$, and if $\widehat{\bar{v}}_{H,F}$ is consistent and essentially unbiased for $\bar{v}_{H,F}$.

Proof: From the properties of a compound ratio estimator, we know that our estimator is consistent and essentially unbiased for $y_{F,H}/\bar{v}_{H,F}$ (Appendix E). By the argument in the main text given in Section 2, leading to Equation 5, this quantity is equal to N_H. ■

Result C.9 *The adjusted basic scale-up estimator given by*

$$\widehat{N}_H = \frac{\widehat{y}_{F,H}}{\widehat{\bar{d}}_{U,F}} \frac{1}{\widehat{\phi}_F} \frac{1}{\widehat{\delta}_F} \frac{1}{\widehat{\tau}_F} \tag{C.42}$$

is consistent and essentially unbiased for N_H if there are no false positive reports, and if each of the individual estimators is consistent and essentially unbiased.

Proof: From the results in Online Appendix E, we know that this compound ratio estimator will be consistent and essentially unbiased for $y_{F,H}/(\bar{d}_{U,F}\ \phi_F\ \delta_F\ \tau_F)$. The denominator is $\bar{v}_{H,F}$ by construction, leaving us with $y_{F,H}/\bar{v}_{H,F}$. By the argument in the main text given in Section 2, leading to Equation 5, this quantity is equal to N_H.

■

Result C.10 *The adjusted scale-up estimator*

$$\widehat{N}_H = \frac{\widehat{y}_{F,H}}{\widehat{\bar{d}}_{F,F}} \, \frac{1}{\widehat{\delta}_F} \, \frac{1}{\widehat{\tau}_F} \tag{C.43}$$

is consistent and essentially unbiased for N_H if there are no false positives, and if each of the individual estimators is consistent and essentially unbiased.

Proof: From the results in Online Appendix E, we know that this compound ratio estimator will be consistent and essentially unbiased for $y_{F,H}/(\bar{d}_{F,F} \; \delta_F \; \tau_F)$. The denominator is $\bar{v}_{H,F}$ by construction, leaving us with $y_{F,H}/\bar{v}_{H,F}$. By the argument in the main text given in Section 2, leading to Equation 5, this quantity is equal to N_H.

■

Result C.11 *The adjusted scale-up estimator*

$$\widehat{N}_H = \frac{\widehat{y}_{F,H}}{\widehat{\bar{d}}_{F,F}} \, \frac{1}{\widehat{\delta}_F} \, \frac{1}{\widehat{\tau}_F} \, \widehat{\eta}_F \tag{C.44}$$

is consistent and essentially unbiased for N_H if each of the individual estimators is consistent and essentially unbiased.

Proof: From the results in Online Appendix E, we know that this compound ratio estimator will be consistent and essentially unbiased for $(y_{F,H} \; \eta_F)/(\bar{d}_{F,F} \; \delta_F \; \tau_F)$. The numerator is $y^{+}_{F,H}$ by construction and the product of the denominators is $\bar{v}_{H,F}$ by construction, leaving us with $y^{+}_{F,H}/\bar{v}_{H,F}$. By the argument in Online Appendix A this quantity is equal to N_H. ■