How many people do you know?: Efficiently estimating personal network size [*]

Tyler H. McCormick[†‡] Matthew J. Salganik[§‡]

Tian Zheng[†‡]

[†] Department of Statistics, Columbia University, New York, New York, 10027

[§] Department of Sociology, Princeton University, Princeton, New Jersey, 08544

[‡] These authors contributed equally to this work.

September 17, 2008

Abstract

In this paper 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

[*] The authors thank Peter Killworth, Russell Bernard and Chris McCarty for sharing their survey data as well as Andrew Gelman, Thomas DiPrete, and Delia Baldassari for their constructive comments. They also gratefully acknowledge the support of the National Science Foundation through grant number DMS-0532231 and a Graduate Research Fellowship, the Institute for Social and Economic Research and Policy and the Applied Statistics Center at Columbia University.

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 to be asked about are chosen properly, the simple scale-up degree estimates can enjoy the same bias-reduction as that from the our more complex latent non-random mixing model.

Keywords: Social Networks; Survey Design; Personal Network Size; Negative Binomial Distribution; Latent Non-random Mixing Model

1 Introduction

Social networks have become an increasingly common framework for understanding and explaining social phenomena. Yet, 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 paper we add to the toolkit of researchers interested in network phenomena by developing methodology to address two fundamental questions posed in the seminal paper of Pool and Kochen (1978): first, we would like to know for any individual, 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).[1]

[1] At first glance it may seem that any method which could estimate the degree distribution could also estimate the degree of an individual, but that is not the case. For example, the scale-up method of Killworth et al. (1998b), which will be described in greater detail later in this paper, provides estimates of the degree distribution, but cannot be used to estimate the degree of a specific individual. It is also the case that some methods that allow a researcher to estimate the social network size of an individual are so costly that they cannot

Recently, the second question, that of degree distribution, has received the most attention because of interest in so-called "scale-free" networks (Barabási, 2003). Some networks, particularly technological networks, appear to have power-law degree distributions (i.e., $p(d) \sim d^{-\alpha}$ for some constant α), and a number of mathematical and computational studies have found that this extremely skewed degree distribution may affect the dynamics of processes happening on the network including the spread of diseases and the evolution of group behavior (Pastor-Satorras and Vespignani, 2001; Santos et al., 2006). However, the actual functional form of the degree distribution of the acquaintanceship network is not known, and that question has become so central to some researchers that Killworth et al. (2006) went so far as to declare that estimating the degree distribution is "one of the grails of social network theory."

While estimating the degree distribution is certainly important for understanding how networks affect the dynamics of social processes, the ability to quickly estimate the personal network size of an individual is probably of greater long-term 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 "meatgrinder" that completely removes people from their social contexts. Having a survey instrument which allows for the collection of social content would allow researchers to address a 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.

This paper develops a method to estimate both individual network size and degree distribution in a population using a battery of questions that can

be employed on the scale required to estimate the degree distribution in a population.

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 known to suffer from three shortcomings: transmission errors, barrier effects and recall error. In Section 3 we propose a latent non-random mixing model which resolves these problems. As a byproduct of the latent non-random mixing model we also obtain new information about the mixing patterns in the acquaintanceship network that we believe will be of substantive value to the social science community. 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. Most importantly we show that future researchers can achieve improved network size estimates without complex statistical computation if the names asked about are chosen properly. We conclude with a discussion of the limitations of this method, specifically how additional demographic information for first names (currently collected but not released by the Census Bureau) could improve network size estimates.

2 Previous research

The most straightforward method for estimating the personal network size of a respondent would be to simply ask them how many people they "know." Although we are not aware of any direct evidence that this procedure works poorly, we suspect it would not be very accurate because of the well-documented problems with self-reported social network data (Killworth and Bernard, 1976; Bernard et al., 1984; Brewer, 2000; Butts, 2003). A number of more clever attempts have been made to address these questions and we will review them here. Because of space constraints we will not, however, review the larger literature on network data collection in general; interested readers should see Marsden (1990, 2005).

In the literature, we have identified four main methods attempting to estimate individual personal network size—the reverse small-world method, the summation method, the diary method, and finally the phonebook method/scale-up method—and these methods will now be described with strengths and weaknesses summarized in Table 1.[2]

One of the earliest methods for estimating personal network size was the *reverse small-world* method (Killworth and Bernard, 1978; Killworth et al., 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, the required number of targets is quite large; most studies use 500 targets which at a rate 15 seconds per target would take more than 2 hours to complete. Also, this procedure of searching in one's social network for an appropriate contact is difficult to model (Watts et al., 2002) and therefore is hard to embed within a statistical framework that would allow for formal inference and estimation of sampling uncertainty (i.e., standard errors).[3]

[2] The random subgraph method described in Granovetter (1976) is excluded because it only allows for the estimation of the *average* personal network size in a population.

[3] The reverse small-world procedure bears some resemblance to the "name generator" procedures which are frequently employed in network research (Burt, 1984; Campbell and Lee, 1991; Marsden, 2005). In these procedures, respondents are asked "name generator" questions which elicit names and then "name interpreters" to probe the respondent's relationship with the named individual. For example, the 1985 and 2004 General Social Survey included social network modules where respondents were asked to list people with whom "they discussed important matters." This approach tends to produce "core" discussion networks which are much smaller (around 3 people) than the total number of acquaintances (Burt, 1984; Marsden, 1987; McPherson et al., 2006). Others have employed different name generators which yield weaker ties, for example Fischer (1982).Ultimately name generators depend

Method	embeddable in survey	statistical modeling
reverse small-world method	no	no
summation method	yes	no
diary method	no	no
phonebook/scale-up method	yes	yes

Table 1: Strengths and weaknesses of methods for estimating person network size.

An additional procedure that cannot be modeled statistically, but which can be embedded in a survey is the *summation method* (McCarty et al., 2001). In this method, 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. McCarty et al. (2001) propose 16 relation types which when added together should yield the total personal network size.[4] Unfortunately, since it is not possible to construct a list of mutually exclusive groups, this procedure will lead to double counting (e.g., someone who is a coworker can also be a neighbor) and respondents may not be able to answer these questions accurately.

In addition to the reverse small-world method and the summation method, there are two methods that were originally proposed by Pool and Kochen which have had substantial impact on later work. The *diary method* required subjects to keep a daily record of all known people encountered over the span of 100 days. This method, while yielding very rich and accurate data, requires too much cooperation and time to be employed in routine sample surveys.[5]

on respondents enumerating alters one by one and thus will never be able to quickly measure an acquaintance network that could number in the hundreds or thousands.

[4] The 16 categories are: immediate family, other birth family, family of spouse or significant other, coworkers, people at work but don't work with directly, best friends/confidantes, people known through hobbies/recreation, people from religious organization, people from other organization, school relations, neighbors, just friends, people known through others, childhood relations, people who provide a service, and, lastly, other (McCarty et al., 2001).

[5] Gurevich (1961) employed this method for his dissertation research on social networks, but because of the logistical difficulties with employing the diary method, he was only able to study 27 people.

Later efforts have attempted to reduce the burden on respondents by using data on contacts that are recorded automatically; for example, Christmas card mailing lists (Hill and Dunbar, 2003), email logs (Kossinets and Watts, 2006), or cell-phone records (Onnela et al., 2007). Because these methods are not embeddable within the standard sampling survey framework, however, their general applicability in the context of this paper is limited.

The second method proposed by Pool and Kochen, however, has the potential to be employed in a survey framework and is amenable to statistical modeling. This method—the *phone book method*—has also received the most subsequent development. In its original form, 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 this procedure was then developed further as the scale-up method (Killworth et al., 1998b).

We believe the scale-up method holds the greatest potential for getting accurate estimates quickly with reasonable estimates of uncertainty. The scale-up method, however, 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 will describe the scale-up method and these three issues in detail. Section 2.2 presents an earlier model by Zheng et al. (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 Δ such that $\delta_{ij} = 1$ if person i knows person j.[6] The personal network size or degree of person i is then $d_i = \sum_j \delta_{ij}$.

The most direct method to estimate the personal network size of an individual, d_i, would be to collect a simple random sample of n other population members and ask person i if she knows each of these others. Inference for individual degrees is then based on the fact that the number of person i's acquaintances among these n randomly sampled individuals from the population, follows approximately a binomial distribution with size n and probability d_i/N. In 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 et al. (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 respondents 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 the respondent 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 you report knowing 3 women who gave birth this represents about one-millionth of all women who gave birth within the

[6] What it means to "know" someone is itself a complex issue. Throughout this paper we will assume McCarty et al.'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."

last year. We could then use this information to estimate that you know about one-millionth of all Americans,

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

The accuracy 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 that 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. (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)

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

Our simple degree estimate $\hat{d}_i$ then has standard error

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

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

$$\mathrm{SE}(d_i) \approx \sqrt{\hat{d}_i} \cdot \sqrt{\frac{1-\frac{\sum_{k=1}^K N_k}{N}}{\frac{\sum_{k=1}^K N_k}{N}}} \approx \sqrt{\frac{1-\frac{3.6\text{ million}}{300\text{ million}}}{\frac{3.6\text{ million}}{300\text{ million}}}}\sqrt{250} = 143.5.$$

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 61.5. In Figure 1, we plot $\mathrm{SE}(\hat{d}_i)/\sqrt{\hat{d}_i}$ 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 percent (or approximately 60 million in a population of 300 million). After roughly 20 percent adding additional subpopulations to the survey still decreases the estimated standard error but at a slower rate.[7]

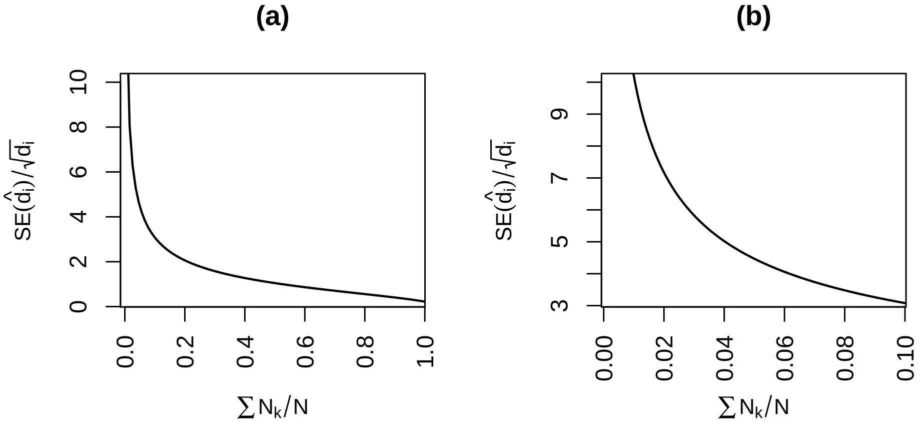

Figure 1: Standard error of the simple scale-up degree estimate (scaled by the square root of the true degree) plotted against fractional subpopulation size. As we increase the fraction of population represented by survey subpopulations, the precision of the estimate improves, with diminishing improvements after about 20%.

The original studies using the scale-up method relied on 29 subpopulations including some defined by first name (e.g., Michael, Christina), occupation (e.g., postal worker, pilot, gun dealer), ethnicity (e.g., Native American), or medical condition (e.g., diabetic, on kidney dialysis); a complete list can be found in McCarty et al. (2001).[8]

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

[7] The above error estimates depend on subpopulation sizes only through their sum and thus one obtains the same reduction in estimated error by asking one question about a large subpopulation as would be obtained by asking several questions about smaller subpopulations. This, however, is not true in practice since, as we will shown in this paper, one single large subpopulation introduces more recall error and bias due to social structure than multiple smaller subpopulations would.

[8] The survey actually included 32 subpopulations, but for three of them researchers lacked an estimated population size. For these three groups, the scale-up method was used in reverse to estimate their group sizes given estimated individual degrees.

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 et al., 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, they have prevented previous researchers from using the scale-up method to estimate the degree of any particular individual.[9]

A final source of error is that responses to these questions are prone to recall error. To understand this problem, consider how you would answer the question, "How many people do you know named Michael?" For many people

[9] However, since researchers believed that these biases likely canceled out for sufficiently large samples, they have still used the scale-up method to estimate the degree distribution (McCarty et al., 2001).

this is not an easy question and there is evidence that people cannot answer accurately (Killworth et al., 2003). 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 2, 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 et al., 2006).

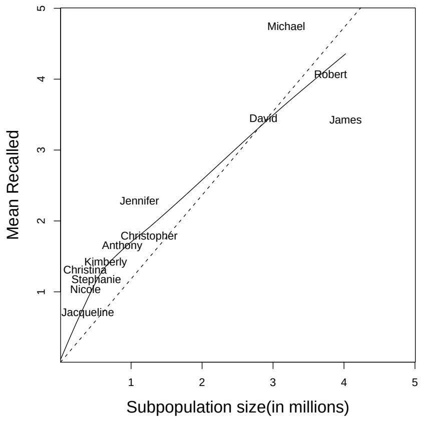

Figure 2: Mean number recalled as a function of subpopulation size for 12 names. If respondents recall perfectly, then we would expect the mean number recalled to increase linearly as the subpopulation size increases. The best-fit line and loess curve show that this was not the case suggesting that there is recall error.

2.2 The Zheng et al. (2006) Model with Overdispersion

Before presenting our model for estimating person network size using "How many X's do you know?" data, it is important to review the multilevel overdispersed Poisson model of Zheng et al. (2006) which, rather than treating non-random mixing (i.e., barrier effects) as an impediment to network size estimation, treated it as something important to estimate for its own sake. Zheng et al. (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), on average, as popular as the rest of the population, then $b_k \approx N_k/N$.

The responses to many of the questions in the McCarty et al. (2001) data did not follow a Poisson distribution, however. In fact, most of the responses show overdispersion, that is, excess variance given the mean. 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 someone in prison than others. To model this increased variance Zheng et al. (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} can be modeled with a negative binomial distribution,

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

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

Despite being developed to estimate ω_k, the Zheng et al. model also produces personal network size estimates, d_i. However, these estimates are still susceptible to the problems of transmission effects, barrier effects, and recall bias that plague personal network size estimation based on the "How many X's do you know?" data.

3 A new statistical method for degree estimation

We now develop a new statistical procedure to address the three known problems with estimating individual degree using the "How many X's do you know?" data: transmission effects, barrier effects, and recall bias. Transmission errors, while probably the most difficult to quantify, are also the easiest to eliminate. We will limit our analysis to the 12 subpopulations defined by first names that were asked about in McCarty et al. (2001). These 12 names, half male and half female, are presented in Figure 3 with their age profiles. Though McCarty et al.'s definition of knowing someone (see footnote 6 on page 8) 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's unlikely that you know someone, but don't know his/her first name.[10] Even though using only first names controls transmission errors, it does not address bias from barrier effects or recall bias. In the remainder of this section, we propose a latent non-random mixing model

[10] A definition of "know" that requires respondents to know an acquaintance's name would address this issue. One could define "know", for example, as was done in the 2006 General Social Survey: "people that you are acquainted with (meaning that you know their name and would stop and talk at least for a moment if you ran into the person on the street or in a shopping mall)." However, simply changing the definition of "know" may not have a strong effect on how respondents actually answer the questions.

to address these two issues.

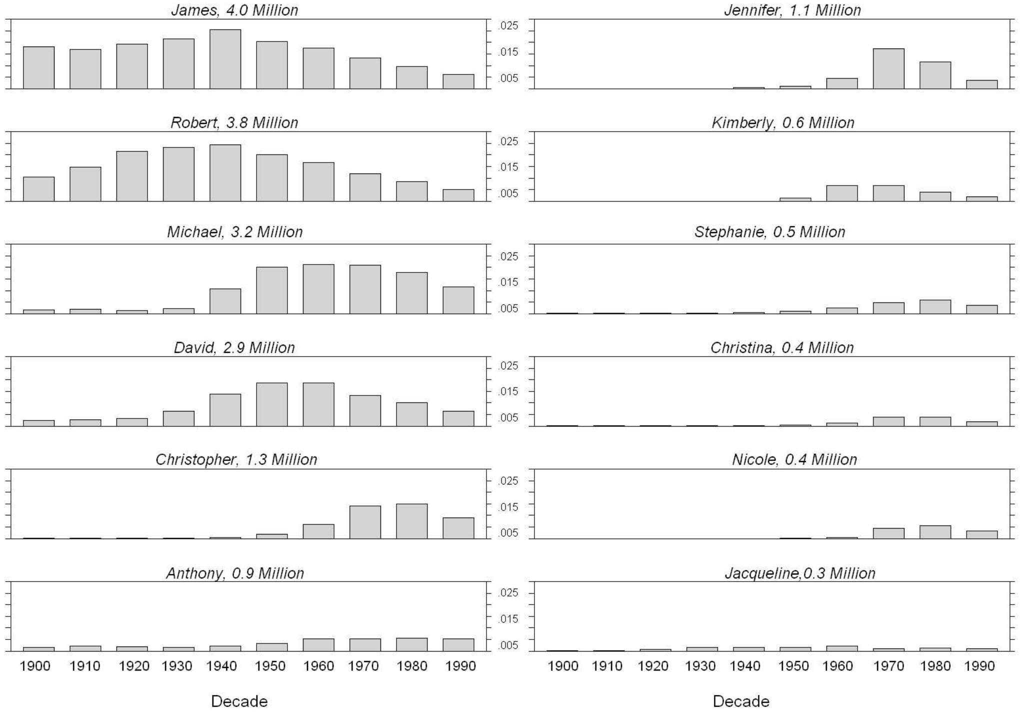

Figure 3: Age profiles for the 12 names used in the analysis (data source: Social Security Administration). The heights of the bars represent the percent of American newborns in a given decade with a particular name. The total subpopulation size is given across the top of each graph. The male names chosen by McCarty et al. are much more popular than the female names. These age profiles are the information required to construct the matrix of $\frac{N_{ak}}{N_a}$ terms.

3.1 Latent non-random mixing model

We begin by considering the impact of barrier effects, or non-random mixing, on degree estimation. Figure 4 gives a graphical representation of a hypothetical 30-year old male with the shaded oval representing the breadth of the individual's network. Following standard network terminology (Wasserman and Faust, 1994), we refer to the respondent as *ego* and the people to whom he can form ties as *alters*. In this case, the respondent's alters are divided in 8 alter groups based on age and gender, though one could divide the network based on additional characteristics as well. This example captures the well-documented tendency for people to form ties to alters that are similar (McPherson et al., 2001). In this case, 30 percent of ego's network is made up of individuals in the most similar alter category (males 21-40) while only 1 percent of his ties are to the more socially distant alter category of females 61 and older.

If we ignore non-random mixing and ask this respondent how many Michaels he knows, we will overestimate the size of his network using the scale-up method because Michael tends to be a more popular name among younger males (Figure 3). If we asked how many Roses he knows, in contrast, we would underestimate the size of his network since Rose is a name that is more common with older females. In both cases, the properties of the estimates are affected by the demographic profiles of the names that are used in the estimate. The simple scale-up method, however, does not account for this problem.

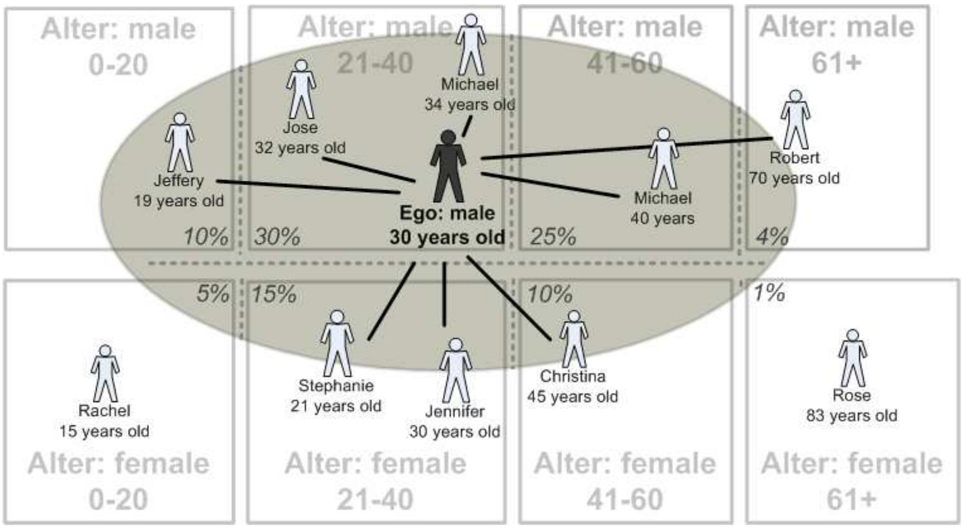

Figure 4: Non-random mixing by age and gender in an example ego's network. In the latent non-random mixing model $m(e, a)$ allows the propensity of ties to vary based on characteristics of both the alters and the egos. Here, by *ego*, we refer to a survey respondent and by *alter*, we refer to a member of subgroup a with whom the respondent could potentially form ties.

We model the non-random mixing sketched in Figure 4 using a negative binomial model. Specifically, we model explicitly the propensity for a respondent in ego-group e to know members of alter group a. The model is then

$$\begin{aligned} y_{ik} &\sim \text{Neg-Binom}(\mu_{ike}, \omega'_k) \qquad (7) \\ \text{where } \mu_{ike} &= d_i \sum_{a=1}^{A} m(e,a)\, \frac{N_{ak}}{N_a} \qquad (8) \\ & \qquad (9) \end{aligned}$$

and d_i is the degree of person i, e is the ego group that person i belongs to, N_{ak}/N_a is the relative size of name k within alter-group a (for example, 4% of males between ages 21 and 40 are named Michael), and $m(e,a)$ is the mixing coefficient between ego-group e and alter-group a that we will define in the following:

$$m(e,a) = \text{E}\left(\frac{d_{ia}}{d_i = \sum_{a=1}^{A} d_{ia}} \middle| i \text{ in ego group } e \right) \qquad (10)$$

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, $\sum_{a=1}^{A} m(e,a) = 1$.

Therefore, the number of people that person i knows in group 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 are in group k, (N_{ak}/N_a), and the mixing rate between people in group e and people in group a, ($m(e,a)$). Additionally, if we do not observe non-random mixing, then $m(e,a) = N_a/N$ and μ_{ike} in (7) reduces to $d_i b_k$ in (6).

In addition to μ_{ike}, the latent non-random mixing model also depends on the overdispersion, ω'_k, which represents the variation 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 non-random mixing between the defined alter and ego groups. Explicitly modeling this variation should cause a reduction in overdispersion parameter ω'_k when compared to ω_k in (6) and Zheng et al. (2006). The term ω'_k is still present in the latent non-random mixing model, however, since there is still residual overdispersion based on additional ego and alter characteristics that could effect their propensity to form ties.

Fitting the model in practice requires choosing the number of ego groups, E, and alter groups, A. In this case, we classified egos by gender and by three age categories—youth (18-24), adult (25-64) and senior (65+)—giving a total of six ego groups. Constructing the alter groups requires information on the demographic profiles of the names used. The only systematic source of this data that we could find was from the Social Security Administration (SSA) which provides decade-by-decade tables of the proportion of births in a decade made up of children with a particular name.[11] We use these tables as a proxy for the relative proportion of individuals with a particular name in the population.[12] Based on this information we opted for eight alter categories based on crossing gender (2 categories) and age (4 categories: 0-20, 21-40, 41-60, 61+). Together this gives us a total of 48 mixing parameters, $m(e, a)$, to estimate

[11] The Census Bureau collects respondent's first names so they have all the information that would be needed to compile the demographic profile of first names, but they do not routinely release this information and our efforts to acquire this information were ultimately unsuccessful.

[12] By using the SSA data we are making three implicit assumptions. First we assume that life expectancy is uncorrelated with an individual's first name, since the data provide information only about the number of individuals born with a certain name and not the number alive at the time of the survey. This may not be true in all cases. If a particular name is more common with individuals living in poverty, for example, the life expectancy for individuals with that name is likely overall lower. Second, we assume that individuals with a particular first name are not predisposed to any behavior such as immigration or emigration that would disproportionately alter the number of individuals in the population with that name at the time of data collection. We do not believe that this assumption is problematic for the names in the McCarty et al. survey. Third, we must assume that the factors that influenced individuals' decisions about registering for social security are uncorrelated with first names, even during the early years of Social Security (which was established in 1935) when registration was not universal. Again, there are conditions under which this conditions could be violated, but there is little reason to believe that is the case here. Having data on the demographic profiles of the first names from the Census Bureau would reduce the need for such assumptions.

(6 ego groups by 8 alter groups). We believe that this represents a reasonable compromise between parsimony of parameters and obtaining detailed information. Our model is flexible in this regard in that we could further stratify based on available information of either the respondents or the demographic profiles of names. This flexibility is especially important for survey researchers who often have much more demographic information about respondents than is currently available for the names.

3.2 Correction for recall error

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

$$\frac{5}{4.8 \text{ million}/300 \text{ million}} \approx 300 \text{ people.} \tag{11}$$

However, 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 et al., 2006). Consistent with the approach taken in Killworth et al. (2003), we could choose to address this issue in two ways which, though ultimately equivalent, suggest two distinct modeling strategies.

First, we could assume that the respondent is inaccurately recalling the number of people named Michael she knows from her true network. Under this framework, any correction we propose should increase the numerator in (11). 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 5 Michaels reported and applying some function to produce a corrected response (presumably some number greater than 5), which would then be used to fit the proposed model. It is difficult, however, to speculate about the nature of this function in any detail.

Another approach would be to assume that respondents are not recalling from their actual network, but rather from a *recalled network* which 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 et al., 2006; McCormick and Zheng, 2007) suggests that the calibration curve, $f(\cdot)$ should impose less correction for smaller subpopulations and a progressively greater correction as the popularity of the subpopulation increases.

We devise such a calibration curve as described in the Appendix and apply it to our model as follows:

$$
\begin{aligned}
y_{ik} &\sim \text{Neg-Binom}(\mu_{ike}, \omega'_k) \\
\text{where} & \\
\mu_{ike} &= d_i f\left(\textstyle\sum_{a=1}^{A} m(e,a)\,\frac{N_{ak}}{N_a}\right).
\end{aligned} \tag{12}
$$

3.3 Model Fitting Algorithm

We use a multilevel model and Bayesian inference to estimate d_i, $m(e,a)$, and ω'_k in the latent non-random 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 et al. (2006) postulate 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 $\log(m(e,a))$ 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, $p(1/\omega'_k) \propto 1$. Since ω'_k is constrained to (1,∞), the inverse falls on (0,1). The Jacobian for the transformation is ω'^{-2}_k. Finally, we give noninformative uniform priors to the hyperparameters μ_d, $\mu_{m(e,a)}$, σ_d and $\sigma_{m(e,a)}$. The joint posterior density can then be expressed as

$$
\begin{aligned}
p(d, m(e,a), \omega', \mu_d, \mu_{m(e,a)}, \sigma_d, \sigma_{m(e,a)}|y) \;\propto\; & \prod_{k=1}^{K}\prod_{i=1}^{N} \binom{y_{ik}+\xi_{ik}-1}{\xi_{ik}-1} \left(\frac{1}{\omega'_k}\right)^{\xi_{ik}} \left(\frac{\omega'_k-1}{\omega'_k}\right)^{y_{ik}} \\
& \times \prod_{i=1}^{N} \left(\frac{1}{\omega'_k}\right)^{2} N(\log(d_i)|\mu_d, \sigma_d) \\
& \times \prod_{e=1}^{E} N(\log(m(e,a))|\mu_{m(e,a)}, \sigma_{m(e,a)})
\end{aligned} \tag{13}
$$

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

Adapting Zheng et al. (2006), we use a Gibbs-Metropolis algorithm in each iteration v.

1. For each i, update d_i using a Metropolis step with jumping distribution $\log(d_i^*) \sim N(d_i^{(v-1)}$,(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 $\log(m(e, a)^*) \sim N(m(e, a)^{(v-1)}$, (jumping scale of $m(e, \cdot))^2)$. Then, rescale so that the row sum is one.

3. Update $\mu_d \sim N(\hat{\mu}_d, \sigma_d^2/n)$ where $\hat{\mu}_d = \frac{1}{n}\Sigma_{i=1}^{n} d_i$

4. Update $\sigma_d^2 \sim \text{Inv-}\chi^2(n-1, \hat{\sigma}_d^2)$, where $\hat{\sigma}_d^2 = \frac{1}{n} \times \Sigma_{i=1}^{n}(d_i - \mu_d)^2$.

5. Update $\mu_{m(e,a)} \sim N(\hat{\mu}_{m(e,a)}, \sigma_{m(e,a)}^2/n)$ for each e where $\hat{\mu}_{m(e,a)} = \frac{1}{A}\Sigma_{a=1}^{a} m(e, a)$

6. Update $\sigma_{m(e,a)}^2 \sim \text{Inv-}\chi^2(a-1, \hat{\sigma}_{m(e,a)}^2)$, for each e where $\hat{\sigma}_{m(e,a)}^2 = \frac{1}{A} \times \Sigma_{a=1}^{A}(m(e, a) - \mu_{m(e,a)})^2$

7. For each k, update ω_k' using a Metropolis step with jumping distribution $\omega_k'^* \sim N(\omega_k'^{(v-1)}$,(jumping scale of $\omega_k')^2)$.

4 Results

To fit the model we used data from McCarty et al. (2001) which consisted of survey responses from 1,370 adults living in the United States who were contacted via random digit dialing.[13] We obtained approximate convergence

[13] More specifically, we used survey 1 (796 respondents, January 1998) and survey 2 (574 respondents, January 1999). Sometimes responses were categorized, in which cases we used the central value in the bin (e.g., imputing 7.5 for the response 5-10). To correct for responses that were suspiciously large (e.g,, a person claiming to know over 50 Michaels), we truncated all response at 30, a procedure which affects only 0.25% of the data. We also inspected the

of our algorithm ($\hat{R}_{max} < 1.1$; see Gelman et al. (2003)) using three parallel chains with 2000 iterations per chain. We used the first half of each chain for burn-in and thin the chain every ten iterations. As we will demonstrate using a simulation study in Section 5.2 (see Figure 11) , the latent non-random mixing model estimates with more accuracy both the degree distribution and the individual degrees of the respondents, a major improvement over previous methods. Next we will present those estimates.

4.1 Personal network size estimates

We estimated a mean network size of 611 (median = 472) and the distribution of network sizes in presented in Figure 5. The solid line in Figure 5 is 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 quiet well.[14] 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 et al., 2007). The fit is clearly poor, a result consistent with previous work showing that another social network—the sexual contact network—is also poorly approximated by the power-law distribution (Hamilton et al., 2008). Together these results suggests that some of the interest around power-law degree distributions in social networks may be misplaced.

The estimated distribution is also presented separately for males and fe-

data using scatterplots which revealed a respondent who was coded as knowing seven people in each subpopulation. We removed this case from the dataset.

[14] Though we state in Section 3.3 that the prior distribution for the degree is lognormal, the observed lognormal distribution in Figure 5 is not an artifact of our model. We confirmed this claim by performing additional simulation studies showing that the model will recover the true nature of the population distribution even if that distribution is not lognormal. In these experiments we first truncated the distribution estimates presented in Figure 5 at the median. We then used these (now not lognormally distributed) values to generate data using random draws from a negative binomial distribution. We fit our model to this artificial data and are able to accurately recover the data-generating degree distribution.

males in Figure 6. Overall, we estimate that the degree distribution for males is similar to the distribution of females, though males have slightly larger networks on average. Amongst male respondents, we estimate a median degree of approximately 500 (mean 640) and we expect 90 percent of males to have degree between 172 and 1581. For females we estimate a median degree of 452 (mean 590) with 90 percent of females expected to have degrees between 157 and 1488.

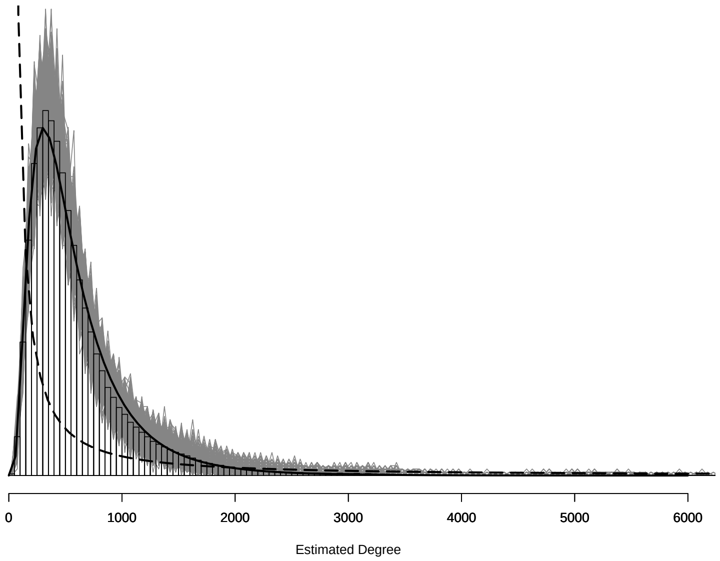

Figure 5: Estimated degree distribution from the fitted model. The median is about 470 and the mean is about 610. The shading represents posterior simulation draws to indicate inferential uncertainty in the histograms. The overlay line is a log-normal distribution with parameters estimated via maximum likelihood using the observed data ($\hat{\mu}_{mle} = 6.2$ and $\hat{\sigma}_{mle} = 0.68$). The dashed line is a power-law density with scaling parameter estimated via maximum likelihood ($\hat{\alpha}_{mle} = 1.28$)

Figure 7 compares the estimated degree from the latent non-random mixing model to estimates from the method of Zheng et al. (2006), a previous method that did not explicitly deal with the three known problems with estimating individual degrees using “How many X’s do you know?” data. In general, the estimates from the latent non-random mixing model tend to be slightly smaller

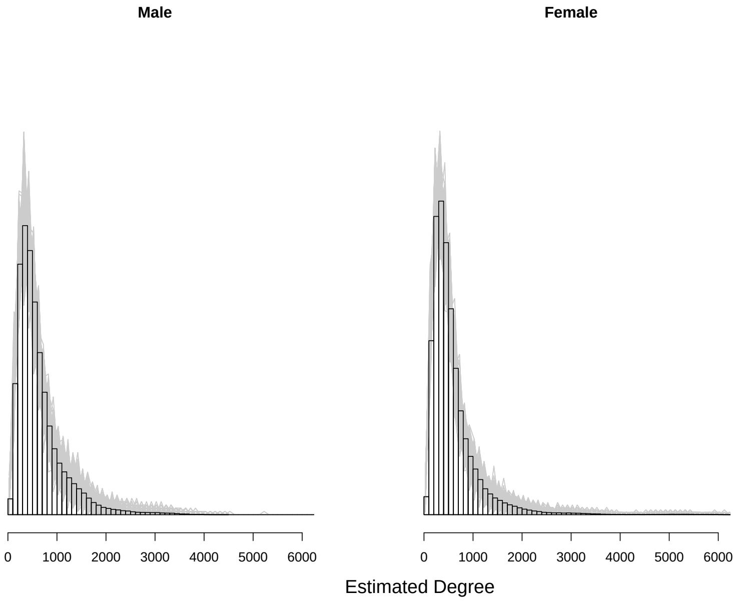

Figure 6: Estimated degree distribution by gender. Shading represents simulation draws from the posterior and implies inferential uncertainty. The average degree for males is slightly larger than for females. For males the median is about 500 (mean 640) while the median for females is about 452 (mean 590).

with an estimated median degree of 472 (mean 611) compared to an estimated median degree of 610 (mean 750) in Zheng et al. (2006). The difference in degree estimates is likely due to non-random mixing in the acquaintance network that was not addressed in the procedure of Zheng et al. (2006). Figure 7 also reveals a clear advantage of the latent non-random mixing estimates, that the correction in degree estimation differs among ego groups due to non-random mixing. For example, although there were six males names and six female names used in the McCarty et al. data, these names were of very different sizes so that the combined set of names queried included 16.1 million males but only 3.3 million females (see Figure 3). The Zheng et al. model does not account for this fact and thus overestimates the degree of male respondents. The latent non-random mixing model, however, incorporates this information and thereby produces the corresponding corrections. Similarly, Figure 3 demonstrates that the names used by McCarty et al. were most popular amongst adult respondents, which explains why Figure 7 shows the largest correction on the degree estimates of adult respondents.

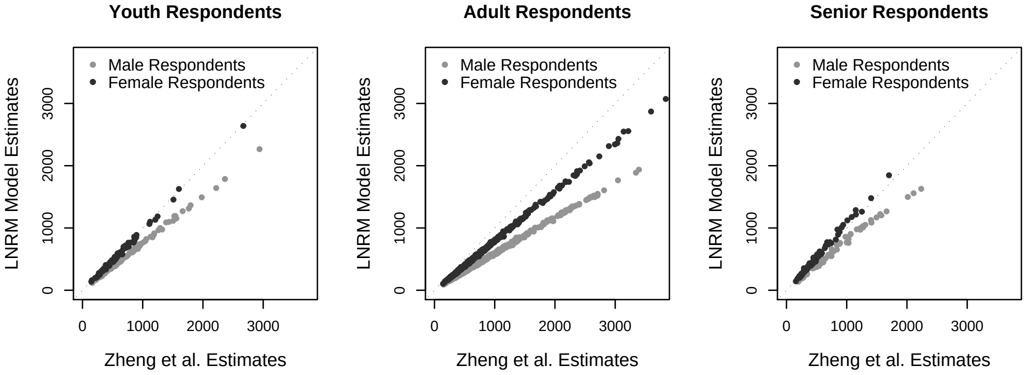

Figure 7: Comparison of the estimates from Zheng et al. and the latent non-random mixing (LNRM) model broken down by age and gender: grey points represent males and black points females. Whereas the Zheng et al. estimates do not account for non-random mixing, the latent non-random mixing model considers non-random mixing within each ego group individually. Since our model has six ego groups, there are six distinct patterns in the figure.

4.2 Mixing estimates

Though we developed this procedure to obtain good estimates of personal network size, it also gives us information about the mixing rates in the population. As far as we know, this is the first survey-based approach to estimate such information which is thought to be important for the spread of disease (Del Valle et al., 2007) and information (Volz, 2006). These results are presented in Figure 8. As mentioned in the previous section, the mixing matrix represents the proportion of the network of a person in ego group e that is made up of alter group a. In this case we have six respondent categories and eight alter categories, demonstrating that the number of groups can be manipulated to suit individual research questions.

In general, Figure 8 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 that is consistent with the large sociological literature on homophily—the tendency for people to form ties to those who are similar (McPherson et al., 2001). This trend is especially apparent for 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 non-random mixing model could be used to estimate the extent of network-based segregation, an approach that could have many advantages over traditional measures of residential segregation (Echenique and Fryer, 2007).

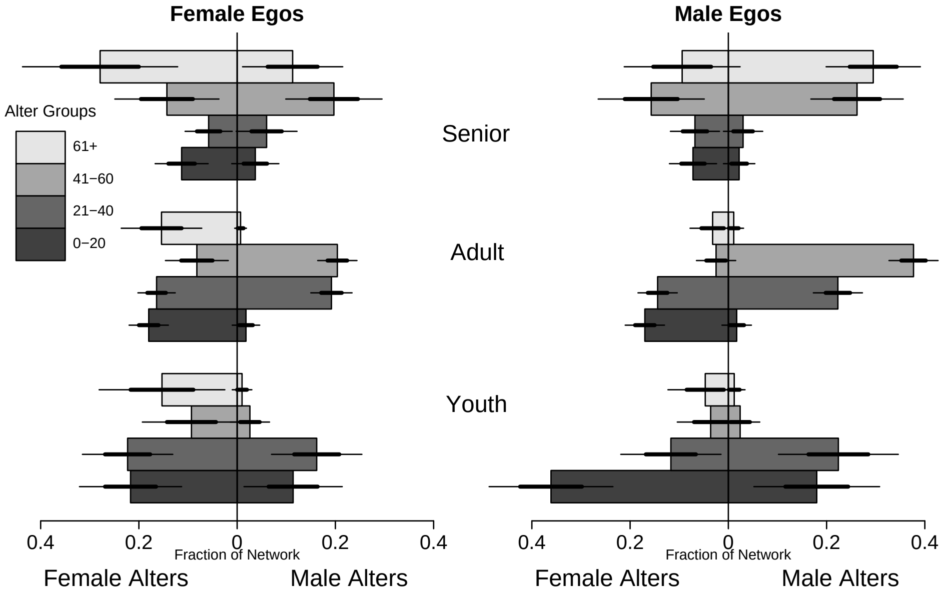

Figure 8: Barplot of the mixing matrix. Each of the six stacks of bars represents the network of one ego group. Each stack describes the proportion of the given ego group's ties that are formed with all of the alter groups; thus, the total proportion within each stack is 1. For each individual bar, a shift to the left indicates an increased propensity to know female alters. Thick lines represent +/- one standard error (estimated from the posterior) while thin lines are +/- two standard errors.

4.3 Overdispersion

Another way to assess the latent non-random 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 non-random 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 et al. (2006) since Zheng et al. (2006) does not model non-random mixing.

This prediction turned out to be correct. With the exception of Anthony, all of the estimated overdispersion estimates from the latent non-random mixing model are lower than those presented in Zheng et al. (2006). To judge the magnitude of the difference we create 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 non-random mixing explicitly in the latent non-random mixing model. In the denominator, an ω_k value of one corresponds to no overdispersion. The ratio for group k, therefore, is the proportion of overdispersion encountered in Zheng et al. (2006) that is explicitly modeled in the latent non-random mixing model. The standardized difference was on average 0.213 units lower for the latent non-random mixing model estimates than for the Zheng et al. estimates, indicating that roughly 21 percent of the overdispersion found in Zheng et al. (2006) can be explained by non-random mixing base due to age and gender. If appropriate ethnicity or other demographic information about the names were available, we expect this reduction to be even larger.

5 Designing future surveys

In the previous sections we analyzed existing data in a way that resolves 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 allow researchers to capitalize on the simplicity of the scale-up estimates while enjoying the same bias-reduction as in the latent non-random mixing model. The findings in this section, therefore, offer an efficient and easy-to-apply degree estimation method that is accessible to a wide range of applied researchers who may not have the training or experience necessary to fit the latent non-random mixing model.

In Section 5.1, we derive the requirement for selecting first names for the scale-up method so that the estimator is equivalent to the degree estimator from fitting a latent non-random mixing model using MCMC computation.[15] 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 30 then 20% of the people with the names used must also be females under 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 non-random mixing model accounts for barrier effects due to some demographic factors by estimating degree

[15] Those using the simple scale-up method will not, however, be able to estimate the social mixing parameters.

differentially based on characteristics of the respondent and of the potential alter population. If, however, there were conditions where the simple scale-up estimator was expected to be equivalent to the latent non-random mixing model, then the simple estimator would enjoy the same reduction of bias from barrier effects as the more complex latent non-random mixing model estimates. In this section we derive such conditions.

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

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

On the other hand, the scale-up estimator assumes

$$\begin{aligned} \mathrm{E}\left(\sum_{k=1}^{K} y_{ike}\right) &= \sum_{k=1}^{K} \mu_{ike} = d_i \sum_{a=1}^{A} m(e,a) \left[\sum_{k=1}^{K} \frac{N_{ak}}{N_a}\right] \\ &\equiv d_i \frac{\sum_{k=1}^{K}\sum_{a=1}^{A} N_{ak}}{N}, \ \forall e. \end{aligned} \tag{14}$$

With the third equality in (14), the Killworth et al. scale-up estimator (2) is in expectation equivalent to that of the latent non-random mixing model. This equality holds if

$$m(e,a) = \frac{N_a}{N}, \ \forall a, \ \forall e, \tag{15}$$

or

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

In other words, the two estimators are equivalent if there is random mixing (15) or if the combined set of names represents a “scaled-down” version of the population (16). Since random mixing is not a reasonable assumption for the acquaintances network of Americans, 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 ages 21 and 40 ($\frac{N_a}{N}$) then 15% of the people asked about must also be males between ages 21 and 40 ($\frac{\sum_{k=1}^{K} N_{ak}}{\sum_{k=1}^{K} N_k}$).

In actually choosing a set of names to satisfy the scaled-down condition, we found it more convenient to work with a rearranged form of (16):

$$\frac{\sum_{k=1}^{K} N_{ak}}{N_a} = \frac{\sum_{k=1}^{K} N_k}{N}, \forall a. \tag{17}$$

In order to find a set of names that satisfy (17) we found it helpful to create Figure 9 that displays the relative popularity of many names over time by gender; lighter color represent greater popularity. Given the trendiness of names demonstrated in Figure 9, we tried to select a set of names such that the popularity across alter categories ended up balanced. For example, if we consider selecting names to address barrier effects due to age, the names selected should be popular in a particular time period then unpopular in the remaining time periods.

Consider selecting three names from Figure 9: Walter, Bruce and Kyle. Walter was most popular amongst those born from 1910-1940, but relatively unpopular otherwise, whereas Bruce was popular during the middle of the century and Kyle near the end. Thus, the effects of names at any one time period will be canceled out by the popularity of names in the other time periods, preserving the required equality in the sum (17).[16]

When choosing what names to use, in addition to satisfying (17), it is also important consider the overall popularity of the names. For efficiency's sake

[16] Note that this only works well if the names Walter, Bruce, and Kyle have similar overall popularity because if one is much more popular than the others, it will dominate the combined set of names. The McCarty et al. names for example have gender profiles that are reasonably balanced but the male names are overall much more popular than the female names, see Figure 3.

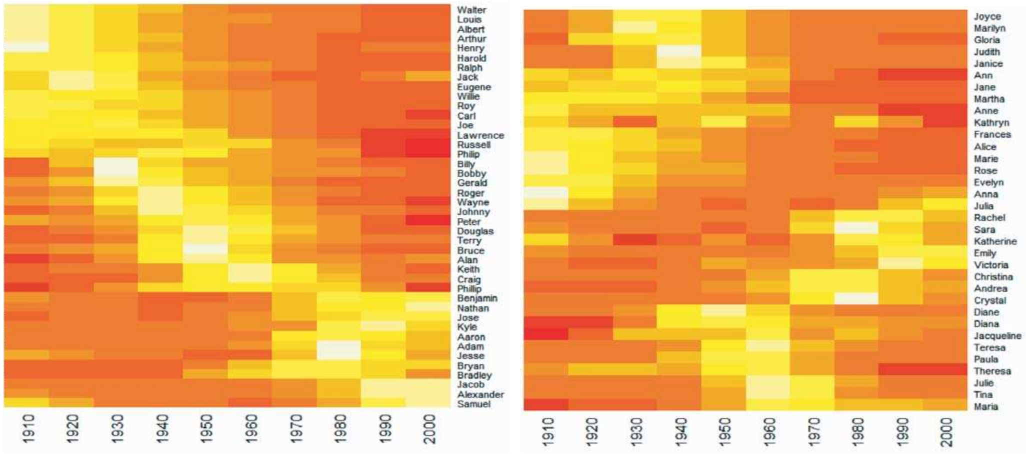

Figure 9: Heat maps of additional male and female names based on data from the Social Security Administration. Lighter color indicates higher popularity. Three distinct popularity profiles are clear in the male names. The clustering of female popularity profiles is still present, but less clear for the male names.

we suggest using names that are not too obscure, but we also suggest avoiding names that are too popular in order to minimize the recall problems described in Section 3.2. Generally, we found that names comprising approximately 0.1 to 0.2 percent of the population were easiest to work with because at this level there are few recall problems and the average response ranges from 0.6-1.3 from the respondents. Choosing names that are not commonly associated with nicknames will also help to minimize transmission errors.

5.2 Simulation Study

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

The name-age profiles presented in Figure 9, are all of the desired popularity (between 0.1 and 0.2 percent of the population). We have plotted the figures separately for male and female names and clustered names based on age profiles. In the figure for male names, there are three general clusters, roughly corresponding to the younger, middle-aged, and older respondents. For women the pattern is less clear, which is expected since female names tend to be more trendy for short time periods.

We demonstrate our guidelines by selecting two sets of names directly from these figures. The first set of names in Table 2—the *good names*—were selected using the procedure described in the previous section. This represents our best attempt to choose a set of names that will produce reasonable estimates from the simple scale-up estimator. As a point of comparison we also selected another set of the names—the *bad names*—that were popular with individuals born in the first decades of the twentieth century. As a final point of comparison, we will also use the set of 12 names from the McCarty et al. data.

Figure 10 provides a visual check of the scaled-down condition (15) for these three sets of names by plotting the combined demographic profiles for each set compared to that of the overall population. The figure reveals clear problems with the McCarty et al. names and the bad names. In the bad names, for example, a much larger fraction of the subpopulation of alters is made up of older individuals than in the population overall (as expected given our method of selection). Thus, we expect that scale-up estimates based on

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

Table 2: 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.

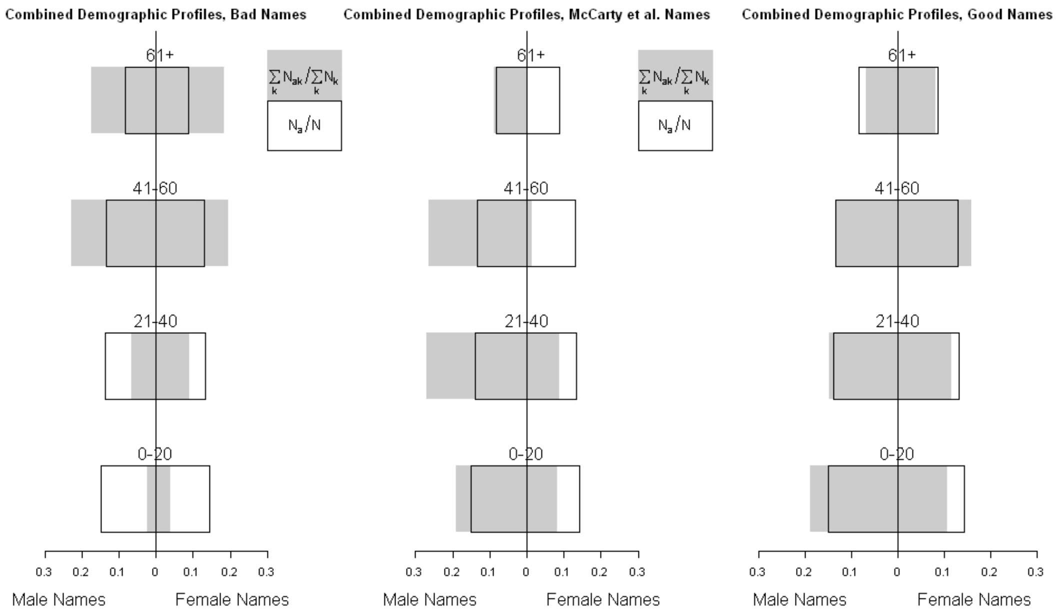

Figure 10: Combined demographic profiles for three sets of names (shaded bars) and population proportion of the corresponding category (solid lines).

the bad names will over-estimate the degree of older respondents.

We assessed this prediction via a simulation study that fit the latent non-random mixing model to the McCarty et al. data and then used these estimated parameters (degree, overdispersion, mixing matrix) to generate a negative binomial sample of size 1,370. We then fit the scale-up estimate, the latent non-random mixing model and the Zheng et al. model to this simulated data to see how these estimates could recover the known data-generating parameters.

Figure 11 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 age of the respondent. For the bad names (Table 2) individual degree is systematically over-estimated for older individuals and under-estimated for younger individuals in all three models, but the latent non-random mixing model showed the least age bias in estimates. This over-estimation of the degree of older respondents was expected

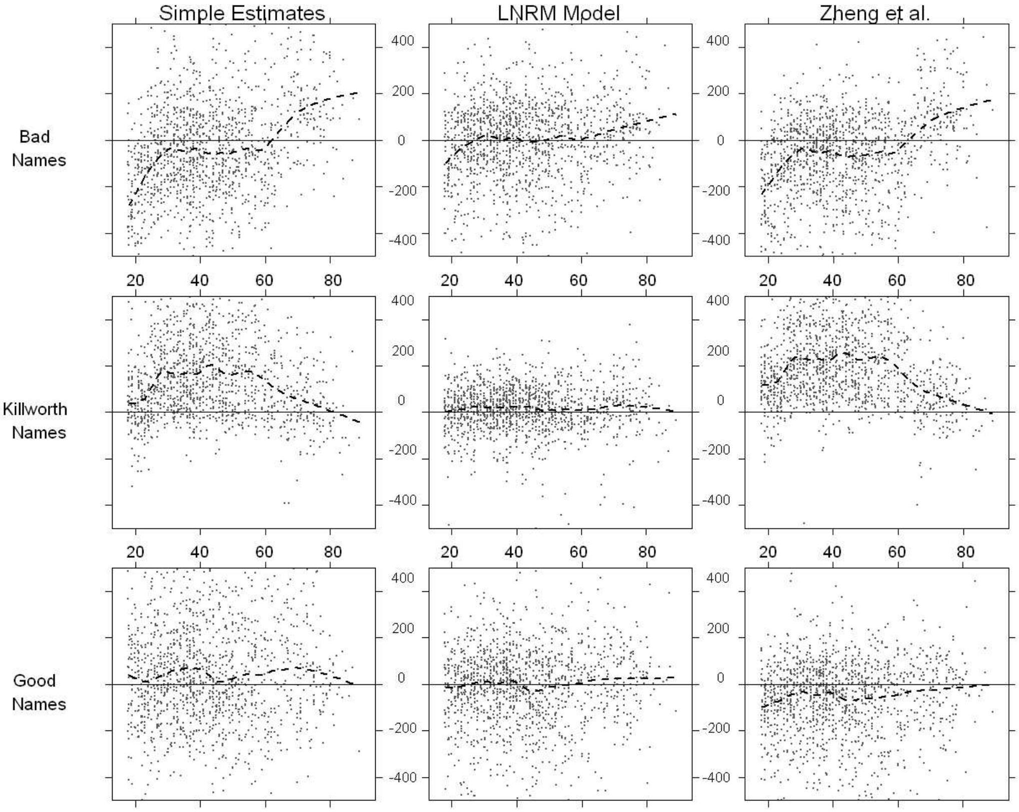

Figure 11: A comparison of the performance of the latent non-random mixing model, the Zheng et al. overdispersion model, and the Killworth et al. scale-up method. In each panel the difference between the estimated degree and the known data-generating degree is plotted against age. Three different sets of names were used: a set of names that do not satisfy the scaled-down condition (bad names), the names used in the McCarty et al. survey, 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 non-random mixing model. With the McCarty et al. names, the scale-up estimate and the Zheng et al. estimates show age bias, but the estimates from the latent non-random mixing model are excellent. With the good names, all three procedures preform well.

given the combined demographic profiles of the set of bad names (Figure 10). For the names from the McCarty et al. (2001) survey the scale-up estimator and the Zheng et al. model over-estimate the degree of the younger members of the population, again as expected given the combined demographic profiles of this set of names (Figure 10). The latent non-random mixing model, however, produced 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 non-random mixing model preformed better than existing methods when names were not chosen according to the scaled-down condition, suggesting that it is the best approach from estimating personal network size with most data. However, when the names were 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 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 2.1. Simulation results using 6, 12, and 18 names chosen using the guidelines suggested above agree well (details omitted) with the results from the binomial model in (5). This 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 a 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 2, $\sum N_k$=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. These questions require limited time to answer, impose no more demands on respondents than the average survey question, and can easily be integrated into currently existing surveys. The usefulness of this method has previously been limited, however, by three previously documented problems.

In this paper we have proposed two additional tools for researchers. First, the latent non-random 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 show 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 non-random mixing model. We also provided guidelines for selection such a set of names.

Though the methods presented here account for bias in individual degree estimation in ways that are not present in other methods, they are only as good as the available data on the demographics of first names. Using “How many X’s do you know?” data to estimate person network size requires knowing the number of people in the population with the different first names. In many countries such information may not be available. Further, the scaled-down condition that we proposed can only control for non-random mixing across dimensions for which there are sufficient data. For example, even if the set of names used satisfies the scaled-down condition with respect to age and gender, there still could be a bias in the individual estimates that is correlated with something that is not included in the model, such as race/ethnicity. We therefore believe that improved information about the demographics of different first names, information that is collected but not released by the U.S. Census Bureau, would be a great benefit to social science, and as such we suggest that this information be released to the public.

A potential area for future methodological work involves improving the calibration curve used to adjust for recall bias. The curve is currently fit deterministically based on the twelve 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 8 alter categories, but if there were more, a more automated procedure would be preferable.

In addition to the general benefit to social science from more accurate estimates of personal network size, we think that one of the most interesting and important potential applications of these improved network size estimates is for the study of "hidden" or "hard-to-reach" populations, such as injection drug users, men who have sex with men, and sex workers and their clients. In most countries these are the subpopulations at greatest risk for becoming infected with HIV, but, unfortunately, the sizes of these subpopulations are not known and this hinders efforts to fight the spread of the disease (UNAIDS, 2003). As was shown by Killworth et al. (1998b) and Bernard et al. (1991), estimates of person network size along with "How many X's do you know?" data can be used to estimate the size of hidden populations. For example, if you know 300 people and 2 injection drug users, then we can estimate that there are about 2 million injection drug users in the United States ($\frac{300 \text{ million}}{300} \cdot 2 =$ 2 million). Thus, the improved degree estimates described in this paper should lead to improved estimates of the sizes of hidden populations.

Appendix

A Details of the calibration curve

Killworth et al. (2003) and Zheng et al. (2006) both suggested that the relation between $\beta_k = \log(b_k)$ and $\beta'_k = \log(b'_k)$ (the subpopulation sizes in the actual social network and the recalled social network on the log scale) begins along the $y = x$ line and the slopes decreases to $1/2$ (corresponding to a square root relation on the original scale) as the fractional subpopulation size increases.

Based on this assumption and some boundary conditions, McCormick and Zheng (2007) derived that

$$\beta'_k = b + \frac{1}{2}(\beta_k - b) + \frac{1}{2}(1 - e^{\beta_k - b}),$$

where $b = -7$. For details on this derivation, the readers are referred to McCormick and Zheng (2007). Therefore the calibration curve between b_k and b'_k used in (12) is then

$$\begin{aligned} f(x) = \exp(g(\log(x))) &\rightarrow e^b \text{ as } x \rightarrow e^b, \\ &= O(\sqrt{x}) \text{ as } x \rightarrow \infty, \end{aligned}$$

where

$$g(x) = b + \frac{1}{2}(x - b) + \frac{1}{2a}\left(1 - e^{-a(x-b)}\right).$$

References

Barabási, A. L. (2003). *Linked.* Plume.

Barton, A. H. (1968). Bringing society back in: Survey research and macromethodology. *American Behavorial Scientist*, 12(2):1–9.

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.

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.

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.

Brewer, D. D. (2000). Forgetting in the recall-based elicitation of person and social networks. *Social Networks*, 22:29–43.

Burt, R. S. (1984). Network items and the General Social Survey. *Social Networks*, 6:293–339.

Butts, C. T. (2003). Network inference, error, and informant (in)accuracy: a Bayesian approach. *Social Networks*, 25:103–140.

Campbell, K. E. and Lee, B. A. (1991). Name generators in surveys of personal networks. *Social Networks*, 13:203–221.

Clauset, A., Shalizi, C., and Newman, M. (2007). Power-law distributions in empirical data. *arXiv:0706.1062*.

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

Del Valle, S. Y., Hyman, J. M., Hethcote, H. W., and Eubank, S. G. (2007). Mixing patterns between age groups in social networks. *Social Networks*, 29(4):539–554.

Echenique, F. and Fryer, R. G. (2007). A measure of segregation based on social interactions. *Quaterly Journal of Economics*, 122(2):441–485.

Fischer, C. S. (1982). *To Dwell Among Freinds.* University of Chicago Press.

Freeman, L. C. and Thompson, C. R. (1989). Estimating acquaintanceship volume. In Kochen, M., editor, *The Small World*, pages 147–158. Ablex Publishing.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition.* Chapman & Hall/CRC.

Granovetter, M. (1976). Network sampling: Some first steps. *American Journal of Sociology*, 81(6):1287–1303.

Gurevich, M. (1961). *The Social Structure of Acquaintanceship Networks.* PhD thesis, MIT.

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.

Hill, R. A. and Dunbar, R. I. M. (2003). Social network size in humans. *Human Nature*, 14(1):53–72.

Killworth, P. D. and Bernard, H. R. (1976). Informant accuracy in social network data. *Human Organization*, 35(3):269–289.

Killworth, P. D. and Bernard, H. R. (1978). The reverse small-world experiment. *Social Networks*, 1(2):159–192.

Killworth, P. D., Bernard, H. R., and McCarty, C. (1984). Measuring patterns of acquaintanceship. *Current Anthropology*, 23:318–397.

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.

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.

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.

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.

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.

Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88–90.

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. (1999). *Sampling: Design and Analysis.* Duxbury Press.

Marsden, P. V. (1987). Core discussion networks of Americans. *American Sociological Review*, 52:122–131.

Marsden, P. V. (1990). Network data and measurement. *Annual Review of Sociology*, 16:435–463.

Marsden, P. V. (2005). Recent developments in network measurement. In Carrington, P. J., Scott, J., and Wasserman, S., editors, *Models and Methods in Social Network Analysis*, chapter 2, pages 8–30. Cambridge University Press.

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.

McCormick, T. H. and Zheng, T. (2007). Adjusting for recall bias in "how many X's do you know?" surveys. Joint Statistical Meetings: Salt Lake City, Utah.

McPherson, M., Smith-Lovin, L., and Brashears, M. (2006). Social isolation in America: Changes in core discussion networks over two decades. *American Sociological Review*, 71:353–375.

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.

Milgram, S. (1967). The small world problem. *Psychology Today*, 1:62–67.

Onnela, J., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Science, USA*, 104(18):7332–7336.

Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203.

Pool, I. and Kochen, M. (1978). Contacts and influence. *Social Networks*, 1:5–51.

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, USA*, 103(9):3490–3494.

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.

UNAIDS (2003). *Estimating the Size of Popualtions at Risk for HIV*. Number 03.36E. UNAIDS, Geneva.

Volz, E. (2006). Tomography of random social networks. *Working Paper*.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis*. Cambridge University Press.

Watts, D., Dodds, P. S., and Newman, M. (2002). Identity and search in social networks. *Science*, 296:1302–1305.

Zheng, T., Salganik, M. J., and Gelman, A. (2006). How many people do you know in prison?: Using overdispersion in count data to estiamte social structure in networks. *Journal of the American Statistical Association*, 101:409–423.