Survey methods for estimating the size of weak-tie personal networks*

Dennis M. Feehan[†], Vo Hai Son[‡], Abu Abdul-Quader[§]

May 19, 2021

Abstract

Researchers increasingly use *aggregate relational data* to learn about the size and distribution of survey respondents’ weak-tie personal networks. Aggregate relational data are collected by asking questions about respondents’ connectedness to many different groups (e.g., “How many teachers do you know?”). This approach can be powerful but, to make use of aggregate relational data, researchers must locate external information about the size of each of these groups from a census or from administrative records (e.g., the number of teachers in the population). This need for external information makes aggregate relational data difficult or impossible to collect in many settings. Here, we show that relatively simple modifications can overcome this need for external data, significantly increasing the flexibility of the method and weakening key assumptions required by the associated estimators. Our methods are appropriate for using a sample survey to study the size and distribution of weak-tie network connections. Our methods can also be used as part of the network scale-up method to estimate the size of hidden populations. We illustrate our approach with two empirical studies: a large simulation study, and original household survey data collected in Hanoi, Vietnam.

*We thank Ashton Verdery and Casey Breen for helpful comments on drafts of this manuscript. We are also grateful to Nga Nguyen and Ali Safarnejad of UNAIDS Hanoi for help during data collection, and to Hanh Bui of UC Berkeley for research assistance. DMF thanks the Berkeley Population Center, which is funded by the Eunice Kennedy Shriver National Institute of Child Health and Development (P2C HD 073964) and the Berkeley Center for the Economics and Demography of Aging, which is funded by the National Institute on Aging (5P30AG012839). The views and opinions expressed in this paper are those of the authors and not of the organizations or agencies where the authors are employed. Replication code and data will be made publicly available upon publication.

[†]Department of Demography, UC Berkeley, feehan@berkeley.edu (corresponding author)

[‡]Vietnam Ministry of Health

[§]US CDC Hanoi

1 Introduction and Background

Many population-level insights about social networks come from surveys that investigate relatively strong ties, such as the ones produced by name generators (e.g., Fischer 1982; Marsden 2005; Perry, Pescosolido, and Borgatti 2018). Comparatively few studies have examined weaker ties, but there are several reasons why weak ties can be important: first, information about weak ties is needed to understand homophily, segregation, and social capital (e.g., DiPrete et al. 2011; Lubbers, Molina, and Valenzuela-García 2019; McCormick, Salganik, and Zheng 2010; Zheng, Salganik, and Gelman 2006); second, weak ties are hypothesized to play a critical role in important social processes, such as social support, information diffusion and social influence, (Granovetter 1973; e.g., Lubbers, Molina, and Valenzuela-García 2019); and, third, reports about weak ties can be instrumentally useful in trying to measure the size and composition of hidden and hard to reach populations through network scale-up and related methods (e.g., Bernard et al. 2010; Feehan and Salganik 2016a; Maltiel et al. 2015).

Researchers have developed several methods to try and estimate the size of weak tie networks from a sample of people, but this has proven to be a challenging problem (de Sola Pool and Kochen 1978; P. D. Killworth et al. 1998; P. D. Killworth et al. 1998; McCarty et al. 1997; McCarty et al. 2001). One promising approach is based on collecting *aggregate relational data*. A growing number of studies has collected aggregate relational data, typically with one of two goals: (1) using aggregate relational data to study of patterns of connectedness and segregation in weak-tie networks themselves (DiPrete et al. 2011; Lubbers, Molina, and Valenzuela-García 2019; McCormick, Salganik, and Zheng 2010; Zheng, Salganik, and Gelman 2006); (2) using aggregate relational data in conjunction with the network scale-up method to estimate the size and composition of hidden populations, such as people who have died, people who use drugs, men who have sex with men, women who have an abortion, and others (Bernard et al. 1989; Bernard et al. 2010; Feehan et al. 2016; P. D. Killworth et al. 1998; Salganik et al. 2011; Sully, Giorgio, and Anjur-Dietrich 2020; Teo et al. 2019).

Collecting aggregate relational data starts from a probability sample, typically respondents to a survey. Respondents are asked a series of questions like “How many people do you know who are in group X?” where X typically ranges over several different kinds of groups, possibly including names, occupations, membership in civic organizations, or other salient characteristics. We call these *probe groups*, because they are used to probe the size and composition of survey respondents’ networks. The idea is that respondents’ reported connections to these probe groups contain information about the structure of the population’s network: for example, survey respondents who have large personal networks will tend to report more connections to these probe groups.

The *known population method* combines aggregate relational data with information about the size of each probe group to produce estimates for the size of respondents’ personal networks. The known population

method has several appealing features: it can easily be added to standard social science surveys; it permits researchers to conduct internal consistency checks that can help detect problems with data collection or methodological assumptions (Bernard et al. 2010; Feehan et al. 2016; P. D. Killworth et al. 1998); it produces a simple and easy to use estimator whose assumptions are clear (Feehan and Salganik 2016a); and it can serve as the basis for more complex models that estimate features of the social network from which the aggregate relational data were drawn (Maltiel et al. 2015; McCormick, Salganik, and Zheng 2010; Zheng, Salganik, and Gelman 2006).

Despite its advantages, the known population method can be difficult to use in practice because it requires researchers to locate a set of probe groups that satisfy several criteria. Standard practice requires that:

- (1) The true size of each probe group is accurately known from censuses or administrative records (P. D. Killworth et al. 1998; McCarty et al. 1997); typically, this information about size needs to also distinguish between children and adults, and it needs to coincide with the geography of the study;
- (2) Survey respondents are able to understand who is and who is not a member of each probe group; so, for example, if researchers find administrative records that have the number of public school teachers in a city, then survey respondents must be asked to report about only public school teachers (i.e., the school teachers in the administrative records, rather than all school teachers);
- (3) There are at least 20 *probe groups* (Bernard et al. 2010);
- (4) The probe groups are just the right size: if they are too small, then many respondents will report no connections to any probe group members (since there a small group will have a small chance of being connected to the survey respondents); on the other hand, if they are too large, then respondent may ‘guesstimate’ the number they know, rather than carefully counting. Currently, the recommendation is that the probe groups should be between about 0.1% to 4% of the population (Bernard et al. 2010);
- (5) Together, the set of probe groups is representative of the general population (in a sense we make precise in Appendix B) (Feehan and Salganik 2016a; Killworth et al. 1990; McCormick, Salganik, and Zheng 2010).

To date, important work in helping to address these challenges has focused on statistical innovations, including approaches that aim to understand possible sources of bias in the known population method analytically (Feehan and Salganik 2016a; Killworth et al. 2003; Salganik et al. 2011); and approaches that propose statistical models that aim to make more efficient use of aggregate relational data (Habecker, Dombrowski, and Khan 2015; Maltiel et al. 2015; McCormick, Salganik, and Zheng 2010; Zheng, Salganik, and Gelman 2006).

In this study, our goal is to help complement these statistical advances by proposing changes in how aggregate relational data are collected. Our strategy is based on the fact that aggregate relational data

are collected from a probability sample (often, a household survey or random digit dialing phone survey). With a probability sample, researchers can directly estimate the size of the probe groups by asking respondents whether or not they are themselves members of the probe groups. As we demonstrate, researchers can then use these survey-based estimates for the size of the probe groups in place of known population sizes, thus removing the need to use only probe groups whose size can reliably be determined from administrative records. We argue that this relatively simple change to the way in which data are collected has the potential to greatly increase the flexibility of the method.

As a secondary goal, we also explore the hypothesized *quantity/quality tradeoff* in reports about personal network members (Feehan et al. 2016). To do so, in our second empirical example – a household survey – we asked respondents to report about aggregate relational connections through more than one network. The idea is that researchers can learn about the size and distribution of network connections through multiple weak-tie networks using a single survey module. This approach builds upon previous work that has used survey experiments to investigate how asking respondents about different types of network may affect the accuracy of respondents’ network reports, as well as the inferences made from using network reports to estimate the size of hidden populations (Feehan et al. 2016), and work that investigates segregation in patterns of network connectivity (DiPrete et al. 2011). In the survey we study here, each respondent was asked about her connections to probe groups through more than one network (instead of randomizing respondents to report about one of several possible networks). In the context of estimating the size of hidden populations, our approach allows us to produce multiple estimates for hidden population size, one estimate for each tie definition respondents are asked about. As we show, the results can reveal interesting patterns that may help improve the method in the future.

The remainder of the paper starts with a discussion of the changes we propose to data collection. We then provide two new methods for producing estimates from the data we propose collecting: one design-based estimator and one model-based estimator. Next, we turn to two empirical tests of our proposed methods. First, we use existing complete network data to simulate many aggregate relational data surveys. We use the results to evaluate and compare the sampling properties of the new estimators to the properties of existing estimators. Second, we used the proposed approach as part of a household survey we conducted in Hanoi, Vietnam. This second study illustrates how our methods can be used to produce network scale-up estimates for the size of hidden and hard-to-reach populations. We conclude with a discussion and an outline of topics for future work.

2 New approach

We propose a simple change to the way in which aggregate relational data are usually collected: in addition to collecting information about survey respondents’ personal network connections to the probe

groups, we propose that researchers also ask about respondents’ own membership in each probe group. This change in data collection means that, instead of requiring external information about the size of these probe groups, we can estimate the total size of the probe groups directly from our sample.

Asking respondents to report about their own membership in the probe groups has several advantages: (1) it can be applied in settings where there are few sources of administrative information, or where sources of administrative information do not match the geography of the study; (2) a much wider range of probe groups can potentially be used, including groups about which no administrative information exists at all; and (3) since probe group size and reported connections to the probe groups are estimated from the same survey respondents using the same question wording, this approach may be more robust in situations where there is ambiguity about what group membership means (for example, who counts as a ‘teacher?’). Thus, by relaxing several of the key requirements of the traditional known population estimator, our approach can yield a more flexible method for estimating the size and distribution of weak-tie personal network connections.

Of course, these advantages come at a cost: first, our proposed approach requires us to ask at least some of respondents about whether or not they are members of the probe groups, which increases respondent burden. We expect that, in many cases, this increase in respondent burden is a price worth paying in order to be able to choose probe groups that are most understandable to survey respondents and most likely to satisfy the conditions required by the estimators.

Second, since the size of the probe groups is now estimated, rather than assumed to be known, we expect that our proposed approach will add sampling variance to our estimate. Appendix B.1.1 contains a mathematical discussion of this issue that confirms this expectation and formally describes how the sampling variance will be increased when probe group sizes are estimates from survey respondents. However, empirical results from the two studies we discuss below suggest that this increase in sampling variance is not a major concern, at least for the sample sizes we investigate here.

Finally, changing how the data are collected will also require researchers to change how they analyze aggregate relational data: the size of the probe groups can no longer be treated as given, but now become an additional quantity to estimate. In the remainder of this section, we focus on this third issue by illustrating how standard statistical estimators can be adapted when probe group sizes are estimated from the survey, rather than being known from external sources.

There are two approaches to producing estimates from aggregate relational data: design-based (e.g., Feehan 2015; Feehan and Salganik 2016a; Habecker, Dombrowski, and Khan 2015) and model-based (e.g. P. D. Killworth et al. 1998; Maltiel et al. 2015; McCormick, Salganik, and Zheng 2010; Zheng, Salganik, and Gelman 2006). There are strengths and weaknesses to each of these approaches, and our goal here is not to advocate for one or the other; instead, we illustrate the usefulness of our proposed changes in data

collection by developing estimators that can be used under each perspective.

Design-based estimators

The design-based approach assumes that characteristics of the population are fixed, but observed only for people who are sampled and interviewed (see, e.g., Särndal, Swensson, and Wretman 2003). Uncertainty in an estimate comes from the random selection of survey respondents, and population-level inferences are based on analyzing this random selection mechanism. Design-based estimates typically use survey weights to account for complex sampling designs. Uncertainty intervals are intended to capture sampling variation – i.e., the variation that is expected to result from only observing a random sample of respondents, instead of a census of everyone.

Formally, in the design-based framework, we start with a probability sample s taken from a frame population F . For each sampled respondent $i \in s$, there is an associated weight w_i . Feehan and Salganik (2016a) showed that the original known population estimator of P. D. Killworth et al. (1998) can be written in the design-based framework as

$$\hat{d}^{\text{kp}} = \frac{\sum_{i \in s} \sum_{A_j \in \mathcal{A}} w_i y_{i,A_j}}{N_{\mathcal{A}}}, \quad (1)$$

where $\hat{d}^{\text{kp}}$ is an estimator for $\bar{d}_{F,F}$, the average number of network connections that frame population members have to other frame population members; $\mathcal{A}$ is the set of all of the probe groups; $A_j \in \mathcal{A}$ is a specific probe group (e.g. policemen, bus drivers, etc.); y_{i,A_j} is the number of connections that respondent i reports to probe group A_j (e.g. how many bus drivers the respondent reports knowing); and $N_{\mathcal{A}}$ is the total size of the set of probe groups.

In the classic known population estimator (Equation (1)), $N_{\mathcal{A}}$ is assumed to be known from some external source, often an administrative database or a census. In our new approach, $N_{\mathcal{A}}$ is not known; thus, a simple modification of Equation (1) is needed. Let z_{i,A_j} be an indicator variable for whether or not respondent $i \in s$ reports being a member of group $A_j \in \mathcal{A}$. Then we can estimate the average size of respondents' networks using

$$\hat{d}^{\text{new}} = \frac{\sum_{i \in s} \sum_{A_j \in \mathcal{A}} w_i y_{i,A_j}}{\sum_{i \in s} \sum_{A_j \in \mathcal{A}} w_i z_{i,A_j}} = \frac{\sum_{i \in s} \sum_{A_j \in \mathcal{A}} w_i y_{i,A_j}}{\hat{N}_{\mathcal{A}}}. \quad (2)$$

Equation (2) substitutes an estimator for the probe group sizes, $\hat{N}_{\mathcal{A}} = \sum_{i \in s} \sum_{A_j \in \mathcal{A}} w_i z_{i,A_j}$ for the known value $N_{\mathcal{A}}$ but it is otherwise the same as Equation (1). Appendix A makes explicit the conditions that must hold in order for the estimator to be design consistent and essentially unbiased, and Appendix B

provides two different derivations to motivate the design-based estimator in Equation (2). In particular, this analysis reveals that a *probe group condition* is needed by both the new estimator and the known population estimator. The probe group condition says that the average network size of probe group members should be the same as the average network size in the frame population.

Design-based scale-up estimator. Many studies collect aggregate relational data with the goal of using the network scale-up method to estimate the size of a hidden population; we now explain how our approach can be adapted with this goal in mind. Feehan and Salganik (2016a) showed that a design-based estimator for hidden population size is given by

$$\hat{N}_H = \frac{\hat{y}_{F,H}}{\hat{d}^{\text{kp}}} = \frac{\sum_{i \in s} w_i y_{i,H}}{\sum_{i \in s} \sum_{A_j \in \mathcal{A}} w_i y_{i,A_j}} \times N_{\mathcal{A}}, \quad (3)$$

where $y_{i,H}$ is then number of connections that survey respondent i reports to people in the hidden population H (e.g., “how many people do you know who inject drugs?”). To use the network scale-up estimator with data collected using our proposed approach, we replace the known population estimator $\hat{d}^{\text{kp}}$ in Equation (3) with the new estimator $\hat{d}^{\text{new}}$ from Equation (2):

$$\hat{N}_H^{\text{new}} = \frac{\hat{y}_{F,H}}{\hat{d}^{\text{new}}} = \frac{(\sum_{i \in s} w_i y_{i,H}) \times (\sum_{i \in s} \sum_{A_j \in \mathcal{A}} w_i z_{i,A_j})}{\sum_{i \in s} w_i y_{i,A_j}}. \quad (4)$$

Appendix A derives the precise conditions that need to hold in order for Equation (4) to be design consistent and essentially unbiased for the hidden population size.

Model-based estimators

Under the model-based approach, researchers develop a stochastic model that is intended to describe or approximate how the observed data were generated. Then, data are used to make inferences about values of model parameters and the data generating mechanism. Uncertainty intervals are based on the variation expected under the model.

The original model scale-up model, proposed by P. D. Killworth et al. (1998), assumed each respondent i ’s reported connections to probe group A_j , y_{i,A_j} , are given by

$$y_{i,A_j} | d_i, p_j \sim \text{Binomial}(d_i, p_j),$$

where $p_j = \frac{N_{A_j}}{N}$ is the prevalence of probe group A_j and d_i is respondent i ’s network size. By further assuming that connections to all subgroups are independent and by taking probe group prevalences p_j

to be known, P. D. Killworth et al. (1998) derived a maximum likelihood estimator for respondent i 's personal network size:

$$\hat{d}_i^{\text{ML}} = \frac{\sum_{A_j \in \mathcal{A}} y_{i,A_j}}{\sum_{A_j \in \mathcal{A}} N_{A_j}}.$$

Averaging these estimated $\hat{d}_i^{\text{ML}}$ values across n respondents results in an estimator¹ for the average network size, $\hat{\bar{d}}^{\text{ML}} = n^{-1} \sum_{i=1}^n \hat{d}_i^{\text{ML}}$.

Subsequent research has expanded the original model of P. D. Killworth et al. (1998), to make it more flexible with the goal of using aggregate relational data to estimate structural features of weak tie networks (McCormick, Salganik, and Zheng 2010; Zheng, Salganik, and Gelman 2006) and to produce scale-up estimates of hidden population size (Maltiel et al. 2015).

Our goal here is to illustrate how a model-based approach can be adapted to our strategy for collecting aggregate relational data. We first present a relatively simple base model for aggregate relational data when probe group sizes are known (Figure 1, left panel); we will call this the KP (known population) model. This KP model is closely related to previously developed models. Next, we show how to modify the model to add the components needed to account for the fact that probe group size is not known, but instead estimated from respondents' reported probe group memberships (Figure 1, right panel).

KP model for probe groups whose size is known.

The KP model can be described in three parts. First, we model the size of respondent i 's personal network, d_i , using a log-normal distribution; this approach is similar to previous aggregate relational data models (Maltiel et al. 2015; McCormick, Salganik, and Zheng 2010; e.g., Zheng, Salganik, and Gelman 2006):

$$\begin{aligned}\mu &\sim \text{Uniform}(1, 8) \\ \sigma &\sim \text{Uniform}(0.25, 2) \\ d_i | \mu, \sigma &\sim \text{Log-Normal}(\mu, \sigma),\end{aligned}$$

where μ is the mean degree (on the log scale) and σ is the standard deviation in degree (on the log scale)².

Next, the number of reported connections to each probe group is modeled as a function of the probe group size and the respondent's degree:

¹Note that, when sampling weights are ignored, $\hat{\bar{d}}^{\text{ML}}$ is the same as Equation (3); in fact, this maximum likelihood derivation originally motivated the average degree estimator in Equation (3).

²Our prior for σ is the same as Maltiel et al. (2015); however, for μ , we expanded the $\text{Uniform}(3, 8)$ prior of Maltiel et al. (2015) because in our second empirical example, some of the tie definitions we consider are much stronger than the ones that have been previously used in scale-up studies. *A priori*, we expect these stronger tie definitions to result in smaller network sizes.

Figure 1: Graphical description of (a) the KP (known population) model, with probe group sizes assumed to be known; and (b) the new model, with probe group sizes estimated from respondent reports. The plates with green borders show the part of model that differs when using our new data collection strategy. The plates with purple borders contain the part of the model used to estimate the size of K hidden populations; it is not included when network size is the only quantity of interest.

$$y_{i,A_g} | d_i, p_g \sim \text{Poisson}(d_i p_g),$$

where y_{i,A_g} is the number of connections that respondent i reports to probe group A_g ; p_g is the prevalence of probe group A_g ; g ranges from 1 to G ; and G is the total number of probe groups.

Finally, when the goal is to estimate the size of one or more hidden populations, in addition to personal network size, the model has a third component. Reported connections to hidden population h can be modeled as

$$p_h \sim \text{Uniform}(0, 1)$$

$$Y_{i,h} | d_i, p_h \sim \text{Poisson}(d_i p_h),$$

where p_h is the prevalence of the hidden population. Thus, reported connections to the hidden population are modeled in the same way as reported connections to the probe groups, except that information about respondent's membership in the hidden population is not used.

Incorporating unknown probe group sizes

In order to adapt this model-based approach to situations where the size of the probe groups is not known, we propose modeling the total number of survey respondents who report being members of probe

group g , $Z_g = \sum_{i \in s} z_{i,g}$ as a function of the size of the sample $n = |s|$, and the population prevalence of probe group, p_g :

$$p_g \sim \text{Uniform}(0, 1)$$

$$Z_g \mid p_g, n \sim \text{Binomial}(n, p_g).$$

We adopt an uninformative $\text{Uniform}(0, 1)$ prior for p_g .

This model appears to work well in the empirical examples we study below; however, we emphasize that our goal here is keep the model relatively simple by adding just enough complexity to illustrate how estimating the size of the probe groups could be incorporated into a model-based framework. Future work could make this model more elaborate by accounting for possible overdispersion in reported connections to each group and by using a model-based adjustment for errors in recalling reported connections to each group (Maltiel et al. 2015; McCormick, Salganik, and Zheng 2010; e.g., Zheng, Salganik, and Gelman 2006).

3 Simulation study

We created a simulation study as the first empirical test of the estimators proposed above. The goal of the simulation study is to investigate the sampling properties of our new estimators and the previous estimators in controlled environments where the true population values are known. We study the properties of both the design and model-based degree estimators.

In order to conduct our simulation study, we start with the fb100 dataset, which is a collection of complete Facebook friendship relationships recorded from the first 100 schools to participate in the social network Facebook³ (Traud, Mucha, and Porter 2012). The fb100 dataset is useful because (1) it contains 100 different networks, allowing us to explore a range of real-world patterns of connectivity without resorting to a hypothetical model of network structure; and (2) in addition to network structure, it contains information about several characteristics for each person in the 100 school populations. We use these characteristics to explore the effect of choosing different types of probe groups.

Specifically, for each school, we create candidate probe groups out of the following characteristics: **status** (distinguishes students, staff, and faculty); **gender**; **year**; **dorm**; and **major**. We treat these characteristics as categorical variables, and convert each level of each categorical variable into an indicator. For example, there is an indicator variable for whether or not each person in the school network is a student; an indicator variable for whether or not each person in the school network is a faculty member; and so forth. We then

³More information about the Facebook100 dataset can be found in Traud, Mucha, and Porter (2012), and at <https://archive.org/details/oxford-2005-facebook-matrix>.

make a list of all of the indicator variables whose prevalence is between 0.1% and 10% of the population of the school’s network⁴. We call the groups that satisfied this size requirement the *candidate probe groups* for each school. In our simulation study, we investigate many combinations of candidate groups to ensure our results are not overly dependent on one randomly selected candidate group.

In each of the 100 networks in the fb100 dataset, we choose $G = 20$ different sets of candidate groups, with each set of candidate groups having a random subset of 10 of the possible probe groups for the school. We simulate $M = 250$ surveys of using each of the possible sets of candidate groups, meaning that we simulated a total of $100 \times 20 \times 250 = 500,000$ surveys. Each simulated survey has a sample size of $N = 500$.

For each school, the simulation proceeds as follows:

- For g in $1, \dots, G = 20$:
 - Draw a set of probe groups $\mathcal{A}_g$ by randomly selecting 10 groups from the set of candidate probe groups for the school
 - For m in $1, \dots, M = 250$:
 - * Simulate a survey sample by choosing a simple random sample of size $N = 500$ people from the school network
 - Simulate questions about whether or not each respondent is a member of each of the 10 groups in $\mathcal{A}_g$
 - Simulate aggregate relational data questions (ie, questions about network connections to members of each of the 10 groups in $\mathcal{A}_g$)

Recall that, in order to produce consistent and essentially unbiased estimates of average personal network size, the set of probe groups must satisfy the *probe group condition* (see Appendix A); this condition says that the average network size of probe group members should be the same as the average network size in the population. Note that, according to the design of this simulation, there is no guarantee that the probe groups will satisfy the probe group condition; instead, the simulation mimics the reality most researchers face, in which they must try to pick probe groups without knowing for sure whether or not the probe group condition is satisfied.

Simulation results

We first investigate the design-based estimator. For each simulated survey, we calculated two estimates for the average network size: one based on the known population method (which uses information about the

⁴Bernard et al. (2010) recommends that probe groups be between 0.1% and 4% of the population; however, most fb100 networks have a few thousand nodes, which is considerably smaller than most real-world populations where aggregate relational data would be collected. Therefore, to avoid having only groups that are extremely small in absolute number, we consider candidate probe groups that are up to 10% of the network.

sizes of the probe groups; Equation (1)) and one based on our new estimator (which does not; Equation (2)). We summarize the accuracy of each estimate by calculating its relative error:

$$\text{rel-err}(\hat{d}) = \frac{\hat{d} - \bar{d}}{\bar{d}}, \quad (5)$$

where $\hat{d}$ is an estimate for the average network size, and $\bar{d}$ is the true average network size. The relative error will be close to zero when the estimate is accurate, and it will be far from zero when it is inaccurate.

The left-hand panel of Figure 2 and Table 1 compare the distribution of relative errors across all 500,000 simulated surveys for the known population method and for the new estimator. Two findings emerge from the figure: first, the distributions of relative errors from both methods appear to be centered around 0; and, second, the distribution of relative errors from the known population method is more tightly concentrated around 0 than the distribution of relative errors from the new method. Table 1 confirms these observations using quantitative summaries of the two distributions of relative errors shown in the left-hand panel of Figure 2. The table reveals that the average of the relative errors is very close to zero in both cases, suggesting that both estimators have little to no bias. However, the standard deviation of the relative errors – which captures sampling variation – is slightly larger for the new estimator. This is to be expected: since the new method is based on estimating an additional quantity – i.e., the size of the probe groups – it has additional sampling variance (Appendix B.1.1).

The right-hand panel of Figure 2 shows the relationship between the relative errors for each method; each point in the scatterplot shows the result from one of the 500,000 simulated surveys. The figure shows that the relative errors are highly correlated (the correlation coefficient $\rho = 0.78$), meaning that, for a given set of probe groups and a given sample, the known population estimator and the new estimator produce errors that tend to be in the same direction and that tend to have the same magnitude.

Next, we investigated model-based estimates for the average network size from the same simulations. The computational burden required to fit the model for every survey we simulated was prohibitive; therefore, we fit the model to a subset of 10 of the simulated surveys from each of the 100 universities in our sample. The results shown here summarize these 1,000 model fits. We evaluate the model-based point estimates by using the median of the posterior distribution for the estimated average of the degree distribution; in other words, we take as our estimator for $\bar{d}$ the posterior median of the parameter μ in the models above.

Parallel to the results for the design-based estimators, we investigate the relative errors of the model-based fits for (1) the KP model, which requires information about probe group sizes (Figure 1a); and (2) the new model, which estimates the size of the probe groups (Figure 1b). The left-hand panel of Figure 3 shows the distribution of relative errors for the KP model and for the new model. The distributions look qualitatively similar, with both showing evidence of a slight positive bias. This finding is confirmed by

Figure 2: **(a)** The distribution of relative errors in estimated average network size for the known population method (left) and for the new estimator (right). Table 1 confirms that the sampling variation is higher for the new method. **(b)** Comparison of the relative error in estimated average network size using the known population estimator (x axis) and the new estimator (y axis).

Table 1: Mean and standard deviation in the relative error across all simulations for estimating the average network size using the known population estimator and the new estimator (Equation (2)). These results suggest that both estimators have very small bias, but that the sampling variance for the new estimator is slightly higher than for the known population estimator.

Estimator	Mean(rel-err)	SD(rel-err)
Known population (design-based)	-0.01	0.10
New (design-based)	0.00	0.12

Table 2: Mean and standard deviation in the relative error across model-based estimates for the average network size using the known population estimator and the new estimator (Equation (2)).

Estimator	Mean(rel-err)	SD(rel-err)
Known population (model-based)	0.12	0.16
New (model-based)	0.06	0.16

Table 2, which reveals that the KP estimator has a moderate positive bias, though with a lot of sample to sample variation. Again, these results are not surprising: model-based estimates would be expected to be unbiased on data generated by the model, but not necessarily on data from a real-world network. The table also suggests that estimating the probe group size from the respondents may slightly reduce the bias, though the sample to sample variation is again considerable. The right-hand panel of Figure 3 compares the relative error for the KP model to the relative error for the new model. Similar to the design-based estimators, there is a high correlation between these two sets of relative errors ($\rho = 0.86$), and the slight reduction in bias from the new model’s estimates is revealed by the fact that points are shifted downward from the diagonal $y = x$ line.

Figure 3: **(a)** The distribution of relative errors in estimated average network size for the model with probe group sizes assumed known (left) and with probe group sizes simultaneously estimated (right). **(b)** Comparison of the relative error in estimated average network size using the model with probe group sizes assumed known (x axis) and the model with probe group sizes simultaneously estimated (y axis). The posterior median estimate of the average degree is used as the point estimate when calculating relative error in fits from each model.

Finally, we assess how similar relative errors from the new design-based estimator are to relative errors from the new model-based estimator. Figure 4 shows the relationship between the relative error in the design-based estimate and the relative error in the model-based posterior median estimate for μ across

the 1,000 surveys in our subsample; on average, relative errors from the model-based estimator are about 0.06 higher than relative errors from the design-based estimator and the correlation coefficient is $\rho = 0.89$. These findings suggests that, in a given survey, the accuracy of the estimate from the new design-based estimator is similar to the accuracy of the estimate from the new model-based estimator.

Figure 4: Comparison between relative error in the design-based estimate (x axis) and the model based estimate (y axis) for the subsample of 1,000 simulated surveys we fit the model to. Model-based estimates are the posterior median of the parameter μ . The strong positive relationship suggests that the two approaches produce qualitatively similar results when applied to these simulated surveys.

To wrap up, the simulation study shows that our proposed change in data collection requires relatively simple modifications to both design-based and model-based approaches to estimating average personal network size. In the case of the design-based estimators, our simulation results suggest that estimating probe group size from the respondents adds a modest amount of sampling variation to the estimates. For model-based estimators, our simulation results suggest that there is not a cost, and that there may even be modest benefits to estimating probe group sizes from respondents.

4 Household survey in Hanoi, Vietnam

The simulation study discussed in Section 3 helps to illustrate the behavior of our proposed data collection strategy and our new estimators in networks where the true average network size is known. Of course, this simulation cannot provide direct evidence about how well our proposed methods work in a real world survey. Thus, we now turn to our second empirical illustration: a household survey we conducted in Hanoi, Vietnam. The goal of the survey was to use the network scale-up method to estimate the size

of four hidden populations at elevated risk to HIV/AIDS: female sex workers, men who have sex with men, people who use drugs, and people who inject drugs. For our purposes, this study serves as a field test of estimating probe group size from survey respondents’ reported membership in the probe groups. This survey also allows us to illustrate how design and model-based scale-up estimators—which aim to use aggregate relational data to estimate the size of a hidden population, rather than just the average network size – can be adapted when probe group sizes are estimated from the sample.

A secondary goal of the survey was to investigate the *quantity/quality tradeoff* which is hypothesized to be at play when asking respondents to report about people they are connected to through network ties of varying strength (Feehan et al. 2016). When moving from estimates based on a weak tie definition to estimates based on a stronger tie definition, the quantity/quality hypothesis predicts that two important things will happen: first, since weak tie networks tend to be bigger than strong tie networks, we expect to learn about fewer people through strong tie networks; thus, sampling error should increase when estimates are based on strong tie networks, compared to weak tie networks. Second, since we expect people to have more accurate information about others who they are connected to through strong ties, we expect that nonsampling errors will decrease in strong tie networks, compared to weak tie networks.

Previous empirical work has found support for the quantity/quality tradeoff hypothesis when survey respondents report about groups that are not expected to be stigmatized (Feehan et al. 2016; Feehan and Cobb 2019). In our survey, we wished to investigate what effect tie strength might have on reports about groups that are potentially stigmatized. Thus, our survey asked each respondent questions about her connections to probe groups under four different tie definitions (Table 4). These tie definitions were designed to be nested inside of one another, and to lie on a continuum from very weak ties (any people the respondent knows at all) to relatively strong ties (people the respondent shared a meal with in the past month).

Probe groups

Our new approach to data collection gave us great flexibility in considering different types of probe groups for our study. After pilot testing several different possibilities, we decided to use occupations as our probe groups (Table 3). Importantly, we were able to focus on how effectively respondents would be able to understand and reply to questions about the probe groups, rather than being constrained to groups whose totals existed in administrative records. For example, motorcycle taxis called *xe om* are common in Hanoi. It is clear to people in Hanoi who *xe om* drivers are, so they could answer a survey question about how many *xe om* drivers they were connected to. But we were not able to locate any administrative data about the total number of *xe om* drivers in the city. Even if some administrative data existed, official tallies might not include *xe om* drivers in the informal sector. So it would not have been possible to use *xe om* drivers as a probe group in conjunction with the known population method.

Table 3: Probe groups used in the Hanoi survey. Respondents were asked about their own membership in each probe group, and also their network members in each probe group.

Probe groups
Doctor
Lawyer
Primary teacher
Bank worker
Taxi driver
Xe om driver
Hairdresser
Nurse
Coffee shop owner
Farmer (rice grower only)
Builder/mason
Security guard
Student

But our approach enabled us to use xe om drivers as a probe group despite the lack of administrative information: by asking survey respondents if they are themselves xe om drivers, we can estimate the total number of xe om drivers in Hanoi from our probability sample.

Estimating probe group sizes from the survey allowed us to circumvent similar challenges for several other occupations, for which we also found no Hanoi-specific administrative data. Previous experience designing network scale-up studies suggests that this situation is not unusual: it can be difficult to get accurate administrative data about the number of people in different occupations, particularly for a specific geographic area and for occupations that are often largely informal. Even when administrative data can be located, it can be difficult to ascertain how complete and accurate it is.

Results

We partnered with Vietnam’s national statistical agency, GSO, to obtain a probability sample of households in Hanoi as part of the Vietnam Health and Living Standards (VHLS) survey. The VHLS was collected using a complex, multi-stage design that produced a representative sample of adults living in Hanoi. We obtained our respondents from a randomly selected subset of the households interviewed as part of VHLS in Hanoi. To account for the complex sampling design, all of the design-based analyses use the sampling weights and sampling uncertainty is estimated using the rescaled bootstrap (Feehan and Salganik 2016b; Rao and Wu 1988; Rao, Wu, and Yue 1992). Our model-based estimates do not use the sampling weights, which is consistent with previous studies that have adopted the model-based approach in this context (e.g., P. D. Killworth et al. 1998; Maltiel et al. 2015; McCormick, Salganik, and Zheng 2010; Zheng, Salganik, and Gelman 2006). Appendix D provides additional information about the

Table 4: Tie definitions used in the Hanoi survey. All respondents were asked about all four tie definitions; the text in the Table uses the example of taxi drivers. Respondents were first asked whether they knew any group members; if respondents said yes, then they were asked more detailed questions about the number of taxi drivers known using each tie definition.

Tie name	Wording
Any	How many taxi drivers do you KNOW who are age 18-54; who currently live in Hanoi; who you know by sight AND name; and who also know you by sight and name? (Note: You should not consider famous people that you know about, but who do not know about you.)
Interaction	How many taxi drivers have you interacted with – through face to face conversation, chat, SMS, or email – in the past 12 months?
Meal 12mo	Among [Interaction alters], how many did you share a meal or drink (coffee, tea, beer, etc) with in the last 12 months?
Meal 1mo	Among [Meal 12mo alters], how many did you share a meal or drink with in the last month?

sampling design.

Our survey instrument had three parts: first, each respondent was asked some basic demographic questions, including whether or not she was personally a member of any of the occupational groups listed in Table 3⁵. Responses to these survey questions enable us to estimate the size of the probe groups. Next, each survey respondent was asked a number of questions about her connections to people in the probe groups through each of the four network ties. Responses to these questions allowed us to estimate the average network size of each respondent in each of the four networks using the design-based estimator (Equation (2)) and the model (Figure 1). Finally, each survey respondent was asked additional aggregate relational data questions about hidden populations whose size we wished to estimate. Again, we use the responses to these questions together with adapted design-based scale-up estimator (Equation (4)) and the adapted model-based estimator (Figure 1) to produce two sets of estimated hidden population sizes.

About 26% of respondents reported being a member of one of the probe groups⁶. Figure 5 shows the resulting size estimates. Several of the probe groups are estimated to be quite small, around ten to twenty thousand people. But these small groups do not necessarily pose a problem for the estimators: the denominator of the design-based estimator (Equation (2)) reveals that estimating the size of respondents’ personal networks requires the total probe group size, and not the size of each individual probe group; Figure 5 reveals that this quantity is estimated to be much larger, approximately a fifth of the adult population in Hanoi.

After reporting about their own membership in the probe groups, respondents were asked how many

⁵Note that if it is possible for respondents to be in more than one of the probe groups, they should be asked about membership in each group. In our case, no survey respondents reported being members of more than one of these occupations.

⁶No respondents reported being doctors, so we omitted doctors from the probe groups in the analysis below.

Figure 5: Estimated size of each probe group population, and estimated total size of all probe group populations (in blue). Note that the y axis is on a log scale. Group sizes vary considerably, and many of the probe groups are estimated to be quite small, with 95% confidence intervals including zero. Nonetheless, the total size of the probe groups was estimated to be reasonably large (about 1 million people).

connections they had to each probe group using each of the four tie definitions shown in Table 4. Figure 6 summarizes the responses by showing the average number of reported connections for each probe group and for each tie definition. (Note that the y axis of the figure is on a log scale.) Two important features emerge from Figure 6: first, for all groups, the average number of reported contacts decreases as tie definitions move from weaker to stronger. This is consistent with what we would expect from the design of the tie definitions, and from the way the questions were asked in the survey: respondents first reported ‘any contact’ and were then asked about each of the stronger tie definitions in turn. Second, the order of the probe groups is quite stable across tie definitions: for each pair of tie definitions, we calculated the rank correlation coefficient between the average number of reported connections to each probe group; the smallest of these correlation coefficients was 0.89. Thus, groups that have relatively large numbers of reported connections for one tie also tend to have relatively large numbers of reported contacts for another.

Figure 6: Average number of reported connections to each probe group for each of the four network ties used in the Hanoi survey. Note that the y axis is on a log scale. Tie definitions are ordered from weakest (any contact) to strongest (shared a meal in the last month). For each probe group, the average number of reported connections decreases as the tie definitions increase in strength. The relative ordering of the average number of reported connections to each probe group was highly consistent across tie definitions.

Figure 7 shows the estimated average network size among adult residents of Hanoi for each tie definition using the design-based and model-based estimators. These estimated network sizes are based on combining the information shown in Figures 5 and 6 using the design-based estimator (red) and the model-based

estimator (blue). The design-based estimates have 95% confidence intervals, which capture sampling uncertainty and were calculated using the rescaled bootstrap (Feehan and Salganik 2016b; Rao and Wu 1988; Rao, Wu, and Yue 1992). Model-based estimates are 95% credible intervals based on MCMC draws from the model posterior. Three salient features emerge from Figure 7: first, as expected, estimated average network size decreases as the tie definition becomes stronger. Second, the estimated average personal network sizes from the design-based estimator are very similar to the estimated average personal network sizes from the model-based estimator. This was somewhat surprising: it was not obvious, *ex ante*, that the two estimates would agree. The design-based estimate uses the sampling weights, together with the relatively simple estimator in Equation (2); the model-based estimator, on the other hand, requires more complexity to estimate, and it disregards the sampling weights altogether. Nonetheless, Figure 7 reveals that, in this case, the estimates are highly similar. Third, the estimated uncertainty in each set of estimates is also quite similar between the two methods. Again, it was not obvious that this would be the case: the design-based intervals show an estimate of the sampling variance, incorporating the sampling weights and the sampling design, while the model-based intervals show uncertainty estimated using the stochastic assumptions encoded in the model.

At the end of the survey, respondents were asked about their connections to four hidden populations that are at elevated risk of HIV/AIDS: female sex workers; men who have sex with men; people who inject drugs; and people who use drugs. These questions were framed in the same way that the questions about connections to the probe groups were; however, rather than reporting their responses to the interviewer, respondents were asked to self-administer the responses to these more sensitive questions using a separate form. (These groups are socially stigmatized in many populations, and the goal was to try to mitigate the effects of social desirability bias; see Appendix E.2 for more detail). We also wanted to compare patterns of reported connections to the potentially stigmatized hidden populations to the patterns of connections to groups we did not expect to be stigmatized. Thus, as part of the self-administered part of the questionnaire, respondents were also asked to report their connections to two additional groups: women who smoke and women who had a baby in the past 12 months.

Figure 8 shows the estimated sizes of each group using the model-based and design-based estimators. Again, design-based estimates are shown with 95% confidence intervals, which capture sampling uncertainty and were calculated using the rescaled bootstrap (Feehan and Salganik 2016b; Rao and Wu 1988; Rao, Wu, and Yue 1992). Model-based estimates have 95% credible intervals based on MCMC draws from the model posterior.

We use differences in size estimates across the four different network ties shown in Figure 8 to assess the predictions of the quantity/quality tradeoff hypothesis. The results in Figure 8 are generally consistent with the first prediction of the quantity/quality tradeoff hypothesis: in these estimates, the relative

Figure 7: Estimated average personal network size of adult residents of Hanoi, Vietnam. Estimates based on our new estimator were produced for four different tie definitions. Uncertainty estimates for the design-based estimator use the rescaled bootstrap, and account for the complex sample design. Model-based uncertainty estimates come from the 95% credible interval around the posterior median value of μ .

sampling error increases as the tie definition get stronger (Appendix Table 7). However, Figure 8 shows that support for the second prediction – that nonsampling error decreases as tie definition gets stronger – is less clear. For one of these groups – men who have sex with men (MSM) – we see modest increases in estimated size as the tie definition gets stronger. As discussed in Appendix E.2, this pattern is consistent with what we would expect to see if people are more likely to be aware that their strong-tie network connections are MSM, which is in turn consistent with the quantity/quality tradeoff hypothesis’s prediction that reports from stronger ties will have lower nonsampling error. But for the other three groups – female sex workers, people who inject drugs, and people who use drugs – the highest estimates come from the weakest tie definition. Taking people who use drugs as an example, we can think of three plausible explanations for these patterns: first, when asked to report about connections to these groups, people may be more likely to speculatively include weak tie connections as drug users; second, people may be more aware that weak tie connections are drug users because people who use drugs make an effort to hide their status from stronger-tie connections; third, stigma associated with drug use may lead respondents to be unwilling to disclose that close network connections are drug users, but more willing to disclose that weaker tie network connections are drug users. In Appendix E.2, we discuss these possibilities in greater detail. Of course, it is also possible that these factors interact with each other to produce the patterns observed here.

One way to better understand the mechanism behind the cross-tie patterns in size estimates from Figure 8 would be to compare the estimates to high quality reference data on the size of these hidden populations. Unfortunately, for most of the groups shown in Figure 8, no high quality reference estimates were available. However, we were able to find approximate comparison estimates for the groups we did not expect to be stigmatized: women who smoke and women with a baby under age 1. We approximate the number of women who had a baby under 1 by simply taking GSO’s population projection for people aged 0-4 in Hanoi and dividing it by 5; this is roughly the number of babies under age 1, which is in turn roughly the number of women who had a baby in the past 12 months. For women who smoke, we were able to obtain point estimates from an epidemiological study that estimated the prevalence of smoking among women in Hanoi. (See Appendix E.1 for more information on how we calculated these comparison estimates.) The horizontal dotted lines in Figure 8 show these alternate estimates in comparison to the scale-up estimates from each tie definition. The scale-up estimates agree well with the approximate comparison for women who smoke, and are somewhat lower than the rough number of women who had a baby in the past year.

5 Discussion

We propose a simple change in how aggregate relational data can be collected on a survey: rather than requiring that the size of probe groups be known from administrative records or census data, we show

Figure 8: Network scale-up estimates for the size of four hidden populations (Female Sex Workers; Men who have sex with men; People who inject drugs; and People who use drugs), and two other populations (Women who smoke and women who had a baby in the past 12 months). The scale-up estimator has been modified to use our new estimator for average personal network size. Each tie definition can be used to produce a separate estimate for the size of each hidden population. Dotted lines show estimates from alternative sources for two groups: women who had a baby in the past year, and women who smoke.

that researchers can estimate the size of probe groups directly from survey respondents. By eliminating the need to find data about the probe groups, this change can greatly increase the flexibility of these methods. We show how to adapt existing design- and model-based estimators for network size to use data collected in this way.

To test our new approach in a controlled setting where true population values were known, we conducted a simulation study. The results of the simulation study suggest that estimating the size of probe groups from information about respondents performs well: it appears to add little or no bias. The new approach increases sampling uncertainty a small amount for the design-based estimator, and it may reduce errors a little bit for the model-based estimator.

To test our new approach in the field, we conducted a household survey in Hanoi, Vietnam. We showed how our approach enabled occupations to be used as probe groups, despite the lack of administrative data on the size of each occupational group. We also experimented with asking survey respondents to simultaneously report about more than one different type of network tie. Asking about different ties enabled us to simultaneously estimate the size of several different networks; it also allowed us to use the scale-up method to produce multiple estimates for the size of four hidden populations.

Our results show that asking about ties of different strength can produce informative patterns, and we interpret these cross-tie patterns in size estimates in light of the previously hypothesized quantity/quality tradeoff. The results suggest mixed support for the predictions of the quantity/quality tradeoff hypothesis: in all cases, sampling error increased as respondents reported about stronger ties. However, we also expected that respondents would omit fewer hidden population members from reports about stronger network ties; if this were the case, we would have expected to see that stronger ties produced larger size estimates for each hidden population. We only find this pattern for one of the groups, MSM. This suggests that further work is needed to fully understand the tradeoffs that are at play when respondents are asked to report about hidden populations that may be stigmatized. In Appendix E.2, we develop formal models that can be useful in helping to reason about these potential tradeoffs.

The results of our study allow us to update some of the recommendations for future studies that collect aggregate relational data. First, our findings demonstrate that researchers need not be restricted to probe groups whose size can be found in administrative or census data. We recommend that, rather than focusing on probe groups for whom administrative data are available, researchers instead focus on finding (i) probe groups that are most likely to satisfy the probe group condition – i.e., probe groups whose members are typical of frame population members, in terms of their average network size; and (ii) probe groups that are easy for survey respondents to understand and report about.

Best practices currently suggest that 20 probe groups are needed to effectively estimate personal network size using the known population method (Bernard et al. 2010); in our study, we found that a smaller

number of probe groups could also work well: we used 10 probe groups in the simulation study and 13 probe groups in the Hanoi survey. Future work could explore which factors are most important in choosing an effective number of probe groups; we expect that the optimal number will depend on the tie definition, as well as the size and nature of the probe groups.

Our study had several important limitations, and these limitations help outline potential directions for future research. In our empirical studies, we only used probe groups whose size was estimated from the survey respondents. Future studies could investigate the potential benefits of combining questions about probe groups whose size is known with questions about probe groups whose size is estimated from respondents. This approach could also permit internal consistency checks to be used with our new approach to data collection (see Appendix B.2 for a discussion of this topic).

In the Hanoi survey, we estimated the size of four hidden populations at elevated risk of HIV using a scale-up survey alone. Previous research has shown that the conditions required by the scale-up estimator can be relaxed if researchers also obtain parallel samples from among members of each hidden population (e.g., Feehan and Salganik 2016a; Salganik et al. 2011). These data would be especially valuable here because they could potentially help explain the mechanisms behind the cross-tie pattern of estimates shown in Figure 8. Future studies could combine information from a probability sample of the population with additional samples of members of the hidden population for multiple ties to help provide empirical evidence about possible mechanisms driving differences in tie-specific size estimates.

It would also be valuable to develop study designs to disentangle mechanisms that could lead to the cross-tie patterns we found in our Hanoi study results. Our results suggest that different tie definitions can produce different estimates for the size of a hidden population. The context-specific relationship between the tie definition and the hidden population being studied may play an important role in determining which network tie will produce the best size estimate. Future studies could explore novel approaches to empirically quantifying the effect that stigma might have as a function of tie strength; for example, future work might explore incorporating list experiments (e.g. Blair and Imai 2012; Glynn 2013) or randomized response designs (e.g. Warner 1965) to scale-up studies.

In this study, we used a relatively simple model to illustrate how to adapt model-based estimators to our methods. Future work could expand the model here to incorporate innovations that have been previously developed in the scale-up literature, such as regression-based adjustments for recall errors (e.g. McCormick, Salganik, and Zheng 2010) and estimated barrier effects (Maltiel et al. 2015).

Finally, the point estimates from our simulation study and from our field test in Hanoi both revealed that estimates from the model- and design-based approaches were strikingly similar. This was somewhat surprising: the Hanoi study was based on a complex sampling design and the model-based estimator ignored the survey weights. It would be useful to better understand when the model-based estimators

can safely ignore the sampling design and, more generally, how to incorporate survey weights into complex models.

References

- Bernard, H. R., T. Hallett, A. Iovita, E. C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M. J. Salganik, T. Saliuk, O. Scutelnicu, G. A. Shelley, P. Sirinirund, S. Weir, and D. F. Stroup. 2010. “Counting Hard-to-Count Populations: The Network Scale-up Method for Public Health.” *Sexually Transmitted Infections* 86(Suppl 2):ii11–15. doi: 10.1136/sti.2010.044446.
- Bernard, H. Russell, Eugene C. Johnsen, Peter D. Killworth, and Scott Robinson. 1989. “Estimating the Size of an Average Personal Network and of an Event Subpopulation.” Pp. 159–75 in *The Small World*, edited by M. Kochen. Norwood, NJ: Ablex Publishing.
- Blair, Graeme, and Kosuke Imai. 2012. “Statistical Analysis of List Experiments.” *Political Analysis* 20(1):47–77.
- Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. “Stan : A Probabilistic Programming Language.” *Journal of Statistical Software* 76(1). doi: 10.18637/jss.v076.i01.
- de Sola Pool, Ithiel, and Manfred Kochen. 1978. “Contacts and Influence.” *Social Networks* 1(1):5–51. doi: 10.1016/0378-8733(78)90011-4.
- Deville, Jean-Claude, and Carl-Erik Särndal. 1992. “Calibration Estimators in Survey Sampling.” *Journal of the American Statistical Association* 87(418):376–82. doi: 10.1080/01621459.1992.10475217.
- DiPrete, Thomas A., Andrew Gelman, Tyler McCormick, Julien Teitler, and Tian Zheng. 2011. “Segregation in Social Networks Based on Acquaintanceship and Trust.” *American Journal of Sociology* 116(4):1234–83.
- Feehan, Dennis M. 2015. “Network Reporting Methods.” PhD thesis, Princeton University.
- Feehan, Dennis M., and Curtiss Cobb. 2019. “Using an Online Sample to Estimate the Size of an Offline Population.” *Demography* 56(6):2377–92. doi: 10.1007/s13524-019-00840-z.
- Feehan, Dennis M., and Matthew J. Salganik. 2016a. “Generalizing the Network Scale-Up Method: A New Estimator for the Size of Hidden Populations.” *Sociological Methodology* 46(1):153–86.
- Feehan, Dennis M., and Matthew J. Salganik. 2016b. “Surveybootstrap: Tools for the Bootstrap with Survey Data.”
- Feehan, Dennis M., Aline Umubyeyi, Mary Mahy, Wolfgang Hladik, and Matthew J. Salganik. 2016. “Quantity Versus Quality: A Survey Experiment to Improve the Network Scale-up Method.” *American Journal of Epidemiology* kwv287. doi: 10.1093/aje/kwv287.
- Fischer, Claude S. 1982. *To Dwell Among Friends*. University of Chicago Press.

- Glynn, A. N. 2013. "What Can We Learn with Statistical Truth Serum?: Design and Analysis of the List Experiment." *Public Opinion Quarterly* 77(S1):159–72. doi: 10.1093/poq/nfs070.
- Granovetter, Mark S. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 1360–80.
- Habecker, Patrick, Kirk Dombrowski, and Bilal Khan. 2015. "Improving the Network Scale-up Estimator: Incorporating Means of Sums, Recursive Back Estimation, and Sampling Weights." *PloS One* 10(12):e0143406.
- Killworth, P. D., E. C. Johnsen, H. R. Bernard, G. Ann Shelley, and C. McCarty. 1990. "Estimating the Size of Personal Networks." *Social Networks* 12(4):289–312.
- Killworth, P. D., E. C. Johnsen, C. McCarty, G. A. Shelley, and H. R. Bernard. 1998. "A Social Network Approach to Estimating Seroprevalence in the United States." *Social Networks* 20(1):23–50.
- Killworth, P. D., C. McCarty, H. R. Bernard, G. A. Shelley, and E. C. Johnsen. 1998. "Estimation of Seroprevalence, Rape, and Homelessness in the United States Using a Social Network Approach." *Evaluation Review* 22(2):289–308.
- Killworth, Peter D., Christopher McCarty, H. Russell Bernard, Eugene C. Johnsen, John Domini, and Gene A. Shelly. 2003. "Two Interpretations of Reports of Knowledge of Subpopulation Sizes." *Social Networks* 25(2):141–60. doi: 10.1016/S0378-8733(02)00040-0.
- Lubbers, Miranda J., José Luis Molina, and Hugo Valenzuela-García. 2019. "When Networks Speak Volumes: Variation in the Size of Broader Acquaintanceship Networks." *Social Networks* 56:55–69. doi: 10.1016/j.socnet.2018.08.004.
- Lumley, T. 2010. *Complex Surveys: A Guide to Analysis Using R*. Vol. 565. Wiley.
- Lumley, Thomas. 2019. *Survey: Analysis of Complex Survey Samples*.
- Maltiel, Rachael, Adrian E. Raftery, Tyler H. McCormick, and Aaron J. Baraff. 2015. "Estimating Population Size Using the Network Scale up Method." *Annals of Applied Statistics* 9(3):1247–77.
- Marsden, Peter V. 2005. "Recent Developments in Network Measurement." Pp. 8–30 in *Models and Methods in Social Network Analysis*, edited by P. J. Carrington, J. Scott, and S. Wasserman. Cambridge University Press.
- McCarty, C., P. D. Killworth, H. R. Bernard, E. C. Johnsen, and G. A. Shelley. 2001. "Comparing Two Methods for Estimating Network Size." *Human Organization* 60(1):28–39.
- McCarty, Chris, H. Russell Bernard, Peter D. Killworth, Gene A. Shelley, and Eugene C. Johnsen. 1997. "Eliciting Representative Samples of Personal Networks." *Social Networks* 19(4):303–23. doi: 10.1016/S0378-8733(96)00302-4.

- McCormick, Tyler H., Matthew J. Salganik, and Tian Zheng. 2010. “How Many People Do You Know?: Efficiently Estimating Personal Network Size.” *Journal of the American Statistical Association* 105(489):59–70. doi: 10.1198/jasa.2009.ap08518.
- Perry, Brea L., Bernice A. Pescosolido, and Stephen P. Borgatti. 2018. *Egocentric Network Analysis: Foundations, Methods, and Models*. Vol. 44. Cambridge university press.
- Rao, J. N. K., and Norma P. Pereira. 1968. “On Double Ratio Estimators.” *Sankhyā: The Indian Journal of Statistics, Series A* 83–90.
- Rao, J. N. K., and C. F. J. Wu. 1988. “Resampling Inference with Complex Survey Data.” *Journal of the American Statistical Association* 83(401):231–41. doi: 10.2307/2288945.
- Rao, JNK, CFJ Wu, and K. Yue. 1992. “Some Recent Work on Resampling Methods for Complex Surveys.” *Survey Methodology* 18(2):209–17.
- Salganik, Matthew J., Maeve B. Mello, Alexandre H. Abdo, Neilane Bertoni, Dimitri Fazito, and Francisco I. Bastos. 2011. “The Game of Contacts: Estimating the Social Visibility of Groups.” *Social Networks* 33(1):70–78. doi: 10.1016/j.socnet.2010.10.006.
- Särndal, C. E., B. Swensson, and J. Wretman. 2003. *Model Assisted Survey Sampling*. Springer Verlag.
- Sully, Elizabeth, Margaret Giorgio, and Selena Anjur-Dietrich. 2020. “Estimating Abortion Incidence Using the Network Scale-up Method.” *Demographic Research* 43:1651–84.
- Teo, Alvin Kuo Jing, Kiesha Prem, Mark IC Chen, Adrian Roellin, Mee Lian Wong, Hanh Hao La, and Alex R. Cook. 2019. “Estimating the Size of Key Populations for HIV in Singapore Using the Network Scale-up Method.” *Sex Transm Infect* sextrans–2018.
- Traud, Amanda L., Peter J. Mucha, and Mason A. Porter. 2012. “Social Structure of Facebook Networks.” *Physica A: Statistical Mechanics and Its Applications* 391(16):4165–80.
- Valliant, Richard, Jill A. Dever, and Frauke Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. Springer.
- Warner, Stanley L. 1965. “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias.” *Journal of the American Statistical Association* 60(309):63–69. doi: 10.1080/01621459.1965.10480775.
- Wolter, Kirk. 2007. *Introduction to Variance Estimation*. Second. New York: Springer.
- Zheng, Tian, Matthew J. Salganik, and Andrew Gelman. 2006. “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* 101(474):409–23. doi: 10.2307/27590705.

Appendixes

A Formal results for design-based estimators

We now formally state the conditions that must hold for the proposed design-based estimators for average network size, and for hidden population size, to be design-consistent and essentially unbiased. We will rely on the framework developed in Feehan and Salganik (2016a).

For convenience, we first repeat a previously known result that we make use of in our derivations. This result says that ratio estimators and compound ratio estimators are consistent and essentially unbiased.

Lemma A.1 (Compound ratio estimators). *Suppose that $\hat{y}_1, \dots, \hat{y}_n$ are estimators that are design-consistent and unbiased for $Y_1, \dots, Y_n$, respectively. Then the compound ratio estimator*

$$\hat{R} = \frac{\hat{y}_1, \dots, \hat{y}_k}{\hat{y}_{k+1} \dots \hat{y}_n}$$

is design-consistent and essentially unbiased for $R = (Y_1 \dots Y_k) / (Y_{k+1} \dots Y_n)$.

Proof. This result can be obtained by linearizing the estimator using a Taylor series expansion. See, for example, Rao and Pereira (1968), Wolter (2007), or Appendix E of Feehan and Salganik (2016a) for more details. $\square$

By *essentially unbiased*, we mean that the bias of the estimator gets small very quickly as the sample size grows. For example, it is well known that for ratio and compound ratio estimators like the ones discussed in Lemma A.1, the bias divided by the standard error goes to zero like $n^{-1/2}$ under simple random sampling (Särndal, Swensson, and Wretman 2003, Sec 5.6; Rao and Pereira 1968). In practice, applications of ratio estimators in applied settings have found that the bias is typically negligible unless the denominator is extremely small, which we do not expect to be the case in our setting (Feehan and Salganik (2016a), Appendix E discusses this topic in greater detail).

Throughout this Appendix, we use the notation $y_{A,B}$ or $y(A,B)$ to refer to the number of reports made by group A about connections to group B . For example, if F is the frame population and A_1 is bus drivers, then y_{F,A_1} would be the total number of connections to bus drivers that members of the frame population would report if they were all interviewed. $\bar{y}_{A,B}$ or $\bar{y}(A,B)$ is the average number of connections to group B , per member of group A . To continue our example, $\bar{y}_{F,A_1}$ would be the average number of connections that people in the frame population report to bus drivers. Similarly, we use $d_{A,B}$ to refer to the actual number of connections between group A and group B ; this may, of course, differ from the reported number of connections because reports about connections can be inaccurate.

Now we are ready to formally show the conditions under which the design-based average degree estimator will be consistent and essentially unbiased for the average degree. We will see that this result relies upon three key conditions. Like the known population estimator, the estimator below relies upon a *reporting condition*, which says that respondents' reports about their connections to people in the probe groups are accurate in aggregate; and a *probe alter condition*, which says that members of the probe groups are typical of members of the frame population, in terms of their average personal network sizes. The estimator below also relies upon an additional condition that is required to use respondents' reports to estimate the size of the probe groups: we call this condition the *probe membership condition*; it says that, in aggregate, respondents accurately report whether or not they are members of the probe groups.

Theorem A.1 (Design-based average degree). *Suppose that we have a probability sample s and, for each $i \in s$, an associated probability of inclusion π_i and weight $w_i = \frac{1}{\pi_i}$. Suppose also that we have a multiset of probe groups $\mathcal{A}$, chosen so that $A_j \subset F$ for each $A_j \in \mathcal{A}$. Assume that the network relationship respondents are asked to report about is symmetric, so that person a is connected to person b , b is also connected to a . Let $y_{i,\mathcal{A}} = \sum_{A_j \in \mathcal{A}} y_{i,A_j}$ be the total number of connections respondent i reports to the probe groups, and let $x_i = \sum_{A_j \in \mathcal{A}} x_{i,A_j}$ be the number of probe groups that respondent i reports being a member of. Finally, suppose the following three conditions hold:*

- $x_{\mathcal{A}} = \sum_{i \in F} x_i = N_{\mathcal{A}}$ (probe membership condition)
- $y_{F,\mathcal{A}} = d_{F,\mathcal{A}}$ (reporting condition)
- $\bar{d}_{\mathcal{A},F} = \bar{d}_{F,F}$ (probe alter condition)

Then the estimator

$$\hat{d}_{F,F} = \frac{\sum_{i \in s} \sum_{A_j \in \mathcal{A}} y_{i,A_j}}{\sum_{i \in s} w_i x_i} \quad (6)$$

is design-consistent and essentially unbiased for $\bar{d}_{F,F}$.

Proof. The proof closely follows the logic of Feehan and Salganik (2016a), Result B.3. The probe membership condition says that, in a census, the sum of the respondents' reported memberships in the probe groups must equal the actual size of the probe groups. When this probe membership condition is satisfied, $\hat{N}_{\mathcal{A}} = \sum_{i \in s} w_i x_i$ is a Horvitz-Thompson estimator for the total size of the probe groups, $N_{\mathcal{A}}$. As a Horvitz-Thompson estimator, $\hat{N}_{\mathcal{A}}$ is design-consistent and unbiased for $\sum_{i \in F} x_i$ (e.g., Särndal, Swensson, and Wretman 2003, Sec. 2.8); and, by the probe membership condition, $\sum_{i \in F} x_i = N_{\mathcal{A}}$. Similarly, $\hat{y}_{F,\mathcal{A}}$ is a Horvitz-Thompson estimator that is consistent and unbiased for $y_{F,\mathcal{A}}$. By Lemma A.1, we have

$$\frac{\hat{y}_{F,\mathcal{A}}}{\hat{N}_{\mathcal{A}}} \rightsquigarrow \frac{y_{F,\mathcal{A}}}{N_{\mathcal{A}}}.$$

The reporting condition implies that

$$\frac{y_{F,\mathcal{A}}}{\widehat{N}_{\mathcal{A}}} = \frac{d_{F,\mathcal{A}}}{N_{\mathcal{A}}}.$$

By the symmetry of the network relationship, $d_{F,\mathcal{A}} = d_{\mathcal{A},F}$. So $\frac{d_{F,\mathcal{A}}}{N_{\mathcal{A}}} = \bar{d}_{\mathcal{A},F}$. Finally, the probe alter condition implies that $\bar{d}_{\mathcal{A},F} = \bar{d}_{F,F}$. $\square$

Given the estimator for the average size of respondents' personal networks, we can now construct an estimator for the size of a hidden population:

$$\widehat{N}_H = \frac{\widehat{y}_{F,H}}{\widehat{d}_{F,F}},$$

where $\widehat{d}_{F,F}$ is the estimator for average degree discussed in Result A.1. It turns out there is no need for a new formal result here: Feehan and Salganik (2016a), Results C.10 and C.11 show that this estimator is design-consistent and essentially unbiased under the assumption that (i) reported connections to hidden population members are accurate, in aggregate; (ii) the average network size of hidden population members is the same as the average network size of people in the frame population. For an in-depth discussion, please see Feehan and Salganik (2016a).

B Two ways to motivate the estimator

In this appendix, we will show that the design-based estimator we introduce can be understood in two different ways: it can be seen as a simple extension of the original Killworth estimator; or, it can be motivated by a simple model.

Modified Killworth estimator

The first way to motivate the estimator introduced in the main text is by analogy with the known population estimator. Our modification is to estimate the size of the probe alters from the sample.

Estimator

$$\widehat{d}(F, F) = \frac{\widehat{y}(F, \mathcal{A})}{\widehat{N}_{\mathcal{A}}},$$

where $\widehat{y}(F, \mathcal{A})$ comes from the ARD questions on the survey, and $\widehat{N}_{\mathcal{A}}$ comes from the questions about survey respondents' membership in the probe alter groups.

The most important conditions that this estimator relies upon are the same as the known population estimator – in particular, this estimator requires the probe alter condition, which says that

$$\bar{d}(\mathcal{A}, F) = \bar{d}(F, F).$$

In words, this estimator assumes that members of the probe alter groups are typical of frame population members, in terms of their connections to frame population members.

A new condition

The second approach to motivating the new estimator starts from a new condition. This new derivation makes use of two facts.

Useful facts. Suppose X and Z are groups of people and suppose that $X_1 \subseteq X$ and $X_2 \subseteq X$ partition X , i.e., $X_1 \cup X_2 = X$ and $X_1 \cap X_2 = \phi$. Then

- (i) $\bar{y}(Z, X) = \bar{y}(Z, X_1) + \bar{y}(Z, X_2)$.
- (ii) $\bar{y}(X, Z) = p_{X_1|X} \bar{y}(X_1, Z) + (1 - p_{X_2|X}) \bar{y}(X_2, Z)$,

where $p_{X_1|X} = \frac{|X_1|}{|X|}$ is the size of X_1 relative to the size of X and $p_{X_2|X} = \frac{|X_2|}{|X|}$ is the size of X_2 relative to the size of X .

Now recall that F is the frame population for a survey and assume that $\mathcal{A}$ is a multiset of probe groups. Assume also that members of the probe groups are eligible to respond to the survey, so that $A_j \subseteq F$ for all $A_j \in \mathcal{A}$. Using facts (i) and (ii), we can write

$$\begin{aligned} \bar{y}(F, F) &= \bar{y}(F, \mathcal{A}) + \bar{y}(F, F - \mathcal{A}) \\ &= [p \bar{y}(\mathcal{A}, \mathcal{A}) + (1 - p) \bar{y}(F - \mathcal{A}, \mathcal{A})] + \\ &\quad [p \bar{y}(\mathcal{A}, F - \mathcal{A}) + (1 - p) \bar{y}(F - \mathcal{A}, F - \mathcal{A})], \end{aligned} \tag{7}$$

where $p = p_{\mathcal{A}|F} = \frac{N_{\mathcal{A}}}{N_F}$.

Equation (7) reveals that we can estimate $\bar{y}(F, F)$ if we have estimates for the following five quantities

- $p = p_{\mathcal{A}|F}$ – the prevalence of probe group membership among members of the frame population; we can estimate this directly
- $\bar{y}(\mathcal{A}, \mathcal{A})$ – the average number of reported connections from probe group members to other probe group members; we can estimate this directly
- $\bar{y}(F - \mathcal{A}, \mathcal{A})$ – the average number of reported connections from people who are not probe group members to people who are probe group members; we can estimate this directly

- $\bar{y}(\mathcal{A}, F - \mathcal{A})$ – the average number of connections from people who are probe group members to people who are not probe group members; we can estimate this indirectly (explained below)
- $\bar{y}(F - \mathcal{A}, F - \mathcal{A})$ – the average number of connections from people who are not probe group members to other people who are not probe group members; we need an additional condition to estimate this (explained below)

Two of these quantities require additional explanation. We start with $\bar{y}(\mathcal{A}, F - \mathcal{A})$, the average number of connections between people who are probe group members and people who are not probe group members.

Estimating $\bar{y}(\mathcal{A}, F - \mathcal{A})$ using symmetry

$\bar{y}(\mathcal{A}, F - \mathcal{A})$ is the average number of reported connections from people who are members of the probe groups to people who are not members of the probe groups. Although we do interview members of the probe groups, $\mathcal{A}$, we cannot directly estimate $\bar{y}(\mathcal{A}, F - \mathcal{A})$ because we do not ask questions about all of respondents' network members who are in any group outside of $\mathcal{A}$ ⁷. However, we can estimate $\bar{y}(\mathcal{A}, F - \mathcal{A})$ by exploiting the symmetry of the relationship that respondents are asked to report about. Since the relationship is symmetric, the total number of connections $y(\mathcal{A}, F - \mathcal{A})$ must equal $y(F - \mathcal{A}, \mathcal{A})$. Thus,

$$\begin{aligned}\bar{y}(\mathcal{A}, F - \mathcal{A}) &= \frac{y(\mathcal{A}, F - \mathcal{A})}{N_{\mathcal{A}}} \\ &= \frac{N_{F - \mathcal{A}}}{N_{\mathcal{A}}} \bar{y}(F - \mathcal{A}, \mathcal{A}) \\ &= \frac{(1 - p)}{p} \bar{y}(F - \mathcal{A}, \mathcal{A}).\end{aligned}$$

Since we can directly estimate p and $\bar{y}(F - \mathcal{A}, \mathcal{A})$ from the data we collect, we can estimate $\bar{y}(\mathcal{A}, F - \mathcal{A})$ via

$$\hat{\bar{y}}(\mathcal{A}, F - \mathcal{A}) = \frac{(1 - \hat{p})}{\hat{p}} \hat{\bar{y}}(F - \mathcal{A}, \mathcal{A}).$$

Estimating $\bar{y}(F - \mathcal{A}, F - \mathcal{A})$ by introducing an assumption

$\bar{y}(F - \mathcal{A}, F - \mathcal{A})$ is the average number of reported connections from people who are not members of the probe groups to other people who are not members of the probe groups. In order to estimate $\bar{y}(F - \mathcal{A}, F - \mathcal{A})$, we need to introduce a new condition. The condition we use here seems reasonable in the absence of any additional substantive information about the population being studied; of course, if researchers have additional substantive information that is informative about $\bar{y}(F - \mathcal{A}, F - \mathcal{A})$, then that information could be used to develop alternative assumptions.

⁷If it were feasible to ask this question, then we could directly estimate network degree by summing up responses to each category; this is how the summation method works (McCarty et al. 2001), for example.

We start by defining a new quantity $\lambda(A, B)$, which describes the probability that a randomly chosen member of a generic group A will be connected to a specific randomly chosen member of generic group B ⁸:

$$\lambda(A, B) \approx \frac{\bar{y}(A, B)}{N_B} = \frac{y(A, B)}{N_A N_B}.$$

Equation (7) showed that there are four relevant types of reports that we need to estimate average degree:

- $\mathcal{A} \rightarrow \mathcal{A}$
- $F - \mathcal{A} \rightarrow \mathcal{A}$
- $\mathcal{A} \rightarrow F - \mathcal{A}$
- $F - \mathcal{A} \rightarrow F - \mathcal{A}$

The first and last types of reports both involve a group reporting other members of the same group; the second and third types of report involve reporting between different groups, i.e., $\mathcal{A}$ and $F - \mathcal{A}$. The condition we introduce is intended to capture the intuition that we would expect within-group reporting to be different from between-group reporting. Specifically, we consider

$$\begin{aligned} \frac{\lambda(\mathcal{A}, \mathcal{A})}{\lambda(\mathcal{A}, F - \mathcal{A})} &= \frac{\bar{y}(\mathcal{A}, \mathcal{A})}{\bar{y}(\mathcal{A}, F - \mathcal{A})} \frac{N_{F - \mathcal{A}}}{N_{\mathcal{A}}} \\ &= \frac{\bar{y}(\mathcal{A}, \mathcal{A})}{\bar{y}(\mathcal{A}, F - \mathcal{A})} \frac{(1 - p)}{p}. \end{aligned} \tag{9}$$

Equation (9) captures the extent to which a member of $\mathcal{A}$ is more likely to report being connected to another member of $\mathcal{A}$, as opposed to a member of $F - \mathcal{A}$.

An analogous quantity captures the extent to which a member of $F - \mathcal{A}$ is more likely to report being connected to another member of $F - \mathcal{A}$, as opposed to a member of $\mathcal{A}$:

$$\frac{\lambda(F - \mathcal{A}, F - \mathcal{A})}{\lambda(F - \mathcal{A}, \mathcal{A})} = \frac{\bar{y}(F - \mathcal{A}, F - \mathcal{A})}{\bar{y}(F - \mathcal{A}, \mathcal{A})} \frac{p}{(1 - p)}.$$

The condition we introduce in order to estimate $\bar{y}(F - \mathcal{A}, F - \mathcal{A})$ is a *balance condition*, which says that

⁸Strictly speaking, the definition of λ is slightly more complex, since we do not allow self-connections in our network. This means that we need to adjust the denominator in cases where there is overlap between group A and group B . The complete definition of λ is given by

$$\lambda(A, B) = \frac{\bar{y}(A, B)}{N_B - p_{A|B}}, \tag{8}$$

where $p_{A|B} = \frac{N_{A \cap B}}{N_B}$. In this paper, we will only consider situations in which $p_{A|B}$ is equal to 1 or 0. Further, since $N_B \approx N_B - p_{A|B}$ for any reasonably large N_B , this adjustment in the denominator will make almost no difference in most applied settings.

$$\frac{\lambda(\mathcal{A}, \mathcal{A})}{\lambda(\mathcal{A}, F - \mathcal{A})} = \frac{\lambda(F - \mathcal{A}, F - \mathcal{A})}{\lambda(F - \mathcal{A}, \mathcal{A})}. \quad \text{Balance condition}$$

Under this assumption, some algebra reveals that $\bar{y}(F - \mathcal{A}, F - \mathcal{A})$ can be estimated using

$$\hat{\bar{y}}(F - \mathcal{A}, F - \mathcal{A}) = \frac{(1 - \hat{p})}{\hat{p}} \hat{\bar{y}}(\mathcal{A}, \mathcal{A}).$$

Putting it all together

Combining our results so far we have

$$\begin{aligned} \hat{\bar{y}}(F, F) &= \hat{\bar{y}}(\mathcal{A}, \mathcal{A}) + \hat{\bar{y}}(F - \mathcal{A}, \mathcal{A}) + \frac{(1 - \hat{p})}{\hat{p}} \hat{\bar{y}}(F - \mathcal{A}, \mathcal{A}) + \frac{(1 - \hat{p})}{\hat{p}} \hat{\bar{y}}(\mathcal{A}, \mathcal{A}) \\ &= \left[1 + \frac{(1 - \hat{p})}{\hat{p}} \right] \left[\hat{\bar{y}}(\mathcal{A}, \mathcal{A}) + \hat{\bar{y}}(F - \mathcal{A}, \mathcal{A}) \right] \\ &= \frac{\hat{\bar{y}}(F, \mathcal{A})}{\hat{p}} \\ &= \frac{\hat{\bar{y}}(F, \mathcal{A})}{\hat{N}_{\mathcal{A}}}. \end{aligned}$$

Thus, we see that this second perspective is equivalent to the first: we are able to estimate the average size of respondents' networks if the average number of connections from probe group members is equal to the average number of connections among all frame population members, i.e., if $\bar{d}(\mathcal{A}, F) = \bar{d}(F, F)$. Again, we call this the *probe group condition*. Qualitatively, it says that members of the probe groups have networks that are typical of frame population members, in terms of average personal network size.

B.1 Increased sampling variation under the new estimator

In this part of the appendix, we use the example of the network scale-up estimator to discuss how we expect sampling variation to be affected by the requirement that we estimate the size of the probe groups.

The basic scale-up estimator is

$$\hat{N}_H = \frac{\hat{y}_{F,H}}{\hat{\bar{d}}_{F,F}} = N_{\mathcal{A}} \frac{\hat{y}_{F,H}}{\hat{y}_{F,\mathcal{A}}}. \quad (10)$$

As a ratio estimator, we can approximate the sampling variance of Equation (10) with

$$\text{var}(\hat{N}_H) \approx N_{\mathcal{A}}^2 \left(\frac{1}{y_{F,\mathcal{A}}} \right)^2 \left[\text{var}(\hat{y}_{F,H}) + \left(\frac{y_{F,H}}{y_{F,\mathcal{A}}} \right)^2 \text{var}(\hat{y}_{F,\mathcal{A}}) - 2 \left(\frac{y_{F,H}}{y_{F,\mathcal{A}}} \right) \text{cov}(\hat{y}_{F,H}, \hat{y}_{F,\mathcal{A}}) \right]. \quad (11)$$

(See, e.g., Särndal, Swensson, and Wretman (2003) for a discussion of variance estimation for ratio estimators.)

B.1.1 Sampling variance when estimating probe group sizes from survey respondents

In our Hanoi study, we estimated the size of the known populations from the survey respondents. An adapted scale-up estimator uses the new design-based estimator for average degree (Equation (2)). Substituting this into the scale-up estimator, we have

$$\hat{N}_H = \frac{\hat{y}_{F,H}}{\hat{d}_{F,F}} = \hat{N}_A \frac{\hat{y}_{F,H}}{\hat{y}_{F,A}}.$$

The difference is that now the total size of the probe groups, N_A , is estimated rather than known.

Note that asking about whether or not the survey respondents are in the probe groups enables us to estimate the size of each $A_j \cap F$; that is, as we discuss in the main paper, if we ask our survey respondents if they are high school teachers, we can use the responses to estimate the number of high school teachers in the frame population.

As a compound ratio estimator, we can approximate the sampling variance of this new estimator $\hat{N}_H$ with

$$\text{var}(\hat{N}_H) \approx \left(N_A \frac{y_{F,H}}{y_{F,A}} \right)^2 \left[\frac{\text{var}(\hat{N}_A)}{N_A^2} + \frac{\text{var}(\hat{y}_{F,H})}{y_{F,H}^2} + \frac{\text{var}(\hat{y}_{F,A})}{y_{F,A}^2} + 2 \frac{\text{cov}(\hat{y}_{F,H}, \hat{N}_A)}{y_{F,H} N_A} - 2 \frac{\text{cov}(\hat{y}_{F,H}, \hat{y}_{F,A})}{y_{F,H} y_{F,A}} - 2 \frac{\text{cov}(\hat{N}_A, \hat{y}_{F,A})}{N_A y_{F,A}} \right]. \quad (12)$$

Equation (12) shows that the sampling variance of the size estimate will depend on

- N_A - the total size of the probe groups
- $y_{F,A}$ - the number of connections to the probe groups
- $y_{F,H}$ - the number of connections to the hidden population
- $\text{var}(\hat{N}_A)$ - the sampling variance of the estimated size of the probe groups (based on the total number of reported probe groups the respondents report being in)
- $\text{var}(\hat{y}_{F,H})$ - the sampling variance of reported connections to the hidden population
- $\text{var}(\hat{y}_{F,A})$ - the sampling variance of the total reported connections to the probe groups; this should be related to the variance in respondents' personal network sizes
- $\text{cov}(\hat{y}_{F,H}, \hat{y}_{F,A})$ - the covariance between the two estimators
- $\text{cov}(\hat{y}_{F,H}, \hat{N}_A)$ - the covariance between the two estimators
- $\text{cov}(\hat{y}_{F,A}, \hat{N}_A)$ - the covariance between the two estimators

The ‘extra’ variance due to the fact that we estimate the size of the set of probe groups is

$$\left(N_{\mathcal{A}} \frac{y_{F,H}}{y_{F,\mathcal{A}}}\right)^2 \left[\frac{\text{var}(\hat{N}_{\mathcal{A}})}{N_{\mathcal{A}}^2} + 2 \frac{\text{cov}(\hat{y}_{F,H}, \hat{N}_{\mathcal{A}})}{y_{F,H} N_{\mathcal{A}}} - 2 \frac{\text{cov}(\hat{N}_{\mathcal{A}}, \hat{y}_{F,\mathcal{A}})}{N_{\mathcal{A}} y_{F,\mathcal{A}}} \right]. \quad (13)$$

B.2 Empirical checks

An appealing feature of the known population method is that it permits researchers to conduct internal consistency checks that can help reveal problems with data collection and assess how well the probe groups satisfy the probe group condition. However, the internal consistency checks for the known population method rely on the fact that the size of the probe groups $\mathcal{A}$ are known; we now describe an approach to generalizing the internal consistency checks to our new data collection strategy, in which we have sample-based information about the size of the groups in $\mathcal{A}$.

Let $\mathcal{A}$ be the set of probe groups, and consider a partition of $\mathcal{A}$ into subsets⁹ $\mathcal{A}^\bullet$ and $\mathcal{A}^\circ$. The idea will be pretend that we don’t have any direct information about $N_{\mathcal{A}^\circ}$ (i.e., the size of $\mathcal{A}^\circ$), and to try to estimate it using the rest of the information we have available to us, ie: (i) information about the size of $\mathcal{A}^\bullet$; (ii) information about respondents’ connections to people in $\mathcal{A}^\bullet$; and (iii) information about respondents’ connections to people in $\mathcal{A}^\circ$. We call this a ‘hold-out’ estimate of $N_{\mathcal{A}^\circ}$, and we will denote it by $\hat{N}_{\mathcal{A}^\circ}^{HO}$. $\hat{N}_{\mathcal{A}^\circ}^{HO}$ can be defined as:

$$\begin{aligned} \hat{N}_{\mathcal{A}^\circ}^{HO} &= \frac{y(F, \mathcal{A}^\circ)}{\hat{d}_{\mathcal{A}^\bullet}} \\ &= \frac{\text{Number of reported connections to hold-out set}}{\text{average network size based on included set}} \\ &= \frac{y(F, \mathcal{A}^\circ)}{\hat{y}(F, \mathcal{A}^\bullet)} \hat{N}_{\mathcal{A}^\bullet}. \end{aligned}$$

A second, independent estimate of $N_{\mathcal{A}^\circ}$ comes from information about survey respondents’ membership in the groups in $\mathcal{A}^\circ$:

$$\hat{N}_{\mathcal{A}^\circ} = \sum_{i \in s} w_i \left(\sum_{g \in \mathcal{A}^\circ} z_{i,A_g} \right),$$

where g indexes the probe groups in $\mathcal{A}^\circ$, and z_{i,A_g} is an indicator variable for whether or not respondent i reports being a member of probe group A_g . $\hat{N}_{\mathcal{A}^\circ}$ is a standard, Horvitz-Thompson estimator for the size of $\mathcal{A}^\circ$ based on our probability sample of survey respondents.

We have two separate estimates for the size of $\mathcal{A}^\circ$: $\hat{N}_{\mathcal{A}^\circ}^{HO}$ and $\hat{N}_{\mathcal{A}^\circ}$. These estimates are based on separate pieces of information; thus, we can compare them to see how much they differ. We would expect these

⁹Note that, technically, $\mathcal{A}$, $\mathcal{A}^\bullet$, and $\mathcal{A}^\circ$ are multisets.

estimates to differ if there were problems in the data collection, or if the probe group assumption was violated.

C Stan code for model-based estimates

We fit our models using `stan` (Carpenter et al. 2017). For the known population models fit to the simulations, we ran each model for 2,000 iterations (including 1,000 warmup iterations). Across all fits, the maximum $\hat{R}$ statistic was 1.01. For the new models fit to the simulations, we ran each model for 4,000 iterations (including 2,000 warmup iterations). Across all fits, the maximum $\hat{R}$ statistic was 1.03. For the Hanoi data, for each tie definition we ran the model with 4 chains and 10,000 iterations (including 2,000 warmup iterations). Across all fits, the maximum $\hat{R}$ statistic was 1.02.

Code for model with known probe group sizes

```
//  
// Random degree model with known popn estimation (i.e., assuming  
// probe gp sizes are known) for Hanoi paper  
// added hidden populations  
//  
  
data {  
  // the size of the entire population (not the sample)  
  real<lower=0> N;  
  
  // the size of the sample  
  int<lower=0> n;  
  
  // the number of probe groups  
  int<lower=0> G;  
  
  // the number of hidden groups  
  // (can be set to 0, if there are no hidden groups)  
  int<lower=0> K;  
  
  // reported membership in groups  
  // (not needed when probe group size is known)  
  //int<lower=0> tot_Z[G];  
  
  // probe group sizes, treated as known  
  // needed when probe group size is known  
  real<lower=0,upper=1> p_g[G];  
  
  // reported connections to groups (the ARD)
```



```

int<lower=0> Y[n,G];

// reported connections to hidden groups (also ARD)
int<lower=0> X[n,K];

}

parameters {

    // the prevalence of the probe groups
    // (not needed when probe group size is known)
    //vector<lower=0,upper=1>[G] p_g;

    // the prevalence of the hidden groups
    vector<lower=0,upper=1>[K] p_k;

    // the mean and variance of the log-degree distribution
    // NB: be sure the constraints are right here - this can cause
    //      bugs if they are not!
    real<lower=1,upper=8> mu;
    real<lower=0.25,upper=2> sigma;

    // the degree for each survey respondent
    // note that, under this model, the degree is not discrete - so it might
    // be more accurate to think about it as an expected degree
    vector<lower=0>[n] d;

}

model {

    // priors on probe group prevalences: uniform on (0,1)
    // NB: if these are changed, be sure to change constraints on parameters too
    //p_g ~ uniform(0,1);

    // priors on hidden population prevalences: uniform on (0,1)
    // NB: if these are changed, be sure to change constraints on parameters too
    if(K > 0) {
        p_k ~ uniform(0,1);
    }

    ////////////
    // Log-Normal degrees
    // NB: these priors are on the log scale
    // NB: if these are changed, be sure to change constraints on parameters too

```

```

mu ~ uniform(1,8);
sigma ~ uniform(0.25,2);

d ~ lognormal(mu, sigma);

// model for respondents' group memberships, given Nk's
// (not needed when probe group size is known)
//for(g in 1:G) {
// tot_Z[g] ~ binomial(n, p_g[g]);
//}

// model for probe group ARD
for(g in 1:G) {

    ////////////
    // Poisson model
    Y[,g] ~ poisson(d * p_g[g]);

}

// model for hidden group ARD
if(K > 0) {
    for(k in 1:K) {

        ////////////
        // Poisson model
        X[,k] ~ poisson(d * p_k[k]);

    }
}

}

generated quantities {
    real<lower=0> mean_degree;
    real<lower=0> var_degree;

    // this is the mean and variance of the log-normal degree distribution
    mean_degree = exp(mu + (sigma^2)/2);
    var_degree = exp((sigma^2) - 1) * exp(2*mu + sigma^2);
}

```

Code for model with estimated probe group sizes

```

//
// Random degree model with probe group estimation for Hanoi paper
// added hidden populations
//

data {
  // the size of the entire population (not the sample)
  real<lower=0> N;

  // the size of the sample
  int<lower=0> n;

  // the number of probe groups
  int<lower=0> G;

  // the number of hidden groups
  // (can be set to 0, if there are no hidden groups)
  int<lower=0> K;

  // reported membership in groups
  int<lower=0> tot_Z[G];

  // reported connections to groups (the ARD)
  int<lower=0> Y[n,G];

  // reported connections to hidden groups (also ARD)
  int<lower=0> X[n,K];
}

parameters {

  // the prevalence of the probe groups
  vector<lower=0,upper=1>[G] p_g;

  // the prevalence of the hidden groups
  vector<lower=0,upper=1>[K] p_k;

  // the mean and variance of the log-degree distribution
  // NB: be sure the constraints are right here - this can cause
  // bugs if they are not!
  real<lower=1,upper=8> mu;
  real<lower=0.25,upper=2> sigma;

  // the degree for each survey respondent

```

```

// note that, under this model, the degree is not discrete - so it might
// be more accurate to think about it as an expected degree
vector<lower=0>[n] d;

}

model {

  // priors on probe group prevalences: uniform on (0,1)
  // NB: if these are changed, be sure to change constraints on parameters too
  p_g ~ uniform(0,1);

  // priors on hidden population prevalences: uniform on (0,1)
  // NB: if these are changed, be sure to change constraints on parameters too
  if(K > 0) {
    p_k ~ uniform(0,1);
  }

  ////////////
  // Log-Normal degrees
  // NB: these priors are on the log scale
  // NB: if these are changed, be sure to change constraints on parameters too
  mu ~ uniform(1,8);
  sigma ~ uniform(0.25,2);

  d ~ lognormal(mu, sigma);

  // model for respondents' group memberships, given Nk's
  for(g in 1:G) {
    tot_Z[g] ~ binomial(n, p_g[g]);
  }

  // model for probe group ARD
  for(g in 1:G) {

    ////////////
    // Poisson model
    Y[,g] ~ poisson(d * p_g[g]);

  }

  // model for hidden group ARD
  if(K > 0) {
    for(k in 1:K) {

```

```

//////////
// Poisson model
X[,k] ~ poisson(d * p_k[k]);

}
}

}

generated quantities {
  real<lower=0> mean_degree;
  real<lower=0> var_degree;

  // this is the mean and variance of the log-normal degree distribution
  mean_degree = exp(mu + (sigma^2)/2);
  var_degree = exp((sigma^2) - 1) * exp(2*mu + sigma^2);
}

```

D Sampling design for Hanoi study

The 2017 Vietnam Health and Living Standards Survey (VHLSS) used a stratified, multi-stage design to obtain a nationally representative sample of households across Vietnam. As a result of this design, the VHLSS sampled 140 enumeration areas in Hanoi. The sample for our study was obtained by randomly sub-sampling 46 of these 140 enumeration areas. In each enumeration area, 20 households were randomly selected for the VHLSS, including 5 backups. After respondents completed the VHLSS interview, they were asked to also participate in our scale-up survey. A household member aged 18-55 was selected to respond to the survey. Figure 9 shows the unweighted age and sex distribution of our respondents.

The randomness of sampling, together with nonresponse, frame errors, and other factors lead to discrepancies between the size of age-sex groups implied by the base weights and population totals from the projection (Figure 10, panel a). As is standard in household surveys, we used calibration to adjust the weights provided by GSO in order to match population margins by age and sex group (Deville and Särndal 1992; Lumley 2010; see, eg, Valliant, Dever, and Kreuter 2013). We obtained population projections for Hanoi in 2017 from the General Statistical Office’s Population Projections for 2009-2049, dated February 2011 (pg. 88). We rescaled age-sex distributions to match the population size projected by GSO, based on the raw survey weights. The population projections were based on 5-year age groups. We approximated the number of people aged 18-19 by multiplying the projected number of 15-19 year-olds by 0.4. We dropped one survey response for whom gender was not recorded. Calibration was conducted with the R package `survey` (Lumley 2019).

Figure 9: Unweighted age and sex distribution of survey respondents.

Figure 10 shows the relationship between population projection-based estimates for the size of age-sex groups (x axis) and estimates based on the survey weights (y axis). The left-hand panel shows the weights before calibration, and the right-hand panel shows the calibrated weights.

E Additional results from Hanoi study

In this Appendix, we present additional results from the Hanoi survey, including tables that show exact values for several of the survey results.

Figure 10: **(a)** Relationship between population projection (x axis) and implied population size from unadjusted survey weights (y axis). **(b)** Relationship between population projection (x axis) and implied population size from calibrated survey weights (y axis). Calibration ensures that totals match the population projections in each age/sex group. Grey lines show 20 percent above and below equality.

Table 5: Estimated size of each of the probe groups, based on respondents reported membership in each group. 95% confidence intervals capture estimated sampling uncertainty, and are based on the rescaled bootstrap.

Probe group	Estimated size	95% Confidence interval
Coffeeshop owner	8,809	(0 - 30,949)
Nurse	10,719	(0 - 26,957)
Lawyer	15,567	(0 - 37,312)
Xe om driver	17,496	(0 - 47,207)
Hairdresser	19,266	(4,155 - 39,606)
Primary teacher	33,491	(5,798 - 68,759)
Bank worker	36,078	(6,398 - 75,599)
Security guard	37,836	(3,934 - 80,111)
Taxi driver	79,332	(28,564 - 140,786)
Builder/Mason	184,044	(95,896 - 291,745)
Student	363,858	(245,369 - 504,082)
Rice farmer	469,667	(251,600 - 708,828)
Total		
All probe groups	1,276,163	(1,008,964 - 1,548,140)

Figure 11: Relationship between the estimated size of each probe group (x axis) and the average number of connections survey respondents report to each group (y axis). Each point is a probe group, and each panel is a tie definition.

Figure 12: Average number of reported connections to each of the hidden populations, and to women who smoke and women who had a baby in the past 12 months. This figure is analagous to Figure 6.

Table 6: Design-based and model-based size estimates for the four hidden populations, and for women who smoke and women who had a baby in the past year. Design-based estimates have a 95% confidence interval that captures estimated sampling uncertainty, based on the rescaled bootstrap. Model-based estimates have a 95% credible interval based on the assumptions of the model.

Quantity	Estimate source	Tie definition estimate (95% interval)			
		Any contact	Conversational contact	Shared a meal in last 12 mo.	Shared a meal in last month
Women who had a baby in the past 12mo	Model	107,530 (92,021 - 125,031)	109,344 (93,156 - 127,302)	87,755 (73,520 - 103,877)	94,574 (77,070 - 115,462)
Women who smoke	Model	10,219 (8,175 - 12,668)	8,400 (6,448 - 10,748)	6,573 (4,491 - 9,303)	8,652 (5,295 - 13,338)
People who use drugs	Model	17,354 (14,266 - 20,975)	9,754 (7,607 - 12,363)	2,212 (1,156 - 3,837)	2,798 (1,147 - 5,547)
People who inject drugs	Model	9,690 (7,728 - 12,054)	2,390 (1,531 - 3,567)	695 (206 - 1,674)	1,119 (253 - 3,060)
Men who have sex with men	Model	1,384 (831 - 2,165)	1,032 (526 - 1,829)	1,453 (647 - 2,788)	1,956 (675 - 4,379)
Female Sex Workers	Model	5,468 (4,137 - 7,116)	3,262 (2,214 - 4,627)	129 (5 - 715)	289 (11 - 1,564)
Female Sex Workers	Design	3,803 (1,526 - 6,502)	2,072 (389 - 4,222)	0 (0 - 0)	0 (0 - 0)
Men who have sex with men	Design	1,555 (637 - 2,664)	1,283 (312 - 2,373)	2,132 (335 - 4,248)	3,489 (0 - 6,895)
People who inject drugs	Design	7,987 (2,014 - 16,602)	2,307 (631 - 5,062)	799 (0 - 1,984)	746 (0 - 1,910)
People who use drugs	Design	14,062 (6,965 - 23,149)	7,687 (3,827 - 11,804)	1,747 (607 - 3,294)	1,745 (361 - 3,536)
Women who smoke	Design	9,285 (5,033 - 14,505)	8,522 (4,031 - 14,307)	7,754 (1,717 - 16,692)	9,785 (1,125 - 23,648)
Women who had a baby in the past 12mo	Design	96,526 (74,662 - 121,377)	99,066 (76,028 - 123,706)	75,816 (54,672 - 100,020)	76,552 (53,720 - 105,713)

In Table 7 we compare the estimated uncertainty for each population across the four tie definitions. To do so, we define the *relative interval width* (RIW) to be the ratio of the width of the uncertainty estimate to the point estimate:

$$\text{relative interval width} = RIW(\hat{N}_H) = \frac{\hat{N}_H^{\text{upper}} - \hat{N}_H^{\text{lower}}}{\hat{N}_H},$$

where $\hat{N}_H$ is a particular estimate and $\hat{N}_H^{\text{upper}}$ and $\hat{N}_H^{\text{lower}}$ are the upper and lower boundaries of the associated uncertainty interval.

The relative interval width is appealing because it adjusts for the size of the group being estimated; thus, comparing the relative interval width for different estimates compares the estimated relative uncertainty of the estimates.

Table 7: Relative interval width for estimated population sizes in Hanoi. We define the relative interval width to be the ratio of the width of the uncertainty interval, divided by the point estimate. Across estimates, higher relative interval widths imply a higher level of relative uncertainty. Two estimates are 0; these are denoted by a dash.

Quantity	Estimate source	Tie definition - Relative interval width			
		Any contact	Conversational contact	Shared a meal in last 12 mo.	Shared a meal in last month
Women who had a baby in the past 12mo	Model	0.31	0.31	0.35	0.41
Women who smoke	Model	0.44	0.51	0.73	0.93
People who use drugs	Model	0.39	0.49	1.21	1.57
People who inject drugs	Model	0.45	0.85	2.11	2.51
Men who have sex with men	Model	0.96	1.26	1.47	1.89
Female Sex Workers	Model	0.54	0.74	5.48	5.37
Female Sex Workers	Design	1.31	1.85	–	–
Men who have sex with men	Design	1.30	1.61	1.84	1.98
People who inject drugs	Design	1.83	1.92	2.48	2.56
People who use drugs	Design	1.15	1.04	1.54	1.82
Women who smoke	Design	1.02	1.21	1.93	2.30
Women who had a baby in the past 12mo	Design	0.48	0.48	0.60	0.68

E.1 Reference estimates

The Hanoi Preventative Medicine Center estimated the prevalence of women who currently smoke in Hanoi by age: for ages 18-44, they estimated 0.57% (95% CI: 0.21% - 1.51%) and for ages 45-69, they estimated 0.42% (95% CI: 0.15% - 1.12%)¹⁰. We applied these estimated rates to the proportion of women in Hanoi in each age group to estimate that there are 11,016 total women who smoke in the city.

To estimate the number of women who had babies in the past 12 months, we took the projected population of Hanoi aged 0-4 of both sexes for 2017, which is 629,362. We then divided this by 5 to estimate the number of children under age 1 as 125,872. We use this number as an approximation for the expected number of women with children under age 1, noting that it is not expected to be exactly correct, due to errors in the projection, women who have twins, infant mortality, migration, and other factors.

E.2 Sensitivity framework

To reason about patterns in estimated hidden population sizes across tie definitions, we use the sensitivity framework of Feehan and Salganik (2016a), which can be applied to the design-based estimator. The framework decomposes the design-based scale-up estimand into the true hidden population size and several adjustment factors¹¹:

$$\hat{N}_H \cdot \frac{\eta}{\delta \tau} \rightsquigarrow N_H, \quad (14)$$

where the squiggly arrow means that the quantity on the left-hand side is consistent and essentially unbiased for the quantity on the right-hand side; $\hat{N}_H = \frac{y_{F,H}}{d_{F,F}}$ is the design-based scale-up estimand, N_H is the true hidden population size, and η , δ , and τ are adjustment factors. The adjustment factors capture conditions that the estimator relies upon; when these conditions are satisfied, each adjustment factor is equal to 1, and the estimator is unbiased. The first adjustment factor $\delta = \frac{\bar{d}_{H,F}}{\bar{d}_{F,F}}$ is the ratio of the average network size among hidden population members ($\bar{d}_{H,F}$) to the average network size among all frame population members ($\bar{d}_{F,F}$). When $\delta > 1$, hidden population members have bigger networks, on average, than frame population members. On the other hand, when $\delta < 1$, hidden population members have smaller networks, on average, than frame population members.

The remaining two adjustment factors capture the extent to which reported connections are accurate in aggregate. $\eta = \frac{y_{F,H}^+}{y_{F,H}}$ is the ratio of the number of reported connections to hidden population members that actually lead to hidden population members ($y_{F,H}^+$) to the total number of reported connections to

¹⁰Personal communication with Nga Nguyen (UNAIDS Hanoi), January 11, 2018.

¹¹Note that we omit the term for the frame ratio from Feehan and Salganik (2016a) because, in our study, we asked aggregate relational data questions about connections to people in the frame population. In other words, we estimate $\bar{d}_{F,F}$ rather than $\bar{d}_{F,U}$ (which some previous scale-up studies have estimated).

hidden population members ($y_{F,H}$). When there are false positive reports—that is, when some respondents report that others are hidden population members when they are not—then $\eta < 1$. And $\tau = \frac{y_{F,H}^+}{d_{F,H}}$ is the ratio of the number of reported connections from frame population members to the hidden populations, to the actual number of connections from frame population members to the hidden population.

E.3 Formal models for the true positive rate (τ)

We developed formal models to help think through possible explanations for the cross-tie patterns shown in Figure 8. Like all formal models, these are simplifications of the actual mechanism that leads to a given survey response. But they can be a valuable tool for reasoning about different possible factors that lead to survey responses in a structured way.

Throughout this section, we focus on the parameter τ because we hypothesize that τ is the most important adjustment factor affecting differences in size estimates as a function of tie strength. This hypothesis arose during discussions we had with local stakeholders and interviewers as part of trying to choose the tie definitions to use during survey design. These discussions suggested that respondents might be reluctant to admit to associating with members of some of the hidden populations due to social stigma. If this were true, then it could result in an unwillingness to report connections to members of the hidden populations, leading to a value less than 1 for τ .

It is important to note that, while we suspect τ is the most important adjustment factor, this is a conjecture: it could alternately be the case, for example, that hidden population members’ average network size differs from the average network frame population network size as a function of tie definition, making the parameter δ more relevant. We note that similar models could be developed for other parameters. Ultimately, the best way to resolve this uncertainty would be to collect data about the size of the adjustment factors.

We start from the discussion in Feehan (2015) (Appendix C.2). We assume that the strength of a particular tie definition can be assigned a value $t \in T$, where T is a range of possible tie strengths $T = [t_{\min}, t_{\max}]$. Lower values of t imply stronger ties, so that $t_{\min}$ is the strongest possible tie. The actual numerical value of t is arbitrary; all that matters is the ordering of tie strengths (similar to the concept of utility).

To try to understand what might explain the patterns in size estimates across tie definitions shown in Figure 8, we consider two scenarios. In each scenario, our goal is to understand if it is reasonable to expect that a specific tie strength will be optimal in the sense that it will produce the highest value of τ . In both scenarios, we show that this is a reasonable expectation. The implication is that choosing the best type of tie to ask about can make a difference in the accuracy of the estimates; thus, researchers should treat tie strength as a critical study design parameter.

Scenario 1 - τ is a function of awareness

The first scenario models τ as a function of awareness. This is the scenario that led to the original quantity/quality tradeoff hypothesis. The idea is that we expect respondents to know more about people who they are connected to through strong ties as opposed to weak ties, and we expect respondents to make a good faith effort to report as accurately as they are able to.

We formalize this by assuming that transmission error, τ , is related to tie strength according to:

$$\tau(t) = \alpha(t),$$

where $\alpha(t)$ is a function we will call awareness. We assume that $\alpha(t)$ has a few important properties. First, we assume the range of $\alpha(t)$ is $[0, 1]$, which is the set of possible values for τ . Second, since we expect respondents to be more aware of other people they are connected to through stronger ties, we assume that $\alpha'(t) < 0$. Finally, we also expect that moving from a very strong tie to a moderately strong tie will make more of a difference to τ than moving from a moderately strong tie to a weak tie. Thus, we also assume that $\alpha''(t) < 0$.

Since, by assumption, $\alpha(t)$ is strictly nonincreasing in t , it follows that the researcher should pick as strong tie as possible in order to maximize $\tau(t)$. Of course, in practice, we expect this to be balanced by the fact that stronger tie definitions will produce more sampling error; this predicted tension between sampling error and hypothesized awareness originally motivated the quantity/quality tradeoff hypothesis proposed in Feehan et al. (2016).

Scenario 2 - τ is a function of awareness and stigma

In many cases, we expect the first scenario to be reasonable. However, for some stigmatized hidden populations, it seems plausible that respondents may be reluctant to admit that they are connected to a hidden population member through a strong tie. In this case, the accuracy of respondents' reports is a more complex function of how much they know about their network members, and how much they are willing to accurately report about their network members. This second scenario is intended to describe this situation more formally. We now assume that transmission error, τ , is related to tie strength according to two factors:

$$\tau(t) = \alpha(t)[1 - \sigma(t)],$$

where $\alpha(t)$ has the same construction as before, and $\sigma(t)$ is a new factor that is related to stigma. Similar to $\alpha(t)$, $\sigma(t)$ also has range $[0, 1]$. Thus, $\tau(t)$ is near one when awareness is high (i.e., $\alpha(t)$ is near 1) and

stigma is low (i.e., $\sigma(t)$ is near 0). We hypothesize that stigma will prevent people from being willing to report about hidden population members who are close ties – thus, we assume that $\sigma' > 0$. Again, we also assume that the effect of tie strength diminishes as tie strength weakens, so that $\sigma''(t) < 0$.

Together, these assumptions imply that

$$\frac{d\tau(t)}{dt} = \alpha'(t)\sigma(t) + \alpha(t)\sigma'(t).$$

We will now show that, under this model, we still expect there to be a tie strength at which transmission error is maximized. We are interested in finding a tie strength $t = t^*$ that maximizes $\tau(t)$. The first order condition tells us that, at such a maximum, $\left. \frac{d\tau(t)}{dt} \right|_{t=t^*} = 0$, meaning that

$$\frac{\alpha'(t^*)}{\alpha(t^*)} = -\frac{\sigma'(t^*)}{\sigma(t^*)}.$$

The second order condition says that, for τ^* to be a maximum, we must have

$$\left. \frac{d^2\tau(t)}{dt^2} \right|_{t=t^*} = \alpha''(t)\sigma(t) + \alpha'(t)\sigma'(t) + \alpha'(t)\sigma'(t) + \alpha(t)\sigma''(t) < 0.$$

Dividing through by $\alpha(t)$, this becomes

$$\begin{aligned} \frac{\alpha''(t)}{\alpha(t)}\sigma(t) + 2\frac{\alpha'(t)}{\alpha(t)}\sigma'(t) + \sigma''(t) &< 0 \\ \frac{\alpha''(t)}{\alpha(t)}\sigma(t) - 2\frac{\sigma'(t)^2}{\sigma(t)} + \sigma''(t) &< 0, \end{aligned}$$

where we obtain the second line by substituting the first-order condition in for $\frac{\alpha'(t)}{\alpha(t)}$. Considering each of the three terms on the left hand side of the inequality one at a time, note that

- $\frac{\alpha''(t)}{\alpha(t)}\sigma(t) < 0$, since $\sigma(t) > 0$, $\alpha(t) > 0$ and, by assumption, $\alpha''(t) < 0$
- $-2\frac{\sigma'(t)^2}{\sigma(t)} < 0$, since $\sigma(t) > 0$ and, by assumption, $\sigma'(t) > 0$
- $\sigma''(t) < 0$, again by assumption

Thus, the expression will always be negative, and the optimum value t^* is indeed a maximum.

The implication of this second model is thus different from the first: in this scenario, the value t^* that produces the most accurate estimates may not be the strongest possible tie that sampling error will allow; instead, the countervailing forces of awareness and willingness to disclose may mean that a tie of intermediate strength is be more accurate.

Parametric example for scenario 2

We now develop a concrete example that illustrates scenario 2 by specifying specific functional forms for $\alpha(t)$ and $\sigma(t)$. Suppose that $\alpha(t) = \exp(-K_1 t)$ and $\sigma(t) = \exp(-K_2 t)$ for constants $K_1, K_2 > 0$. We then have

$$\tau(t) = \exp(-K_1 t) [1 - \exp(-K_2 t)] = \exp(-K_1 t) - \exp(-(K_1 + K_2)t).$$

Figure 13 illustrates this setup with an illustration of $\alpha(t)$, $\sigma(t)$ and $\tau(t)$ for several combinations of parameter values.

We will show how these parametric assumptions allow us to find an expression for the tie strength t^* that will maximize $\tau(t)$. Note that

$$\tau'(t) = -K_1 \exp(-K_1 t) + (K_1 + K_2) \exp(-(K_1 + K_2)t).$$

At a maximum t^* , $\tau'(t^*) = 0$. So we have

$$\begin{aligned} -K_1 \exp(-K_1 t^*) + (K_1 + K_2) \exp(-(K_1 + K_2)t^*) &= 0 \\ K_1 \exp(-K_1 t^*) &= (K_1 + K_2) \exp(-(K_1 + K_2)t^*) \\ \exp(-K_1 t^*) &= \frac{(K_1 + K_2)}{K_1} \exp(-(K_1 + K_2)t^*) \\ -K_1 t^* &= \log \left(\frac{(K_1 + K_2)}{K_1} \right) - (K_1 + K_2)t^* \\ t^* &= \frac{1}{K_2} \log \left(\frac{(K_1 + K_2)}{K_1} \right). \end{aligned}$$

To explicitly confirm that t^* is a maximum¹², we check the second order condition. Using the fact that $\tau''(t) = K_1^2 \exp(-K_1 t) - (K_1 + K_2)^2 \exp(-(K_1 + K_2)t)$, we find that in order for t^* to be a maximum, we must have

$$\begin{aligned} K_1^2 \exp(-K_1 t^*) - (K_1 + K_2)^2 \exp(-(K_1 + K_2)t^*) &< 0 \\ K_1^2 \exp \left(-\frac{K_1}{K_2} \log \left(\frac{(K_1 + K_2)}{K_1} \right) \right) - (K_1 + K_2)^2 \exp \left(-\frac{(K_1 + K_2)}{K_2} \log \left(\frac{(K_1 + K_2)}{K_1} \right) \right) &< 0 \\ K_1^2 \left[\frac{(K_1 + K_2)}{K_1} \right]^{-\frac{K_1}{K_2}} - (K_1 + K_2)^2 \left[\frac{(K_1 + K_2)}{K_1} \right]^{-\frac{(K_1 + K_2)}{K_2}} &< 0 \end{aligned}$$

¹²The fact that t^* is a maximum in this case also follows from the more general discussion in the previous section, since $\alpha'(t) < 0$, $\alpha''(t) < 0$, $\sigma'(t) > 0$, and $\sigma''(t) < 0$.

Or

$$\begin{aligned}
K_1^2 \left[\frac{(K_1 + K_2)}{K_1} \right]^{-\frac{K_1}{K_2}} &< (K_1 + K_2)^2 \left[\frac{(K_1 + K_2)}{K_1} \right]^{-\frac{(K_1 + K_2)}{K_2}} \\
K_1^2 &< (K_1 + K_2)^2 \left[\frac{(K_1 + K_2)}{K_1} \right]^{-1} \\
K_1 &< (K_1 + K_2).
\end{aligned}$$

This will always hold, since we assumed that $K_1, K_2 > 0$. So t^* is indeed a maximum.

To further aid intuition, Figure 13 illustrates the optimal value t^* for several hypothetical values of K_1 and K_2 under this parametric model. The figure reveals that different scenarios can lead to different optimal tie strengths. For example, the bottom-left panel shows a situation in which awareness is strongly related to tie strength and stigma is weakly related to tie strength ($K_1 = 0.9, K_2 = 0.1$). In this case, a strong tie produces the most accurate estimates. On the other hand, the top-right panel shows a situation in which awareness is weakly related to tie strength, while stigma is strongly related to tie strength ($K_1 = 0.1, K_2 = 0.9$). In this case, a stronger tie produces the most accurate estimates.

Figure 13: Illustration of the relationship between awareness ($\alpha(t)$, in blue), lack of stigma ($1 - \sigma(t)$, in red), and the parameter τ (in black) based on the hypothetical parametric model. The dotted vertical line shows the value of tie strength t^* that maximizes the parameter τ . The curves illustrate how different relationships between awareness and lack of stigma can interact to produce different optimal tie strengths. Note that the scale of tie strength (x axis) is arbitrary; only the ordering of tie strengths is relevant. In these plots, the strongest tie is at $t = 0$ and tie strength weakens as t increases.