How Many People Do You Know?: Efficiently Estimating Personal Network Size

Tyler H. McCormick, Matthew J. Salganik, and Tian Zheng

In this article we develop a method to estimate both individual social network size (i.e., degree) and the distribution of network sizes in a population by asking respondents how many people they know in specific subpopulations (e.g., people named Michael). Building on the scale-up method of Killworth et al. (1998b) and other previous attempts to estimate individual network size, we propose a latent non-random mixing model which resolves three known problems with previous approaches. As a byproduct, our method also provides estimates of the rate of social mixing between population groups. We demonstrate the model using a sample of 1,370 adults originally collected by McCarty et al. (2001). Based on insights developed during the statistical modeling, we conclude by offering practical guidelines for the design of future surveys to estimate social network size. Most importantly, we show that if the first names asked about are chosen properly, the estimates from the simple scale-up model enjoy the same bias-reduction as the estimates from our more complex latent nonrandom mixing model.

KEY WORDS: Latent nonrandom mixing model; Negative binomial distribution; Personal network size; Social networks; Survey design.

1. INTRODUCTION

Social networks have become an increasingly common framework for understanding and explaining social phenomena. But despite an abundance of sophisticated models, social network research has yet to realize its full potential, in part because of the difficulty of collecting social network data. In this article we add to the toolkit of researchers interested in network phenomena by developing methodology to address two fundamental challenges posed in the seminal work of Pool and Kochen (1978). First, for an individual, we would like to know how many other people she knows (i.e., her degree, d_i); and second, for a population, we would like to know the distribution of acquaintance volume (i.e., the degree distribution, p_d).

Recently, the second question, of degree distribution, has received the most attention because of interest in so-called "scale-free" networks (Barabási 2003). This interest was sparked by the empirical finding that some networks, particularly technological networks, appear to have power law degree distributions [i.e., $p(d) \sim d^{-\alpha}$ for some constant α], as well as by mathematical and computational studies demonstrating that this extremely skewed degree distribution may affect the dynamics of processes occurring on the network, such as the spread of diseases and the evolution of group behavior (Pastor-Satorras and Vespignani 2001; Santos, Pacheco, and Lenaerts 2006). The degree distribution of the acquaintanceship network is not known, however, and this has become so central to some researchers that Killworth et al. (2006) declared that estimating the degree distribution is "one of the grails of social network theory."

Although estimating the degree distribution is certainly important, we suspect that the ability to quickly estimate the personal network size of an individual may be of greater importance to social science. Currently, the dominant framework for empirical social science is the sample survey, which has been astutely described by Barton (1968) as a "meat grinder" that completely removes people from their social contexts. Having a survey instrument that allows for the collection of social content would allow researchers to address a wide range of questions. For example, to understand differences in status attainment between siblings, Conley (2004) wanted to know whether siblings who knew more people tended to be more successful. Because of difficulty in measuring personal network size, his analysis was ultimately inconclusive.

In this article we report a method developed to estimate both individual network size and degree distribution in a population using a battery of questions that can be easily embedded into existing surveys. We begin with a review of previous attempts to measure personal network size, focusing on the scale-up method of Killworth et al. (1998b), which is promising but is known to suffer from three shortcomings: transmission errors, barrier effects, and recall error. In Section 3 we propose a latent nonrandom mixing model that resolves these problems, and as a byproduct allows for the estimation of social mixing patterns in the acquaintanceship network. We then fit the model to 1,370 survey responses from McCarty et al. (2001), a nationally representative telephone sample of Americans. In Section 5 we draw on insights developed during the statistical modeling to offer practical guidelines for the design of future surveys.

2. PREVIOUS RESEARCH

The most straightforward method for estimating the personal network size of respondents would be to simply ask them how many people they "know." We suspect that this would work poorly, however, because of the well-documented problems with self-reported social network data (Killworth and Bernard 1976; Bernard et al. 1984; Brewer 2000; Butts 2003). Other,

Tyler H. McCormick is Ph.D. Candidate, Department of Statistics, Columbia University, New York, NY 10027 (E-mail: *tyler@stat.columbia.edu*). Matthew J. Salganick is Assistant Professor, Department of Sociology and Office of Population Research, Princeton University, Princeton, NJ 08544 (E-mail: *mjs3@princeton.edu*). Tian Zheng is Associate Professor, Department of Statistics, Columbia University, New York, NY 10027 (E-mail: *tzheng@stat.columbia.edu*). This work was supported by National Science Foundation grant DMS-0532231 and a graduate research fellowship, and by the Institute for Social and Economic Research and Policy and the Applied Statistics Center at Columbia University. The authors thank Peter Killworth, Russ Bernard, and Chris McCarty for sharing their survey data, as well as Andrew Gelman, Thomas DiPrete, Delia Baldassari, David Banks, an associate editor, and two anonymous reviewers for their constructive comments. All of the authors contributed equally to this work.

© 2010 American Statistical Association
Journal of the American Statistical Association
March 2010, Vol. 105, No. 489, Applications and Case Studies
DOI: 10.1198/jasa.2009.ap08518

more clever attempts have been made to measure personal network size, including the reverse small-world method (Killworth and Bernard 1978; Killworth, Bernard, and McCarty 1984; Bernard et al. 1990), the summation method (McCarty et al. 2001), the diary method (Gurevich 1961; Pool and Kochen 1978; Fu 2007; Mossong et al. 2008), the phonebook method (Pool and Kochen 1978; Freeman and Thompson 1989; Killworth et al. 1990), and the scale-up method (Killworth et al. 1998b).

We believe that the *scale-up method* has the greatest potential for providing accurate estimates quickly with reasonable measures of uncertainty. But the scale-up method is known to suffer from three distinct problems: barrier effects, transmission effects, and recall error (Killworth et al. 2003, 2006). In Section 2.1 we describe the scale-up method and these three issues in detail, and in Section 2.2 we present an earlier model by Zheng, Salganik, and Gelman (2006) that partially addresses some of these issues.

2.1 The Scale-Up Method and Three Problems

Consider a population of size N. We can store the information about the social network connecting the population in an adjacency matrix, $\mathbf{\Delta} = [\delta_{ij}]_{N\times N}$, such that $\delta_{ij} = 1$ if person i knows person j. Although our method does not depend on the definition of "know," throughout we assume McCarty et al. (2001)'s definition: "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 then could be based on the fact that the responses would follow a binomial distribution with n trials and probability d_i/N. This method is extremely inefficient in large populations, however, because the probability of a relationship between any two people is very low. For example, assuming an average personal network size of 750 (as estimated by Zheng, Salganik, and Gelman 2006), 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, for example, asking "how many women do you know who gave birth in the last 12 months?" instead of asking if she knows 3.6 million distinct people. The scale-up method uses responses to questions of this form ("How many X's do you know?") to estimate personal network size. For example, if a respondent reports knowing 3 women who gave birth, this represents about 1-millionth of all women who gave birth within the last year. This information then could be used to estimate that the respondent knows about 1-millionth of all Americans,

$$\frac{3}{3.6 \text{ million}} \cdot (300 \text{ million}) \approx 250 \text{ people}. \tag{1}$$

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

$$\hat{d}_i = \frac{\sum_{k=1}^K y_{ik}}{\sum_{k=1}^K N_k} \cdot N, \tag{2}$$

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 equivalent not to asking about 3.6 million *random* people, but rather to asking about women roughly age 18–45. This creates statistical challenges that we address in detail in later sections.

To estimate the standard error of the simple estimate, we follow the practice of Killworth et al. (1998a) by assuming

$$\sum_{k=1}^K y_{ik} \sim \text{Binomial}\left(\sum_{k=1}^K N_k, \frac{d_i}{N}\right). \tag{3}$$

The estimate of the probability of success, $p = d_i/N$, is

$$\hat{p} = \frac{\sum_{i=1}^k y_{ik}}{\sum_{k=1}^K N_k} = \frac{\hat{d}_i}{N}, \tag{4}$$

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 $\hat{d}_i$ then has standard error

$$\begin{aligned}\text{SE}(\hat{d}_i) &= N \cdot \text{SE}(\hat{p}) \\ &= N\sqrt{\frac{1}{\sum_{k=1}^K N_k}\hat{p}(1-\hat{p})\frac{N - \sum_{k=1}^K N_k}{N-1}} \\ &\approx \sqrt{\frac{N - \sum_{k=1}^K N_k}{\sum_{k=1}^K N_k}\hat{d}_i} = \sqrt{\hat{d}_i} \cdot \sqrt{\frac{1 - \sum_{k=1}^K N_k/N}{\sum_{k=1}^K N_k/N}}. \end{aligned} \tag{5}$$

For example, when asking 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

$$\begin{aligned}\text{SE}(\hat{d}_i) &\approx \sqrt{\hat{d}_i} \cdot \sqrt{\frac{1 - \sum_{k=1}^K N_k/N}{\sum_{k=1}^K N_k/N}} \\ &\approx \sqrt{750} \cdot \sqrt{\frac{1 - 3.6 \text{ million}/300 \text{ million}}{3.6 \text{ million}/300 \text{ million}}} \approx 250,\end{aligned}$$

assuming a degree of 750 as estimated by Zheng, Salganik, and Gelman (2006).

If we also had asked respondents about 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. Figure 1 plots $\text{SE}(\hat{d}_i)/\sqrt{\hat{d}_i}$ 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% (or approximately 60 million in a population of 300 million). Although the foregoing standard error depends only on the sum of the subpopulation sizes, there are

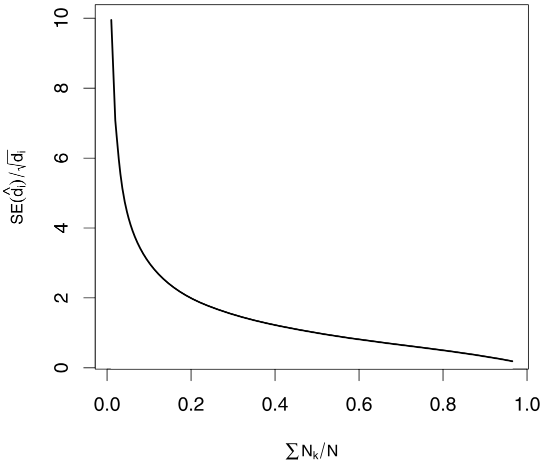

Figure 1. Standard error of the scale-up degree estimate (scaled by the square root of the true degree) plotted against the sum of the fractional subpopulation sizes. As the fraction of population represented by survey subpopulations increases, the precision of the estimate improves. Improvements diminish after about 20%.

other sources of bias that make the choice of the individual subpopulations important, as we show later.

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, 2006). Transmission errors occur when the respondent knows someone in a specific subpopulation but is not aware that the person is actually in that subpopulation; for example, a respondent might know a woman who recently gave birth but might not know that the woman 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 also can be called nonrandom mixing. For example, because 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 based solely on her personal network size and the number of women who have recently given birth. Similarly, an 80-year-old man probably knows fewer such women than would be expected under random mixing. Consequently, estimating personal network size by asking only "how many women do you know who have recently given birth?"—the estimator presented eq. (1)—will tend to overestimate the degree of women in their 30s and underestimate the degree of men in their 80s. Because these barrier effects can introduce a bias of unknown size, they have prevented previous researchers from 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 underrecall the number of people they know in large subpopulations (e.g., people named Michael) and overrecall the number of people they in small subpopulations (e.g., people who committed suicide) (Killworth et al. 2003; Zheng, Salganik, and Gelman 2006).

2.2 The Zheng, Salganik, and Gelman (2006) Model With Overdispersion

Before presenting our model for estimating personal network size using "how many X's do you know?" data, it is important to review the multilevel overdispersed Poisson model of Zheng, Salganik, and Gelman (2006), which, rather than treating nonrandom mixing (i.e., barrier effects) as an impediment to network size estimation, treats it as something important to estimate for its own sake. 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, y_{ik}'s, would follow a Poisson distribution with rate parameter determined by the degree of person i, d_i, and the network prevalence of group k, b_k. Here b_k is the proportion of ties that involve individuals in subpopulation k in the entire social network. If we can assume that individuals in the group being asked about (e.g., people named Michael) are as popular as the rest of the population on average, then $b_k \approx N_k/N$.

The responses to many of the questions in the data of McCarty et al. (2001) do not follow a Poisson distribution, however. In fact, most of the responses show overdispersion, that is, excess variance given the mean. Consider, for example, 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 than others to know someone in prison. To model this increased variance, Zheng, Salganik, and Gelman (2006) allowed individuals to vary in their propensity to form ties to different groups. If these propensities follow a gamma distribution with a mean value of 1 and a shape parameter of $1/(\omega_k - 1)$, then the y_{ik}'s can be modeled with a negative binomial distribution,

$$y_{ik} \sim \text{Neg-Binom}(\text{mean} = \mu_{ik},\ \text{overdispersion} = \omega_k), \qquad (6)$$

where $\mu_{ik} = d_i b_k$. Thus ω_k estimates the variation in individual propensities to form ties to people in different groups and represents one way of quantifying nonrandom mixing (i.e., barrier effects).

Although it was developed to estimate ω_k, the model of Zheng et al. also produces personal network size estimates, d_i. These estimates are problematic for two reasons, however. First, the normalization procedure used to address recall problems (see Zheng, Salganik, and Gelman 2006 for complete details) only shifts the degree distribution back to the appropriate scale; it does not ensure that the degree of individual respondents are being estimated accurately. Second, the degree estimates from the model remain susceptible to bias due to transmission error and barrier effects.

3. A NEW STATISTICAL METHOD FOR DEGREE ESTIMATION

We now describe a new statistical procedure to address the three aforementioned problems with estimating individual degree using "how many X's do you know?" data. Transmission errors, while probably the most difficult to quantify, are also the easiest to eliminate. We limit our analysis to the 12 subpopulations defined by first names that were asked about by McCarty et al. (2001). These 12 names (half male and half female) are presented in Figure 2. Although McCarty et al.'s definition of "knowing" someone does not explicitly require respondents to know individuals by name, we believe that using first names provides the minimum imaginable bias due to transmission errors; that is, it is unlikely that a person knows someone but does not know his or her first name. Even though using only first names controls transmission errors, it does not address bias from barrier effects or recall bias. In this section we propose a latent nonrandom mixing model to address these two issues.

3.1 Latent Nonrandom Mixing Model

We begin by considering the impact of barrier effects, or nonrandom mixing, on degree estimation. Imagine, for example, a hypothetical 30-year-old male survey respondent. If we were to ignore nonrandom mixing and ask this respondent how many Michaels he knows, then we would overestimate his network size using the scale-up method, because Michael tends to be a more popular name among younger males (Figure 2). In contrast, if we were to ask how many Roses he knows, then we would underestimate the size of his network, because Rose is a name that is more common in older females. In both cases, the properties of the estimates are affected by the demographic profiles of the names used, something not accounted for in the scale-up method.

We account for nonrandom mixing using a negative binomial model that explicitly estimates the propensity for a respondent in ego group e to know members of alter group a. Here we are following standard network terminology (Wasserman and Faust 1994), referring to the respondent as *ego* and the people to whom he can form ties as *alters*. The model is then

$$y_{ik} \sim \text{Neg-Binom}(\mu_{ike}, \omega'_k),$$

$$\text{where } \mu_{ike} = d_i \sum_{a=1}^{A} m(e,a) \frac{N_{ak}}{N_a}, \quad (7)$$

where d_i is the degree of person i, e is the ego group to which person i belongs, N_{ak}/N_a is the relative size of name k within alter group a (e.g., 4% of males age 21–40 are named Michael), and $m(e,a)$ is the mixing coefficient between ego group e and alter group a, that is,

$$m(e,a) = \mathrm{E}\left(\frac{d_{ia}}{d_i = \sum_{a=1}^{A} d_{ia}} \middle| i \text{ in ego group } e \right), \quad (8)$$

where d_{ia} is the number of person i's acquaintances in alter group a. That is, $m(e,a)$ represents the expected fraction of the ties of someone in ego group e that go to people in alter group a. For any group e, $\sum_{a=1}^{A} m(e,a) = 1$.

Thus the number of people that person i knows with name k, given that person i is in ego group e, is based on person i's degree (d_i), the proportion of people in alter group a that have name k (N_{ak}/N_a), and the mixing rate between people in group e and people in group a [$m(e,a)$]. In addition, if we do not observe nonrandom mixing, then $m(e,a) = N_a/N$ and μ_{ike} in (7) reduces to $d_i b_k$ in (6).

Along with μ_{ike}, the latent nonrandom mixing model also depends on the overdispersion, ω'_k, which represents the variation

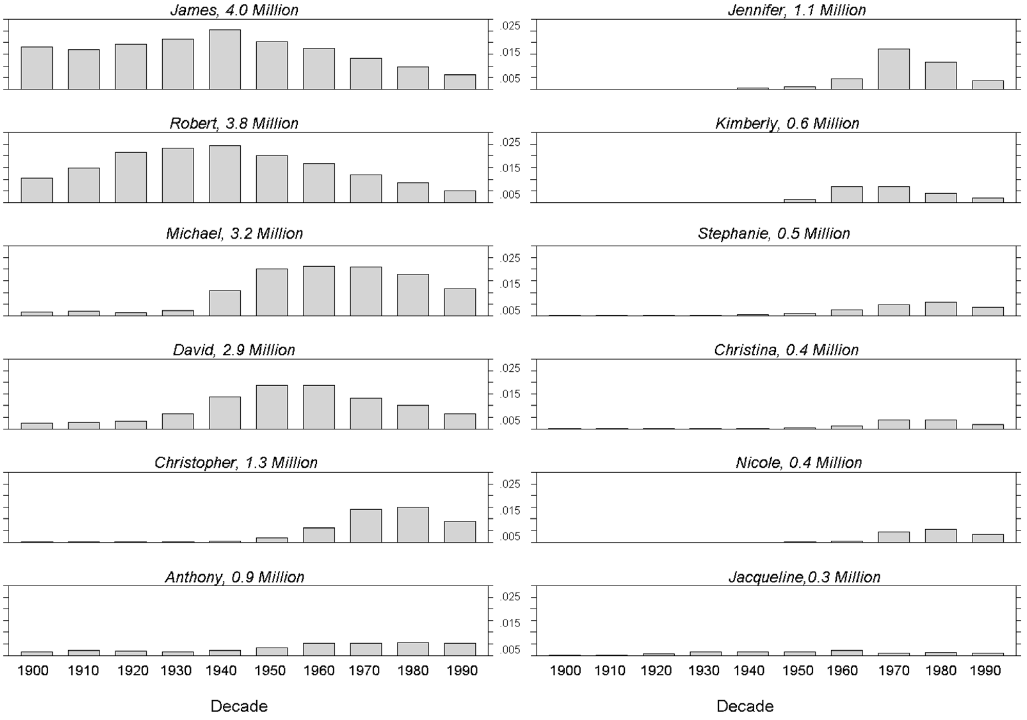

Figure 2. Age profiles for the 12 names used in the analysis (data source: SSA). The heights of the bars represent the percentage of American newborns in a given decade with a particular name. The total subpopulation size is given across the top of each graph. These age profiles are required to construct the matrix of $\frac{N_{ak}}{N_a}$ terms in eq. (7). The male names chosen by McCarty et al. are much more popular than the female names.

in the relative propensity of respondents within an ego group to form ties with individuals in a particular subpopulation k. Using $m(e, a)$, we model the variability in relative propensities that can be explained by nonrandom mixing between the defined alter and ego groups. Explicitly modeling this variation should cause a reduction in overdispersion parameter ω'_k compared with ω_k in (6) and Zheng, Salganik, and Gelman (2006). The term ω'_k is still in the latent nonrandom mixing model, however, because there remains residual overdispersion based on additional ego and alter characteristics that could affect their propensity to form ties.

Fitting the model requires choosing the number of ego groups, E, and alter groups, A. In this case we classified egos into six categories by crossing gender (2 categories) with three age categories: youth (age 18–24 years), adult (age 25–64), and senior (age 65+). We constructed eight alter groups by crossing gender with four age categories: 0–20, 21–40, 41–60, and 61+. Thus to estimate the model, we needed to know the age and gender of our respondents and, somewhat more problematically, the the relative popularity of the name-based subpopulations in each alter group ($\frac{N_{ak}}{N_a}$). We approximated this popularity using the decade-by-decade birth records made available by the Social Security Administration (SSA). Because we are using the SSA birth data as a proxy for the living population, we are assuming that several social processes—immigration, emigration, and life expectancy—are uncorrelated with an individual's first name. We also are assuming that the SSA data are accurate, even for births from the early twentieth century, when registration was less complete. We believe that these assumptions are reasonable as a first approximation and probably did not have a substantial effect on our results. Together these modeling choices resulted in a total of 48 mixing parameters, $m(e, a)$, to estimate (6 ego groups by 8 alter groups). We believe that this represents a reasonable compromise between parsimony and richness.

3.2 Correction for Recall Error

The model in eq. (7) is a model for the actual network of the respondents assuming only random sampling error. Unfortunately, however, the observed data rarely yield reliable information about this network, because of the systematic tendency for respondents to underrecall the number of individuals that they know in large subpopulations (Killworth et al. 2003; Zheng, Salganik, and Gelman 2006). For example, assume that a respondent recalls knowing five people named Michael; then the estimated network size would be

$$\frac{5}{4.8\text{ million}/300\text{ million}} \approx 300\text{ people}. \tag{9}$$

But Michael is a common name, making it likely that there are additional Michaels in the respondent's actual network who were not counted at the time of the survey (Killworth et al. 2003; Zheng, Salganik, and Gelman 2006). We could choose to address this issue in two ways, which, although ultimately equivalent, suggest two distinct modeling strategies.

First, we could assume that the respondent is inaccurately recalling the number of people named Michael that she knows from her true network. Under this framework, any correction that we propose should increase the numerator in eq. (9). This requires that we propose a mechanism by which respondents underreport their true number known on individual questions. In our example, this would be equivalent to taking the five Michaels reported and applying some function to produce a corrected response (presumably some number greater than five), which then would be used to fit the proposed model. It is difficult to speculate about the nature of this function in any detail, however.

Another approach would be to assume that respondents are recalling not from their actual network, but rather from a *recalled network* that is a subset of the actual network. We speculate that the recalled network is created when respondents change their definition of "know" based on the fraction of their network made up of the population being queried such that they use a more restrictive definition of "know" when answering about common subpopulations (e.g., people named Michael) than when answering about rare subpopulations (e.g., people named Ulysses). This means that, in the context of Section 2.2, we no longer have that $b_k \approx N_k/N$. We can, however, use this information for calibration, because the true subpopulation sizes, N_k/N, are known and can be used as a point of comparison to estimate and then correct for the amount of recall bias.

Previous empirical work (Killworth et al. 2003; Zheng, Salganik, and Gelman 2006; McCormick and Zheng 2007) suggests that the calibration curve, $f(\cdot)$, should impose less correction for smaller subpopulations and progressively greater correction as the popularity of the subpopulation increases. Specifically, both Killworth et al. (2003) and Zheng, Salganik, and Gelman (2006) suggested that the relationship between $\beta_k = \log(b_k)$ and $\beta'_k = \log(b'_k)$ begins along the $y = x$ line, and that the slope decreases to $1/2$ (corresponding to a square root relation on the original scale) with increasing fractional subpopulation size.

Using these assumptions and some boundary conditions, McCormick and Zheng (2007) derived a calibration curve that gives the following relationship between b_k and b'_k:

$$b'_k = b_k\left[\frac{c_1}{b_k}\exp\left(\frac{1}{c_2}\left(1-\left[\frac{c_1}{b_k}\right]^{c_2}\right)\right)\right]^{1/2}, \tag{10}$$

where $0 < c_1 < 1$ and $c_2 > 0$. By fitting the curve to the names from the McCarty et al. (2001) survey, we chose $c_1 = e^{-7}$ and $c_2 = 1$. (For details on this derivation, see McCormick and Zheng 2007.) We apply the curve to our model as follows:

$$y_{ik} \sim \text{Neg-Binom}(\mu_{ike}, \omega'_k),$$
$$\text{where } \mu_{ike} = d_i f\left(\sum_{a=1}^{A} m(e, a)\frac{N_{ak}}{N_a}\right). \tag{11}$$

3.3 Model Fitting Algorithm

Here we use a multilevel model and Bayesian inference to estimate d_i, $m(e, a)$, and ω'_k in the latent nonrandom mixing model described in Section 3.1. We assume that $\log(d_i)$ follows a normal distribution with mean μ_d and standard deviation σ_d. Zheng, Salganik, and Gelman (2006) postulated that this prior should be reasonable based on previous work, specifically McCarty et al. (2001), and found that the prior worked well in their case. We estimate a value of $m(e,a)$ for all E ego groups and all A alter groups. For each ego group, e, and each alter group, a, we assume that $m(e,a)$ has a normal prior distribution with mean $\mu_{m(e,a)}$ and standard deviation $\sigma_{m(e,a)}$. For ω'_k, we use independent uniform(0, 1) priors on the inverse scale, $p(1/\omega'_k) \propto 1$. Because ω'_k is constrained to $(1, \infty)$, the inverse falls on $(0, 1)$. The Jacobian for the transformation is $\omega_k^{\prime -2}$. Finally, we give noninformative uniform priors to the hyperparameters μ_d, $\mu_{m(e,a)}$, σ_d, and $\sigma_{m(e,a)}$. Then the joint posterior density can be expressed as

$$\begin{aligned} & p\big(d, m(e,a), \omega', \mu_d, \mu_{m(e,a)}, \sigma_d, \sigma_{m(e,a)} | y\big) \\ & \quad \propto \prod_{k=1}^{K} \prod_{i=1}^{N} \binom{y_{ik} + \xi_{ik} - 1}{\xi_{ik} - 1} \left(\frac{1}{\omega'_k}\right)^{\xi_{ik}} \left(\frac{\omega'_k - 1}{\omega'_k}\right)^{y_{ik}} \\ & \quad \times \prod_{i=1}^{N} \left(\frac{1}{\omega'_k}\right)^2 \mathrm{N}(\log(d_i) | \mu_d, \sigma_d) \\ & \quad \times \prod_{e=1}^{E} \mathrm{N}\big(m(e,a) | \mu_{m(e,a)}, \sigma_{m(e,a)}\big), \end{aligned} \qquad (12)$$

where $\xi_{ik} = d_i f(\sum_{a=1}^{A} m(e,a) \frac{N_{ak}}{N_a}) / (\omega'_k - 1)$.

Adapting Zheng, Salganik, and Gelman (2006), we use a Gibbs–Metropolis algorithm in each iteration v, as follows:

1. For each i, update d_i using a Metropolis step with jumping distribution $\log(d_i^*) \sim \mathrm{N}(d_i^{(v-1)}$, (jumping scale of $d_i)^2)$.
2. For each e, update the vector $m(e, \cdot)$ using a Metropolis step. Define the proposed value using a random direction and jumping rate. Each of the A elements of $m(e, \cdot)$ has a marginal jumping distribution $m(e,a)^* \sim \mathrm{N}(m(e,a)^{(v-1)}$, (jumping scale of $m(e, \cdot))^2)$. Then rescale so that the row sum is 1.
3. Update $\mu_d \sim \mathrm{N}(\hat{\mu}_d, \sigma_d^2/n)$, where $\hat{\mu}_d = \frac{1}{n} \sum_{i=1}^{n} d_i$.
4. Update $\sigma_d^2 \sim \text{Inv-}\chi^2(n-1, \hat{\sigma}_d^2)$, where $\hat{\sigma}_d^2 = \frac{1}{n} \sum_{i=1}^{n} (d_i - \mu_d)^2$.
5. Update $\mu_{m(e,a)} \sim \mathrm{N}(\hat{\mu}_{m(e,a)}, \sigma_{m(e,a)}^2/A)$ for each e where $\hat{\mu}_{m(e,a)} = \frac{1}{A} \sum_{a=1}^{A} m(e,a)$.
6. Update $\sigma_{m(e,a)}^2 \sim \text{Inv-}\chi^2(A-1, \hat{\sigma}_{m(e,a)}^2)$ for each e, where $\hat{\sigma}_{m(e,a)}^2 = \frac{1}{A} \sum_{a=1}^{A} (m(e,a) - \mu_{m(e,a)})^2$.
7. For each k, update ω'_k using a Metropolis step with jumping distribution $\omega_k^{\prime *} \sim \mathrm{N}(\omega_k^{\prime (v-1)}$, (jumping scale of $\omega'_k)^2)$.

4. RESULTS

To fit the model, we used data from McCarty et al. (2001), comprising survey responses from 1,370 adults living in the United States who were contacted via random digit dialing in two surveys: survey 1, with 796 respondents, conducted in January 1998, and survey 2, with 574 respondents, conducted in January 1999. To correct for responses that were suspiciously large (e.g., a person claiming to know more than 50 Michaels), we truncated all responses at 30, a procedure that affected only 0.25% of the data. We also inspected the data using scatterplots, which revealed a respondent who was coded as knowing seven people in each subpopulation. We removed this case from the data set.

We obtained approximate convergence of our algorithm ($\hat{R}_{max} < 1.1$; see Gelman et al. 2003) using three parallel chains with 2,000 iterations per chain. We used the first half of each chain for burn-in and thinned the chain by using every tenth iterate. All computations were performed using custom code written for the software package R (R Development Core Team 2009), which is available on request.

4.1 Personal Network Size Estimates

We estimated a mean network size of 611 (median, 472). Figure 3 presents the distribution of network sizes. In the figure, the solid line represents a log-normal distribution with parameters determined via maximum likelihood ($\hat{\mu}_{mle} = 6.2$ and $\hat{\sigma}_{mle} = 0.68$); the lognormal distribution fits the distribution quite well. This result is not an artifact of our model, as has been confirmed by additional simulation studies (data not shown). Given the recent interest in power laws and networks, we also explored the fit of the power law distribution (dashed line) with parameters estimated via maximum likelihood ($\alpha_{mle} = 1.28$) (Clauset, Shalizi, and Newman 2007). The fit is clearly poor, a result consistent with previous work showing that another social network—the sexual contact network—also is poorly approximated by the power law distribution (Hamilton, Handcock, and Morris 2008).

Figure 4 compares the estimated degree from the latent nonrandom mixing model with estimates obtained using the method of Zheng, Salganik, and Gelman (2006). In general, the

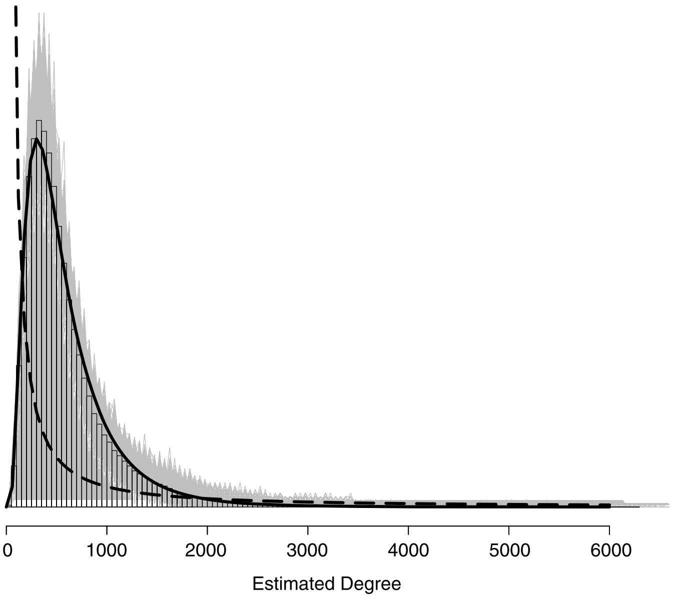

Figure 3. Estimated degree distribution from the fitted model. The median is about 470, and the mean is about 610. The shading represents random draws from the posterior distribution to indicate inferential uncertainty in the histograms. The solid line is a log-normal distribution fit using maximum likelihood to the posterior median for each respondent ($\hat{\mu}_{mle} = 6.2$ and $\hat{\sigma}_{mle} = 0.68$). The dashed line is a power law density with scaling parameter estimated by maximum likelihood ($\hat{\alpha}_{mle} = 1.28$).

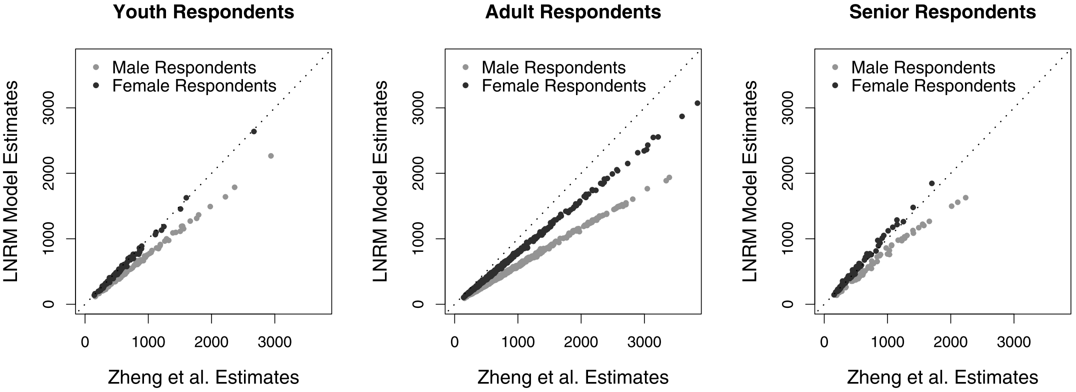

Figure 4. Comparison of the estimates from Zheng et al. and the latent nonrandom mixing model broken down by age and gender. Gray points represent males; black points, females. The latent nonrandom mixing model accounts for the fact that the names from McCarty et al. are predominately male and predominantly middle-aged, and thus produces lower degree estimates for respondents in these groups. Because our model has six ego groups, there are six distinct patterns in the figure.

estimates from the latent nonrandom mixing model tend to be slightly smaller, with an estimated median degree of 472 (mean, 611), compared with that of 610 (mean 750) obtained using the method of Zheng, Salganik, and Gelman (2006). Figure 4 also reveals that the differences between the estimates vary in ways that are expected given that the names in the data of McCarty et al. are predominantly male and predominantly middle-aged (see Figure 2). The latent nonrandom mixing model accounts for this fact, and thus produces lower estimates for male respondents and adult respondents than the method of Zheng, Salganik, and Gelman (2006).

4.2 Mixing Estimates

Although our proposed procedure was developed to obtain good estimates of personal network size, it also provides information about the mixing rates in the population, which is considered to affect the spread of information (Volz 2006) and disease (Morris 1993; Mossong et al. 2008). Even though previous work has been done on estimating population mixing rates (see, e.g., Morris 1991), we believe this is the first survey-based approach for estimating such information indirectly.

As mentioned in the previous section, the mixing matrix, $m(e, a)$, represents the proportion of the network of a person in ego group e that is composed of people in alter group a. The estimated mixing matrix presented in Figure 5 indicates plausible relationships within subgroups, with the dominant pattern being that individuals tend to preferentially associate with others of similar age and gender, a finding consistent with the large sociological literature on homophily (the tendency for people to form ties to those who are similar to themselves) (McPherson, Smith-Lovin, and Cook 2001). This trend is especially apparent in adult males, who demonstrate a high proportion of their ties to other males. With additional information on the race/ethinicity of the different names, the latent nonrandom mixing model could be used to estimate the extent of social network–based segregation, an approach that could have many advantages over traditional measures of residential segregation (Echenique and Fryer 2007).

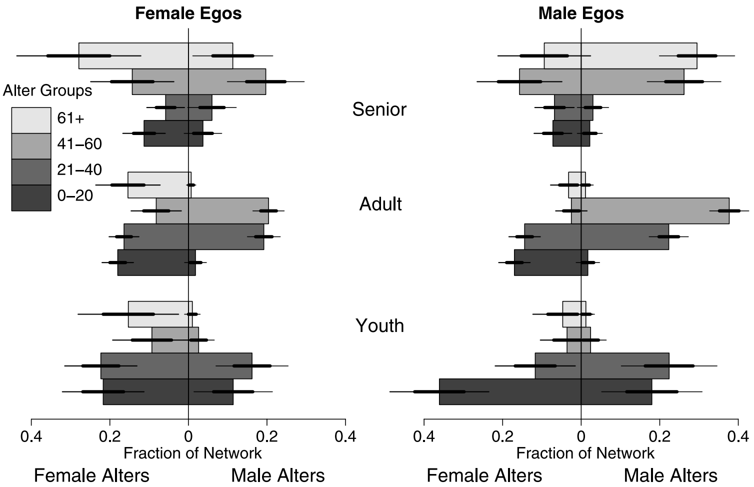

Figure 5. Barplot of the mixing matrix. Each of the six stacks of bars represents one ego group. Each stack describes the proportion of the given ego group's ties that are formed with all of the alter groups; thus the total proportion within each stack is 1. For each individual bar, a shift to the left indicates an increased propensity to know female alters. Thick lines represent ±1 standard error (estimated from the posterior); thin lines, ±2 standard errors.

4.3 Overdispersion

Another way to assess the latent nonrandom mixing model is to examine the overdispersion parameter ω_k', which represents the variation in propensity to know individuals in a particular group. In the latent nonrandom mixing model, a portion of this variability is modeled by the ego group–dependent mean, μ_{ike}. The remaining unexplained variability forms the overdispersion parameter, ω_k'. In Section 3.1 we predicted that ω_k' would be smaller than the overdispersion ω_k reported by Zheng, Salganik, and Gelman (2006), because Zheng, Salganik, and Gelman (2006) did not model nonrandom mixing.

This prediction turned out to be correct. With the exception of Anthony, all of the estimated overdispersion estimates from the latent nonrandom mixing model are lower than those presented by Zheng, Salganik, and Gelman (2006). To judge the magnitude of the difference, we created a standardized difference measure, $\frac{\omega_k'-\omega_k}{\omega_k-1}$. Here the numerator, $\omega_k' - \omega_k$, represents the reduction in overdispersion resulting from modeling nonrandom mixing explicitly in the latent nonrandom mixing model. In the denominator, an ω_k value of 1 corresponds to no overdispersion; thus the ratio for group k is the proportion of overdispersion encountered in Zheng, Salganik, and Gelman (2006) that is explicitly modeled in the latent nonrandom mixing model. The standardized difference was on average 0.213 units lower for the latent nonrandom mixing model estimates, indicating that roughly 21% of the overdispersion found in Zheng, Salganik, and Gelman (2006) can be explained by nonrandom mixing due to age and gender. If appropriate ethnicity or other demographic information about the names were available, then we would expect this reduction to be even larger.

5. DESIGNING FUTURE SURVEYS

In the preceding sections we analyzed existing data in such a way as to resolve the three known problems with estimating personal network size from "how many X's do you know?" data. In this section we offer survey design suggestions that can allow researchers to capitalize on the simplicity of the scale-up estimates while enjoying the same bias reduction as in the latent nonrandom mixing model. Thus this section provides an efficient and easily applied degree estimation method that is accessible to a wide range of researchers who may not wish to fit the latent nonrandom mixing model.

In Section 5.1 we derive the requirement for selecting first names such that the scale-up estimate is equivalent to the degree estimate derived from fitting a latent nonrandom mixing model using Markov chain Monte Carlo computation. The intuition behind this result is that the names asked about should be chosen so that the combined set of people asked about is a "scaled-down" version of the overall population; for example, if 20% of the general population is females under age 30, then 20% of the people with the names used also must be females under age 30. Section 5.2 presents practical advice for choosing such a set of names and presents a simulation study of the performance of the suggested guidelines. Finally, Section 5.3 offers guidelines on the standard errors of the estimates.

5.1 Selecting Names for the Scale-Up Estimator

Unlike the scale-up estimator (2), the latent nonrandom mixing model accounts for barrier effects due to some demographic factors by estimating degree differentially based on characteristics of the respondent and of the potential alter population. But if there were conditions under which the simple scale-up estimator was expected to be equivalent to the latent nonrandom mixing model, then the simple estimator would enjoy the same reduction of bias from barrier effects as the more complex latent nonrandom mixing model estimator. In this section we derive such conditions.

The latent nonrandom mixing model assumes an expected number of acquaintances for an individual i in ego group e to people in group k [as in (7)],

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

In contrast, the scale-up estimator assumes that

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

$$\equiv d_i \frac{\sum_{k=1}^{K} \sum_{a=1}^{A} N_{ak}}{N}, \qquad \forall e. \qquad (13)$$

Equation (13) shows that the scale-up estimator of Killworth et al. (2) is in expectation equivalent to that of the latent nonrandom mixing if either

$$m(e,a) = \frac{N_a}{N}, \qquad \forall a, \forall e \qquad (14)$$

or

$$\frac{\sum_{k=1}^{K} N_{ak}}{\sum_{k=1}^{K} N_k} = \frac{N_a}{N}, \qquad \forall a. \qquad (15)$$

In other words, the two estimators are equivalent if there is random mixing (14) or if the combined set of names represents a "scaled-down" version of the population (15). Because 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% of the population is males between age 21 and 40 ($\frac{N_a}{N}$), then 15% of the people asked about also must be males between age 21 and 40 ($\frac{\sum_{k=1}^{K} N_{ak}}{\sum_{k=1}^{K} N_k}$).

When actually choosing a set of names to satisfy the scaled-down condition, we found it more convenient to work with a rearranged form of (15),

$$\frac{\sum_{k=1}^{K} N_{ak}}{N_a} = \frac{\sum_{k=1}^{K} N_k}{N}, \qquad \forall a. \qquad (16)$$

To find a set of names that satisfy (16), it is helpful to create Figure 6, which 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. Consider, for example, the names Walter, Bruce, and Kyle. These names have similar popularity overall, but Walter was popular in 1910–1940, whereas Bruce was popular during

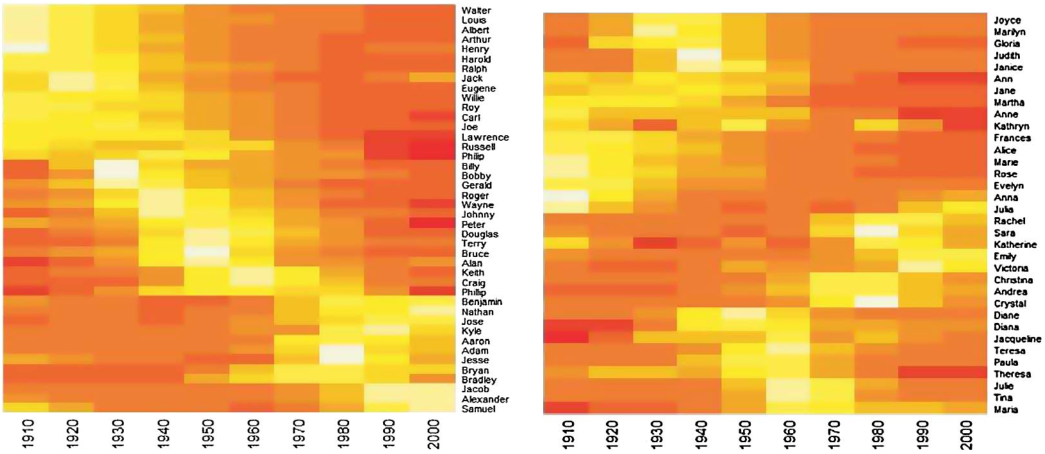

Figure 6. Heat maps of additional male and female names based on SSA data. Lighter color indicates higher popularity.

the middle of the twentieth century, and Kyle was popular near the end of the century. Thus the popularity of the names during any one time period will be balanced by the popularity of other names in the other time periods, preserving the required equality in the sum (16).

When choosing what names to use, besides satisfying eq. (16), we recommend choosing names that compromise 0.1%–0.2% of the population, which will minimize recall errors and yield average responses of 0.6–1.3. Finally, we recommend choosing names not commonly associated with nicknames, to minimize transmission errors.

5.2 Simulation Study

We now demonstrate the use of the foregoing guidelines in a simulation study. Again we use the age and gender profiles of the names as an example. If other information were available, then the general approach presented here would still be applicable.

Figure 6 shows the popularity profiles of several names with the desired level of overall popularity (0.1%–0.2% of the population). We used this figure to select two sets of names (Table 1). We selected the first set—the *good names*—using the procedure described in the previous section to satisfy the scaled-down condition. We also selected a second set of names—the *bad names*—that were popular with individuals born in the first decades of the twentieth century and thus did not satisfy the scaled-down condition. For comparison, we also use the set of 12 names from the data of McCarty et al.

Table 1. A set of names that approximately meet the scaled-down condition (the good names) and a set of names that do not (the bad names)

Good names		Bad names	
Male	Female	Male	Female
Walter	Rose	Walter	Alice
Bruce	Tina	Jack	Marie
Kyle	Emily	Harold	Rose
Ralph	Martha	Ralph	Joyce
Alan	Paula	Roy	Marilyn
Adam	Rachel	Carl	Gloria

Figure 7 provides a visual check of the scaled-down condition (14) for these three sets of names by plotting the combined demographic profiles for each set compared with that of the overall population. The figure reveals clear problems with the McCarty et al. names and the bad names. For example, in the bad names, older individuals represent a much larger fraction of the subpopulation of alters compared with the overall population (as expected given our method of selection). Thus we would expect scale-up estimates based on the bad names to overestimate the degree of older respondents.

We evaluated this prediction using a simulation study in which we fit the latent nonrandom mixing model to the McCarty et al. data and then used these estimated parameters (i.e., degree, overdispersion, and mixing matrix) to generate a negative binomial sample of size 1,370. We then fit the scale-up estimate, the latent nonrandom mixing model, and the model of Zheng et al. to these simulated data to see how these estimates could recover the known data-generating parameters.

Figure 8 presents the results of the simulation study. In each panel the difference between the estimated degree and the known data-generating degree for individual i is plotted against the respondent's age. For the bad names (Table 1), individual degree is systematically overestimated for older individuals and underestimated for younger individuals in all three models, but the latent nonrandom mixing model shows the least age bias in estimates. This overestimation of the degree of older respondents is as expected given the combined demographic profiles of the set of bad names (Figure 7). For the names from the McCarty et al. (2001) survey, the scale-up estimator and the model of Zheng et al. overestimate the degree of the younger members of the population, again as expected given the combined demographic profiles of this set of names (Figure 7). But the latent nonrandom mixing model produces estimates with no age bias. Finally, for the good names—those selected according to the scaled-down condition—all three procedures work well, further supporting the design strategy proposed in Section 5.1.

Overall, our simulation study shows that the proposed latent nonrandom mixing model performed better than existing methods when names were not chosen according to the scaled-down condition, suggesting that it is the best approach to estimating personal network size with most data. But when the names were

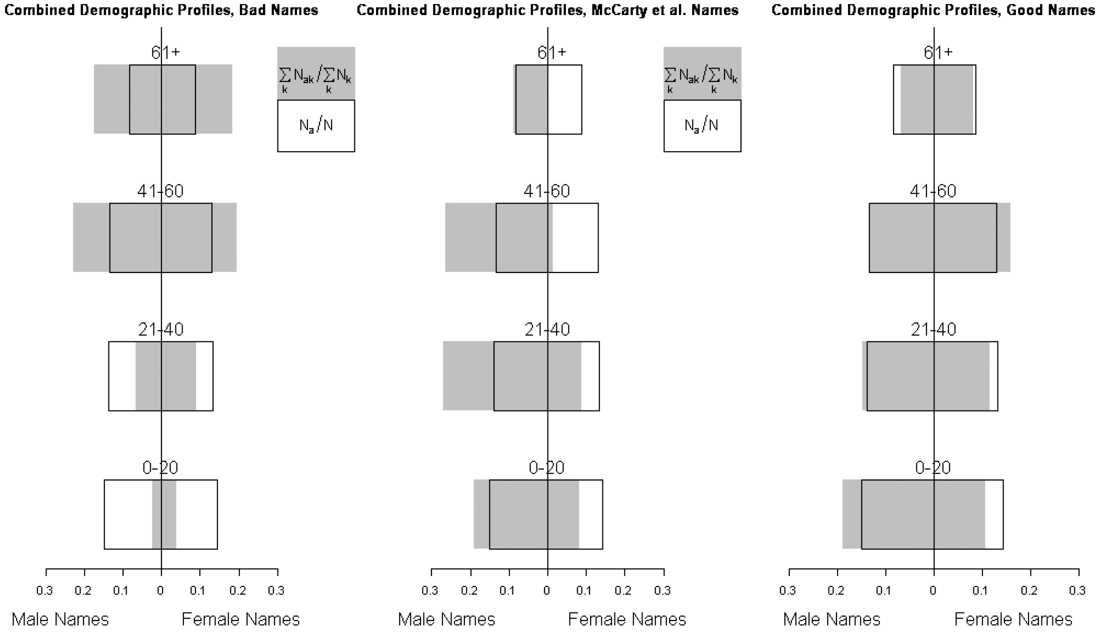

Figure 7. Combined demographic profiles for three sets of names (shaded bars) and population proportion of the corresponding category (solid lines). Unlike the bad names and the names of McCarty et al., the good names approximately satisfy the scaled-down condition [eq. (15)].

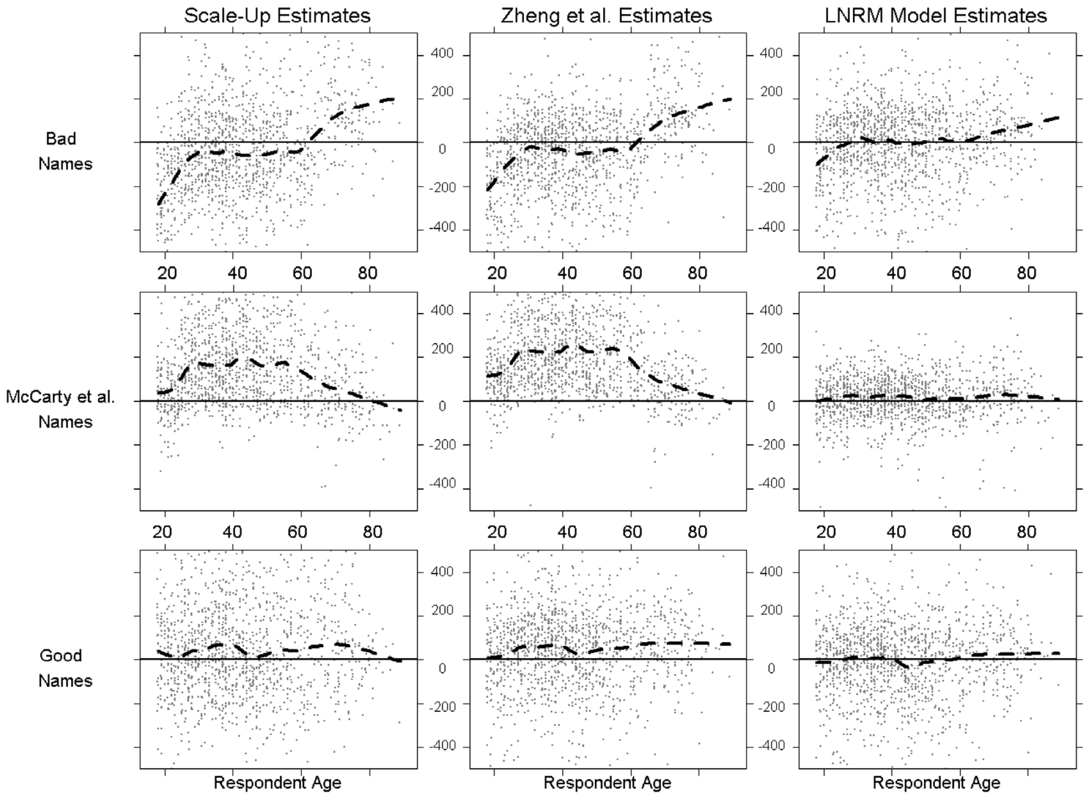

Figure 8. A comparison of the performance of the latent nonrandom mixing model, the Zheng et al. overdispersion model, and the Killworth et al. scale-up method. In each panel the difference between the estimated degree and the known data-generating degree is plotted against age. Three different sets of names were used: a set of names that do not satisfy the scaled-down condition (bad names), the names used in the survey of McCarty et al., and a set of names that satisfy the scaled-down condition (good names). With the bad names, all three procedures show some age bias in estimates, but these biases are smallest with the latent nonrandom mixing model. With the names of McCarty et al., the scale-up estimates and the Zheng et al. estimates show age bias, but the estimates from the latent nonrandom mixing model are excellent. With the good names, all three procedures perform well.

chosen according the scaled-down condition, even the much simpler scale-up estimator works well.

5.3 Selecting the Number of Names

For researchers planning to use the scale-up method, an important issue to consider besides 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 this trade-off, we return to the approximate standard error under the binomial model presented in Section 2.1. Simulation results using 6, 12, and 18 names chosen using the foregoing suggested guidelines agree well with the results from the binomial model in (5) (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 (5) 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 presented in Table 1, $\sum N_k = 4$ million, so a researcher could expect a standard error of 177 for a respondent with degree 500.

6. DISCUSSION AND CONCLUSION

Using "how many X's do you know?"–type data to produce estimates of individual degree and degree distribution holds great potential for applied researchers. Especially, these questions can be easily integrated into existing surveys. But this method's usefulness has been limited by three previously documented problems. In this article we have proposed two additional tools for researchers. First, the latent nonrandom mixing model in Section 3 deals with the known problems when using "how many X's do you know?" data, allowing for improved personal network size estimation. In Section 5 we showed that if future researchers choose the names used in their survey wisely—that is, if the set of names satisfies the scaled-down condition—then they can get improved network size estimates without fitting the latent nonrandom mixing model. We also provided guidelines for selection such a set of names.

Although the methods presented here have advantages, they also have somewhat more strenuous data requirements compared with previous methods. Fitting the latent nonrandom mixing model or designing a survey to satisfy the scaled-down condition requires information about the demographic profiles of the first names used, which may not be available in some countries. If such information were not available, then other subpopulations could be used (e.g., women who have given birth in the last year, men who are in the armed forces); however, then transmission error becomes a potential source of concern. A further limitation to note is that even if the set of names used satisfies the scaled-down condition with respect to age and gender, the subsequent estimates could have a bias correlated with something that is not included, such as race/ethnicity.

A potential area for future methodological work involves improving the calibration curve used to adjust for recall bias. Currently, the curve is fit deterministically based on the 12 names in the McCarty et al. (2001) data and the independent observations of Killworth et al. (2003). In the future, the curve could be dynamically fit for a given set of data as part of the modeling process. Another area for future methodological work is formalizing the procedure used to select names that satisfy the scaled-down condition. Our trial-and-error approached worked well here because there were only eight alter categories, but in cases with more categories, a more automated procedure would be preferable.

A final area for future work involves integrating the procedures developed here with efforts to estimate the size of "hidden" or "hard-to-count" populations. For example, there is tremendous uncertainly about the sizes of populations at highest risk for HIV/AIDS in most countries: injection drug users, men who have sex with men, and sex workers. Unfortunately, this uncertainty has complicated public health efforts to understand and slow the spread of the disease (UNAIDS 2003). As was shown by Bernard et al. (1991) and Killworth et al. (1998b), estimates of personal network size can be combined with responses to questions such as "how many injection drug users do you know?" to estimate the size of hidden populations. The intuition behind this approach is that respondents' networks, should on average be representative of the population. Therefore, if an American respondent were to report knowing 2 injection drug users and was estimated to know 300 people, then we could estimate that there are about 2 million injection drug users in the United States ($\frac{300 \text{ million}}{300} \cdot 2 = 2$ million), and this estimate could be improved by averaging over respondents (Killworth et al. 1998b). Thus the improved degree estimates described herein should lead to improved estimates of the sizes of hidden populations, but future work might be needed to tailor these methods to public health contexts.

[Received September 2008. Revised September 2009.]

REFERENCES

Barabási, A. L. (2003), *Linked*, New York: Penguin Group. [59]

Barton, A. H. (1968), "Bringing Society Back in: Survey Research and Macro-Methodology," *American Behavorial Scientist*, 12 (2), 1–9. [59]

Bernard, H. R., Johnsen, E. C., Killworth, P., and 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. [69]

Bernard, H. R., Johnsen, E. C., Killworth, P. D., McCarty, C., Shelley, G. A., and Robinson, S. (1990), "Comparing Four Different Methods for Measuring Personal Social Networks," *Social Networks*, 12, 179–215. [60]

Bernard, H. R., Killworth, P., Kronenfeld, D., and Sailer, L. (1984), "The Problem of Informant Accuracy: The Validity of Retrospective Data," *Annual Review of Anthropology*, 13, 495–517. [59]

Brewer, D. D. (2000), "Forgetting in the Recall-Based Elicitation of Person and Social Networks," *Social Networks*, 22, 29–43. [59]

Butts, C. T. (2003), "Network Inference, Error, and Informant (In)accuracy: A Bayesian Approach," *Social Networks*, 25, 103–140. [59]

Clauset, A., Shalizi, C., and Newman, M. (2007), "Power-Law Distributions in Empirical Data," *SIAM Review*, to appear. Available at *arXiv:0706.1062*. [64]

Conley, D. (2004), *The Pecking Order: Which Siblings Succeed and Why*, New York: Pantheon Books. [59]

Echenique, F., and Fryer, R. G. (2007), "A Measure of Segregation Based on Social Interactions," *Quaterly Journal of Economics*, 122 (2), 441–485. [65]

Freeman, L. C., and Thompson, C. R. (1989), "Estimating Acquaintanceship Volume," in *The Small World*, ed. M. Kochen, Norwood, NJ: Ablex Publishing, pp. 147–158. [60]

Fu, Y.-C. (2007), "Contact Diaries: Building Archives of Actual and Comprehensive Personal Networks," *Field Methods*, 19 (2), 194–217. [60]

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis* (2nd ed.), New York: Chapman & Hall/CRC. [64]

Gurevich, M. (1961), "The Social Structure of Acquaintanceship Networks," Ph.D. thesis, MIT. [60]

Hamilton, D. T., Handcock, M. S., and Morris, M. (2008), "Degree Distributions in Sexual Networks: A Framework for Evaluating Evidence," *Sexually Transmitted Diseases*, 35 (1), 30–40. [64]

Killworth, P. D., and Bernard, H. R. (1976), "Informant Accuracy in Social Network Data," *Human Organization*, 35 (3), 269–289. [59]

——— (1978), "The Reverse Small-World Experiment," *Social Networks*, 1 (2), 159–192. [60]

Killworth, P. D., Bernard, H. R., and McCarty, C. (1984), "Measuring Patterns of Acquaintanceship," *Current Anthropology*, 23, 318–397. [60]

Killworth, P. D., Johnsen, E. C., Bernard, H. R., Shelley, G. A., and McCarty, C. (1990), "Estimating the Size of Personal Networks," *Social Networks*, 12, 289–312. [60]

Killworth, P. D., Johnsen, E. C., McCarty, C., Shelly, G. A., and Bernard, H. R. (1998a), "A Social Network Approach to Estimating Seroprevalence in the United States," *Social Networks*, 20, 23–50. [60]

Killworth, P. D., McCarty, C., Bernard, H. R., Johnsen, E. C., Domini, J., and Shelly, G. A. (2003), "Two Interpretations of Reports of Knowledge of Subpopulation Sizes," *Social Networks*, 25, 141–160. [60,61,63,69]

Killworth, P. D., McCarty, C., Bernard, H. R., Shelly, G. A., and Johnsen, E. C. (1998b), "Estimation of Seroprevalence, Rape, and Homelessness in the U.S. Using a Social Network Approach," *Evaluation Review*, 22, 289–308. [59,60,69]

Killworth, P. D., McCarty, C., Johnsen, E. C., Bernard, H. R., and Shelley, G. A. (2006), "Investigating the Variation of Personal Network Size Under Unknown Error Conditions," *Sociological Methods & Research*, 35 (1), 84–112. [59-61]

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. [61]

Lohr, S. (1999), *Sampling: Design and Analysis*, Pacific Grove, CA: Duxbury Press. [60]

McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E., and Shelley, G. A. (2001), "Comparing Two Methods for Estimating Network Size," *Human Organization*, 60, 28–39. [59-64,67,69]

McCormick, T. H., and Zheng, T. (2007), "Adjusting for Recall Bias in 'How Many X's Do You Know?' Surveys," in *Conference Proceedings of the Joint Statistical Meetings*, Salt Lake City, Utah. [63]

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001), "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27, 415–444. [61,65]

Morris, M. (1991), "A Log-Linear Modeling Framework for Selective Mixing," *Mathematical Biosciences*, 107 (2), 349–377. [65]

——— (1993), "Epidemiology and Social Networks: Modeling Structured Diffusion," *Sociological Methods and Research*, 22 (1), 99–126. [65]

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. J. (2008), "Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases," *PLoS Medicine*, 5 (3), e74. [60,65]

Pastor-Satorras, R., and Vespignani, A. (2001), "Epidemic Spreading in Scale-Free Networks," *Physical Review Letters*, 86 (14), 3200–3203. [59]

Pool, I., and Kochen, M. (1978), "Contacts and Influence," *Social Networks*, 1, 5–51. [59,60]

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [64]

Santos, F. C., Pacheco, J. M., and Lenaerts, T. (2006), "Evolutationary Dynamics of Social Dilemmas in Structured Heterogenous Populations," *Proceedings of the National Academy of Sciences of the USA*, 103 (9), 3490–3494. [59]

Shelley, G. A., Killworth, P. D., Bernard, H. R., McCarty, C., Johnsen, E. C., and Rice, R. E. (2006), "Who Knows Your HIV Status II? Information Propogation Within Social Networks of Seropositive People," *Human Organization*, 65 (4), 430–444. [61]

UNAIDS (2003), "Estimating the Size of Populations at Risk for HIV," Number 03.36E, UNAIDS, Geneva. [69]

Volz, E. (2006), "Tomography of Random Social Networks," working paper, Cornell University, Dept. of Sociology. [65]

Wasserman, S., and Faust, K. (1994), *Social Network Analysis*, England and New York: Cambridge University Press. [62]

Zheng, T., Salganik, M. J., and Gelman, A. (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, 409–423. [60,61,63-66]