CHAPTER 23: Network-Based Methods for Accessing Hard-to-Reach Populations Using Standard Surveys

Tyler H. McCormick, University of Washington and Tian Zheng, Columbia University

23.1 INTRODUCTION

Standard surveys often exclude members of certain groups, known as *hard-to-reach groups*. One reason these individuals are excluded is difficulty accessing group members. Persons who are homeless are very unlikely to be reached by a survey that uses random-digit dialing, for example. Other individuals can be accessed using standard survey techniques, but are excluded because of issues in reporting. Members of these groups are often reluctant to self-identify because of social pressure or stigma (Shelley et al. 1995). Individuals who are homosexual, for example, may not be comfortable revealing their sexual preferences to an unfamiliar survey enumerator. A third group of individuals is difficult to reach because of issues with both access and reporting (commercial sex workers, for example). Even basic demographic information about these groups is typically unknown, especially in developing nations.

One approach to estimating demographic information about hard-to-reach groups is to reach members of these groups through their social network. Some network-based approaches, such as Respondent-Driven Sampling (RDS), recruit respondents *directly* from other respondents' networks (Heckathorn 1997; Heckathorn 2002), making the sampling mechanism similar to a stochastic process on the social network (Goel and Salganik 2009). RDS (see Chapter 24) affords researchers face-to-face contact with members of hard-to-reach groups, facilitating exhaustive interviews and even genetic or medical testing. The price for an entry to these groups is high, however, as RDS uses a specially designed link-tracing framework for sampling. Estimates from RDS are also biased because of the network structure captured during selection, with much statistical work surrounding RDS being intended

to re-weigh observations from RDS to have properties resembling a simple-random-sample. Though methods such as RDS can be advantageous (researchers interview members of hard-to-reach groups directly, for example), financial and logistical challenges often prevent researchers from employing these methods, especially on a large scale.

In this chapter, we focus on methods that utilize social network structure, but collect data about networks and hard-to-reach groups *indirectly* via standard surveys. *Indirectly* in this context means that survey respondents provide information, through carefully crafted network-based questions, about general population and members of hard-to-reach groups. These methods are easily implemented on standard surveys and require no specialized sampling methodology.

We focus specifically on *Aggregated Relational Data* (ARD), or "How many X's do you know," questions (Killworth et al. 1998). In these questions, "X" defines a population of interest (e.g. How many people who are homeless do you know?). A specific definition of "know" defines the network the respondent references when answering the question. In contrast to RDS, ARD do not require reaching members of the hard-to-reach groups directly. Instead, ARD access hard-to-reach groups indirectly through the social networks of respondents on standard surveys. ARD never affords direct access to members of hard-to-reach populations, making the level of detail achievable though RDS impossible with ARD. Unlike RDS, however, ARD require no special sampling techniques and are easily incorporated into standard surveys. ARD are, therefore, feasible for a broader range of researchers across the social sciences, public health, and epidemiology to implement with significantly lower cost than RDS. The work presented in this chapter draws heavily on related work in the statistics literature. Though we present statistical results, the focus of this chapter is on designing surveys that reduce common sources of bias found in estimates using ARD. In the following sections, we provide background on ARD (Section 23.2) and related methods for deriving network features using ARD questions in standard survey (Section 23.3), including discussions of potential sources of bias using ARD

and methods we profile to address these challenges. More specifically, Section 23.3 provides recommendations for selecting populations that reduce bias in estimating degree, or respondent network size. These degree estimates are necessary for estimating hard-to-reach population sizes. Section 23.3.4 moves beyond estimating sizes of hard-to-count populations with these data and provides survey design recommendations for estimating demographic profiles of such groups. We end with a discussion (Section 23.4).

23.2 NETWORK-RELATED QUESTIONS IN STANDARD SURVEYS

In this section, we discuss methods for asking network-related questions using standard surveys. By standard surveys we mean a design where respondents are sampled randomly without replacement from a sampling frame (including various types of stratified or cluster designs). Asking respondents on a survey about their social network servers to increase the sample size of the survey, including both respondents sampled directly and reports about individuals they are connected to through their social network. The data we discuss in this chapter attain this network information indirectly. They are considerably easier to obtain than complete network data and there are currently limited lines of research using this type of data. A dearth of methods for indirect network data remains, however and the few existing methods estimate very specific characteristics of the network and do not address relationships between groups.

23.2.1 Coverage Methods

Several methods to collect social context of survey respondents have been developed, mostly to estimate the respondent's network size, or *degree*. One of the earliest methods was the *reverse small-world* method in (Killworth and Bernard 1978; Killworth, Bernard, and McCarty 1984; Bernard et al. 1990) which, motivated by the small-world experiments of (Milgram 1967), asked respondents to name someone they would use if they were required to pass a message to a given *target*. By asking respondents

about a large number of such targets, it is possible that a respondent will enumerate a large proportion of his acquaintance network. Unfortunately, however, this procedure required a large number (as many as 500) targets and, thus, remained impractical for most surveys. In contrast, the *summation* method (McCarty et al. 2001) requires fewer categories. Respondents are asked how many people they know in a list of specific relationship types, for example, immediate family, neighborhood, coworkers, etc., and these responses are then summed to yield an overall estimate. These relationship types often overlap, however, so degree estimates suffer from double-counting.

23.2.2 Sampling Methods

(Pool and Kochen 1978) developed the *phone book method* where a respondent was provided randomly selected pages from the phone book and based on the proportion of pages which contained the family name of someone known to respondent, it was possible to estimate the respondent's social network size. The estimation was improved greatly in later work by (Freeman and Thompson 1989) and (Killworth et al. 1990) which instead of providing respondents pages of phone books provided them with lists of last names. The general logic of the phone book procedure was then developed further as the *scale-up* procedure (Killworth et al. 1998) using Aggregated Relational Data (ARD). Aggregated relational data questions ask respondents "How many X's do you know[1]," and are easily integrated into standard surveys. Here, *X*, represents a subpopulation of interest.

Aggregated Relational Data (ARD). Among methods to measure network information indirectly, we find the most promise in Aggregated Relational Data (ARD). ARD are most often used to estimate the size of populations that are difficult to count directly. The *scale-up* method, an early method for ARD, uses ARD questions where the subpopulation size is known (people named Michael, for example) to estimate degree in a straightforward manner. Information about the size of some populations is often

[1] The definition of "know" defines the network of interest, though the methods presented here do not depend on the definition of "know."

available through administrative records, such as the Social Security Administration in the United States. Suppose that you know two persons named Nicole, and that at the time of the survey, there were 358,000 Nicoles out of 280 million Americans. Thus your two Nicoles represent a fraction (2/358,000) of all the Nicoles. Extrapolating to the entire country yields an estimate of (2/358,000) × (280million) = 1,560 people known by you. Then, the size of unknown subpopulations is estimated by solving the given equation for the unknown subpopulation size with the estimated degree. Using this method, ARD has been used extensively to estimate the size of populations such as those with HIV/AIDS, injection drug users, or the homeless (for example (Killworth et al. 1990; Killworth et al. 1998)).

Unlike the previously described methods, ARD allows researchers to choose specific subpopulations of interest without sampling or surveying members of these subpopulations directly. This feature holds potential to learn additional information about these subpopulations and their relationship to the overall network. (Shelley et al. 2006), for example, uses ARD to explore how the structure of the network of seropositive individuals impacts the dissemination of information about their disease status.

Despite the potential value of ARD and the ease of obtaining these data through standard surveys, the literature on learning about network structure from ARD remains underdeveloped. The scale-up method, for example, is easy to implement but does not account for network structure. Consider, for example, asking a respondent how many people named "Rose" she/he knows. If each person was equally likely to know Rose's[2], then this would be equivalent to asking if they know each person on a list of the one-half million Rose's in the U.S. If we were to take all one-half million of these Roses and put their names on a list, then each respondent would have the same chance of knowing each of these one-half million individuals if knowing someone named Rose were entirely random. That is, each respondent on each Rose is a Bernoulli trial with a fixed success probability proportional to the size of this respondent's

[2] This also assumes that one could recall their acquaintanceships with complete accuracy. This assumption is often not valid and we will discuss the issue in further detail in subsequent sections.

network size. Network structure makes these types of independence assumptions invalid. For example, since Rose is most common amongst older females and people are more likely to know individuals of similar age and the same gender, older female respondents are more likely to know a given Rose than older male respondents. Statistical models are needed to understand how these responses change based on homophily, as in this example, and on more complicated network properties. Ignoring social network structure induces bias in the individuals' responses. Since estimates of hard-to-count populations are then constructed using responses to aggregated relational data questions, the resulting estimates are also biased (Killworth et al. 1998; Bernard et al. 1991).

In addition to the applications of the scale-up method using ARD described in the previous section, two substantial steps in modeling ARD will influence our proposed method. Zheng, Salganik, and Gelman (2006) began by noting that under simple random mixing the responses to the "How many X's do you know?" questions would follow a Poisson distribution with rate parameter determined by the degree of the respondent and the network prevalence of the subpopulation. Here the network prevalence is the proportion of ties that involve individuals in subpopulation and should match the proportion of the population comprised of members of the given subpopulation. Under this assumption for example, the expected number of Rose's known by a respondent with degree equal to 500 would be $500 \times (500{,}000/280 \text{ million}) \approx 1$. They apply their method to data from (McCarty et al. 2001) and find that many of the questions in the data did not follow a Poisson distribution. In fact, most of the responses show overdispersion, or greater-than-expected variance. We can interpret the overdispersion as a factor that decreases the frequency of people who know exactly one person of type X, as compared to the frequency of people who know none. As overdispersion increases from its null value of 1, it is less likely for a person to have an isolated acquaintance from that group. For example, consider the responses to the question: "How many males do you know incarcerated in state or federal prison?" The mean of the responses to this question was 1.0, but the variance was 8.0, indicating that some people are much more likely to know more than one individual in prison than others. To model this increased variance (Zheng,

Salganik, and Gelman 2006) allowed individuals to vary in their propensity to form ties to different groups. In a multilevel model, this corresponds to assuming that these propensities follow a gamma distribution with a shape parameter determined by the overdispersion. The responses then can be modeled as a negative binomial distribution so that the expected number of alters known by a respondent in a given subpopulation is the degree of the respondent times the network prevalence, as under the simple model, but now scaled by the overdispersion parameter to estimate the variation in individual propensities to form ties to people in different groups.

23.3 STATISTICAL METHODS FOR ARD

In this section, we discuss methods for extracting network properties using ARD collected using survey questionnaires. The discussion is organized according to the network features under study.

23.3.1 Estimating Personal Network Size.

The scale-up estimate. Consider a population of size *N*. We can store the information about the social network connecting the population in an adjacency matrix $\Delta = [\delta_{ij}]_{N\times N}$ such that $\delta_{ij} = 1$ if person i knows person j. Though the methods discussed here do not depend on the definition of know, throughout this chapter we will assume the (McCarty et al. 2001) definition of know: "that you know them and they know you by sight or by name, that you could contact them, that they live within the United States, and that there has been some contact (either in person, by telephone or mail) in the past 2 years." The personal network size or degree of person i is then $d_i = \sum_j \delta_{ij}$.

One straightforward way to estimate the degree of person i would be to ask if she knows each of n randomly chosen members of the population. Inference could then be based on the fact that the responses would follow a binomial distribution with n trials and probability d_i/N. In a large population, however,

this method is extremely inefficient because the probability of a relationship between any two people is very low. For example, if one assumes an average personal network size of 750 (as estimated by(Zheng, Salganik, and Gelman 2006)), then the probability of two randomly chosen Americans knowing each other is only about 0.0000025 meaning that a respondent would need to be asked about millions of people to produce a decent estimate.

A more efficient method would be to ask the respondent about an entire set of people at once through ARD type survey questions. For example, asking, "How many women do you know who gave birth in the last 12 months?" instead of asking the respondent if she knows 3.6 million distinct people. The *scale-up* method uses responses to ARD questions of this form ("How many X's do you know?") to estimate personal network size. For example, if you report knowing 3 women who gave birth, this represents about one-millionth of all women who gave birth within the last year. We could then use this information to estimate that you know about one-millionth of all Americans,

$$\frac{3}{3.6 \text{ million}} (300 \text{ million}) \approx 250 \text{ people.}$$

The precision of this estimate can be increased by averaging responses of many groups yielding the scale-up estimator (Killworth et al. 1998) of the degree of person *i*

$$\widehat{d_i} = \frac{\sum_{k=1}^{K} y_{ik}}{\sum_{k=1}^{K} N_k} \cdot N$$

where y_{ik} is the number of people that person *i* knows in subpopulation *k*, N_k is the size of subpopulation *k*, and *N* is the size of the population. One important complication to note with this estimator is that asking "How many women do you know who gave birth in the last 12 months?" is not equivalent to asking about 3.6 million *random* people; rather the people asked about are women, probably between the ages of 18 and 45. This creates statistical challenges that are addressed in detail in subsequent sections. To estimate the standard error of the simple estimate, we follow the practice of (Killworth et al. 1998) by assuming

$$\sum_{k=1}^{K} y_{ik} \sim \text{Binomial}\left(\sum_{k=1}^{K} N_k, p = \frac{d_i}{N}\right).$$

The estimate of the probability of success, $p = \frac{d_i}{N}$, is

$$\hat{p} = \frac{\sum_{k=1}^{K} y_{ik}}{\sum_{k=1}^{K} N_k} = \frac{\widehat{d_i}}{N} \tag{1}$$

with standard error (including finite population correction) (Lohr 1999)

$$\text{SE}(\hat{p}) = \sqrt{\frac{1}{\sum_{k=1}^{K} N_k} \hat{p}(1-\hat{p}) \frac{N - \sum_{k=1}^{K} N_k}{N-1}}.$$

The scale-up estimate $\widehat{d_i}$ then has standard error

$$\text{SE}\left(\widehat{d_i}\right) = N \cdot \text{SE}(\hat{p}) = N \sqrt{\frac{1}{\sum_{k=1}^{K} N_k} \hat{p}(1-\hat{p}) \frac{N - \sum_{k=1}^{K} N_k}{N-1}} \approx \sqrt{\widehat{d_i}} \sqrt{\frac{1 - \frac{\sum_{k=1}^{K} N_k}{N}}{\frac{\sum_{k=1}^{K} N_k}{N}}}.$$

For example, if we asked respondents about the number of women they know who gave birth in the past year the approximate standard error of the degree estimate is calculated as

$$\text{SE}\left(\widehat{d_i}\right) \approx \sqrt{\widehat{d_i}} \sqrt{\frac{1 - \frac{\sum_{k=1}^{K} N_k}{N}}{\frac{\sum_{k=1}^{K} N_k}{N}}} \approx \sqrt{750} \cdot \sqrt{\frac{1 - \frac{3.6 \text{ million}}{300 \text{ million}}}{\frac{3.6 \text{ million}}{300 \text{ million}}}} \approx 250.$$

assuming a degree of 750 as estimated by (Zheng, Salganik, and Gelman 2006). If in addition, we also asked respondents the number of people they know who have a twin sibling, the number of people they know who are diabetics, and the number of people they know who are named Michael, we would have increased our aggregate subpopulation size, $\sum_{k=1}^{K} N_k$, from 3.6 million to approximately 18.6 million and in doing so decreased our estimated standard error to about 100. In Figure 23.1 (McCormick, Salganik, and Zheng 2010), we plot $\text{SE}\left(\widehat{d_i}\right)/\sqrt{\widehat{d_i}}$ against $\sum_{k=1}^{K} N_k / N$. The most drastic reduction in estimated error comes in increasing the survey fractional subpopulation size to about 20 percent (or approximately 60 million in a population of 300 million). Though the above standard error depends only on sum of the

subpopulation sizes, we will show that there are other sources of bias that make the choice of the individual subpopulations important as well.

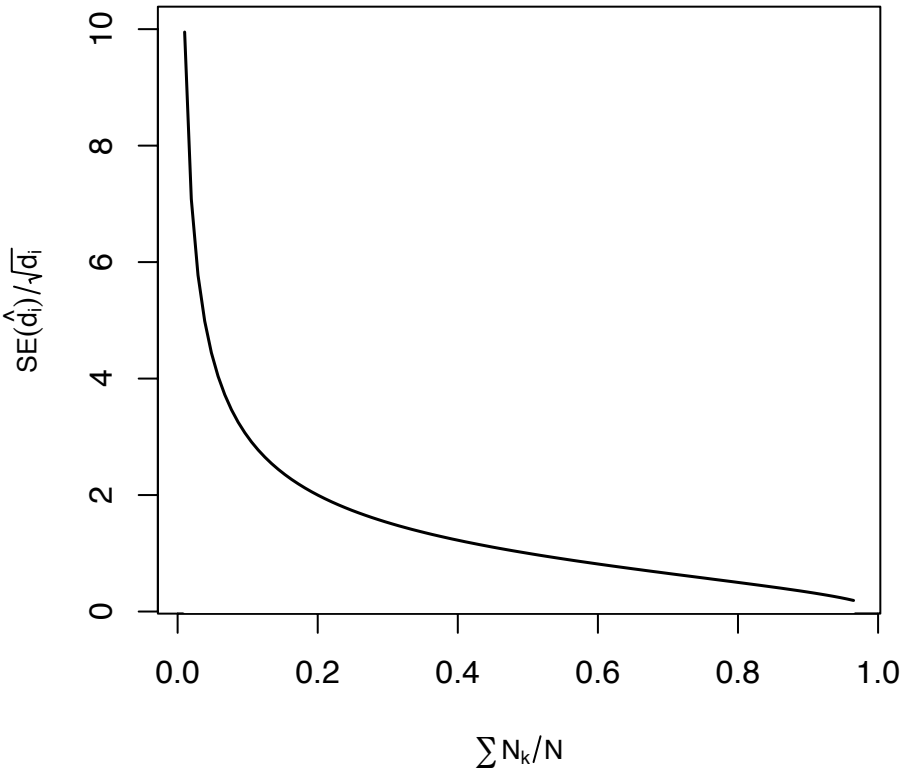

Figure 23.1 Standard error of the scale-up degree estimate (scaled by the square root of the true degree) plotted against the sum of the fractional subpopulation sizes. As we increase the fraction of population represented by survey subpopulations, the precision of the estimate improves, with diminishing improvements after about 20 percent.

Issues with the scale-up estimator. The scale-up estimator using "How many X do you know?" data, is known to suffer from three distinct problems -- transmission errors, barrier effects, and recall problems (Killworth et al. 2003; Killworth et al. 2006) -- when the ARD questions are chosen arbitrarily. *Transmission errors* occur when the respondent knows someone in a specific subpopulation, but is not aware that they are actually in that subpopulation. For example, a respondent might know a woman who recently gave birth, but might not know that she had recently given birth. These transmission errors likely vary from subpopulation to subpopulation depending on the sensitivity and visibility of the

information. These errors are extremely difficult to quantify because very little is known about how much information respondents have about the people they know (Laumann 1969; Killworth et al. 2006; Shelley et al. 2006). *Barrier effects* occur whenever some individuals systematically know more (or fewer) members of a specific subpopulation than would be expected under random mixing, and thus can also be called non-random mixing. For example, since people tend to know others of similar age and gender (McPherson, Smith-Lovin, and Cook 2001), a 30-year old woman probably knows more women who have recently given birth than would be predicted just based on her personal network size and the number of women who have recently given birth. Similarly, an 80-year old man probably knows fewer than would be expected under random mixing. Therefore, estimating personal network size by asking only "How many women do you know who have recently given birth?"—the estimator presented above in (1)--will tend to overestimate the degree of women in their 30's and underestimate the degree of men in their 80's. Because these barrier effects can introduce a bias of unknown size, previous researchers have avoided using the scale-up method to estimate the degree of any particular individual. A final source of error is that responses to these questions are prone to recall error. For example, people seem to under-recall the number of people they know in large subpopulations (e.g., people named Michael) and over-recall the number in small subpopulations (e.g., people who committed suicide) (Killworth et al. 2003; Zheng, Salganik, and Gelman 2006). If people were answering such questions consistently we would expect a linear relationship between the size of the subpopulation and the mean number of individuals recalled. That is, if the size of subgroup doubled, the mean number recalled should also double. This is not the case as can be seen in Figure 23.2 (McCormick, Salganik, and Zheng 2010), which plots the mean number known in each subpopulation as a function of subpopulation size for the 12 names in the (McCarty et al. 2001) data. The figure shows that there was over-recall of small subpopulations and under-recall of large subpopulations, a pattern that has been noted previously (Killworth et al. 2003; Zheng, Salganik, and Gelman 2006).

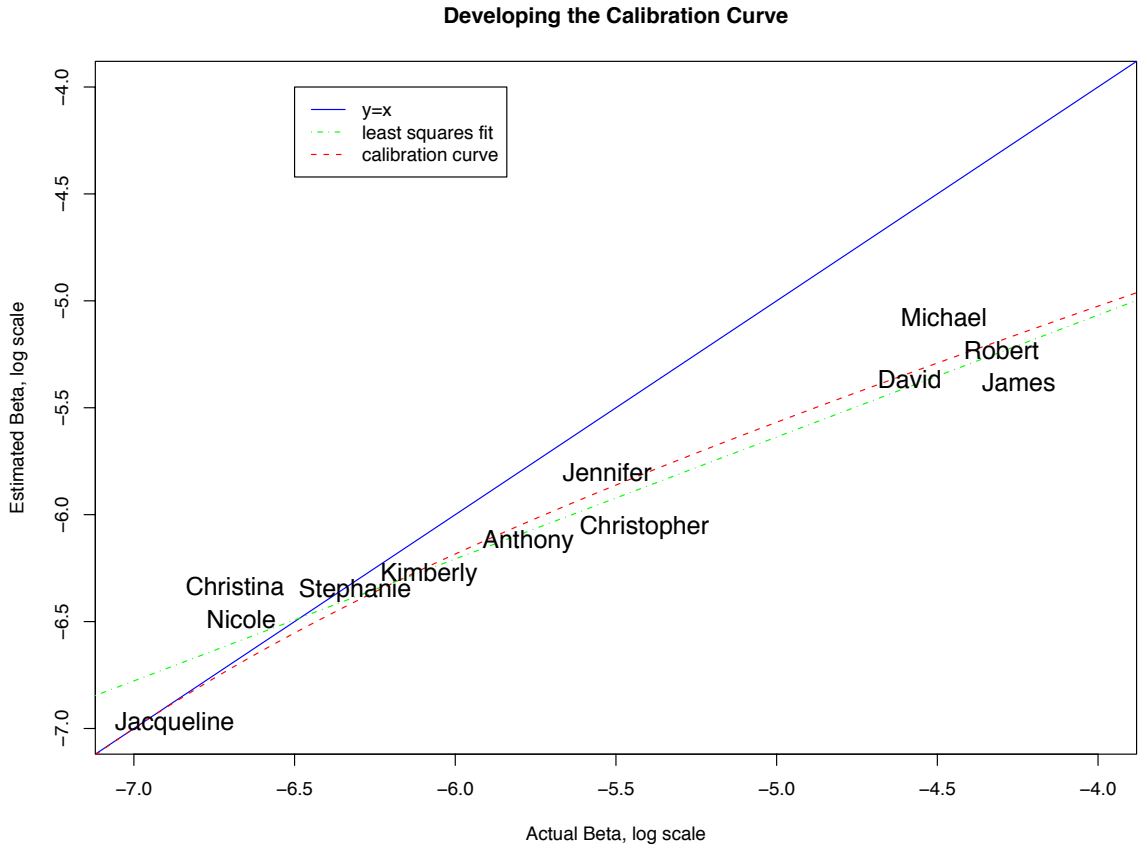

Figure 23.2 Mean number recalled as a function of subpopulation size for 12 names. If respondents recall perfectly, then we would expect the mean number recalled to increase linearly as the subpopulation size increases. The best-fit line and loess curve show that this was not the case suggesting that there is recall error.

Reducing bias in degree estimates. In this section, we review design recommendations for reducing bias in degree estimates described in Section 23.3.1. The intuition behind the recommendations we describe is that the names asked about should be chosen so that the combined set of people asked about should be easy to recall, with perfect transmission of traits being asked (first names), *and* is a "scaled-

down" version of the overall population. For example, if 20 percent of the general population is females under 30 then 20 percent of the people with the names used must also be females under 30.

In ARD, respondents are conceptualized as *egos*, or senders of ties in the network. As discussed in (McCormick, Salganik, and Zheng 2010), we divide the egos into groups based on their demographic characteristics (males 20-40 years old, for example). The individuals who comprise the counts for ARD are the *alters*, or recipients of links in the network. The alters are also divided into groups, though the groups need not be the same for both the ego and the alter groups. The *scale-down* condition was motivated by the *latent non-random mixing* model in (McCormick, Salganik, and Zheng 2010) that assumes an expected number of acquaintances for an individual i in ego group e to people in group k,

$$\mu_{ike} = E\left(y_{ike}\right) = d_i \sum_{a=1}^{A} m(e,a) \frac{N_{ak}}{N_a}.$$

Here, $m(e,a)$ is the *mixing matrix* as in (McCormick, Salganik, and Zheng 2010). The mixing matrix accounts for the propensity for individiuals to know more respondents in some demographic groups than others (a young female respondent will likley know more young females than older males, for example). On the other hand, the scale-up estimator assumes

$$E\left(\sum_{k=1}^{K} y_{ike}\right) = \sum_{k=1}^{K} \mu_{ike} = d_i \sum_{a=1}^{A} m(e,a) \left[\sum_{k=1}^{K} \frac{N_{ak}}{N_a}\right]$$

$$\equiv d_i \frac{\sum_{k=1}^{K} \sum_{a=1}^{A} N_{ak}}{N}, \forall e. \quad (2)$$

(2) shows that the (Killworth et al. 1998) scale-up estimator is in expectation equivalent to that of the latent non-random mixing if either

$$m(e,a) = \frac{N_a}{N}, \forall a, \forall e, \quad (3)$$

or

$$\frac{\sum_{k=1}^{K} N_{ak}}{\sum_{k=1}^{K} N_k} = \frac{N_a}{N}, \forall a. \quad (4)$$

In other words, the two estimators are equivalent if there is random mixing (3) or if the combined set of names represents a "scaled-down" version of the population (4). Since random mixing is not a reasonable

assumption for the acquaintances network in the United States, we need to focus on selecting the names to satisfy the *scaled-down* condition. That is, we should select the set of names such that, if 15 percent of the population is males between ages 21 and 40 $\left(\frac{N_a}{N}\right)$ then 15 percent of the people asked about must also be males between ages 21 and 40 $\left(\frac{\sum_{k=1}^{K} N_{ak}}{\sum_{k=1}^{K} N_k}\right)$. In actually choosing a set of names to satisfy the scaled-down condition, we found it more convenient to work with a rearranged form:

$$\frac{\sum_{k=1}^{K} N_{ak}}{N_a} = \frac{\sum_{k=1}^{K} N_k}{N}, \forall a. \tag{5}$$

In order to find a set of names that satisfy (5) it is helpful to create Figure 23.3 (McCormick, Salganik, and Zheng 2010) that displays the relative popularity of many names over time. From this figure, we tried to select a set of names such that the popularity across alter categories ended up balanced. For example, consider the names Walter, Bruce and Kyle. These names have similar popularity overall, but Walter was popular from 1910-1940, whereas Bruce was popular during the middle of the century and Kyle near the end. Thus, the popularity of the names at any one time period will be balanced by the popularity of names in the other time periods, preserving the required equality in the sum (5).

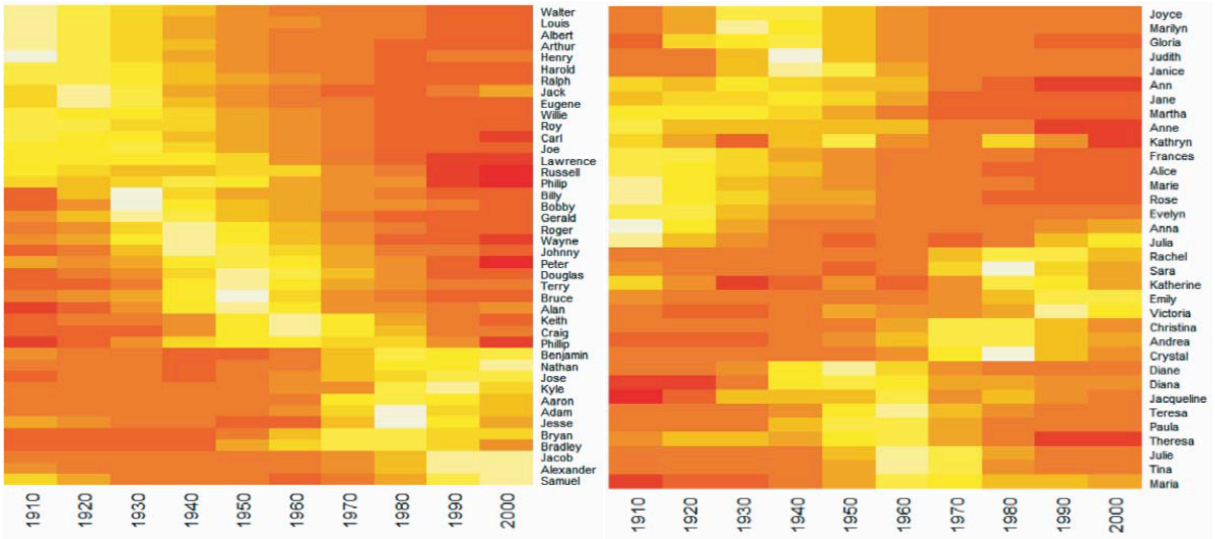

Figure 23.3 Heat maps of additional male and female names based on data from the Social Security Administration. Lighter color indicates higher popularity.

When choosing what names to use, in addition to satisfying (5), we recommend choosing names that compromise 0.1 to 0.2 percent of the population, as these minimize recall errors and yield average responses from 0.6-1.3. Finally, we recommend choosing names that are not commonly associated with nicknames in order to minimize transmission errors.

Selecting the number of names. For researchers planning to use the scale-up method, an important issue to consider in addition to which names to use is how many names to use. Obviously, asking about more names will produce a more precise estimate, but that precision comes at the cost of increasing the length of the survey. To help researchers understand the trade-off, we return to the approximate standard error under the binomial model presented in Section 23.3. Simulation results using 6, 12, and 18 names chosen using the guidelines suggested above agree well with the results from the binomial model in Section 23.3 (results not shown). This agreement suggests that the simple standard error may be reasonable when the names are chosen appropriately. To put the results of (1) into a more concrete context, a researcher who uses names whose overall popularity reaches 2 million would expect a standard error of around $11.6\times\sqrt{500} = 259$ for an estimated degree of 500 whereas with $\sum N_k$=6 million, she would expect a standard error of $6.2\times\sqrt{500} = 139$ for the same respondent. Finally, for the good names, $\sum N_k = 4$ million, so a researcher could expect a standard error of 177 for a respondent with degree 500.

23.3.2 Estimating Non-random Mixing.

In this section, we introduce a missing data perspective for ARD and propose an estimator based on the EM algorithm.

If for a given respondent, i, we could take all the members of the social network with which i has a link and place them in a room, we would compute the mixing rate between the ego and a given alter group, $a = (1, \ldots, A)$, by dividing the room in A mutually exclusive sections and asking alters to stand in their respective group. The estimated mixing rate would then be the number of people standing in a given group divided by the number of people in the room. We could also perform a similar calculation by placing a simple random sample of size n from a population of size N in a room. Then, after dividing the alters into mutually exclusive groups, we could count y_{ia}, or the number of alters respondent i knows in the sample who are in each of the a alter groups. Since we have a simple random sample we can extrapolate back to the population and estimate the degree of the respondent, $\hat{d}_i$, and within alter group degree, $\hat{d}_{ia}$, as

$$\hat{d}_i = \sum_{a=1}^{A} y_{ia}/(n/N) \qquad \text{and} \qquad \hat{d}_{ia} = y_{ia}/(n_a/N_a).$$

Given these two quantities we can estimate the mixing rate between the respondent and an alter group by taking the ratio of alters known in the sample who are in alter group a over the total number known in the sample. This computation is valid because we assumed a simple random sample and, thus, that (in expectation) the demographic distribution of alters in our sample matches that of the population. In ARD, the distribution of the hypothetical alters we sample depends on the subpopulations we select. If we only ask respondents subpopulations which consist of young males, for example, then our hypothetical room from the previous example would contain only the respondent's young, male alters. Estimating the rate of mixing between the respondent and older females would not be possible in this situation. Viewed in this light, ARD is a form of cluster sampling where the subpopulations are the clusters and respondents report the presence/absence of a tie between all alters in the cluster. Since the clusters are no longer representative of the population, our estimates need to be adjusted for the

demographic profiles of the clusters (Lohr 1999). Specifically, if we observe y_{ika} for subpopulations $k = (1, \ldots, K)$ and alter groups $a = (1, \ldots, A)$, then our estimates of $\hat{d}_i$ and $\hat{d}_{ia}$ become

$$\hat{d}_i = \sum_{k=1}^{K} y_{ik} / \left(\sum_{k=1}^{K} N_k / N \right) \qquad and \qquad \hat{d}_{ia} = \sum_{k=1}^{K} y_{ika} / \left(\sum_{k=1}^{K} N_{ak} / Na \right)$$

where N_k is the size of subpopulation k and N_{ak} is the number of members of subpopulation k in alter group a. To estimate the mixing rate, we could again divide the estimated number known in alter group a by the total estimated number known. Under the *scaled-down* condition the denominators in the above expressions cancel and the mixing estimate is the number known in the subpopulations that are in alter group a over the total number known in all K subpopulations. In the examples above, we have assumed the alters are observed so that y_{ika} can be computed easily. This is not the case in ARD, however, since we observe only the aggregate number of ties and not the specific demographic make-up of the recipients. Thus, ARD represent a type of cluster sampling design where the specific ties between the respondent and members of the alter group are *missing*. If we ignore the residual variation in propensity to form ties with group k individuals due to noise, we may assume that the number of members of subpopulation k in alter group a the respondent knows, y_{ika}, follows a Poisson distribution. Under this assumption, we can estimate m_{ia} by imputing y_{ika} as part of an EM algorithm EM. Specifically, for each individual define $y_{ik}^{(com)} = (y_{ika}, \ldots, y_{i1A})^T$ as the complete data vector for each alter group. The complete data log-likelihood for individual i's vector of mixing rates, $m_i = (m_{i1}, \ldots, m_{iA})^T$, is $\ell(m_i; y_{i1}^{(com)}, \ldots, y_{iK}^{(com)})$, which has the form

$$\ell(m_i; y_{i1}^{(com)}, \ldots, y_{iK}^{(com)}) = \sum_{k=1}^{K} \sum_{a=1}^{A} \log\left(Poisson\left(y_{ika}; \lambda_{ika} = d_i m_{ia} \frac{N_{ak}}{N_a}\right)\right). \qquad (6)$$

Using (6) we derive the following two updating steps for the EM:

$$y_{iak}^{(t)} = y_{ik} \left(\frac{m_{ia}^{(t-1)} \frac{N_{ak}}{N_a}}{\sum_{a=1}^{A} m_{ia}^{(t-1)} \frac{N_{ak}}{N_a}} \right)$$

$$m_{ia}^{(t)} = \frac{\sum_{k=1}^{K} y_{ika}^{(t-1)}}{\sum_{k=1}^{K} y_{ik}}.$$

If one sets $m_{ia}^{(0)} = N_a/N$, which corresponds to random mixing in the population, and runs one EM update, this would result in the following *simple ratio estimator* of the mixing rate for individual i:

$$\widehat{m}_{ia} = \frac{\sum_{k-1}^{K} y_{ik}(N_{ak}/N_k)}{\sum_{k-1}^{K} y_{ik}} \quad (7)$$

In our simulation studies (details not shown), this simple estimator produces estimates very close to the converged EM estimates. Additionally, it is easy to show that the simple ratio estimate, $\widehat{m}_{ia}$, is unbiased if $N_{ak}/N_a \neq 0$ for only one alter group a and that for any a there exists a subpopulation, k, such that $N_{ak} = N_a$. We refer to this condition as *complete separability*. Therefore, (7) constitutes a simple estimate for individual mixing rate and can be used to estimate average mixing behaviors of any ego group.

23.3.3 Estimating Demographic Profiles of Hard-T0-Reach Groups Using ARD

In this section, we describe a model presented by (McCormick and Zheng 2012) for estimating latent demographic profiles for hard-to-reach groups. This method will provide information about the demographic make-up of groups which are often difficult to access using standard surveys, such as the proportion of young males who are infected with HIV. Under the set-up of ego groups and alter groups discussed in Section 23.3.1, members of hard-to-reach groups are one type of alter. Thus, the alter groups defined determine the demographic characteristics that can be estimated for the hard-to-reach. The (McCormick and Zheng 2012) method combines estimation and survey-design strategy, making it well-suited for researchers who intend to collect ARD. First one needs to use ARD questions statisfying the *scaled-down* condition in (McCormick, Salganik, and Zheng 2010) for selecting subpopulations to reduce bias in the estimates of respondent degree discussed in Section 23.3.1 and derive the mixing matrix estimates as in Section 23.3.3.

The estimates for respondent degree (Section 23.3.1) and mixing estimates (Section 23.3.3) rely on latent profile information from some "known" populations. Using these estimates, we now further develop a regression-based estimator for unobserved latent profiles. Define $h(\mathrm{a}, k)$ as the fraction of alter group *a* made up of members of group *k*. For each respondent and each unknown subpopulation we now have

$$y_{ik} = \sum_{a=1}^{A} \hat{d}_i \hat{m}_{ia} h(a, k). \tag{7}$$

If we denote the matrix $X_k = \hat{d}\hat{m}_{\cdot k}$ and the vector $h(\cdot, k) = \overleftarrow{\beta_k}$, then (7) can be regarded as a linear regression equation, $\overleftarrow{y_k} = X_k \overleftarrow{\beta_k}$, with the constraint that coefficients, $\overleftarrow{\beta_k}$, are restricted to be non-negative. Lawson and Hanson (1974) propose an algorithm for computing these coefficients. Since the $\hat{m}_{\cdot k}$ sum to one across alter groups, the columns of X_k are collinear. This could produce instability in solving the quadratic programming problem associated with finding our estimated latent profiles. In practice, we have found our estimates perform well despite this feature.

Simulation experiments. Here we present simulation experiments to evaluate our regression-based estimates under four strategies for selecting observed profiles. First, we created profiles which are *completely separable* (defined in Section 23.3.3). Second, we constructed profiles for the names satisfying the *scaled-down* condition presented in Section 23.3.1 using data from the Social Security Administration. These names provide insights into the potential accuracy of our method using actual profiles. As a third case, we include the names from (McCormick, Salganik, and Zheng 2010) which violate the scaled-downed down condition and are almost exclusively popular among older respondents. For the fourth set of names, recall from Section 23.3.3 that the mixing matrix estimates are identifiable only if the matrix of known profiles, $\mathbf{H}_{\mathbf{A}\times\mathbf{K}}$, has rank A. To demonstrate a violation of this condition we selected a set of names with uniform popularity across the demographic groups, or nearly perfect collinearity. There is some correlation in the scaled-down names since several names have similar profiles. The degree of correlation is substantially less than in the flat profiles, however. In each

simulation, (McCormick and Zheng 2012) generated 500 respondents using the Latent Non-random Mixing Model (see McCormick, Salganik, and Zheng 2010) with each of the four profile strategies. Mixing matrix estimates were calculated using the simple estimate derived from the first step of the EM algorithm in Section 23.3.3. We compare our mixing matrix estimates to the estimated mixing matrix from (McCormick, Salganik, and Zheng 2010), which we use to generate the simulated data. We evaluate the latent profiles using six names with profiles known from the Social Security Administration. We repeated the entire process 1,000 times.

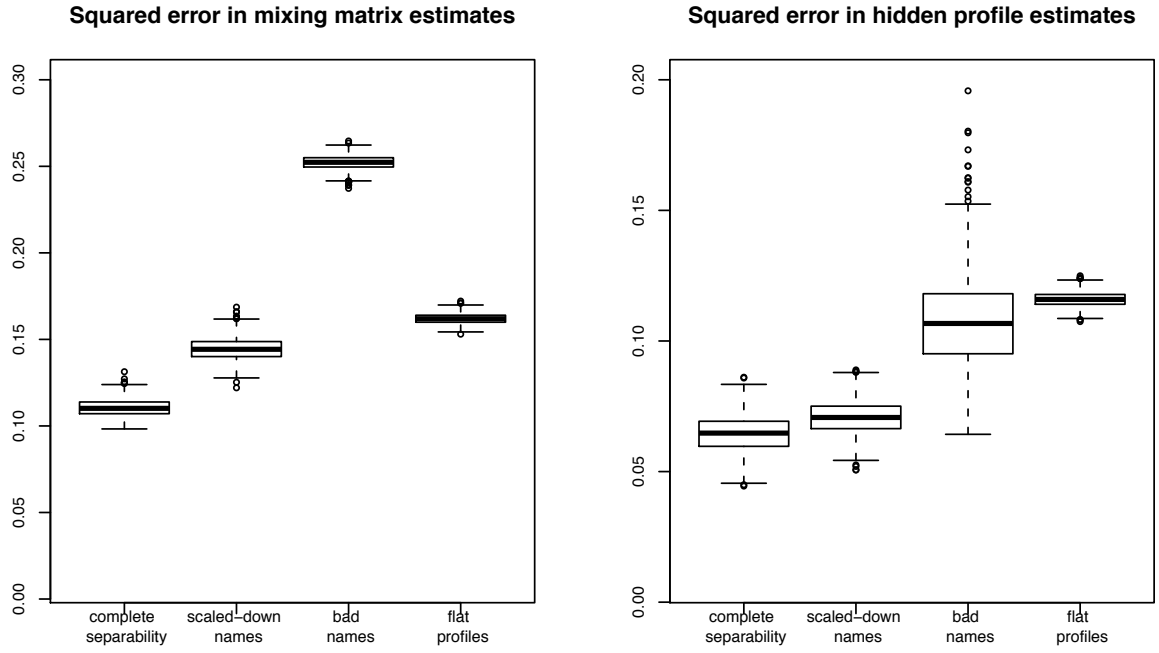

Figure 23.4 Total mean squared error across all elements of the mixing matrix and latent profile matrix. The vertical axis is the sum of the errors across all eight alter groups. We generated 500 respondents using the four profile structures then evaluated our ability to recover the mixing matrix estimated in (McCormick, Salganik, and Zheng 2010), and the known profiles of six additional names. We repeated the simulation 1,000 times. In both cases the ideal profile has the lowest error, followed by the scaled-down names suggested by (McCormick, Salganik, and Zheng 2010).

Figure 23.4 (McCormick and Zheng 2012) presents boxplots of the squared error in mixing matrix and latent profile estimates. In both cases, the ideal, completely separable, profiles have the lowest error. The scaled-down names also perform well, indicating that reasonable estimates are possible even when complete separability is not. The flat profiles perform only slightly worse than the scaled-down names for estimating mixing but significantly worse when estimating latent profiles. The names which violate the scaled-down condition produce poor estimates of both quantities.

23.4 Conclusion and Discussion

In this chapter we present methods for estimating features of hard-to-reach populations using indirectly observed network data. ARD are easy and cheap to collect using standard survey mechansisms. This means that the information needed to estimate sizes of some hard-to-reach populations can be collected using existing surveys designed for other topics. We have focused particularly on survey designs which lead to reliable, but simple, estimates. We belive that the design conditions we propose are critical to the performance of these simple estimators. In cases where data have already been collected, or when it is not possible to develop survey questions in accordance with these guidelines, we suggest using model-based strategies proposed in McCormick et al. (2010) and McCormick and Zheng (2012). We also note that there are many open areas of research in this challenging problem, with contributions to be made both in improving estimation methods as well as verifying and calibrating currently proposed techniques.

REFERENCES

Bernard, H. R., Johnsen, E. C., Killworth, P. D., & Robinson, S. (1991). "Estimating the Size of an Average Personal Network and of an Event Subpopulation: Some Empirical Results." *Social Science Research,* 20, 109–121.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., McCarty, C., Shelley, G. A. & Robinson, S. (1990). "Comparing Four Different Methods for Measuring Personal Social Networks." *Social Networks,*

12, 179–215.

Freeman, L.C., & Thompson, C. R. (1989). “Estimating Acquaintanceship Volume.” In M. Kochen (Ed.), *The Small World*, (pp. 147–158). Ablex Publishing.

Goel, S., & Salganik, M. (2009). “Respondent-Driven Sampling as Markov Chain Monte Carlo.” *Statistics in Medicine,* 28 (17), 2202–2229.

Heckathorn, D. (1997). “Respondent-Driven Sampling: a New Approach to the Study of Hidden Populations.” *Social Problems,* 44 (2), 174–199.

Heckathorn, D. (2002). “Respondent-Driven Sampling II: Deriving Valid Population Estimates From Chain-Referral Samples of Hidden Populations.” *Social Problems,* 49 (1), 11–34.

Killworth, P. D., & Bernard, H. R. (1978). “The Reverse Small-World Experiment.” *Social Networks,* 1 (2), 159–192.

Killworth, P. D., McCarty, C., Johnsen, E. C., Bernard, H. R., & Shelley, G. A. (2006). “Investigating the Variation of Personal Network Size Under Unknown Error Conditions.” *Sociological Methods \& Research,* 35 (1), 84–112.

Killworth, P. D., McCarty, C., Bernard, H. R., Johnsen, E. C., Domini, J., & Shelley, G. A. (2003). “Two Interpretations of Reports of Knowledge of Subpopulation Sizes.” *Social Networks,* 25, 141–160.

Killworth, P D, McCarty, C., Bernard, H. R., Shelley, G. A., & Johnsen, E. C. (1998). “Estimation of Seroprevalence, Rape, and Homelessness in the U.S. Using a Social Network Approach.” *Evaluation Review,* 22, 289–308.

Killworth, P. D., Johnsen, E. C., Bernard, H. R., Shelley, G. A., & McCarty, C. (1990). “Estimating the Size of Personal Networks.” *Social Networks,* 12, 289–312.

Lawson, C. L. & Hanson, R. J. (1974). *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs, NJ.

Killworth, P. D., Bernard, H. R., & McCarty, C. (1984). “Measuring Patterns of Acquaintanceship.” *Current Anthropology,* 23, 318–397.

Laumann, E. O. (1969). “Friends of Urban Men: an Assessment of Accuracy in Reporting Their

Socioeconomic Attributes, Mutual Choice, and Attitude Agreement." *Sociometry,* 32 (1), 54–69.

Lohr, S.L. 1999. *Sampling: Design and Analysis*. Duxbury Press.

McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C., & Shelley, G. A. (2001). "Comparing Two Methods for Estimating Network Size." *Human Organization,* 60, 28–39.

McCormick, T. H., & Zheng, T. (2012). "Latent Demographic Profile Estimation in Hard-to-Reach Groups." *Annals of Applied Statistics*, 6, 1795-1813.

McCormick, T. H., Salganik, M. J., & Zheng, T. (2010). "How Many People Do You Know?: Efficiently Estimating Personal Network Size." *Journal of the American Statistical Association,* 105, 59–70.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology,* 27, 415–444.

Milgram, S. (1967). "The Small World Problem." *Psychology Today,* 1, 62–67.

Pool, I., & Kochen, M. (1978). "Contacts and Influence." *Social Networks,* 1, 5–51.

Shelley, G., Bernard, H. R., Killworth, P. D., Johnsen, E., & McCarty, C. (1995). "Who Knows You HIV Status? What HIV+ Patients and Their Network Members Know About Each Other." *Social Networks,* 17, 189–217.

Shelley, G. E., Killworth, P. D., Bernard, H. R., McCarty, C., Johnsen, E. C., & Rice, R. E. (2006). "Who Knows Your HIV Status II?: Information Propogation Within Social Networks of Seropositive People." *Human Organization,* 65 (4), 430–444.

Zheng, T., Salganik, M. J., & Gelman, A. (2006). "How Many People Do You Know in Prison?: Using Overdispersion in Count Data to Estimate Social Structure." *Journal of the American Statistical Association,* 101, 409–423.