Estimating the Size of Key Affected Populations in Concentrated HIV/AIDS Epidemics Using the Network Scale Up Method*

Rachael Maltiel
Expedia, Inc.

Adrian E. Raftery
University of Washington and University College Dublin

Tyler H. McCormick
University of Washington

September 26, 2013

Abstract

We develop methods for estimating hard-to-reach populations from data collected using network-based questions on standard surveys. Such data arise by asking respondents how many people they know in a specific group (e.g. people named Michael, intravenous drug users). The Network Scale up Method (NSUM) is a tool for producing population size estimates using these indirect measures of respondents' networks. Killworth et al. (1998a,b) proposed maximum likelihood estimators of population size for a fixed effects model in which respondents' degrees or personal network sizes are treated as fixed. We extend this by treating personal network sizes as random effects, yielding principled statements of uncertainty. This allows us to generalize the model to account for variation in people's propensity to know people in particular subgroups (barrier effects), such as their tendency to know people like themselves, as well as their lack of awareness of or reluctance to acknowledge their contacts' group memberships (transmission bias). NSUM estimates also suffer from recall bias, in which respondents tend to underestimate the number of members of larger groups that they know, and conversely for smaller groups. We propose a data-driven adjustment method to deal with this. Our methods perform well in simulation studies, generating improved estimates and calibrated uncertainty intervals, as well as in back estimates of real sample data. We apply them to data from a study of HIV/AIDS prevalence in Curitiba, Brazil. Our results show that when transmission bias is present, external information about its likely extent can greatly improve the estimates.

*Research supported by NICHD grants R01 HD054511 and R01 HD070936 and by a Science Foundation Ireland ETS Walton visitor award, grant reference 11/W.1/I2079.

Keywords: Aggregated relational data; Barrier effect; HIV/AIDS; Recall bias; Social network; Transmission bias.

1 Introduction

The problem of estimating the size of hard-to-reach subpopulations arises in many contexts. In countries with concentrated HIV/AIDS epidemics, the sizes of key affected populations are important for estimating and projecting the epidemic. Concentrated AIDS epidemics are defined as epidemics where AIDS is largely concentrated within particular at-risk groups, such as intravenous drug users (IDU), female sex workers (FSW), and men who have sex with men (MSM). Estimates of the sizes of these groups are also needed to appropriately distribute resources and prevention programs to contain the AIDS epidemic.

The Network Scale Up Method (NSUM) has been proposed as a way to estimate the size of hard-to-reach subpopulations. The NSUM was first proposed by Bernard et al. (1989, 1991) following the 1985 Mexico City earthquake in an attempt to use respondents' knowledge about their social contacts to estimate the number of people that died in the earthquake. Bernard and colleagues realized that the information an individual possesses about others in his or her social network could be used to estimate populations that are currently difficult to size.

Respondents are asked questions of the type "How many X do you know?," where X ranges over different subpopulations of both known and unknown size. Known subpopulations could include people named Michael, diabetics, and women who gave birth to a baby, while unknown subpopulations are typically the groups of interest, such as female sex workers. To standardize what it means to know someone, the McCarty et al. (2001) survey defines it as follows: "For the purposes of this study, the definition of knowing someone is 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 survey can be applied to anyone in the overall population of interest. Respondents do not have to admit to belonging to any particular group, unlike in most other survey methods. "How many X do you know?" questions can easily be integrated into almost any survey, allowing the method to be implemented with limited cost.

Previous statistical work in this area refers to "How many X do you know?" data as aggregated relational data. These questions are widely used on surveys such as the General Social Survey to measure connectivity patterns between individuals. Statistical work in this area includes Zheng et al. (2006) who used aggregated relational data to estimate social structure through overdispersion, McCormick et al. (2010) who developed methods for estimating individuals' personal network size and rates of mixing between groups in the population, and McCormick and Zheng (2012) who estimated the demographic composition

of hard-to-reach populations. While we focus here on estimating the sizes of population groups, the previous work focused primarily on estimating features of the population social network and the dynamics of interactions between population groups.

In its simplest form, the NSUM is based on the idea that for all individuals, the probability of knowing someone in a given subpopulation is the size of that subpopulation divided by the overall population size. For example, if a respondent knows 100 people total and knows 2 intravenous drug users, then it is inferred that 2% of the total population are intravenous drug users. This assumption corresponds to a binomial model for the number of people in a given subpopulation that the respondent knows. However, the total number of people known by a respondent, also called his or her degree or personal network size, also needs to be estimated. A person's degree is estimated by asking the respondents about the number of contacts he or she has in several subpopulations of known size, such as twins, people named Nicole, or women over 70, using the same assumption that an individual should know roughly their degree times the proportion of people in a given subpopulation. The size of the unknown subpopulation is then estimated using responses to questions about the number of people known in the unknown subpopulation combined with the degree estimate, leading to the scale-up estimator (Killworth et al. 1998a,b). The estimator can be improved by increasing the number of respondents and the number of known subpopulations asked about.

The scale-up estimator suffers from several kinds of bias (Killworth et al. 2003, 2006; McCormick et al. 2010). It does not take account of the different propensities of people to know people in different groups, such as people's tendency to know people like themselves; these are called barrier effects. Transmission bias arises when a respondent does not count his or her contact as being in the group of interest, for example because the respondent does not know that the contact belongs to the group. This bias may be particularly large when a group is stigmatized, as is the case of most of the key affected populations in which we are interested. Recall bias refers to the tendency for people to underestimate the number of people they know in larger groups because they forget some of these contacts, and to overestimate the number of people they know in small or unusual groups.

McCormick et al. (2010) proposed strategies for improving degree estimation. Efficiently estimating respondent degree was the focus of that work, however, and so it did not address estimating population size. Further, the McCormick et al. (2010) method requires additional information about the demographic composition of populations with known size. This information is not always available when estimating population group size. Similarly, McCormick and Zheng (2007) proposed a calibration curve to adjust for recall bias that was later incorporated into McCormick et al. (2010). We use a similar approach to addressing recall issues,

but adjust our approach to ensure compatibility with our model for size estimation.

Some attempts have been made to correct for transmission bias in size estimates. These consist of estimating the probability that a respondent counts a contact that belongs to the group of interest as being a member of the group, and then dividing the NSUM size estimate by the estimated probability. Ezoe et al. (2012) surveyed men who have sex with men, the population of interest, to find out how many people in the MSMs' networks knew about their group status. Salganik et al. (2011b)'s implementation of NSUM estimates in Curitiba, Brazil included a game of contacts method where the researchers surveyed heavy drug users to estimate the proportion of their network that are aware of their drug use status. The game of contacts method involves asking heavy drug users about the number of people they know with certain names and then asking if those contacts are aware of the respondent's drug use status as well as the contacts' own drug use status. This allows for an estimate of the proportion of drug users that NSUM survey respondents would be aware of within their own social network. The success of these methods remains to be determined.

Note that Zheng et al. (2006)'s model involved a parameter denoted by b_k, defined as the prevalence parameter or the proportion of total links that involve group k, and they provided a way of estimating it. It is tempting to interpret this as the proportion of the population in group k, and hence as providing a population size estimate for group k. However, this would be incorrect, particularly for populations for which transmission bias is a major concern, such as the hard-to-reach populations that are our main focus. If Zheng et al. (2006)'s prevalence parameter b_k were used to estimate the size of hard-to-reach populations, it would tend to give substantially biased estimates.

In this paper, we develop a Bayesian framework for population group size estimation using the NSUM. We first build a random degree model with a random effect for degree which incorporates variability and uncertainty across individuals' network sizes. We then build on this basic model to adjust for barrier and transmission effects, both separately and combined, resulting in four models altogether. The overall goal is to provide size estimates with reduced bias and error, as well as to assess the uncertainty of the estimates.

In Section 2 we introduce the four models: the random degree model, the barrier effect model, the transmission effect model, and the combined barrier and transmission effect model. We also propose a method for adjusting for recall bias. In Section 3, we show results from several simulation studies, confirming the need to account for biases and the success of our methods in correcting for them. We also show that adjusting for barrier effects using our methods yields better size estimates than the Killworth et al. (1998a,b) estimates for the known populations in the dataset used by McCarty et al. (2001). We will also show the estimates produced by our model on the Curitiba, Brazil data of Salganik et al. (2011a,b).

Lastly, in Section 4, we will discuss additional research needed to make NSUM estimation a viable, accurate method to estimate the size of hard-to-reach populations.

2 Models

Previous size estimates based on "How many X's do you know?" data have been computed using the network scale-up estimator. Let y_{ik} be the number of people known by individual i, $i = 1, \ldots, n$, in group k, $k = 1, \ldots, K$, with group K being of unknown size. (Note that there can be more than one group of unknown size, but we are using one to simplify the exposition.) Let d_i be the number of people that respondent i knows, also called the degree or personal network size. Also, let N_k be the size of group k, and let N be the total population.

The scale-up estimates are based on the assumption that $y_{ik} \sim \text{Binom}(d_i, \frac{N_K}{N})$, or that the number of people known by individual i in group k follows a binomial distribution. We refer to this as the scale-up model. From this model, Killworth et al. (1998a,b) derived the maximum likelihood estimator of d_i as

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

Conditional on estimates $\widehat{d_i}$ of d_i, the maximum likelihood estimator of N_K, the size of the unknown population, is then

$$\widehat{N}_K = N \frac{\sum_{i=1}^{n} y_{iK}}{\sum_{i=1}^{n} \widehat{d_i}}. \tag{2}$$

Equations (1) and (2) are commonly referred to as the scale-up estimates.

Our proposed models build on the scale-up model. We first model degree as a random effect, leading to regularized estimates of degree. We refer to this as our random degree model. We then extend the random degree model to take account of the fact that respondents have different propensities to know members of different groups. For example, people are generally more likely to know people that are similar to them in terms of age, sex, education, race and other characteristics, than to know people who are not. We account for this nonrandom mixing of individuals with an additional random effect, to yield what we call the barrier effects model. We also separately extend the random degree model to account for lack of awareness of or reluctance to acknowledge contacts' group memberships, to yield what we call the transmission bias model. The quality of estimates from this model can be greatly improved by external information on information transmission. Lastly, our combined model accounts for both barrier effects and transmission bias. The models build on each other, as described in Figure 1.

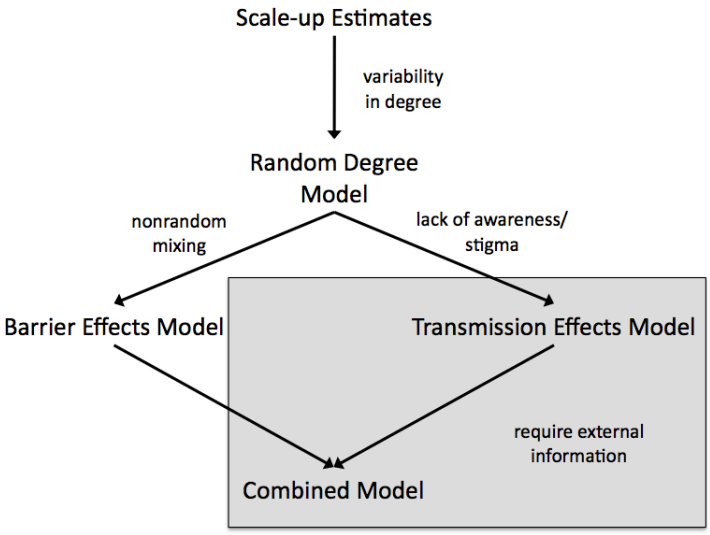


Figure 1: Our four models build on the basic Killworth et al. (1998a,b) scale-up model, accounting for nonrandom mixing or barrier effects, and transmission bias.

2.1 Random degree model

Our first extension of the Killworth et al. (1998a,b) scale-up model is to introduce a random effect for degree, to regularize estimates of degree. If an individual responded that he or she knew a large number of people in a given subpopulation, this would drive up the estimate of the individual's degree d_i. To reduce the sensitivity of estimates to extreme values of d_i, we incorporate degree estimation into our hierarchical modeling framework and achieve regularization through partial pooling.

We call the resulting model our random degree model. It assumes that

$$y_{ik} \sim \text{Binom}\left(d_i, \frac{N_k}{N}\right),$$
$$d_i \sim \text{Log Normal}(\mu, \sigma^2).$$

We choose a log normal distribution for d_i based on the observed distribution of scale-up estimates of degree $\widehat{d}_i$. We found the log normal distribution to have the best fit to estimates of $\widehat{d}_i$ across multiple datasets, including data from the United States, Ukraine, Moldova, Kazakhstan, and Brazil (McCarty et al. 2001; Paniotto et al. 2009; Salganik et al. 2011a).

We estimate the parameters of the random degree model in a Bayesian manner, using

the prior distributions

$$\begin{aligned}\pi(N_K) &\propto \frac{1}{N_K} 1_{N_K \leq N},\\ \mu &\sim \mathrm{U}(3, 8),\\ \sigma &\sim \mathrm{U}(\tfrac{1}{4}, 2).\end{aligned}$$

Our prior for N_K has been used previously for Bayesian estimation of population size with little prior information (Jeffreys 1961; Raftery 1988). The priors for μ and σ were arrived at from the values we saw in fitting the scale-up $\widehat{d}_i$ estimates to several datasets across multiple regions. Our prior for μ allows for mean degrees within a data set ranging from 20 to 3,000, which is consistent with previous research on social networks and the NSUM (McCarty et al. 2001; McCormick et al. 2010). Our prior on σ allows for 95% of degrees to fall in the multiplicative range 1.6 times to 55 times in either direction from the mean, which seemed to more than fully cover the range of results from scale-up estimates across multiple data sets.

2.2 Barrier effects model

Nonrandom mixing, or barrier effects, occur because respondents have different tendencies to know people in different groups, depending on their own characteristics. For example, we might expect a 65-year-old male respondent to know more people named Walter than a 20-year-old female respondent, because Walter was a more common name 65 years ago. This leads to overdispersion in the distribution of the number of people known in a given population relative to what one would expect if the binomial assumption held.

We can model overdispersion in the binomial probabilities as follows. In the Killworth et al. (1998a,b) scale-up and random degree models, the probability that respondent i knows someone in group k is assumed to be constant across respondents, and equal to N_k/N. To model overdispersion, we instead allow this probability, now denoted by q_{ik}, to vary randomly across respondents, following a Beta distribution (Zheng et al. 2006; McCormick et al. 2010). The model then becomes

$$\begin{aligned}y_{ik} &\sim \mathrm{Binom}(d_i, q_{ik}),\\ d_i &\sim \text{Log Normal}(\mu, \sigma^2),\\ q_{ik} &\sim \mathrm{Beta}(m_k, \rho_k).\end{aligned}$$

Here we use the nonstandard parameterization of the Beta distribution according to which $X \sim \mathrm{Beta}(m, \rho)$ if it has the probability density function $f_X(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$, $m = \frac{\alpha}{\alpha+\beta}$ and $\rho = \frac{1}{1+\alpha+\beta}$ (Skellam 1948; Mielke Jr 1975; Diggle et al. 2002, Chapter 9). Then m_k is

the prior mean of q_{ik}, and ρ_k determines its dispersion. We set $E[q_{ik}] = m_k = \frac{N_k}{N}$. We use the prior distributions

$$\pi(m_K) \propto \frac{1}{m_K},$$
$$\rho_k \sim \mathrm{U}(0,1),$$

with the priors for μ and σ remaining the same as in the random degree model.

2.3 Transmission bias model

Transmission bias occurs when a respondent is unaware of or reluctant to acknowledge the group membership status of his or her contacts. For example, if a respondent is not aware that a contact is an intravenous drug user, he or she would not count that contact when responding to a question about the number of intravenous drug users known. We can think of the transmission bias, denoted by τ_k, as the proportion of respondents' contacts in group k that the respondents report. For example, if 50% of intravenous drug users disclose their status to their contacts and if respondents report all the IDUs that they know, then $\tau_K = 0.5$ for the subpopulation K of IDUs. Thus, we can add τ_k to our model as a multiplier of the binomial proportion, since a respondent would mention knowing only a proportion τ_k of their true contacts in group k on average. This yields the transmission bias model

$$y_{ik} \sim \mathrm{Binom}\left(d_i, \tau_k \frac{N_k}{N}\right),$$
$$d_i \sim \text{Log Normal}(\mu, \sigma^2).$$

We specify the additional prior

$$\tau_K \sim \mathrm{Beta}(\eta_K, \nu_K),$$

with the priors for N_K, μ, and σ remaining the same as in the random degree model. For the transmission effect, we assume τ_k to be 1 for the known populations $k = 1, \ldots, K-1$, and to be less than or equal to one for the groups of unknown size, in line with the definition of transmission bias. This means that we are assuming that respondents are aware of and prepared to acknowledge contacts' group membership status for the known groups. This assumption is reasonable as the known populations are typically less stigmatized, making it less likely for respondents to be unaware of or reluctant to acknowledge their contacts' membership statuses. Our simulation results indicated the desirability of using external information about τ_K in the form of an informative prior, which will be discussed further in Section 3.1.

2.4 Combined model

Previous research indicates both barrier and transmission effects to be present in these data (McCarty et al. 2001; Kadushin et al. 2006; McCormick et al. 2010; Salganik et al. 2011a). For a model to produce unbiased estimates, we need to adjust for both sources of bias. Thus, we can combine our barrier and transmission models to get a combined model that accounts for both barrier and transmission effects. Our model is thus

$$\begin{aligned} y_{ik} &\sim \text{Binom}(d_i, \tau_k q_{ik}), \\ d_i &\sim \text{Log Normal}(\mu, \sigma^2), \\ q_{ik} &\sim \text{Beta}(m_k, \rho_k), \end{aligned}$$

with priors the same as in the previous models.

2.5 Recall bias adjustment

Since respondents are asked to say quickly how many people they know in certain groups, it is common for them to forget contacts in large groups or to overcount contacts in small groups. For example, a respondent might know 15 or 20 people in a large group and might forget to mention a few while quickly answering a survey. In addition, small subpopulations can be memorable, such as people who died in a car accident. Respondents might count someone in a small subpopulation as someone they know even if the contact does not actually fall under the definition of "know" in NSUM surveys.

Previous research has suggested methods to adjust for recall bias based on the relationship between respondents' recalled ties and the sizes of known groups of interest (Killworth et al. 2003; Zheng et al. 2006; McCormick and Zheng 2007; McCormick et al. 2010). Our exploratory work suggests a linear relationship between the two on the log scale. This leads to the following model to incorporate recall bias as well as barrier effects and transmission bias:

$$\begin{aligned} y_{ik} &\sim \text{Binom}(d_i, e^{r_k} \tau_k q_{ik}), \\ r_k &\sim N(a + b \log N_k, \sigma_r^2), \\ d_i &\sim \text{Log Normal}(\mu, \sigma^2), \\ q_{ik} &\sim \text{Beta}(m_k, \rho_k). \end{aligned}$$

The additional parameters a, b, and σ_r have uniform flat priors, namely $a \sim U(0, 15)$, $b \sim U(0, 1)$, and $\sigma_r \sim U(0, 1)$. The quantity N_k would be calculated just as in the barrier and combined models, where $N_k = N \cdot m_k$.

However, this model involves a large number of parameters and is quite computationally demanding. For models estimating one unknown subpopulation, the random degree model has $n+3$ parameters, the barrier model has $n+K+2$ parameters, and the transmission model has $n+4$ parameters. This full model has $n+2K+n\cdot K+7$ parameters - a large increase from the simpler models. This increase in parameters, coupled with the limited information about recall bias present in the data, makes inference for this model difficult and, in our judgment, not a worthwhile investment. Instead, we approximate a recall adjustment through a post processing method. This method is computationally very efficient and makes effective use of information available through populations with known size. This method is also easier to implement and, thus, improves the likelihood that the method will be used in practice. The barrier and transmission combined model similarly has $n+K+n\cdot K+4$ parameters, however the relationship between barrier and transmission effects makes a similar post processing approach difficult in this case.

We outline our recall adjusted modeling strategy below. We find that this strategy performs well in practice in our data experiments. We first estimate a linear relationship (on the log scale) between the estimates and the true subpopulation sizes using back estimates. For a data set with $K-1$ known subpopulations, back estimates estimate the k^{th} subpopulation, $k = 1, \ldots, K-1$, treating it as unknown, and treating all other $K-2$ known subpopulations as known to produce the estimate. This can be done for all $K-1$ known subpopulations and then compared to the true, known sizes of those subpopulations for estimation method evaluation. To account for the variability in our estimate of $\hat{N}_k$ as well, we approximate the relationship using the errors-in-variables model

$$\log(\hat{N}_k) = a + b\log(N_k) + \delta_k + \varepsilon_k, \tag{3}$$

where $\hat{N}_k$ is the posterior mean and s_k the posterior standard deviation of the size of the k^{th} subpopulation, computed without knowledge of the true N_k, $\delta_k \sim N(0, s_k^2)$, and $\varepsilon_k \sim N(0, \sigma_\varepsilon^2)$. The model (3) is estimated by maximum likelihood (Ripley and Thompson 1987).

We then adjust for recall bias as follows. Let $Y_K^{[t]}$ denote the t-th value simulated from the posterior distribution of $\log(N_K)$, where t indexes MCMC iterations. We then replace each $Y_k^{[t]}$ with a randomly drawn value

$$\frac{Y_K^{[t]} - a}{b} + Z,$$

where $Z \sim N(0, \sigma_\varepsilon^2/b^2)$ to adjust for recall bias, based on the relationship shown in Equation (3). In our analyses, we have generally found a to be around 6.7, b to be around 0.5, and σ_ϵ to be around 0.35. Our strategy differs from that of McCormick and Zheng (2007) and McCormick et al. (2010) because we apply our adjustment after a complete run of

our sampler. The correction for recall cannot, therefore, influence the path of the sampler as in McCormick and Zheng (2007) and McCormick et al. (2010). The strategy is instead more similar to that employed by Zheng et al. (2006), who adjusted a normalization constant (necessary to preserve identifiability) after sampling to adjust for recall issues. Our proposed method propagates uncertainty from responses to size estimates, however, which is not a feature of the Zheng et al. (2006) approach.

3 Results

We estimated all the models using Markov chain Monte Carlo (MCMC). For all models, μ and σ were sampled from using closed form Gibbs steps while we used random walk Metropolis steps with normal proposals for all the other parameters. Derivations of all Gibbs and Metropolis steps are included in the Appendix. When possible, we used scale-up estimates as starting points for the parameters.

The MCMC algorithms were implemented using the methodology described in Raftery and Lewis (1996), using an initial chain to estimate the conditional posterior standard deviation of each parameter given the other parameters, and then using 2.3 times this value as the standard deviation in the normal proposal. We used the Raftery-Lewis diagnostic to determine the number of iterations needed for the MCMC. In general, our chains behaved well, converging in less than 30,000 iterations. Our combined model, though, required over 150,000 iterations. We also checked the Gelman-Rubin diagnostic on all models on the Curitiba data set, discussed below (Gelman and Rubin 1992). For N_K, our population size of interest, the Gelman-Rubin diagnostic was close to 1 in all models. For the other parameters, the Gelman-Rubin diagnostic was under 1.015 in the random degree, barrier, and transmission models and under 1.1 for 99.5% of 10,416 parameters in the combined model.

One difficulty in verifying NSUM estimation results is that we do not know the true size of hard-to-reach subpopulations. Thus, we first ran several simulations to verify the need for and improvement from our models that adjust for biases when present. We tested our models on data containing no bias, barrier effects, and transmission bias for three types of simulations and we report the results in Section 3.1. Secondly, we computed back estimates on the data from McCarty et al. (2001), or estimates of known subpopulations to be compared to the true size, to assess the efficacy of our models, detailed in Section 3.2. Lastly, in Section 3.3 we give results from estimating all our models on data from the Curitiba study (Salganik et al. 2011a,b).

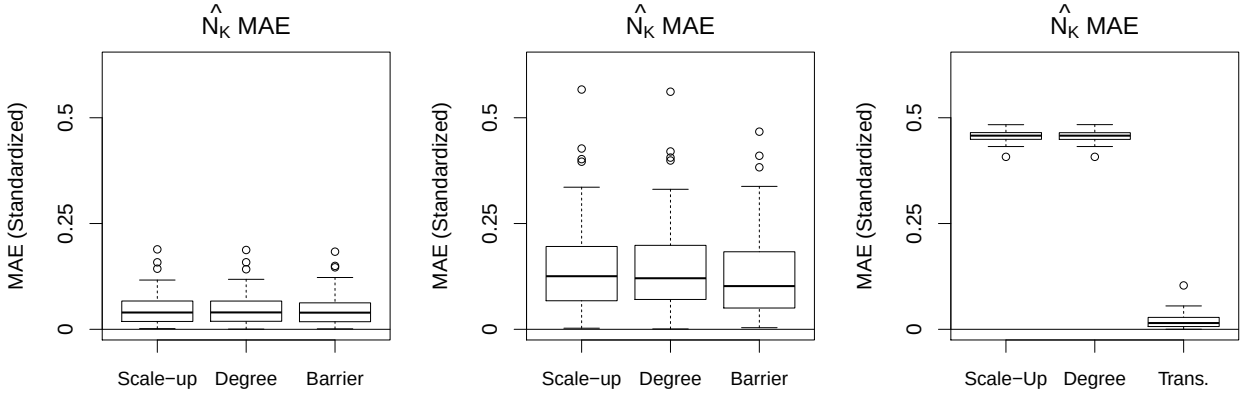
3.1 Simulation studies

For our simulations, we created data sets containing various levels of bias: no bias, barrier effects bias, and transmission bias. In the no bias simulation, the data followed the assumptions of our random degree model: the respondents' degrees followed a log normal distribution while the number of people known in each group followed a binomial distribution based on the respondent's degree and the proportion of the total population in a given group. In the data with barrier effects, we added a beta random effect to the binomial proportion. For the data with transmission bias, we instead added a multiplier τ_K to the binomial proportion.

The no bias and barrier effect simulations were based on data from McCarty et al. (2001) while the transmission bias simulation was based on data from Salganik et al. (2011a). While the McCarty et al. (2001) data is a well understood, commonly used dataset, we had more detailed information on transmission bias for the prior in the Salganik et al. (2011a) Curitiba data set, making it a better choice on which to base a transmission bias simulation. For all simulations, we used a sample size of 500 and simulated 100 data sets. We estimated the size of one unknown population; for the McCarty et al. (2001) based simulations, the unknown population had size 500,000 (based on scale-up estimates of the unknown groups in the McCarty et al. (2001) data set) while for the Salganik et al. (2011a) based simulations, the unknown population had size 65,000 (based on the scale-up estimates of heavy drug users in Curitiba). When barrier effects were present in the data, we used values for the barrier effect parameters estimated in the McCarty et al. (2001) data set by the barrier effect model. For transmission effects, we used $\tau_K = 0.54$ based on the estimate of transmission bias from Salganik et al. (2011b) using the game of contacts method. We also obtained our transmission effect prior of Beta(0.542, 0.011) by fitting a beta distribution to the bootstrapped estimates of the transmission bias τ_K. (Salganik et al. (2011b) had both a transmission bias parameter, to measure respondents' awareness of contacts' status, and a population parameter, to measure differences in the size of networks of people in the population of interest versus people in the general population. We have combined these two parameters for our transmission bias parameter as they are not identifiable in our models.)

Across our simulations, we measured mean absolute error (MAE) to see how much error occurred in estimates when using different models based on different assumptions. Figure 2 depicts the MAE scaled by the true size of the unknown population, with the point estimate being the mean of the posterior of N_K, while the numbers are reported in Table 1 as well. We see that when there is no bias in the data, the scale-up estimates and random degree model produce estimates with little error. The barrier effects model is also able to estimate size with minimal error, even though the barrier effects that the model includes are not present

in the data. When barrier effects are present in the data, the barrier effects model produces an MAE that is 12% lower than the scale-up estimates or the random degree model.



(a) No bias (b) Barrier effects (c) Transmission bias

Figure 2: Simulation study: Mean absolute error (MAE) of posterior means of N_K divided by the true size of N_K. Each panel corresponds to a different simulation setup. The three boxplots in each panel correspond to different estimates: scale-up estimates, random degree model estimates, and estimates from either the barrier effects model or the transmission bias model. Each boxplot shows the distribution of the MAEs across 100 simulated datasets.

Table 1: Standardized mean absolute error, dividing mean absolute error by the true subpopulation sizes, and coverage over the 100 simulations across data set designs and estimation models: scale-up model, random degree (Degree) model, barrier effects model, and transmission bias (Trans.) model.

Data	No bias			Barrier bias			Transmission bias		
Model	Scale-up	Degree	Barrier	Scale-up	Degree	Barrier	Scale-up	Degree	Trans.
MAE	0.046	0.046	0.046	0.145	0.145	0.128	0.457	0.457	0.019
MAE SE	0.003	0.003	0.003	0.012	0.012	0.010	0.001	0.001	0.002
80% Coverage	-	84%	83%	-	27%	87%	-	0%	100%
95% Coverage	-	97%	97%	-	48%	94%	-	0%	100%

We see the largest different in estimates when transmission bias is present in the data. When transmission bias is not accounted for in the model estimates, the MAE is large, while the transmission model results in estimates with minimal error.

Our credible interval coverage, shown in Table 1, also indicates the importance of using a model that correctly adjusts for bias in the data. We see appropriate coverage for both the random degree and barrier models when there is no bias in the data. When there are barrier effects or transmission effects in the data, the random degree model results in under-coverage while the appropriate model shows appropriate interval coverage of the true value.

In particular, we see no coverage for the random degree model when transmission effects are present. While failing to account for barrier effects present in data results in error in estimates and under-coverage, the results are much more extreme when failing to account for transmission effects. We believe accurate assessment of transmission bias to be the highest priority in improving NSUM size estimates.

Through our simulations, we were also able to see the importance of the choice of priors for the transmission effect model. To contrast our transmission bias simulation using the informative prior based on Salganik et al. (2011b)'s game of contacts results, we also ran a simulation using an uninformative Uniform(0,1) prior on τ_K. We found that for τ_K, the posterior distribution was very similar to the prior. Table 2 gives the 95% interval end points and median for the τ_K prior as well as the average interval endpoints and medians for the τ_K posterior for both the informative and uninformative simulations, where the posterior values are averaged over the estimates from the 100 simulation posteriors of τ_K.

Table 2: Comparison of prior and posterior 95% credible interval quantiles and medians for the uninformative and informative prior transmission bias simulations, averaging over the posterior samples for the 100 simulated data sets. We see that the posterior of τ_K aligns very closely with the prior, showing the need for an informative prior to produce accurate size estimates. In addition, we see an incorrect point estimate for prevalence using the uninformative prior (true prevalence is 3.6%) and a wide range of uncertainty.

	Transmission bias τ_K			Prevalence		
	2.5%	Median	97.5%	2.5%	Median	97.5%
Uninformative Prior						
Prior	0.025	0.500	0.975	5.5×10^{-5}%	0.06%	68.8%
Posterior	0.075	0.513	0.973	2.0%	3.9%	30.1%
Informative Prior						
Prior	0.438	0.542	0.644	5.5×10^{-5}%	0.06%	68.8%
Posterior	0.438	0.542	0.644	3.0%	3.6%	4.5%

The close match between the prior and posterior of τ_K has major implications for the posterior estimates of N_K as well. Table 2 shows the 95% credible interval points and medians of N_K averaged over the 100 simulations for both the informative and uninformative prior as well. The estimate of N_K from the transmission bias model is roughly equal to the estimate of N_K from the random degree model divided by τ_K. Our estimates from the transmission bias model were very close to the estimates in the random degree model divided by the prior expected value τ_K. Thus, the error in the prior expectation of the transmission bias will lead to a corresponding error in the estimate of N_K. Our uninformative prior has an expected transmission bias, τ_K, of 50% (as compared to the true 54%) and we do indeed see an overestimate of the median prevalence in Table 2 when using the uninformative prior: the

true prevalence is 3.6% as opposed to the estimate of 3.9% with the noninformative prior.

In addition, if there is a lot of uncertainty in the prior of τ_K, the posterior interval for N_K will also be wide. Figure 2c shows the need to account for transmission bias to produce an unbiased estimate, but Table 2 indicates that an informative prior is needed to account for transmission bias. This indicates the need for methods to estimate transmission bias.

3.2 McCarty Back Estimates

To further assess our methods, we fit back estimates using the random degree and barrier effect models for the 29 known subpopulations in the McCarty et al. (2001) data set and compared them to the known values. We were unable to test models adjusting for transmission bias as we do not have informative priors for the populations in the data set, although it is reasonable to believe that these subpopulations have minimal transmission bias. The McCarty et al. (2001) data set was obtained through random digit dialing within the United States. It contains responses from 1,375 adults from two surveys: survey 1 with 801 responses conducted in January 1998 and survey 2 with 574 responses conducted in January 1999. The McCarty et al. (2001) data set has been analyzed in numerous articles, evaluating methods to estimate degrees in addition to methods to estimate hard-to-reach populations (Killworth et al. 2003; Zheng et al. 2006; McCormick et al. 2010). Since previous research has indicated recall bias to be present in the McCarty data set, we adjusted for recall bias as described in Section 2.5.

Figure 3 shows scale-up point estimates and random degree model and barrier effects model 80% and 95% credible intervals of the posterior of the size estimates of the McCarty et al. (2001) data set as well as scale-up point estimates along the x axis, compared to the true sizes along the y axis. The black diagonal line is the $x = y$ line where the true size and the estimate are equal, which is the goal. We see generally that our estimates are close to the true subpopulation size and our credible intervals cover the true subpopulation size.

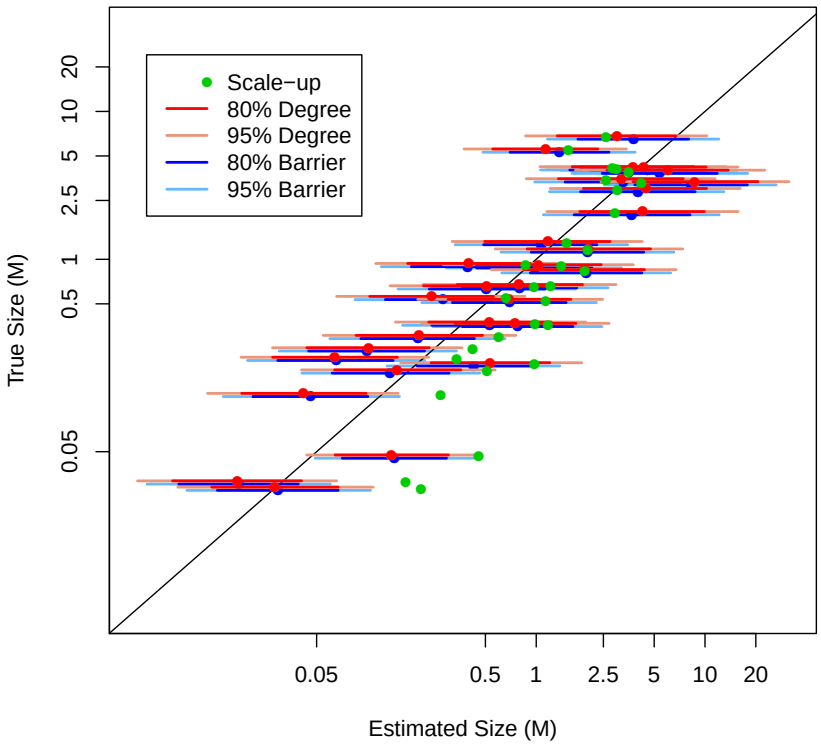


Figure 3: Back estimates and 80% and 95% credible intervals for the McCarty data sets using the random degree and barrier effect models and scale-up estimates. The x axis shows the estimates while the y axis shows the true subpopulation size. The black diagonal line shows the goal where the estimates and true subpopulation sizes are equal.

Table 3 shows the mean absolute error (MAE) and coverage of credible intervals for the estimation methods over the 29 back estimates of the subpopulations in the McCarty et al. (2001) data set. We see that the barrier model produces estimates with the smallest average absolute error, as we would hope given the barrier effects present in the McCarty data set. We also see that both the random degree and barrier effects models result in appropriate credible interval coverage.

3.3 Curitiba Results

The Curitiba dataset consists of 500 adult residents of Curitiba, Brazil and was collected through a household-based random sample in 2010 by Salganik et al. (2011a). One aim of this study was to size the hard-to-reach populations relevant to concentrated HIV/AIDS epidemics. In addion, a game of contacts survey was conducted to estimate transmission

Table 3: Mean absolute error (MAE), standardized by dividing all absolute errors by the true subpopulation sizes, and credible interval coverage for scale-up estimates and random degree and barrier model estimates over the 29 back estimates.

	Model Estimates		
	Scale-up	Degree	Barrier
MAE	1.49	1.48	0.93
80% Coverage	-	72%	66%
95% Coverage	-	97%	93%

bias for heavy drug users (Salganik et al. 2011b). From these game of contacts data, we were able to obtain an informative prior for transmission bias, allowing us to fit all of our models to the Curitiba data set and to assess our models' performance on relevant data. As in our simulations, we used a Beta(0.542, 0.011) prior for transmission bias based on the game of contacts estimate of transmission bias.

We did not adjust for recall bias as the study design did not produce the information needed to do this. A recall adjustment relies on the known and unknown subpopulations being similar in size, as the adjustment is based on a regression of the known subpopulations. A recall adjustment aims to account for the fact that people have a hard time remembering too many contacts in one group in a short time period. Thus, it is reasonable to believe that there will be a different amount of recall error in a subpopulation where a respondent actually knows 10 people then a subpopulation where a respondent actually knows 20 people. For Curitiba, the largest known subpopulation produces a scale-up estimated prevalence of 3.1% while the average subpopulation produces a scale-up estimated prevalence of 1.4%. The scale-up estimate for heavy drug users, our unknown subpopulation of interest, is 3.6%. When we applied recall adjustments to our estimates of heavy drug users, the resulting adjusted prevalences were far too high to be reasonably believed; the regression did not have the necessary data to produce an adjustment on a subpopulation this large. Thus, it is important for researchers to design surveys that use known subpopulations covering the full range of expected possible values for the subpopulation of interest to be able to accurately adjust for recall bias.

The estimates of prevalence of heavy drug users in Curitiba from our models are shown in Figure 4. While there is limited uncertainty in the estimates from the random degree model, the estimates and their uncertainty are likely underestimated due to the transmission bias in the data. The barrier model results in a smaller estimate while the transmission model results in a larger estimate of heavy drug user prevalence. The uncertainty in the combined model seems reasonable and is smaller than in the transmission model (and the transmission prior) with a value between the separate barrier and transmission model estimates. This

compares to the estimates obtained by Salganik et al. (2011a) of 3.3% with a 95% confidence interval from 2.7% to 4.1% without accounting for transmission bias and an estimate of 6.3% with a 95% confidence interval from 4.5% to 8.0% when accounting for transmission bias.

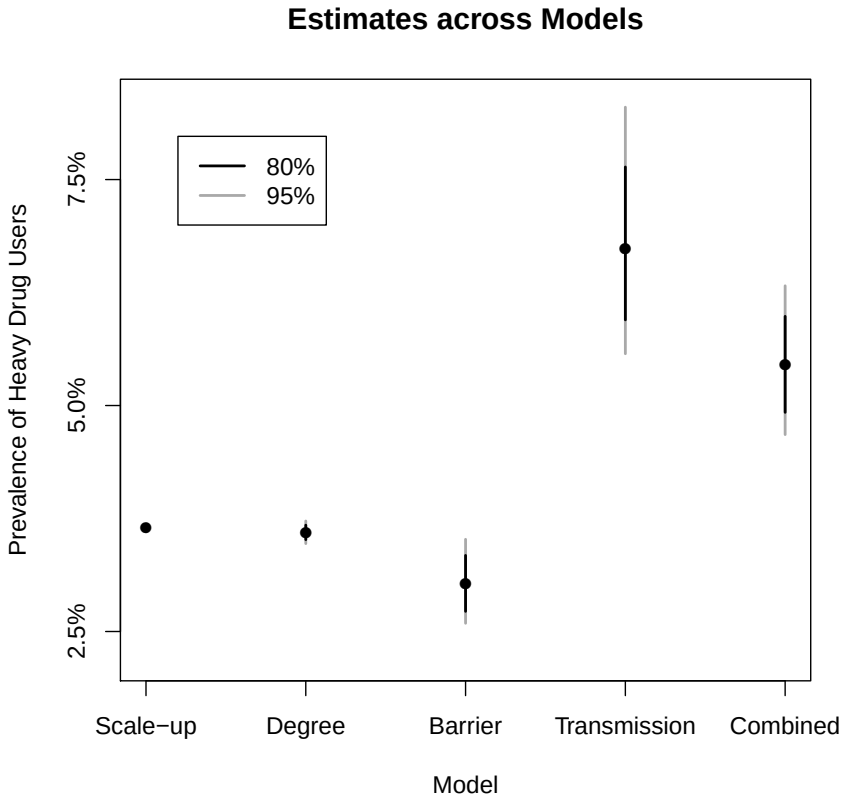


Figure 4: Posterior estimates and credible intervals for the prevalence of heavy drug users in Curitiba based on the random degree, barrier, transmission, and combined models.

4 Discussion

Indirectly observed social network data are one tool for estimating the size of hard-to-reach populations. With knowledge of the true size of a handful of subpopulations, data can be collected to then estimate the size of hard-to-reach subpopulations that currently evade researchers. These techniques can be used to provide accurate size estimates to improve public health efforts related to AIDS in concentrated epidemics as well as other subpopulations that are currently difficult to size. NSUM surveys do not require large resources and can be carried out by adding questions to other surveys already being conducted for other purposes.

Currently the most used method for size estimation from these data is the Killworth et al. (1998a,b) scale-up estimate, but this does not provide estimates of uncertainty and can suffer from barrier effects, transmission bias and recall bias. In this paper we have proposed ways of overcoming these limitations. First we proposed a Bayesian model, called the random degree model, that regularizes estimation of degree and yields estimates of uncertainty about population size. Then we extended the model to incorporate barrier effects, transmission bias, and recall bias, and also proposed a more efficient postprocessing method for accounting for recall bias.

We found that the barrier effects model performs better than the scale-up estimates or the random degree model. This makes sense because barrier effects, or nonrandom mixing, are a pervasive feature of social networks. We also found that adjusting for transmission bias is extremely important when this bias is present. However, data typically do not contain much information about transmission bias, and so it is important to use or generate external information about transmission bias if possible. Finally, we found that adjusting for recall bias can improve estimates and the assessment of their uncertainty.

As seen in simulations in Section 3.1, it is important to adjust for bias in estimates through our proposed models to minimize error in estimates and produce appropriate coverage of credible intervals. While nonrandom mixing can be accounted for using our models that adjust for barrier effects without external information, adjusting for transmission effects does require external information. As seen in our simulations, since the posterior closely aligns with the prior for the transmission bias effect, an informative, accurate prior is needed to appropriately adjust estimates. While researchers have started to find methods to estimate for transmission effects, further work is needed in this area before NSUM can produce estimates of hard-to-reach populations with an acceptable level of error. The game of contacts of Salganik et al. (2011b) is one way of doing this. The future utility of the NSUM will depend crucially on the development and use of ways to estimate transmission bias.

In addition, we observed how recall bias can be adjusted for only when known subpopulations are chosen to cover the size range of the unknown subpopulation. While the size of the unknown subpopulation is of course unknown before estimation, researchers should aim to use external sources to cover possible sizes of the group of interest as best as possible.

In this work, we have presented models to adjust for known biases in the NSUM method. We have shown the importance of adjusting for these biases to produce estimates with minimal error and estimates of uncertainty with appropriate coverage. We believe transmission bias, in particular, needs further research to provide informative priors to appropriately account for this bias common in subpopulations with unknown sizes.

References

Bernard, R. H., Johnsen, E., Killworth, P., and Robinson, S. (1989), "Estimating the Size of an Average Personal Network and of an Event Subpopulation," in *The Small World*, ed. Kochen, M., New Jersey: Ablex Press, pp. 159–175.

— (1991), "Estimating the Size of an Average Personal Network and of an Event Subpopulation: Some Empirical Results," *Social Science Research*, 20, 109–121.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002), *Analysis of Longitudinal Data*, Oxford University Press, USA.

Ezoe, S., Morooka, T., Noda, T., Sabin, M. L., and Koike, S. (2012), "Population Size Estimation of Men Who Have Sex with Men through the Network Scale-Up Method in Japan," *PLoS ONE*, 7, e31184.

Gelman, A. and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical science*, 457–472.

Jeffreys, H. (1961), *Theory of Probability*, Oxford, U.K.: Oxford University Press, 3rd ed.

Kadushin, C., Killworth, P., Bernard, H., and Beveridge, A. (2006), "Scale-up Methods as Applied to Estimates of Heroin Use," *Journal of Drug Issues*, 36, 417.

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

Killworth, P., McCarty, C., Bernard, H., Shelley, G., and Johnsen, E. (1998b), "Estimation of Seroprevalence, Rape, and Homelessness in the United States using a Social Network Approach," *Evaluation Review*, 22, 289–308.

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

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, 84–112.

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

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

McCormick, T. H. and Zheng, T. (2007), "Adjusting for Recall Bias in 'How many Xs do you know?' Surveys," in *Proceedings of the Joint Statistical Meetings*, Washington, D.C.: American Statistical Association.

— (2012), "Latent demographic profile estimation in hard-to-reach groups," *The Annals of Applied Statistics*, 6, 1795–1813.

Mielke Jr, P. (1975), "Convenient Beta Distribution Likelihood Techniques for Describing and Comparing Meteorological Data." *Journal of Applied Meteorology*, 14, 985–990.

Paniotto, V., Petrenko, T., Kupriyanov, V., and Pakhok, O. (2009), "Estimating the Size of Populations with High Risk for HIV Using the Network Scale-up Method," Analytical report, Kiev International Institute of Sociology.

Raftery, A. E. (1988), "Inference and prediction for the binomial N parameter: A hierarchical Bayes approach," *Biometrika*, 75, 223–228.

Raftery, A. E. and Lewis, S. M. (1996), "Implementing MCMC," in *Markov Chain Monte Carlo in Practice*, eds. W.R. Gilks, D. S. and Richardson, S., London: Chapman and Hall, pp. 115–130.

Ripley, B. D. and Thompson, M. (1987), "Regression techniques for the detection of analytical bias," *Analyst*, 112, 377–383.

Salganik, M., Fazito, D., Bertoni, N., Abdo, A., Mello, M., and Bastos, F. (2011a), "Assessing Network Scale-up Estimates for Groups Most at Risk of HIV/AIDS: Evidence From a Multiple-Method Study of Heavy Drug Users in Curitiba, Brazil," *American Journal of Epidemiology*, 174, 1190–1196.

Salganik, M. J., Mello, M. B., Abdo, A. H., Bertoni, N., Fazito, D., and Bastos, F. I. (2011b), "The Game of Contacts: Estimating the Social Visibility of Groups," *Social Networks*, 33, 70–78.

Skellam, J. (1948), "A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable Between the Sets of Trials," *Journal of the Royal Statistical Society. Series B (Methodological)*, 10, 257–261.

Zheng, T., Salganik, M., and Gelman, A. (2006), "How Many People Do You Know in Prison?" *Journal of the American Statistical Association*, 101, 409–423.