The Role of Scaling and Estimating the Degree Ratio in the Network Scale-up Method

Ian Laga, Jess Kunke, Tyler McCormick, Xiaoyue Niu

May 9, 2023

Abstract

The Network Scale-up Method (NSUM) uses social networks and answers to "How many X's do you know?" questions to estimate hard-to-reach population sizes. This paper focuses on two biases associated with the NSUM. First, different populations are known to have different average social network sizes, introducing degree ratio bias. This is especially true for marginalized populations like sex workers and drug users, where members tend to have smaller social networks than the average person. Second, large subpopulations are weighted more heavily than small subpopulations in current NSUM estimators, leading to poor size estimates of small subpopulations. We show how the degree ratio affects size estimates, provide a method to estimate degree ratios without collecting additional data, and demonstrate that rescaling size estimates improves the estimates for smaller subpopulations. We demonstrate that our adjustment procedures improve the accuracy of NSUM size estimates using simulations and data from two data sources.

Keywords: Size estimation, popularity factor, degree ratio, key populations, aggregated relational data.

1 Introduction

The Network Scale-up Method (NSUM) has emerged as a popular and efficient way to estimate the size of hard-to-reach populations like female sex workers, drug users, and men who have sex with men. These hard-to-reach populations are of critical importance to solving several global health problems, including meeting UNAIDS HIV-related targets (UNAIDS, 2021). These populations are at higher risk of contracting and spreading HIV than the general population while simultaneously suffering from marginalization and negative social stigma. The NSUM estimates the size of these populations using survey questions of the form "How many X's do you know," where X includes subpopulations with known sizes and subpopulations of interest with unknown sizes, like female sex workers (Bernard et al. (1989)). These survey responses are known as aggregated relational data (ARD). While some research on ARD concerns the estimation of network structures (Breza et al., 2020), we focus on the role ARD play in the NSUM to estimate hard-to-reach subpopulation sizes. Researches have proposed several modeling improvements to better capture the complexity of the underlying aggregated relational data, including those by Zheng et al. (2006), Maltiel et al. (2015), Teo et al. (2019), and Laga et al. (2023). These approaches aim to either better understand underlying network properties or improve population size estimates from NSUM models by incorporating underlying network properties into the model.

There has also been significant progress towards improving NSUM estimators by multiplying size estimates by bias adjustment factors. Traditional NSUM models often produce biased results when respondents are more or less likely to know people from certain populations (barrier effects), do not know everything about their social contacts (transmission error), or cannot accurately recall everyone in their social network (recall error). Salganik et al. (2011) proposed the game of contacts, which involves interviewing members of the hard-to-reach population in order to estimate the transmission error. One then divides the NSUM population size estimate by the estimated transmission error in order to correct for the fact that respondents do not always know what groups individuals in their social network belong to. Haghdoost et al. (2018) provide a detailed review of methods to estimate scaling factors related to the transmission error. Feehan and Salganik (2016) also show that their proposed generalized scale-up estimators is equal to the basic scale-up estimator multiplied by three adjustment factors and propose several approaches to correct for these factors.

We briefly review the NSUM subpopulation size estimator proposed in Killworth et al. (1990), which we refer to as the MLE estimator (see McCormick (2020) or Laga et al. (2021) for a comprehensive review). Assuming that we know the true degrees, d_i, or are able to consistently estimate them from the ARD, the subpopulation size estimate for some

subpopulation H is given by:

$$\hat{N}_H = N \frac{\sum_{i=1}^{n} y_{iH}}{\sum_{i=1}^{n} d_i}, \tag{1}$$

where y_{iH} denotes the number of people respondent i reports knowing in subpopulation H, d_i is the degree of respondent i, N is the total population size, and N_H is the population size of H.

We propose a joint approach to solve two problems of the MLE estimator: (1) the subpopulation size estimator is biased when the average social network size of members of the subpopulation are larger or smaller than the general population, and (2) the estimator implicitly weights known subpopulations with larger size more than known subpopulations with smaller size. Our approach conveniently relies on only the original ARD, allowing researchers to obtain more accurate size estimates without collecting additional data like those needed for the game of contacts and the generalized scale-up estimator.

The rest of this paper is organized as follows. First, Section 2 provides additional background information about the degree ratio and presents the bias of the MLE estimator under certain conditions. Then, in Section 3, we introduce our approach to estimating the degree ratio using only the original ARD responses. Section 4 discusses scaling MLE estimates to correct for different known subpopulation sizes. We combined the two correction procedures in Section 5. We demonstrate the utility of both of these approaches with

both simulated (Section 6) and two real (Section 7) ARD surveys. Finally, we close with a discussion in Section 8.

2 Background

We first review model properties of the MLE estimator. Feehan and Salganik (2016) show that the MLE estimator is equivalent to their generalized scale-up estimator only when multiplied by three adjustment factors. The degree ratio adjustment factor arises because some populations have larger or smaller social network sizes on average than other populations. The authors define the degree ratio, δ_F, as

$$\delta_F = \frac{\text{avg \# connections from a member of H to F}}{\text{avg \# connections from a member of F to the rest of F}} = \frac{\bar{d}_{H,F}}{\bar{d}_{F,F}},$$

where F refers to the frame population, i.e. the collection of individuals who may be included as respondents in the ARD survey. Thus, if the degree ratio is 0.5 (i.e. there are only half as many links from H to F as there are from F to F), then the MLE estimator is one half the size of the generalized scale-up estimator. Since the MLE estimator implicitly assumes the average degrees of all subpopulations are identical, the estimator misattributes the small number of links to a small subpopulation size, rather than to small degrees. In order to estimate these adjustment factors, the authors propose collecting an additional

ARD survey given to members of the hard-to-reach population, called enriched ARD.

While we recognize the utility of enriched ARD, there are three significant limitations. First, enriched ARD is often prohibitively expensive to collect. The low cost and easy implementation of the NSUM are two of its key benefits. Collecting enriched ARD therefore undermines this advantage, making only well-funded studies able to collect the additional data. Second, it is impossible to collect enriched ARD on impossible-to-reach subpopulations such as individuals who died in an earthquake. Finally, it is inconvenient or impossible to collect enriched ARD for previous ARD studies. Therefore, the methods proposed in Feehan and Salganik (2016) can only be used for ARD moving forward and cannot correct for the biases in existing ARD surveys that did not already collect enriched ARD. Instead, we propose the first method to estimate the degree ratio using only the original ARD responses, allowing researchers to easily correct for bias introduced by the degree ratio.

We present two related findings connecting the bias of the MLE estimator to the degree ratio. For the following results, we assume perfect link reporting (i.e. no transmission error or recall error), that the respondents represent a simple random sample S of size n from the general population of size N, and that the frame population F is the general population. In this case, the inclusion probability for each respondent i is $\pi_i = n/N$. We consider two estimators, where either (1) the d_i are fixed and known, or (2) the d_i are estimated using

the L known subpopulations. In the first case, we can represent the MLE estimator as

$$\hat{N}_H = \frac{\sum_{i \in S}(y_{iH}/\pi_i)}{\frac{1}{N}\sum_{i \in S}(d_i/\pi_i)}, \tag{2}$$

while the second case takes the more complicated form given by

$$\hat{N}_H = \frac{\sum_{i \in S}(y_{iH}/\pi_i)}{\frac{1}{N}\sum_{i \in S}\left[\left(\sum_{k=1}^{L} y_{ik}/\sum_{k=1}^{L} N_k\right)/\pi_i\right]}. \tag{3}$$

Using these estimators, we present the following propositions, where the proofs are shown in the Appendix.

Proposition 1. *Consider the size estimate $\hat{N}_H$ in Equation (2), obtained from a survey with perfect link reporting and from a simple random sample of respondents. Then given known d_i, the bias of the unknown size estimate is approximately given by*

$$Bias(\hat{N}_H) \approx N_H\left(1 - \frac{\bar{d}_H}{\bar{d}_F}\right),$$

where $\bar{d}_H$ denotes the average degree of individuals in the hidden subpopulation and $\bar{d}_F$ denotes the average degree of individuals in the frame population.

Proposition 2. *Consider the size estimate $\hat{N}_H$ in Equation (3), obtained from a survey with perfect link reporting and from a simple random sample of respondents. Then given that the d_i are estimated using the remaining L known subpopulations in the survey, the*

bias of the unknown size estimate is approximately given by

$$Bias(\hat{N}_H) \approx N_H \left(1 - \frac{\bar{d}_H \sum_{k=1}^{L} N_k}{\sum_{k=1}^{L} \bar{d}_k N_k}\right),$$

where $\bar{d}_H$ denotes the average degree of individuals in the hidden subpopulation and $\bar{d}_F$ denotes the average degree of individuals in the frame population.

These results show that when the true degrees are known, the bias depends only on the true unknown subpopulation size and the ratio of the average degrees between the unknown subpopulation and the frame population, while the bias of the estimator when the degrees are also estimated additionally depends on the remaining known subpopulation sizes and the average degrees of individuals in each known subpopulation size. Proposition 2 also shows that the accuracy of the unknown size estimate depends on the specific relationship between the average degrees in subpopulations and the sizes of the subpopulations, and relatively large or small subpopulations will introduce more bias when paired with relatively large or small average degrees.

The degree ratio can significantly bias NSUM subpopulation size estimates. Shelley et al. (1995) found that HIV positive respondents and respondents who were dialysis patients had networks which were only about 2/3 the size than those of the average respondent in their survey. Therefore, given perfect responses to ARD questions, the Killworth et al. (1990) MLE estimator would estimate the size of these two subpopulations to be about

2/3 of the true size. The degree ratio may be more influential for even more stigmatized populations like sex workers or for more social populations like priests and doctors. Thus, our proposed approach to estimate the degree ratio is critical to improve estimates from the most commonly used NSUM estimator.

3 Degree Ratio Adjustment

Here we propose a method to correct for the bias present in MLE estimator. It would be sufficient to know $\bar{d}_k$ for all known and unknown subpopulations. However, these average degrees are unknown, making a direct approach impossible. Thus, our goal is to estimate the $\bar{d}_k$ in order to produce an adjustment factor for $\hat{N}_H$.

First, we propose building a degree ratio adjustment based on the results in Proposition 1. Given that the d_i need to be estimated in NSUM studies, it is intuitive to build an adjustment factor with all $\bar{d}_k$ based on the results in Proposition 2. However, we have found in practice that the more complicated form of the bias resulting from estimating d_i leads to poor estimates, despite the sounder theory. The issue arises because building an adjustment factor based on Proposition 2 involves estimating K average degrees (one for the hidden subpopulation and L for the known subpopulations). The difficulty associated with estimating each average degree compounds to form an adjustment factor that does

not work well in practice. On the other hand, while the theoretical bias presented in Proposition 1 ignores the uncertainty of the d_i estimates, the bias is comprised of only two average degrees, the same one for the hidden subpopulation and an overall average degree associated with the entire frame population. Thus, while an adjustment factor based on Proposition 2 should theoretically outperform an adjustment factor based on Proposition 1, the approach we present below relies on the results from the estimator constructed assuming that the d_i are known.

For the remainder of this paper, we let $\delta_H = \bar{d}_H/\bar{d}_F$, where we depart from the original notation from Feehan and Salganik (2016) to emphasize that the degree ratio (i) exists for both the subpopulations with known size and the hard-to-reach subpopulations, and (ii) varies across subpopulation.

To adjust for degree heterogeneity while avoiding sampling members of the hard-to-reach populations, we propose a novel method to estimate the degree ratio using only the original ARD, completely avoiding spending more resources on surveys. The basic assumption of our approach is that individuals who are "close" to a group have similar degrees to members of that group. This assumption, which we call the *degree-closeness* assumption, is supported by the observance of homophily in social networks across a number of different studies. In order to define who is close to a group, we rely on the latent surface

model for networks using ARD (McCormick and Zheng, 2015), hereafter referred to as the ARD latent surface model.

3.1 Approach

McCormick and Zheng (2015) show that their ARD latent surface model consistently estimates the position of ARD respondents and subpopulations on a hypersphere. The model assumes that the ARD responses for respondent i and subpopulation k follows

$$y_{ik}|d_i, \beta_k, \zeta, \eta_k, \theta_{(\boldsymbol{z}_i, \boldsymbol{\nu}_k)} \sim \text{Poisson}\left(d_i \beta_k \left(\frac{C_{p+1}(\zeta) C_{p+1}(\eta_k)}{C_{p+1}(0) C_{p+1}(\sqrt{\zeta^2 + \eta_k^2 + 2\zeta\eta_k cos(\theta_{(\boldsymbol{z}_i, \boldsymbol{\nu}_k)})})}\right)\right),$$

where d_i is the degree of respondent i, β_k is the prevalence of subpopulation k, ζ and η_k are scaling factors, $\boldsymbol{z}_i$ is the latent position of respondent i of the latent surface, $\boldsymbol{\nu}_k$ is the latent position of the center of subpopulation k, $\theta_{(\boldsymbol{z}_i, \boldsymbol{\nu}_k)}$ is the angular distance between the respondent's position and the center of subpopulation k, p is the dimension of the latent surface, and $C_{p+1}(\cdot)$ is the normalizing constant of the von-Mises Fisher distribution.

Thus, after fitting the ARD latent surface model to any ARD, we are able to calculate the distance between all survey respondents and all subpopulations included in the survey. Based on this formulation, we say that an individual i is closer than individual j to subpopulation k if $\theta_{(\boldsymbol{z}_i, \boldsymbol{\nu}_k)} < \theta_{(\boldsymbol{z}_j, \boldsymbol{\nu}_k)}$. For simplicity, we denote the distance matrix as

$\boldsymbol{\Theta}$ with indices $\theta_{i,k}$, where $\theta_{i,k}$ represents the estimated distance between respondent i and subpopulation k.

Based on $\boldsymbol{\Theta}$, we estimate δ_k for all subpopulations. First, we estimate $\bar{d}_F$ as the mean estimated degree all respondents in the ARD survey, i.e.,

$$\hat{\bar{d}}_F = \frac{1}{n}\sum_{i=1}^{n} \hat{d}_i, \quad \text{where} \quad \hat{d}_i = N \cdot \frac{\sum_{k \in known} y_{ik}}{\sum_{k \in known} N_k}$$

are obtained from the first stage of the Killworth et al. (1998) two-stage procedure obtained with all known subpopulations.

In order to estimate the numerator of the degree ratio, $\bar{d}_H$, we rely on the degree-closeness assumption. Since we do not know whether respondents are a member of subpopulation k, we are unable to estimate $\bar{d}_k$ directly. Instead, we estimate $\bar{d}_k$ by averaging the degrees of respondents that are close to subpopulation k. One could take several approaches to obtain these estimates. One approach is to calculate a weighted average of $\hat{d}_i$ based on their distances $\boldsymbol{\Theta}_{1:n,k}$, where $1:n,k$ indexes over the distances between all respondents and subpopulation k and the weights are larger for respondents with smaller distances. Another approach is to use the P smallest distances for each subpopulation and average the degrees of these P respondents, where P is a user-specified value. Similarly, we could define a separate P_k for each subpopulation. In our studies, we found that the weighted approach is more robust to outliers with extremely large or small degrees, while

the separate P_k approach offers the most accurate results when P_k is chosen accurately. In order to balance the robustness of the weighting procedure with the potential accuracy of using the smallest P_k distance, in Section 3.2 we propose classifying the respondents as either close or not based on the estimated distances. Then, $\hat{\bar{d}}_k$ is calculated as the average degree of respondents who are classified as close to subpopulation k.

To summarize, our proposed approach to estimate δ_k for all subpopulations given only the original ARD is as follows:

1. Fit the ARD latent surface model on the ARD, $\boldsymbol{Y}$, with M posterior samples.
2. Estimate the distance matrix $\hat{\boldsymbol{\Theta}}$ as the posterior mean of the distance matrices $\boldsymbol{\Theta}^m$ for sample m.
3. For each subpopulation k:
   (a) Classify $n_{close,k}$ respondents as close or not by each column of $\hat{\boldsymbol{\Theta}}$. Assign these respondents to class C_k.
   (b) Calculate $\hat{\delta}_k = \left(\frac{1}{n_{close,k}} \sum_{i \in C_k} \hat{d}_i\right) / \left(\frac{1}{n} \sum_{i=1}^{n} \hat{d}_i\right)$.

Alternatively, if desired, the median or maximum a posteriori could be used to estimate $\hat{\boldsymbol{\Theta}}$, consistent with the Bayesian framework.

Given the sociological interest in the degree ratio for different subpopulations, we recommend estimating $\hat{\delta}_k$ for all subpopulations of interest in the ARD survey. However, given that the true sizes are known for all known subpopulations, it only makes sense to adjust the size estimates of the unknown subpopulations by $\hat{N}_H^{adj} = \hat{N}_H^{MLE}/\hat{\delta}_H$.

The question remains, how do we classify respondents as close?

3.2 Classifying Respondents as Close

We propose a classification function based on the range of estimated distances for each subpopulation. Specifically, denote the minimum and maximum of the k^{th} column of $\hat{\mathbf{\Theta}}$ as $\min_k$ and $\max_k$. Then for a given grouping factor Δ_k, a respondent is classified as close and belongs to class C_k if $\hat{\theta}_{i,k} < [\min_k + \Delta_k(\max_k - \min_k)]$. Thus, we partition the range of estimated distances for each subpopulation as close or not, i.e. the closest $X\%$ of distances, rather than classifying the closest $X\%$ of respondents. This approach naturally handles different distributions of estimated distances across the subpopulations.

We recommend choosing Δ_k via either visualizations of the clusters as a function of the estimated distance or using leave-one-out predictive performance with the subpopulations with known N_k. Specifically, we propose the following two approaches.

Visualizations: First, we propose choosing the number of clusters based on visual-

izations like those in Figure 1. For each subpopulation, the user may choose a Δ_k which produces a reasonable and intuitive division between the close and distant respondents. In many situations, there is no obvious choice for Δ_k. Despite this, we still find visualizations to be one of the best approaches to choosing Δ_k, as it provides a straightforward way to vary Δ_k for each subpopulation k, including the subpopulations with unknown size.

Predictive performance: Alternatively, we propose using predictive performance to select Δ_k. Specifically, in order to estimate Δ_k for subpopulation k, one should select the Δ_k which results in the best predictive performance of the remaining known subpopulations (for an unknown subpopulation u, all known subpopulations will be used), where the original MLE size estimates are obtained by $\hat{N}_k = N(\sum_{i=1}^{n} y_{ik})/(\sum_{i=1}^{n} \hat{d}_i)$, where the $\hat{d}_i$ are still estimated using all known subpopulations. While any metric may be used, we recommend metrics that standardize by size, such as the mean absolute percent error (defined as (actual - predicted) / abs(actual)), which does not weight the predictive accuracy of large subpopulations more than the predictive accuracy of small subpopulations. For this reason, we do not recommend using metrics like root mean squared error. Thus, the metric is optimized for $\Delta_k \in (0, 1)$, such that the approximate leave-one-out predictive performance is minimized.

4 Scaling Procedure

From the numerator of Equation (1), the larger y_{iH} values typically associated with larger subpopulations have more influence on $\hat{N}_H$ than the smaller y_{iH} values associated with smaller subpopulations. To correct for this, we propose rescaling size estimates using equally weighted subpopulations, regardless of subpopulation size. We show that this scaling is trivial to apply in practice but can lead to substantially more reasonable size estimates.

While traditional NSUM size estimates, including those from the Killworth et al. (1990) MLE estimator, directly produce interpretable values on the correct scale, these methods can still benefit from additional scaling. Bayesian estimators that do not fix N_k for the known subpopulations, like those of Zheng et al. (2006) and Laga et al. (2023), instead scale size estimates after MCMC sampling using some function involving all or a subset of the known sizes. We have found that the size estimates of the Overdispersed model proposed by Zheng et al. (2006) are almost identical to the leave-one-out size estimates of the Killworth et al. (1990) MLE under certain scaling conditions. However, we have also found that slight modifications of the weighting procedure can significantly improve the accuracy of size estimates of relatively small subpopulations while only slightly decreasing the accuracy of size estimates of relatively large subpopulations. Based on these findings

and the fact that hard-to-reach populations of interest are typically smaller than most known subpopulations used in ARD surveys, we propose rescaling Killworth et al. (1990) MLE size estimates by correcting for different known subpopulation sizes.

We propose scaling all subpopulation sizes using all n_{known} known subpopulations, weighted equally, instead of proportional to size like traditional NSUM estimators do, because the importance of a group is not necessarily proportional to the population size. This first requires producing leave-one-out size estimates for the known N_k. Next, we define a scaling constant C by

$$C = \frac{1}{n_{known}} \sum_{k \in known} \left(\frac{\hat{N}_k}{N_k} \right). \tag{4}$$

and scale the subpopulation size estimates by $\hat{N}_k^{scaled} = \hat{N}_k / C$. Similarly, we scale the degree estimate for respondent i by $\hat{d}_i^{scaled} = \hat{d}_i \cdot C$. As is standard in current NSUM studies, the corresponding standard errors can be calculated via bootstrap.

5 Combined Procedure

Given the degree ratio adjustment and scaling procedure described above, a natural question is whether and how to combine the two methods. Specifically, we could choose one of the four following approaches: (i) adjust for the degree ratio only, (ii) scale the estimates only, (iii) adjust for the degree ratio first and then scale the estimates, or (iv) scale the

estimates first and then adjust for the degree ratio. Based on the analysis of two real ARD studies (Section 7), we propose researchers implement the fourth option to scale the estimates first and then adjust for the degree ratio. To summarize, the proposed methodology to estimate the unknown size of a subpopulation is presented in Algorithm 1.

Algorithm 1: Scaling and degree ratio adjustment combined procedure

Result: Scaled and adjusted N_H estimates

Set n_{known} equal to the number of subpopulations with known sizes N_k;

Estimate $\hat{d}_i = N\dfrac{\sum_{k\in known} y_{ik}}{\sum_{k\in known} N_k}$ for all respondents i using all known subpopulations;

Get approximate leave-one-out size estimates for known subpopulations,

$\hat{N}_k = N\dfrac{\sum_{i=1}^{n} y_{ik}}{\sum_{i=1}^{n} \hat{d}_i}$;

Calculate $C = \dfrac{1}{n_{known}} \sum_{k\in known} \left(\frac{\hat{N}_k}{N_k}\right)$;

Scale approximate leave-one-out size estimates for known subpopulations and unknown subpopulation, $\hat{N}_k^{scaled} = \hat{N}_k/C$ and $\hat{N}_H^{scaled} = \hat{N}_H/C$;

Scale estimated degrees, $\hat{d}_i^{scaled} = \hat{d}_i \cdot C$;

Choose Δ_H and C_H for the unknown subpopulation via either visualization or leave-out-one predictive performance;

Calculate $\hat{\delta}_H = \left(\frac{1}{n_{close,u}} \sum_{i\in C_H} \hat{d}_i^{scaled}\right) / \left(\frac{1}{n} \sum_{i=1}^{n} \hat{d}_i^{scaled}\right)$;

Adjust scaled unknown subpopulation size estimate, $\hat{N}_H^{scaled,adjusted} = \dfrac{\hat{N}_H^{scaled}}{\hat{\delta}_H}$;

Intuitively, scaling first and then adjusting for the degree ratio appears to result in the

most accurate estimates because the degree ratio approach relies on stronger assumptions than the scaling approach. Specifically, the scaling approach is designed to correct for the fact that empirical MLE size estimates underestimate large subpopulations and overestimate small subpopulations, regardless of the underlying reason. On the other hand, the degree ratio approach works well when the remaining bias is due to the difference in average degrees and when we are able to accurately estimate that bias based on the behavior of the other subpopulations. For example, respondents often over-report and under-report how many people they know in small and large subpopulations, respectively, introducing recall error. By first correcting for other biases like this recall error through scaling, the degree ratio approach is more likely to estimate the actual degree ratio. However, empirical evidence still suggests that the scaling corrects for some of the degree ratio, based on the results presented later in Table 1. If the scaling was independent of the true underlying degree ratio, we would expect the improvements of the degree ratio adjustment to improve the size estimates equally well regardless of order.

While the combined procedure could be used to also modify the population size estimates for the known subpopulations, in a practical setting this does not benefit the researcher because the true population sizes are already known. However, we still encourage researchers to estimate the $\hat{\delta}_k$ if the degree ratio is of scientific importance for the known

subpopulations. One example would be for the "priest" subpopulation associated with the Rwanda Biomedical Center/Institute of HIV/AIDS, Disease Prevention and Control Department (RBC/IHDPC) et al. (2012) NSUM study. While we remove the "priest" subpopulation from our analysis below in order to demonstrate that our procedure still works after removing highly influential data points, the combined procedure did significantly improve the NSUM size estimates for "priest," and the corresponding degree ratio estimate supports the previous findings of McCarty et al. (2001) that ARD surveys given to clergy yielded larger average network sizes than ARD surveys given to a representative sample.

While we proposed the above methodology as a general approach to correct for the degree ratio, in practice we recommend using caution or avoiding adjusting the size estimates for subpopulations corresponding to names like "Michael" or "Kristina" in the McCarty et al. (2001) ARD survey. While the popularity of certain names may be related to age and similar demographics, we find that empirically the association is less pronounced, and thus they have less influence on the degree ratio, than other behavioral factors. Applying the degree ratio correction in settings with weak associations risks correcting for spurious relationships in the data rather than for true signals. Furthermore, the NSUM typically more accurately estimates the size of name related subpopulations than other subpopulations. We demonstrate in Section 7 that our proposed methodology significantly improves size es-

timates for all non-name related subpopulations. In practice, one may also exclude certain other non-name related subpopulations that are believed to violate the degree-closeness assumptions.

6 Simulation Study

We simulate a network from a stochastic block model (SBM) with 800 respondents and 8 groups. We set each group size to be 100. The within-group connectivity (i.e. the diagonal of the connectivity matrix) is given by 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, in order to have a range of connectivity. All between-group connectivity probabilities are 0.05. In this design, the first group with within-group connectivity of 0.05 is just as likely to form ties to members of different groups and they are to form ties to members of their own group.

In order to evaluate the model performance, we implement Algorithm 1, where the true degrees are used since we do not work with large enough populations to have accurate degree estimates, unlike for real applications. Specifically, we perform a leave-one-out procedure, where we estimate the scaled and adjusted subpopulation for each subpopulation sequentially, treating each successive subpopulation as unknown. For this simulation study and for both real datasets, we ran the MCMC sampler for 30,000 iterations, removed the

first 2000 points, and thinned each chain by 10, resulting in 2,800 posterior samples.

We fit the ARD latent surface model with $p = 3$ and fixed subpopulation 4, 5, and 7 on the latent sphere, well separated from each other. From this model fit, we calculated $\boldsymbol{\Theta}$ using the posterior mean of the estimated distances. We plot these estimated distances against the true degree in Figure 1. The shape of the points denotes whether the respondent is a member of that subpopulation. We find that on average, the model is able to separate the respondents based on the estimated distance. Given the small with-in group connectivity, the model does a relatively poor job of estimating the distances for group 1 compared to the other groups.

Next, we choose the grouping factors Δ_k via predictive performance. We refer again to Figure 1, where we note the color indicates whether respondents are classified as close or not. For example, for the second facet, we observe that the classification divides the true members into two separate groups instead of classifying them together. In this instance, using a larger Δ_k chosen by visualization may do a better job of separating the data. On the other hand, for the third facet, we see that that the classification actually correctly classifies all members of the group as a single class. In practice, we do not observe the true group membership, so we must only rely on the vertical distribution of the points.

Finally, the results for this simulation study error are shown in Figure 2, where the

original Killworth et al. (1990) MLE estimates are shown in pink, our scaled and adjusted estimates in green, and blue arrows are added on subpopulations where our adjusted estimates have smaller absolute relative error. In this study, we outperform the Killworth et al. (1998) estimator for six of the eight subpopulations. The only population where the scaled and adjusted estimate is worse than the original estimate is for the population that violates the degree-closeness assumption since the within-group connectivity is equal to the between-group connectivity. The percent reduction in average mean absolute percent error is presented in Table 1. The scaled and adjusted size estimates are significantly closer to the truth than the original Killworth et al. (1990) MLE estimates. It is clear in this simulation study that our proposed method is able to consistently correct for the degree ratio and substantially improve the existing Killworth et al. (1990) MLE estimates when the degree-closeness assumption holds.

7 Network Scale-up Method Studies

In this section, we apply our scaling procedure to two real ARD surveys. We show that despite its simplicity, the proposed scaling procedure substantially improves size estimates. We also show that estimating and adjusting the degree ratio further improves size estimates. We follow the same procedure outlined in the simulation study to evaluate the performance

of our proposed methods, except we now estimate the degrees for all respondents using the remaining known subpopulations before estimating the subpopulation sizes. This matches the procedure used when estimating the unknown subpopulations, where only the ARD responses in the known subpopulations are used to estimate the degrees and subpopulation sizes for the unknown subpopulations.

Note that the final results presented below are after removing the subpopulations corresponding to names. Twelve of the 29 known subpopulations and 13 of the 22 known subpopulations correspond to names in the McCarty et al. (2001) and Rwanda Biomedical Center/Institute of HIV/AIDS, Disease Prevention and Control Department (RBC/IHDPC) et al. (2012) surveys, respectively. For the Rwanda Biomedical Center/Institute of HIV/AIDS, Disease Prevention and Control Department (RBC/IHDPC) et al. (2012) survey, we present results both with and without the "priest" subpopulation because it appears to be an extreme outlier in the dataset.

7.1 McCarty ARD Study

First, we apply our proposed scaling methods to the ARD first collected and presented in McCarty et al. (2001). This dataset contains responses from 574 respondents about 32 subpopulations, 3 of which are unknown (individuals who are homeless, have been raped,

or are HIV positive). We remove 53 respondents for having 1 or more missing responses (47 of those 53 respondents had only 1 missing response), resulting in 521 respondents. As the primary purpose of this work is to evaluate the performance of our proposed scaled and adjusted estimator compared to the Killworth et al. (1990) MLE estimator, we do not study the effect of removing these respondents with missing data.

We compare the final scaled and adjusted size estimates for the 17 remaining subpopulations against the original Killworth et al. (1990) MLE size estimates in Figure 3a. For 9 of the 17 known subpopulations, our scaled and adjusted estimates are closer to the truth. The calculated grouping factors ranged from 0.70 to 1.97. For 6 of the 9 subpopulations, the improvement is reasonably substantial, compared to relatively small errors introduced for the remaining 8 subpopulations. We also compare the percent reduction in average mean absolute percent error in Table 1. Prior to scaling and adjusting, the Killworth et al. (1990) MLE significantly overestimates the size of relatively small subpopulations, while after scaling and adjusting there is very little relationship between the estimated bias and the size of the subpopulation. For the McCarty ARD survey, our proposed methods greatly outperform the original Killworth et al. (1990) MLE estimator, at only the cost of additional computing time.

7.2 Rwanda Meal ARD Study

Next, we consider the Rwanda Meal ARD survey (Rwanda Biomedical Center/Institute of HIV/AIDS, Disease Prevention and Control Department (RBC/IHDPC) et al., 2012; Feehan et al., 2016). In 2011, researchers collected ARD from 4,669 respondents in Rwanda in order to estimate the size of four key populations, female sex workers (FSW), male clients of sex workers (MCSW), men who have sex with men (MSM), and people who inject drugs (IDU). Rwanda Biomedical Center/Institute of HIV/AIDS, Disease Prevention and Control (RBC/IHDPC) and their partners require accurate size estimates of these unknown subpopulations in order to plan and implement efficient HIV prevention strategies for current HIV cases and understand the trend of HIV cases across time. As discussed in Laga et al. (2021) and McCormick (2020), the NSUM can efficiently and cheaply estimate the size of multiple key populations simultaneously by adding additional questions to an existing survey, making the NSUM a natural choice for the RBC/IHDPC to better understand these key populations.

One of the primary motivations of the survey was to compare the results of NSUM size estimates between two definitions of whether a respondent "knows" someone (Feehan et al., 2016). The first definition, called the *acquaintance* definition, quantifies the "people the respondent has had some contact with — either in person, over the phone, or on

the computer in the previous 12 months." The *meal* definition restricts the acquaintance definition, quantifying the "people the respondent has shared a meal or drink with in the past 12 months, including family members, friends, coworkers, or neighbors, as well as meals or drinks taken at any location, such as at home, at work, or in a restaurant." Feehan et al. (2016) were able to show that estimates from the meal definition were consistently closer to the known sizes than estimates from the acquaintance definition. While the authors were unable to confidently extend this finding to subpopulations with unknown size (e.g. FSW), it is not unlikely that these estimates for unknown subpopulations would also be more accurate.

In order to use the dataset least prone to errors, for our analysis, we consider only the dataset collected from the meal definition. Given that the meal definition implies a stronger relationship between the respondent and their social connections, it is reasonable to assume that the respondent knows more about each person they recalled, reducing the transmission error. Furthermore, given that the pool of potential connections is smaller, respondents should have an easier time recalling everyone in a given subpopulation, also reducing recall error. In order to show that our proposed method accurately accounts for the bias introduced by differences in average network sizes between groups, it is vital to use a dataset that faces smaller biases from other sources.

In this study, we analyzed responses from 2405 respondents about 22 known subpopulations and the 4 key populations. Only one respondent was removed for a missing response to how many people they know who are Muslims. The calculated grouping factors ranged from 0.64 to 2.43, with a median of 0.86. The estimated distances for the known subpopulations are plotted against the estimated degrees in Figure 3b, where the color denotes whether the respondent is classified as close for each subpopulation. Compared to the SBM simulation results, a larger percentage of respondents are classified as being close to the corresponding subpopulation. Note that while not shown here, when estimating all subpopulations, only 8 respondents are classified as being close to the priest subpopulation, which the Killworth et al. (1990) MLE estimator estimates most poorly.

We also compare the relative error of the Killworth et al. (1990) MLE and our adjusted estimates in Figure 4b, where the original Killworth et al. (1990) MLE estimates are shown in pink and our adjusted estimates are shown in green. We add blue arrows on subpopulations where the relative error of our adjusted estimate is smaller than the relative error of the original Killworth et al. (1990) MLE estimate. Visually, it is clear that adjusting the estimate via our approach significantly improves the overall performance of the Killworth et al. (1998) estimator. Numerically, our adjusted estimates perform better in four of the nine non-name and non-priest known subpopulations. However, the adjusted estimators

perform significantly better than the Killworth et al. (1990) MLE estimator for those four subpopulations, while only performing slightly worse for the remaining subpopulations. Specifically, we find the mean absolute percent error drops from approximately 2.4 to 1.2 for the four subpopulations where our adjusted estimates perform better, while only increasing from approximately 0.29 to 0.59 for the remaining five subpopulations. Furthermore, our approach improved the size estimates of the four smallest subpopulations, suggesting that our approach will likely improve the size estimates of the hard-to-reach subpopulations as well. We present numerical results in Table 1, where scaling and adjusting reduced the average mean absolute percent error 30%. With the "priest" subpopulation included, the percent reduction in average mean absolute percent error is significantly better for all proposed methods when compared to the Killworth et al. (1990) MLE estimates, reducing the average mean absolute percent error by 64% across all non-name subpopulations. The scaled and adjusted size estimates are substantially better when including the priest subpopulation because priests have significantly larger social networks than other members of the population, emphasizing that our proposed methods works especially well when there are clear differences in social network sizes across subpopulations (McCarty et al., 2001).

8 Discussion

We have demonstrated through both simulations and through two data examples that in the presence of homophily, our proposed scaling and degree ratio adjustment can substantially reduce the bias of the Killworth et al. (1990) MLE estimator. However, it is important to recognize that while our degree-closeness assumption does not hold for all subpopulations, it is more likely to hold for the marginalized hard-to-reach populations we are primarily interested in studying. Previous studies have shown that these marginalized populations show substantial homophily. We rely on the performance of our proposed approach with respect to the known subpopulations out of necessity since we are unable to access the quality of NSUM estimators for the unknown subpopulations.

The key novelty of this paper is that our proposed methods solve two very difficult problems related to popular NSUM estimators *without* using auxiliary data. Methods that use auxiliary data may intuitively perform better than our combined procedure and we encourage researchers to use additional data when available. However, collecting additional data is often impossible, necessitating an approach that recycles the available data. The scaling procedure in particular can be performed in a few seconds and significantly improve initial NSUM size estimates.

Furthermore, as we presented in this work, NSUM models should be evaluated using

performance metrics that do not favor large subpopulations. Metrics like root mean squared error are dominated by these populations like "people who have diabetes" or "people are twins" so that the accuracy of size estimates corresponding to populations like "people who were murdered" or "people who committed suicide" are not influential.

As with all methods used to estimate the size of hard-to-reach subpopulations, it is difficult to understand, model, and account for all sources of bias. In some cases, accounting for one source of bias may result in worse estimates if the other sources of bias are ignored. Continued research is needed to understand how the different NSUM biases interact together. Is it sufficient to account for each form of bias independently? We believe the NSUM holds an important role in providing accurate, quick, and affordable size estimates and urge future researchers to continue developing this promising method.

References

Bernard, H. R., Johnsen, E. C., Killworth, P. D., and Robinson, S. (1989). Estimating the size of an average personal network and of an event subpopulation. In *The Small World*, pages 159–175. Ablex Press.

Breza, E., Chandrasekhar, A. G., McCormick, T. H., and Pan, M. (2020). Using aggregated

relational data to feasibly identify network structure without network data. *American Economic Review*, 110(8):2454–84.

Feehan, D. M. and Salganik, M. J. (2016). Generalizing the network scale-up method: a new estimator for the size of hidden populations. *Sociological methodology*, 46(1):153–186.

Feehan, D. M., Umubyeyi, A., Mahy, M., Hladik, W., and Salganik, M. J. (2016). Quantity versus quality: A survey experiment to improve the network scale-up method. *American journal of epidemiology*, 183(8):747–757.

Haghdoost, A., Gohari, M. A., Mirzazadeh, A., Zolala, F., and Baneshi, M. R. (2018). A review of methods to estimate the visibility factor for bias correction in network scale-up studies. *Epidemiology and health*, 40.

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(4):289–312.

Killworth, P. D., McCarty, C., Bernard, H. R., Shelley, G. A., and Johnsen, E. C. (1998). Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. *Evaluation Review*, 22(2):289–308.

Laga, I., Bao, L., and Niu, X. (2021). Thirty years of the network scale-up method. *Journal of the American Statistical Association*, 116(535):1548–1559.

Laga, I., Bao, L., and Niu, X. (2023). A correlated network scale-up model: Finding the connection between subpopulations. *Journal of the American Statistical Association*.

Maltiel, R., Raftery, A. E., McCormick, T. H., and Baraff, A. J. (2015). Estimating population size using the network scale up method. *The Annals of Applied statistics*, 9(3):1247.

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(1):28–39.

McCormick, T. H. (2020). The network scale-up method. *The Oxford Handbook of Social Networks*, page 153.

McCormick, T. H. and Zheng, T. (2015). Latent surface models for networks using aggregated relational data. *Journal of the American Statistical Association*, 110(512):1684–1695.

Rwanda Biomedical Center/Institute of HIV/AIDS, Disease Prevention and Control Department (RBC/IHDPC), School of Public Health (SPH) University of Rwanda, UNAIDS, and ICF International (2012). Estimating the size of populations through a household survey.

Salganik, M. J., Mello, M. B., Abdo, A. H., Bertoni, N., Fazito, D., and Bastos, F. I. (2011). The game of contacts: estimating the social visibility of groups. *Social Networks*, 33(1):70–78.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Shelley, G. A., Bernard, H. R., Killworth, P., Johnsen, E., and McCarty, C. (1995). Who knows your hiv status? what hiv+ patients and their network members know about each other. *Social Networks*, 17(3-4):189–217.

Teo, A. K. J., Prem, K., Chen, M. I., Roellin, A., Wong, M. L., La, H. H., and Cook, A. R. (2019). Estimating the size of key populations for HIV in Singapore using the network scale-up method. *Sexually Transmitted Infections*, 95(8):602–607.

UNAIDS (2021). End inequalities. end aids. global aids strategy 2021-2026.

Zheng, T., Salganik, M. J., and Gelman, A. (2006). How many people do you know in prison? Using overdispersion in count data to estimate social structure in networks. *Journal of the American Statistical Association*, 101(474):409–423.

9 Tables and Figures

Table 1: Percent reduction in average mean absolute percent error (MAPE) for the scaled and adjusted size estimates for the SBM simulation, McCarty, and Rwanda Meal studies. Percent reduction is calculated by $100 * (MAPE^{MLE} - MAPE^{scaled,adjusted})/(MAPE^{MLE})$

Data Set	**Subpopulations**	**Scaled and Adjusted**
SBM Simulation	All	40%
McCarty	All	52%
	Non-names	59%
Rwanda Meal	All	21%
	Non-names, with priest	64%
	Non-names, no priest	30%

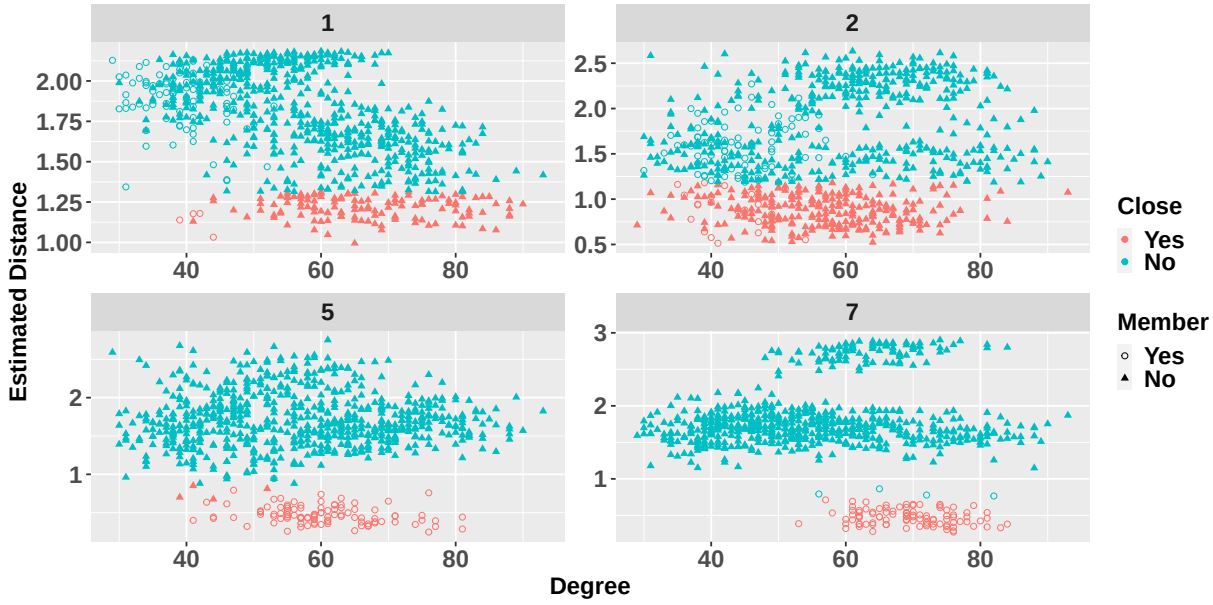


Figure 1: Estimated latent distance to a subset of the subpopulations plotted against the true degree on the x-axis. Each facet represents a different subpopulation. Members of the subpopulation are plotted as open circles, while non-members are plotted as closed triangles. Respondents classified as close are shown in pink, otherwise in cyan.

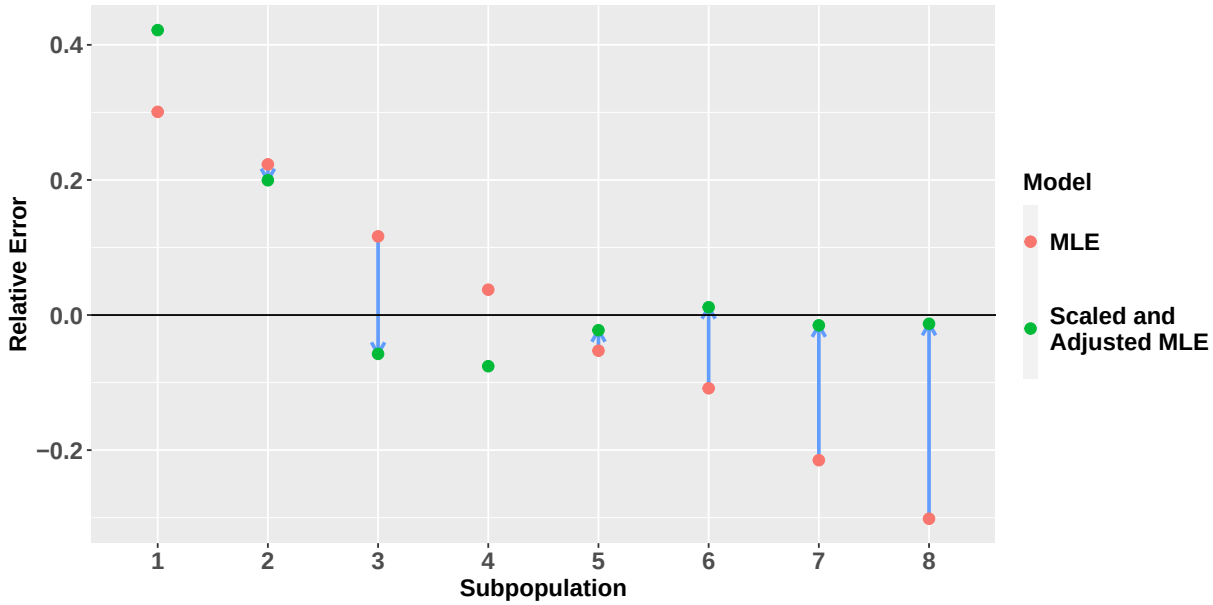
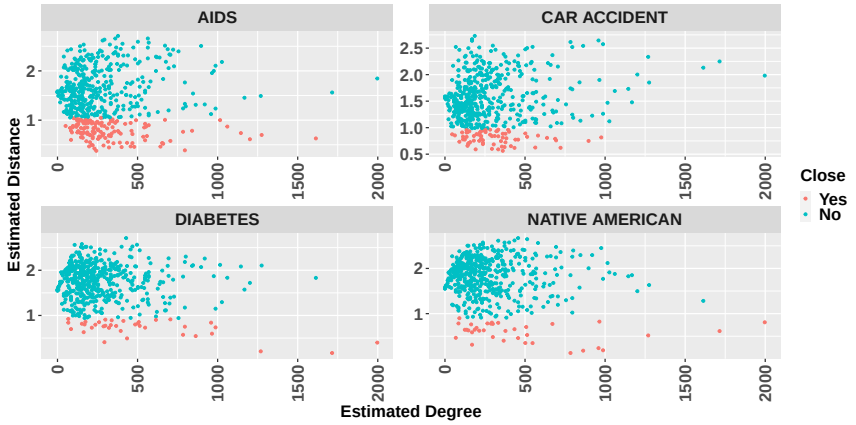


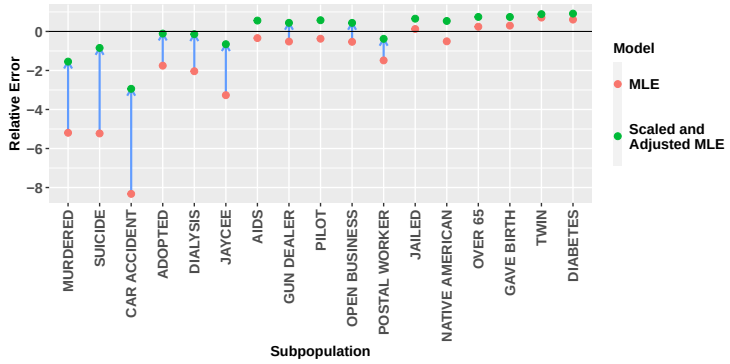
Figure 2: Relative error point estimates for SBM size estimates from the raw Killworth et al. (1990) MLE and scaled and adjusted Killworth et al. (1990) MLE estimates. Relative error is calculated by $100 * (Truth - estimates)/Truth$. Subpopulations are ordered from smallest estimated $\hat{\delta}_k$ to largest. Arrows are added on subpopulations where the scaled and adjusted estimates have smaller relative error.



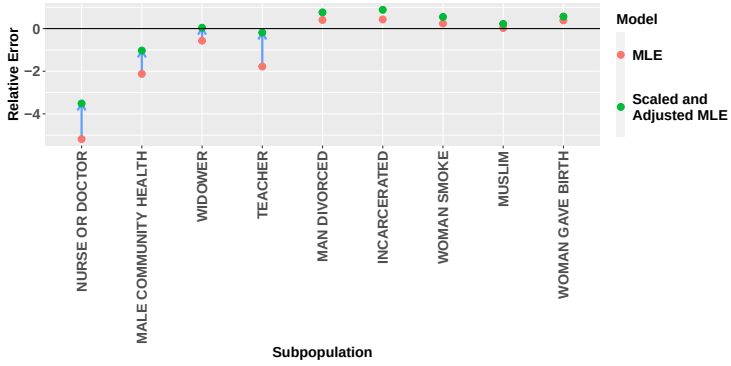
(a)

(b)

Figure 3: Estimated latent distance plotted against the estimated degree on the x-axis for the McCarty study (a) and for the Rwanda Meal study (b). Each facet represents a different subpopulation. Respondents classified as close to the subpopulation are plotted in pink, otherwise plotted in cyan. For visualization only, the estimated degrees are from using all known subpopulations.



(a)



(b)

Figure 4: Relative error point estimates for non-name subpopulations of McCarty (a) and non-name and non-priest subpopulations of Rwanda Meal (b) size estimates from the original Killworth et al. (1990) MLE and scaled and adjusted Killworth et al. (1990) MLE estimates. Relative error is calculated by $100*(Truth-estimates)/Truth$. Subpopulations are ordered alphabetically. Arrows are added on subpopulations where the scaled and adjusted estimates have smaller relative error.

Appendices

A Proofs

A.1 Proof of Proposition 1

Proof. Given perfect link reporting,

$$E\left[\sum_{i\in S}(y_{iH}/\pi_i)\right] = \sum_{i=1}^{N} y_{iH} = \sum_{j\in H} d_j = \text{sum of degrees in subpopulation H} \tag{5}$$

and

$$E\left[\sum_{i\in S}(d_i/\pi_i)\right] = \sum_{i=1}^{N} d_i = \text{sum of degrees in the frame population}$$

The last equality in Equation (5) follows directly from the fact that the number of links from the frame population to members of H in an undirected graph is equal to the number of links from the members of H to the frame population, i.e. the sum of their degrees. Therefore, the numerator is an unbiased estimate of the sum of the degrees of members of H, and the denominator is an unbiased estimate of the average degree size in the frame population. Then, since the ratio of two unbiased estimators is approximately unbiased (Särndal et al., 2003), the expected value of the estimator in Equation (2) is approximately

given by

$$E\left[\hat{N}_H\right] \approx \frac{\sum_{j\in H} d_j}{\frac{1}{N}\sum_{i=1}^{N} d_i} = N_H \frac{\bar{d}_H}{\bar{d}_F} = \delta_H N_H,$$

where $\bar{d}_H$ represents the average degree of members of H and $\bar{d}_F$ represents the average degree of all members of the frame population. □

A.2 Proof of Proposition 2

Proof. Given perfect link reporting, the numerator remains unchanged from Proposition 1, where

$$E\left[\sum_{i\in S}(y_{iH}/\pi_i)\right] = \sum_{i=1}^{N} y_{iH} = \sum_{j\in H} d_j = \text{sum of degrees in subpopulation H.}$$

For the denominator, we have

$$
\begin{aligned}
E\left\{\frac{1}{N}\sum_{i\in S}\left[\left(\sum_{k=1}^{L} y_{ik}/\sum_{k=1}^{L} N_k\right)/\pi_i\right]\right\} &= \frac{1}{N\sum_{k=1}^{L} N_k} E\left(\sum_{i\in S}\sum_{k=1}^{L} y_{ik}/\pi_i\right) \\
&= \frac{1}{N\sum_{k=1}^{L} N_k} E\left(\sum_{k=1}^{L}\sum_{i\in S} y_{ik}/\pi_i\right) \\
&= \frac{1}{N\sum_{k=1}^{L} N_k}\left(\sum_{k=1}^{L}\sum_{i=1}^{N} y_{ik}\right) \\
&= \frac{1}{N\sum_{k=1}^{L} N_k}\left(\sum_{k=1}^{L}\sum_{j\in k} d_j\right) \\
&= \frac{1}{N\sum_{k=1}^{L} N_k}\left(\sum_{k=1}^{L} N_k\bar{d}_k\right)
\end{aligned}
$$

Then, since the ratio of two unbiased estimators is approximately unbiased (Särndal et al., 2003), the expected value of the estimator in Equation (3) is approximately given by

$$
E\left[\hat{N}_H\right] \approx \frac{N_H\bar{d}_H\sum_{k=1}^{L} N_k}{\sum_{k=1}^{L}\bar{d}_k N_k},
$$

where $\bar{d}_k$ represents the average degree of members of subpopulation k. □