Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

A Correlated Network Scale-up Model: Finding the Connection Between Subpopulations

Ian Laga, Le Bao & Xiaoyue Niu

To cite this article: Ian Laga, Le Bao & Xiaoyue Niu (2023): A Correlated Network Scale-up Model: Finding the Connection Between Subpopulations, Journal of the American Statistical Association, DOI: 10.1080/01621459.2023.2165929

To link to this article: https://doi.org/10.1080/01621459.2023.2165929

View supplementary material

Accepted author version posted online: 06 Jan 2023.

Submit your article to this journal

Article views: 54

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

Check for updates

A Correlated Network Scale-up Model: Finding the Connection Between Subpopulations

Ian Laga, Le Bao, and Xiaoyue Niu*

*Corresponding author: Xiaoyue Niu, Xiaoyue@psu.edu

Abstract

Aggregated relational data (ARD), formed from "How many X's do you know?" questions, is a powerful tool for learning important network characteristics with incomplete network data. Compared to traditional survey methods, ARD is attractive as it does not require a sample from the target population and does not ask respondents to self-reveal their own status. This is helpful for studying hard-to-reach populations like female sex workers who may be hesitant to reveal their status. From December 2008 to February 2009, the Kiev International Institute of Sociology (KIIS) collected ARD from 10,866 respondents to estimate the size of HIV-related groups in Ukraine. To analyze this data, we propose a new ARD model which incorporates respondent and group covariates in a regression framework and includes a bias term that is correlated between groups. We also introduce a new scaling procedure utilizing the correlation structure to further reduce biases. The resulting size estimates of those most-at-risk of HIV infection can improve the HIV response efficiency in Ukraine. Additionally, the proposed model allows us to better understand two network features without the full network data: 1. What characteristics affect who respondents know, and 2. How is knowing someone from one group related to knowing people from other groups. These features can allow researchers to better recruit marginalized individuals into the prevention and treatment programs. Our proposed model and several existing NSUM models are implemented in the networkscaleup R package.

Keywords: Size estimation, small area estimation, key populations, aggregated relational data.

Accepted Manuscript

1 Introduction

Hard-to-reach populations are groups of people that are not easily sampled by commonly used surveys, potentially due to their stigmatized status (e.g. female sex workers) or their infeasibility to be reached (e.g. people who committed suicide). There is a long history of developing methods to estimate the sizes of hard-to-reach populations, such as direct survey estimates, capture-recapture, and venue-based sampling, but still no method has emerged as the gold standard. UNAIDS/WHO outlined strengths and weaknesses of many of the methods (UNAIDS and WHO, 2010). Direct estimates typically use random surveys of the general population and calculate what percent of respondents belong to the hard-to-reach population, but are inefficient or lead to biased results for small and hard-to-reach populations. For reasonable sample sizes, many of these surveys do not even reach the hard-to-reach populations, making it impossible to estimate the population size. Other methods require working with members of the hard-to-reach populations directly, which can lead to more accurate and precise estimates, but it is often difficult to directly contact populations that desire to stay hidden due to poor treatment and negative social stigma.

Originally motivated by estimating the size of people who have died in the 1985 Mexico City earthquake (Bernard et al., 1989), the network scale-up method (NSUM) avoids the need for samples from the hard-to-reach population entirely, making it more convenient to implement and shining a light on the scale of impossible-to-reach population. NSUM uses aggregated relational data (ARD), which contains the answers to surveys with questions of the form “How many X’s do you know?” The ARD is collected from the general population, rather than from the target population.

The basic premise of the NSUM is that the number of people that respondent i knows in group k, denoted by y_{ik}, follows the scale-up equation given by

$$\frac{y_{ik}}{d_i} = \frac{N_k}{N}, \qquad (1)$$

Accepted Manuscript

where d_i is the degree (or total number of people that respondent i knows), N_k is the size of group k, and N is the total population size (Killworth et al., 1998b). This model assumes that the probability that a member of respondent i's social network belongs to group k is proportional to the prevalence of k in the general population. N_k could be estimated directly given y_{ik}'s, d_i's and N. However, d_i is typically unknown and difficult to estimate directly, requiring the models to estimate both d_i and N_k, either sequentially or simultaneously. In order to first estimate d_i, the ARD also includes questions about "known population" (e.g. people named John or postal workers, where the population sizes N_k are known through census or other means). Note that we often use the terms "group" and "subpopulation" interchangeably. In this manuscript, we primarily use "group" when introducing the model, to recognize that in general not all NSUM applications are related to subpopulations. When referring to our specific application study, we prefer the term "subpopulation."

The most popular basic NSUM model was proposed by Killworth et al. (1998b). It assumes that the data come from a Binomial distribution given by

$$y_{ik} \sim Binom\left(d_i, \frac{N_k}{N}\right). \qquad (2)$$

In order to estimate the unknown degrees and group sizes, the authors propose the following two-stage procedure: Stage 1 estimates the unknown d_i as $\hat{d}_i$ by maximizing the likelihood

$$L(d_i; \boldsymbol{y}) = \prod_{k=1}^{L} \binom{d_i}{y_{ik}} \left(\frac{N_k}{N}\right)^{y_{ik}} \left(1 - \frac{N_k}{N}\right)^{d_i - y_{ik}},$$

with respect to d_i, where L denotes the number of groups with known N_k and the likelihood involves only the known populations. Stage 2 involves maximizing the likelihood which involves only the unknown N_k, denoted by N_u, i.e.

$$L(N_u; \boldsymbol{y}, \boldsymbol{d}) = \prod_{i=1}^{n} \binom{\hat{d}_i}{y_{iu}} \left(\frac{N_u}{N}\right)^{y_{iu}} \left(1 - \frac{N_u}{N}\right)^{\hat{d}_i - y_{iu}},$$

Accepted Manuscript

where n is the number of respondents. This Stage 2 is repeated independently for each unknown group, i.e. there may be any number of N_u. There are other procedures to estimate the unknown N_u, although the general strategy of using responses corresponding to known N_k to estimate d_i and then back-estimating the unknown N_u remains the same.

After being introduced to UNAIDS as a promising method to estimate most-at-risk people for HIV infection, many countries/cities have attempted to implement ARD surveys. One of the largest surveys is the 2009 Ukraine survey, in which the Kiev International Institute of Sociology (KIIS) collected ARD from 10,866 respondents aged 14 and above from December 2008 to February 2009 to estimate the size of 8 HIV-related subpopulations in Ukraine. The ultimate goal of the survey is to improve HIV response efficiency in Ukraine with the help of accurate size estimates. The authors relied on the NSUM to estimate these population sizes since existing methods like multiplier method and capture-recapture were too resource-intensive to obtain accurate estimates for all of Ukraine and would require studies in at least 60 settlements to obtain national size estimates (Paniotto et al., 2009).

Early frequentist models provided a solid foundation for quickly and easily estimating degrees and group sizes from ARD surveys (Killworth et al. (1998a), Killworth et al. (1998b)). Recent Bayesian models have improved size estimates and answered important scientific questions about social networks. Zheng et al. (2006) included additional overdispersion in the model through a negative binomial overdispersion parameter, both better capturing the variability in the data than the existing methods and providing an estimate of the variation in respondents' propensities to know someone in each group. Later, Maltiel et al. (2015) aimed to model the NSUM biases (barrier effects, transmission error, and recall error) directly through the priors, estimating the strength of each bias within the groups. Most recently, Teo et al. (2019) included respondent covariates both about the respondent (e.g. age, gender) and how the respondent felt about each unknown group (e.g. what level of respect do you feel towards female sex workers) to adjust size estimates and study how these covariates influenced the number of people the respondents knew in each group. However, their model ignored the extra variability in the data and it resulted in small uncertainty intervals, similar to the Killworth et al. (1998b)

Accepted Manuscript

estimates. We refer readers to Laga et al. (2021) for a more complete review of the existing NSUM models and ARD properties.

Until now, all models assume that the response biases for a single participant is independent across all groups. However, we conjecture that this is not the case. Zheng et al. (2006) observed that the residuals from their model were correlated, and respondents who knew individuals who had suffered from one negative experience (e.g. suicide or rape) were more likely to know individuals who suffered from other negative experiences. We aim to properly model the correlation structure to further improve NSUM estimates and answer the sociological question of how different groups are related.

In the Ukraine survey, information about the respondents' demographics and their acquaintance to multiple known and unknown subpopulations is collected. To better utilize all auxiliary information and learn more about the connections among subpopulations, we propose a new ARD model that accounts for overdispersion, decomposes the biases, and incorporates respondent characteristics, while also capturing the correlations between subpopulations. Our regression framework allows for more flexibility and ease-of-use than the existing approaches and provides quantitative measures of how covariates influence the number of people known in both known and unknown groups. The correlation estimates from our model provide insight into how social networks form and can hint at how different groups overlap in society. In addition, we propose various measures to assess the reliability of the model estimates.

This paper is organized as follows. First, Section 2 describes the Ukraine dataset. We introduce our proposed NSUM models in Section 3, along with a novel group size scaling method. The benefits and limitations of the models are discussed and our modeling choices are explained. We also show how the overall bias term in our model can be deconstructed into the three NSUM biases. We establish empirical properties of our model in Section 4. We fit our proposed model to the Ukraine dataset in Section 5. We discuss practical advice for future collection and analysis of ARD in Section 6. Final remarks and discussion are found in Section 7.

Accepted Manuscript

2 Ukraine Data

Of the Eastern European countries, Ukraine has the second highest rate of new HIV infections in the WHO European Region, motivating the study of key populations (European Centre for Disease Prevention and Control/WHO Regional Office for Europe, 2017; Paniotto et al., 2009). From December 2008 to February 2009, the Kiev International Institute of Sociology interviewed 10,866 respondents aged 14 and above, asking "How many X's do you know?" questions about 22 known groups and 8 unknown groups (Paniotto et al., 2009). We consider 4 of the 8 unknown subpopulations, women providing sexual services for payment over the last 12 months (FSW), men providing sexual services for payment over the last 12 months (MSW), men who have sex with men (MSM), and people injecting drugs over the last 12 months (IDUs), since these subpopulations belong to the World Health Organization's list of main key population groups vulnerable to HIV (World Health Organization and others, 2016). Examples of the known groups include men aged 20-30, women who gave birth to a child in 2007, and men who served sentences in places of imprisonment in 2007. Supplementary Table 1 lists the known and unknown groups and the sizes of the known groups.

In addition to the ARD ($\boldsymbol{Y}$), respondents were also asked demographic questions about their gender, age, education, nationality, profession, and whether they had access to the internet ($\boldsymbol{Z}$: individual characteristics), as well as "what level of respect" the respondent believed there was in Ukraine for each group on a 1-5 Likert scale, where 1 represents very low level of respect ($\boldsymbol{X}$: individual by group properties). After removing respondents with missing responses, the remaining sample has 9,241 respondents, which is 85.05% of the original dataset[1]. Furthermore, based on the accuracy of leave-one-out size estimates for the known groups, we keep only 11 of the 22 known groups, so for our analysis, n = 9, 241 and K = 15.

Accepted Manuscript

There are significant differences between the distributions of responses across subpopulations. Figure 1 shows frequency barplots for three subpopulations, men named Pavlo, people who died in 2007, and injection drug users (IDUs). For men named Pavlo and people who died in 2007, the average responses are relatively similar (2.11 and 2.98, respectively). However, compared to men named Pavlo, respondents tend to know either 0 people who died in 2007, or several people who died in 2007 – there were roughly 1.4 times as many respondents who knew 0 people who died in 2007 than respondents who knew 0 men named Pavlo, despite Pavlo having a smaller average response. In this situation, the distribution of people who died in 2007 known by the respondents is more overdispersed than the distribution of men named Pavlo known by the respondents. For the hard-to-reach populations, the overdispersion is even more significant. Most respondents know 0 IDUs (92.6%), while some respondents report knowing 40, 50, 60, and even 130 IDUs. The distribution of responses indicates that there are likely significant barrier effects for certain populations like IDUs that violate the random mixing assumption.

3 Models

In this section, we introduce our correlated NSUM model, discuss its properties, and describe the estimation procedure. We first introduce the ARD notation used in the remainder of the manuscript. Given an ARD survey with n respondents about K groups, the number of people that respondent i reports knowing in group k is y_{ik}. Thus, the ARD matrix of responses $\mathbf{Y}$ is an $n \times K$ matrix. The demographics (e.g. age, gender, occupation, etc.) are denoted by $\mathbf{Z}$, where z_{ij} is the information about respondent i for variable j, $j \in \{1, \dots, p\}$. In some surveys, respondents are also asked how they feel about members in group k using questions of the form "what is your level of respect towards group k?" The exact phrasing of the question can vary, but the answers are denoted $\mathbf{X}$ with entries x_{ik} for respondent i and group k. The key distinction between $\mathbf{Z}$ and $\mathbf{X}$ is that $\mathbf{Z}$ is a respondent-level feature while $\mathbf{X}$ contains information about the interaction between the respondents and the groups. All columns of $\mathbf{Z}$ and $\mathbf{X}$ are centered to have mean zero.

3.1 The Correlated NSUM Model

Accepted Manuscript

Our correlated NSUM model is written as, for i in $1,\ldots,n$, and k in $1,\ldots,K$,

$$
\begin{aligned}
&y_{ik} \sim Poisson\left(exp\left\{\delta_i + \rho_k + \boldsymbol{\beta} \boldsymbol{z}_i + \alpha_k x_{ik} + b_{ik}\right\}\right), \\
&\delta_i \sim \mathcal{N}(0, \sigma_\delta^2), \rho_k \sim \mathcal{N}(\mu_\rho, \sigma_\rho^2), \\
&\boldsymbol{b}_i \sim \mathcal{N}_K(\boldsymbol{\mu}, \Sigma_{K \times K}),
\end{aligned}
\tag{3}
$$

where $\Sigma = \mathrm{diag}(\boldsymbol{\tau}) \Omega \mathrm{diag}(\boldsymbol{\tau})$. One key feature of the model is that after accounting for the covariate effects, we allow the biases (b_{ik}) to have group-specific variance (τ_k) and within-person and between-group correlations (Ω). Driven by network features such as homophily, the correlated bias indicates if someone knows more or less of a certain group of people, he/she tends to know more or less of a similar group. The estimated correlation reveals which groups have similarities. We can also separate the biases into the different terms (barrier, transmission, and recall) after all parameters have been estimated, and these details are shown in Supplementary Section 1. After scaling, $d_i = exp(\delta_i)$ represents the degree of respondent i, and $p_k = N_k / N = exp(\rho_k)$ represents the prevalence of group k. Depending on the application, the regression coefficients, $\boldsymbol{\beta}$, could potentially be divided into global ones (β^{global}, those that are constant across groups) and group-specific ones (β_k^{group} , those that vary with groups). This modeling choice allows researchers the flexibility of determining whether each covariate affects the responses in the same way for each group.

Here we treat all the group sizes as unknown and estimate them. As discussed in Feehan et al. (2021), the so-called "known" groups need to satisfy several conditions for them to be reliably treated as "known," including the sizes should be accurately known from census or administrative data, correct identification of memberships, representativeness of the known groups altogether, and several size requirements for each of them. In reality, most known groups do not meet those conditions. As a result, as shown in Feehan et al. (2021), those "known population method" lead to various biases by treating those population sizes as known. Therefore, we choose to treat all group sizes as unknown. The known group sizes are used to help scale the estimates as detailed in the next section.

We complete the formulation with the following priors:

Accepted Manuscript

$$\alpha_k \sim \mathcal{N}(0,100), \beta_{kj} \sim \mathcal{N}(0,100),$$
$$\sigma_\delta \sim Cauchy(0,2.5)I(\sigma_\delta > 0), \mu_\rho \sim \mathcal{N}(0,100),$$
$$\sigma_\rho \sim Cauchy(0,2.5)I(\sigma_\rho > 0),$$
$$\Omega^{1/2} \sim LKJCholesky(2), \tau_{N,k} \sim Cauchy(0,2.5)I(\tau_{N,k} > 0),$$
$$\boldsymbol{\mu} = \log\left(1/\sqrt{1+\boldsymbol{\tau}_N^2}\right), \boldsymbol{\tau} = \sqrt{1+\boldsymbol{\tau}_N^2}.$$

Note that $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ are not sampled, and are only transformations of the sampled parameters $\boldsymbol{\tau}_N = (\tau_{N,1}, \ldots, \tau_{N,K})$. This parameterization is such that $E(b_{ik}) = 1$, a property shared by the Gamma prior in the Zheng et al. (2006) overdispersed model. The half-Cauchy priors on σ_δ and $\tau_{N,k}$ are recommended by Gelman (2006) to restrict the parameters away from very large values.

3.2 Computation

We provide the `networkscaleup` R package for readers to implement our proposed model (Laga et al., 2022). Parameters are estimated using Markov Chain Monte Carlo (MCMC) via Stan. It is important to note that as presented above, the MCMC implementations would have trouble producing unbiased results without prohibitively long sampling chains. The hierarchical form of the model for the bias terms suffers from inefficient MCMC sampling. Specifically, when the $\mathbf{b}_i$ are all close to one another, the diagonal elements of $\Sigma = \text{diag}(\boldsymbol{\tau})\Omega\,\text{diag}(\boldsymbol{\tau})$ will shrink towards 0. Then, since the diagonal elements of Σ are small, $\mathbf{b}_i$ can only take very small MCMC steps, keeping both $\mathbf{b}_i$ close to one another and the diagonal elements of Σ close to 0. To break this dependence between $\mathbf{b}_i$ and Σ, the model can be reparameterized through *parameter expansion*, which allows the size of the residuals to move independently of the variance parameters (Van Dyk and Meng, 2001; Liu, 2003; Gelman, 2004; Gelman et al., 2008). Expanding our model, we reparameterize the bias as

$$\mathbf{b}_i = \boldsymbol{\mu} + \text{diag}(\boldsymbol{\tau})\Omega^{1/2}\boldsymbol{\epsilon}_i, \qquad (4)$$

where $\Omega^{1/2}$ represents the lower-triangular Cholesky decomposition of the correlation matrix Ω and $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, I)$. By sampling $\boldsymbol{\epsilon}_i$ instead of the $\mathbf{b}_i$ directly, the values of $\mathbf{b}_i$

Accepted Manuscript

can comfortably jump around even when $\Omega^{1/2}$ is small because of the reparameterized error terms ϵ_i.

3.3 Scaling

Here we introduce an equally important component of our proposed model, namely the scaling procedure. In order to convert the ρ_k estimates to interpretable group size estimates, Zheng et al. (2006) proposed a scaling strategy that relies on groups of rare names, those believed to have the least biased ARD responses. However, this approach is dataset dependent and may lead to significantly biased results. In their modeling of the McCarty et al. (2001) data, there were several male and female names to use for scaling. In the Ukraine data, there is only one group corresponding to a name, males named "Pavlo." A scaling procedure that depends only on this group would bias the results significantly, as shown in Supplementary Figure 1. This is because the population size for men named "Pavlo" is significantly overestimated, so scaling by "Pavlo" leads to underestimating all other groups. On the other hand, in the McCarty et al. (2001) data, the average bias of the rare female names is similar to the average bias of the remaining groups, which is a necessary condition for this scaling method to work.

3.3.1 Correlated Scaling

We propose a new scaling procedure that relies on scaling each group using correlated groups with known sizes. The idea behind this approach is simply that correlated groups have similar biases, so they should be scaled in a similar way. Specifically, we propose a weighted scaling procedure: the higher the correlation between populations A and B, the larger the weight A has on scaling B. Denoting the m^{th} posterior sample for ρ_k and $\Omega_{i,k}$ as ρ_k^m and $\Omega_{i,k}^m$, respectively, and letting n_{known} represent the number of groups with known size, our scaling procedure is outlined in Algorithm 1 below.

Algorithm 1: Correlated Scaling

Result: Scaled ρ'_k estimates

Accepted Manuscript

Set N_{mc} equal to the number of posterior samples;

for *each k in 1:K* **do**

for *each m in 1:*N_{mc} **do**

Set $\boldsymbol{\omega} = (\Omega^{m}_{k,1}, \ldots, \Omega^{m}_{k,n_{known}})$;

Set negative elements of $\boldsymbol{\omega} = 0$;

Set $\omega_k = 0$;

Scale $\boldsymbol{\omega}$ such that the elements sum to n_{known};

$$C^{m} = \log\left(\frac{1}{n_{known}} \sum_{k \in known} \frac{e^{\rho^{m}_{k}} \omega_k}{N_k / N}\right);$$

$$\rho'^{m}_{k} = \rho^{m}_{k} - C^{m};$$

end

end

This scaling approach inherently corrects biases such as transmission effects without requiring additional surveys like the game of contacts (Salganik et al., 2011). Collecting additional data from each hard-to-reach groups is still the most promising approach, however it is often infeasible, especially when estimating the size of several hard-to-reach groups simultaneously (like in the Ukraine dataset). While ambitious, scaling sizes using correlated groups has the potential to correct for large biases that would otherwise be impossible to account for without these additional datasets.

Accepted Manuscript

4 Simulation Study

In this section, we implement a variety of simulation studies to better understand the properties of our model and scaling procedures. In Section 4.1, we study how ignoring the correlation structure affects other model parameters. In Section 4.2, we demonstrate the utility of the correlated scaling procedure by simulating data from two realistic scenarios.

4.1 Missing Correlations

We simulate data from the correlated model in Equation (3) excluding covariates, where $K = 5$, $\sigma_\delta = 0.7, \rho_k = \log(2.5)$ for all k, $\tau = (2, 1.05, 0.7, 1, 1.2)$, and

$$\Omega = \begin{pmatrix} 1 & 0.9 & 0.8 & -0.05 & 0 \\ 0.9 & 1 & 0.75 & 0 & -0.1 \\ 0.8 & 0.75 & 1 & 0 & 0 \\ -0.05 & 0 & 0 & 1 & 0.85 \\ 0 & -0.1 & 0 & 0.85 & 1 \end{pmatrix}$$

This correlation matrix is designed to replicate the situation in which there are two clusters of groups, for example for stigmatized groups and for unstigmatized groups. In existing ARD surveys, the number of respondents varies widely from around 200 to over 10,000, so several different sample sizes are used. We perform the simulations at five different respondent sample sizes, $n =$ 100, 300, 1000, 3000, and 10,000. We fit the datasets using an uncorrelated version and the correlated parameterization.

When fitting models that do not estimate a group correlation matrix, the proposed correlated scaling procedure is not possible, so we scale using all known group sizes after standardizing by group size (i.e. so larger groups do not have more weight than smaller groups). Specifically, we define a constant for each posterior sample m

$$C^m = \log\left(\frac{1}{n_{known}} \sum_{k \in known} \frac{e^{\rho_k^m}}{N_k / N}\right), \qquad (5)$$

Accepted Manuscript

where *known* represents the set of known groups. For each posterior sample, we scale ρ'_k by $\rho'^m_k = \rho^m_k - C^m$. This scaling procedure yields group size estimates that have an average relative error of zero across all groups for each posterior sample.

We study the distribution of the point estimates from 100 simulations and the results are shown in Figure 2. The points represent the mean of the estimates across the 100 simulations while the 95% intervals represent the upper and lower 2.5% and 97.5% quantiles of the estimates. Across all sample sizes, $\hat{\rho}_k$ estimates are biased when an uncorrelated model is assumed, but the data come from a correlated model. The effects are larger when the group variance τ_k is larger.

To further study the importance of accounting for the correlation structure, we also perform simulation studies for two versions of the Zheng et al. (2006) models and two versions of the Maltiel et al. (2015) models. Details are shown in Supplementary Materials Section 5. We find that for NSUM models that sample random effects directly (Zheng et al. (2006) Poisson model and Maltiel et al. (2015) barrier effects model (sampled version)), the size estimates are biased when the ARD is correlated. In some cases, integrating out the random effects can produce unbiased point estimates (the Zheng et al. (2006) negative binomial model), while in other cases the integration does not improve estimates (the Maltiel et al. (2015) barrier effects model (integrated version)). We conjecture that models which have separate parameters to estimate the mean of the data and the overdispersion can produce unbiased estimates when data are correlated, while models that have parameters that influence both the mean and the variance simultaneously may lead to biased size estimates. In general, it is important to model the correlation directly, both for obtaining reliable inference results and for understanding the network structure.

4.2 Correlated Scaling

We demonstrate the utility of the correlated scaling through two simulation studies. In the first, we include systematic transmission error through correlated covariates. Specifically, we simulate ARD with n = 1000 from the same parameter and hyperparameter setup as before with Ω as the identity matrix. Now, we simulate each row of X as independent multivariate normal random variables, with mean

Accepted Manuscript

$\mu = (0, 0, 0, -2, -2)$ and Σ equal to the correlation matrix used in the missing correlation simulation study. Then, we fit a model that does not include X. This setup simulates the situation where an unobserved respondent-group level covariate explains both the group correlation and a systematic bias like transmission error (i.e. the two columns with mean -2 correspond to groups where members reveal their status to only a small percent of their social network).

Second, we simulate data from a full network model (a stochastic block model) and introduce transmission error, again where n = 1000. The simulation design is intentionally complex in order to best resemble a realistic network. For each simulation, first, a network is simulated from a stochastic block model with group proportions $(0.5, 0.5, 0.25, 0.25)$ and connectivity matrix P provided in Equation (6). Then, for each true link, there is a probability that a respondent does not report the link. The probability of this missing link between respondents is given by matrix T, where $T_{i,j}$ denotes the probability that a respondent in group i correctly reports each link they have to a member of group j, and $inv - logit(\infty) = 1$ for convenience. This design replicates the situation where respondents are likely to accurately recall links from certain groups (e.g. men named Pavlo), while they are likely to underestimate the number of people they know from other groups (e.g. female sex workers). Furthermore, members in these groups or adjacent groups will provide more accurate answers (female sex workers will more likely know the status of other female sex workers *and* drug users).

$$P = \begin{pmatrix} 0.2 & 0.2 & 0.05 & 0.05 \\ 0.2 & 0.2 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.5 & 0.3 \\ 0.05 & 0.05 & 0.3 & 0.5 \end{pmatrix} \qquad T = inv - logit \begin{pmatrix} \infty & \infty & -1 & -1 \\ \infty & \infty & -1 & -1 \\ \infty & \infty & 2 & 2 \\ \infty & \infty & 2 & 2 \end{pmatrix} \qquad (6)$$

The results from the missing covariate and the stochastic block model are shown in Figures 3 and 4, respectively. In both cases, we plot boxplots of the relative error of each estimated and scaled ρ_k, scaling by either all groups or by our proposed correlated scaling. In both simulations, the correlated scaling clearly accounts for the transmission error and produces unbiased results. Thus, including weights in the

Accepted Manuscript

scaling procedure can account for unobserved errors when the overall bias is similar for correlated populations.

5 Ukraine Analysis

In this section, we fit our correlated NSUM model to the Ukraine data in order to better understand the behavior of the key populations. We used RStan to fit both models (Stan Development Team, 2019). The code used is available in our `networkscaleup` R package (Laga et al., 2022). We run three parallel chains for 10,000 iterations each, remove 2,000 iterations for burn-in, and thin each chain by keeping every fifth sample. In all cases the $\hat{R}$ measure of convergence is well below 1.05, indicating convergence. Additional MCMC diagnostics are included in the Supplementary Material. We include only the main results in the main text. Additionally, in Supplementary Material Section 6, we adapt surrogate residuals, first proposed by Liu and Zhang (2018), to the Bayesian setting and use these residuals to further evaluate model fits. We observe no significant lack-of-fit in all of our diagnostic checks, showing the reliability of our model estimates reported below.

5.1 Parameter Estimates

We first show the results of the size estimates for the hidden populations and corresponding 95% uncertainty intervals in Table 1. We also include the original estimates (raw and adjusted) from the Ukraine study. The adjusted size estimates are obtained by multiplying the average estimates in each subpopulation by a weight calculated based on the level of the respect answers, and are believed to be closer to the true sizes than the raw estimates. Our model produced correlated size estimates are very similar to the Ukraine adjusted size estimates with much wider uncertainty intervals. This is a desired outcome, since simpler models often lead organizations to put too much confidence in size estimates, while the uncertainty around hard-to-reach population sizes is typically very large. While a larger uncertainty interval does not imply a more trustworthy size estimate, it is important to accurately propagate the uncertainty in the data collection method and the modeling.

Next, we consider the covariate parameters estimates, α and β. Table 2 includes the group-specific regression coefficients corresponding to age, age^2, and level of

Accepted Manuscript

respect. Age is standardized with the mean age of 43.7 and standard deviation of 19.0. The standardized age is then squared and centered to create Age^2. Level of respect is centered for each group. Overall, the parameter estimates are consistent with the expected results. For example, younger people are more likely to know kids. The chance that someone knows a prisoner peaks at about 34 and the chance that someone knows divorced men peaks at about 37. For the hard-to-reach populations, younger respondents are more likely to know people in all unknown groups, which could potentially provide some guidance for future sampling.

Regarding level of respect, all significant parameters are positive, consistent with the belief that respondents with a more positive perception of a subpopulation will tend to know more people from the subpopulation. The two largest significant parameters for known groups are for prisoners and kids, which are perhaps more likely to have significant barrier effects. The parameters for FSW and MSM are the largest across all groups, which is also consistent with our intuitions.

We also report the global level regression coefficients (gender, education, nationality, profession, and access to internet). The parameter estimates corresponding to male was 0.01, 0.18 for Ukraine, 0.17 for employed, 0.12 for access to internet, 0.18 for secondary education, and 0.21 for vocational education, with a baseline of candidate of sciences or doctor of sciences. Only the 95% credible interval for male included 0. While previous studies have found that men have larger network sizes than women, for the populations in this study, there is not a significant difference between the number of people reported between the gender of the respondents. The network literature has shown that employed and educated individuals typically have larger network sizes. While we are not aware of literature that studies how access to internet affects network sizes, our finding is intuitive since access to internet typically means the individual can reach a broader range of contacts, for example through email.

Finally, we look at the estimated correlations, the key feature of our model. The correlation matrix is shown in Figure 5 and is sorted using a hierarchical clustering algorithm. Our model produces many expected correlations, e.g. respondents who know more men aged 15-17 also know more women 15-17; respondents who know

Accepted Manuscript

more people aged 20 to 30 are less likely to know people who died in 2007. The results also highlight some interesting relationships that are less obvious. First, we find that respondents who know more men named Pavlo are less likely to know young men, and even less likely to know young women. Without census information about the birth records of men named Pavlo in Ukraine, we can guess that Pavlo is more common among older men in Ukraine. Second, respondents who know more/fewer men who got divorced in 2007 also know more/fewer women who gave birth in 2007. This correlation reflects that divorce and birth are both family issues that have similar barrier effects.

The key populations of the Ukraine survey are all highly correlated with one another, prisoners, and, to a lesser extent, divorced men.

It is important to note that our correlation estimates correspond to the correlation of the *reported* number of connections. Therefore, some of the correlation between the hard-to-reach populations may be an indication of a respondent's willingness to answer questions truthfully. That is, if respondents are unwilling to truthfully divulge how many FSW they know, then they may also be unwilling to answer honestly about IDU, MSW, and MSM, leading to two groups of people: those who are willing to report knowing members of hard-to-reach populations and those not willing, potentially increasing the observed correlation.

6 Practical Advice

In this section, we offer guidance on how to better collect and analyze ARD.

Matching target groups with objectives: The first suggestion is to align the questions about the hard-to-reach groups with how the size estimates will be used in practice. In the Paniotto et al. (2009) study, the question involving drug users, for example, is " Do you know people that used drugs by injection for the last 12 months? How many of them?" This question is phrased properly, because government organizations can use the size estimate based on this question to efficiently implement services for current or recent people who used drugs by injection. On the other hand, consider phrasing "Do you know people that *ever* used drugs by injection? How many of them?" In this case, the size estimates cannot be used directly to allocate resources.

Accepted Manuscript

Therefore, it is vital to phrase ARD questions to correspond to the public health objective.

Level of respect: Second, we suggest phrasing "level of respect" questions to maximize the correlation between the level of respect and the ARD. In Ukraine, the level of respect question is phrased as, "What level of respect do following groups have in Ukraine...". The phrasing focuses on how people in Ukraine feel about the groups rather than how the respondent feels, leading to a weaker relationship between the level of respect and ARD responses. However, in Teo et al. (2019), their measure of respondent social acceptability rating of the hard-to-reach populations results in a closer connection between the ARD and the level of respect. For the three shared hard-to-reach groups (FSW, MSM, and IDU), the pairwise correlations in Ukraine between the number of people the respondent reports knowing and their level of respect for the three groups are 0.038, 0.031, and 0.020, respectively. On the other hand, the pairwise correlations in Singapore are 0.040, 0.129, and 0.131, respectively. Thus, the level of respect does seem to be a relatively strong predictor for MSM and IDU in the Singapore data. It is important to ensure that questions are phrased in order to detect as much correlation as possible.

We are not aware of any study that shows how the phrasing of the level of respect question affects the correlation between the ARD responses and the level of respect responses. The phrasing used in the Ukraine study may actually be preferable, and the low correlation is simply a property of the population. This may be an interesting direction of future research.

Similarly, it is important to collect the level of respect questions for all groups. While the Teo et al. (2019) level of respect responses are more correlated with the ARD than in the Ukraine dataset, the authors only include information about the hard-to-reach populations. This results in a loss of information. In our analysis, we are able to account for the level of respect for all groups, further improving the results.

Inclusion of similar groups: Finally, we recommend including more known groups which face similar stigma to hard-to-reach populations. Zheng et al. (2006) find that other populations associated with negative experiences (e.g. prisoners, homicide

Accepted Manuscript

victims, rape victims, people who have committed suicide, and people who were in auto accidents) are highly correlated. While including correlated groups improves the size estimates in our correlated model, understanding how connected different groups are is also important. If groups are identified as being highly correlated with hard-to-reach populations, future researchers can better understand the social networks of members of hard-to-reach populations, making future survey sampling easier and more efficient.

7 Discussion

Aggregated relational data (ARD) is an extremely useful tool not only to estimate population sizes, but also to learn about properties of social networks. Many models have been developed to better capture the behavior of the data and account for the different sources of bias. One major limitation of the models is that uncertainty estimates are far too small, so researchers are too confident in their estimates. We improve upon these models by incorporating covariates and addressing the empirical correlation between groups, and advocating the idea of correlated scaling. Another benefit of our model is that we make very few assumptions about the biases in the model, allowing the data to drive the parameter estimates. The proposed various model diagnostics are also useful in other general settings when there is not a ground-truth to compare with. Satisfactory diagnostic results will increase the credibility of the new NSUM estimates and make them more acceptable by decision makers.

From this Ukraine study, our results can be used to inform government HIV prevention policy. Of the four key populations we considered, we estimate that there are nearly four times as many injection drug users as there are FSW, MSW, and MSM. This is consistent with other studies that estimate IDUs and their sexual partners make up 64% of people living with HIV in Ukraine (Des Jarlais et al., 2009). Combined with estimates of HIV prevalence among key populations, our size estimates illustrate the severity of the HIV epidemic in Ukraine.

Our analysis hints at, but does not explicitly model, the increased risk in individuals that belong to more than one key population. Our correlated model has shown that

Accepted Manuscript

respondents who report knowing more people in one hidden population are more likely to know people in the other hidden populations. World Health Organization and others (2011) estimated that the HIV prevalence among female sex workers who inject drugs is around 43%, while only around 8.5% in female sex workers who do not inject drugs. Based on these relationships, it is clear that is it not sufficient to understand only the behavior of these key populations as a whole, but it is also necessary better understand the relationship between populations in order to effectively lower new HIV infections.

We believe that future ARD models should better exploit the relationship between populations, as illustrated by both our estimated correlation matrix and the covariate effects. It is clear that some individuals are closer to the key populations than others, either because of their age, gender, and similar characteristics, or because of their existing social networks. We would be able to better understand the properties and behavior of the key populations if we were able to survey respondents who were more familiar with the populations of interest.

8 Acknowledgements

The work is supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R01AI136664. Computational efforts were performed on the Hyalite High Performance Computing System, operated and supported by University Information Technology Research Cyberinfrastructure at Montana State University.

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.

Des Jarlais, D. C., Arasteh, K., Semaan, S., and Wood, E. (2009). HIV among injecting drug users: Current epidemiology, biologic markers, respondent-driven sampling, and supervised-injection facilities. *Current Opinion in HIV and AIDS*, 4(4):308.

Accepted Manuscript

European Centre for Disease Prevention and Control/WHO Regional Office for Europe (2017). HIV/AIDS surveillance in Europe 2017 – 2016 data. *ECDC*.

Feehan, D. M., Son, V. H., and Abdul-Quader, A. (2021). Survey methods for estimating the size of weak-tie personal networks. Technical report.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534.

Gelman, A., Van Dyk, D. A., Huang, Z., and Boscardin, J. W. (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17(1):95–122.

Killworth, P. D., Johnsen, E. C., McCarty, C., Shelley, G. A., and Bernard, H. R. (1998a). A social network approach to estimating seroprevalence in the united states. *Social Networks*, 20(1):23–50.

Killworth, P. D., McCarty, C., Bernard, H. R., Shelley, G. A., and Johnsen, E. C. (1998b). 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. (2022). *networkscaleup: Network Scale-Up Models for Aggregated Relational Data*. R package version 0.1-1.

Liu, C. (2003). Alternating subspace-spanning resampling to accelerate markov chain monte carlo simulation. *Journal of the American Statistical Association*, 98(461):110–117.

Accepted Manuscript

Liu, D. and Zhang, H. (2018). Residuals and diagnostics for ordinal regression models: A surrogate approach. *Journal of the American Statistical Association*, 113(522):845–854.

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.

Paniotto, V., Petrenko, T., Kupriyanov, O., and Pakhok, O. (2009). Estimating the size of populations with high risk for HIV using the network scale-up method. *Ukraine: Kiev International Institute of Sociology*.

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.

Stan Development Team (2019). RStan: the R interface to Stan. R package version 2.19.2.

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 and WHO (2010). Guidelines on estimating the size of populations most at risk to HIV. *Geneva, Switzerland: World Health Organization*, page 51.

Van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.

World Health Organization and others (2011). Global HIV/AIDS response: Epidemic update and health sector progress towards universal access: Progress report 2011.

Accepted Manuscript

World Health Organization and others (2016). Consolidated guidelines on HIV prevention, diagnosis, treatment and care for key populations.

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.

Footnotes

1 We recognize the importance of understanding and accounting for the significant amount of missing responses in our data. We examined several missing-data diagnostics and presented key findings in the Supplementary Material Section 4. We found that while there is a relationship between some of the covariates and the frequency of missing responses, this relationship is fairly weak and is subset to only a few of the subpopulations, most notably the subpopulations related to gender and age but not any of the unknown subpopulations. In general, we do not believe that removing the respondents with missing data significantly affects our inference. It may be of interest to explore sophisticated methods to handle missing data in ARD models.

Accepted Manuscript

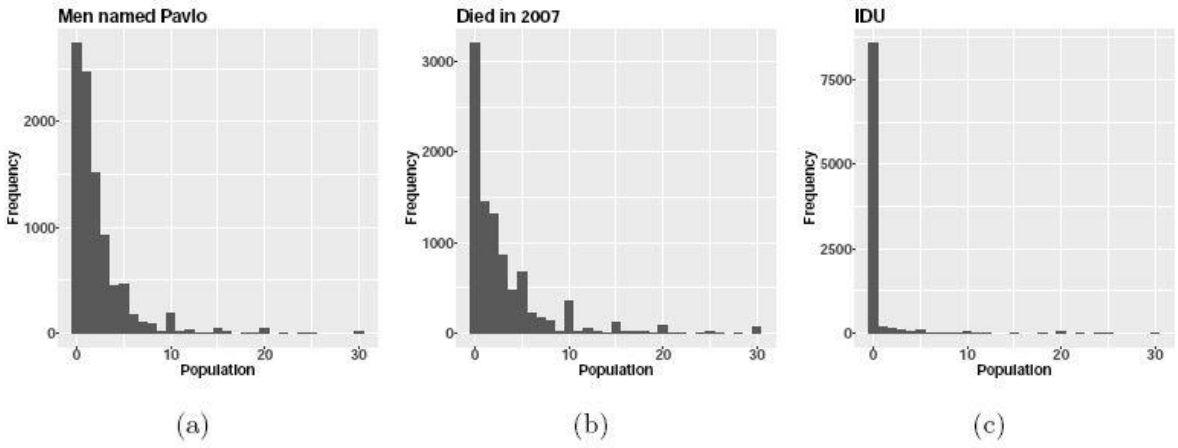

Fig. 1 Barplots of the observed Ukraine ARD for men named Pavlo (a), people who died in 2007 (b), and injection drug users (c). For visualization purposes only, the barplot for 30 represents responses greater than and equal to 30.

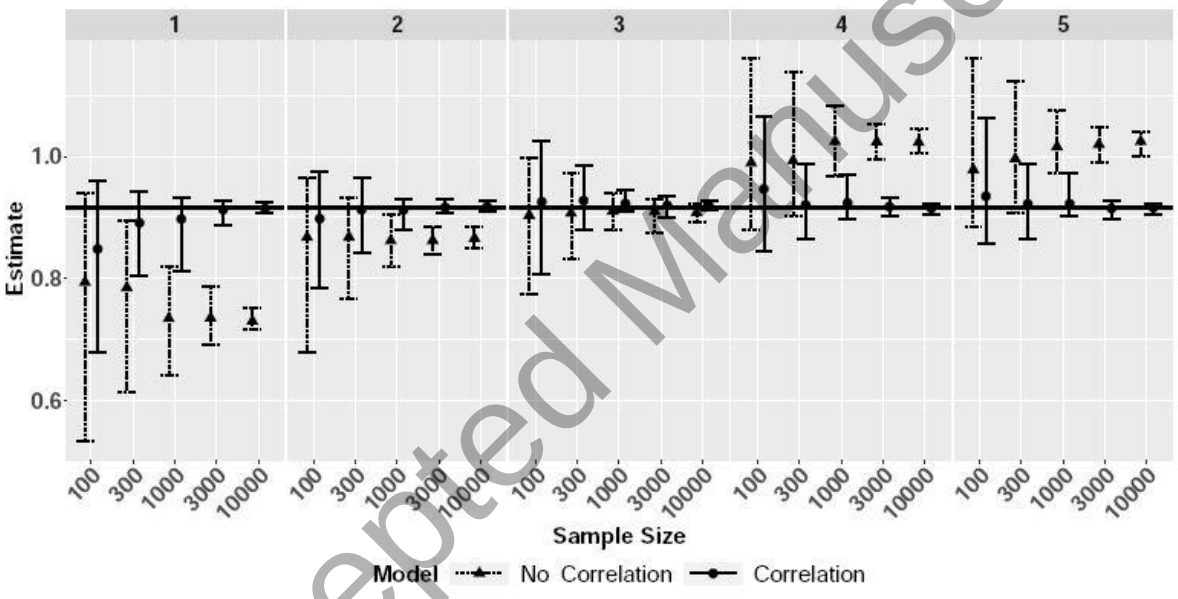

Fig. 2 95% interval of posterior means of ρ across 100 simulations for the missing correlation simulations. The true size is represented by the horizontal black line.

Accepted Manuscript

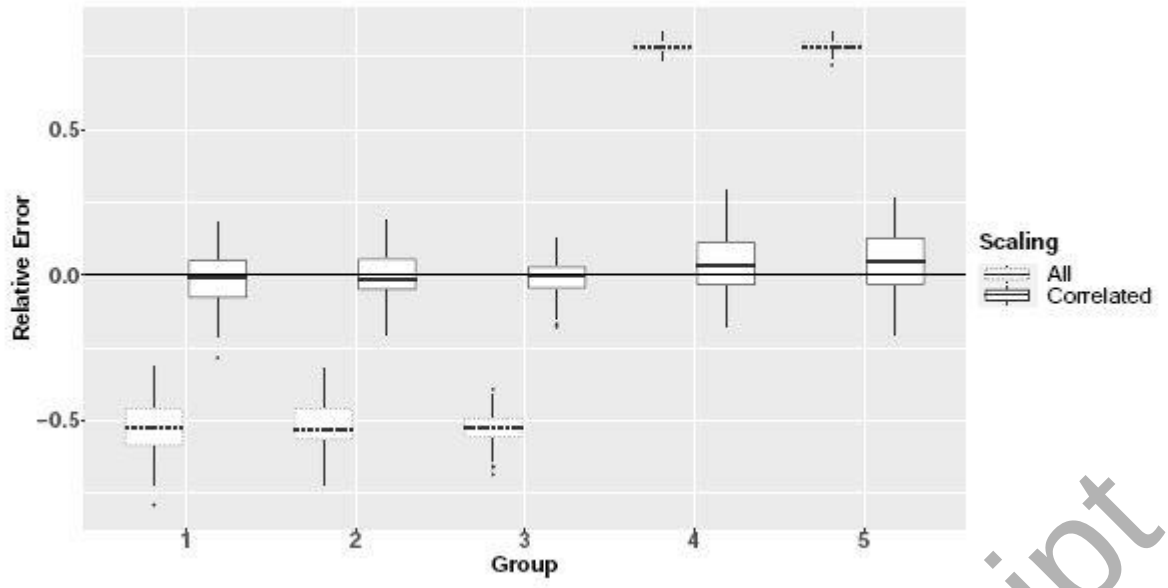

Fig. 3 Boxplot of relative errors of scaled ρ estimates across 100 simulations for missing covariate with transmission error simulations.

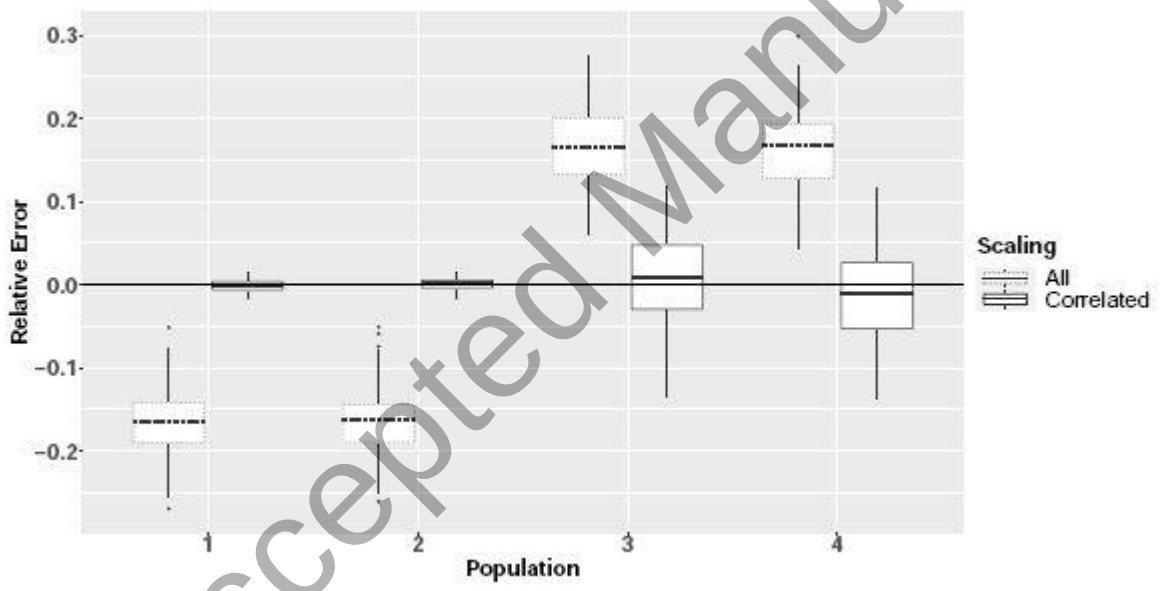

Fig. 4 Boxplot of relative errors of scaled ρ estimates across 100 simulations for SBM with transmission error simulations.

Accepted Manuscript

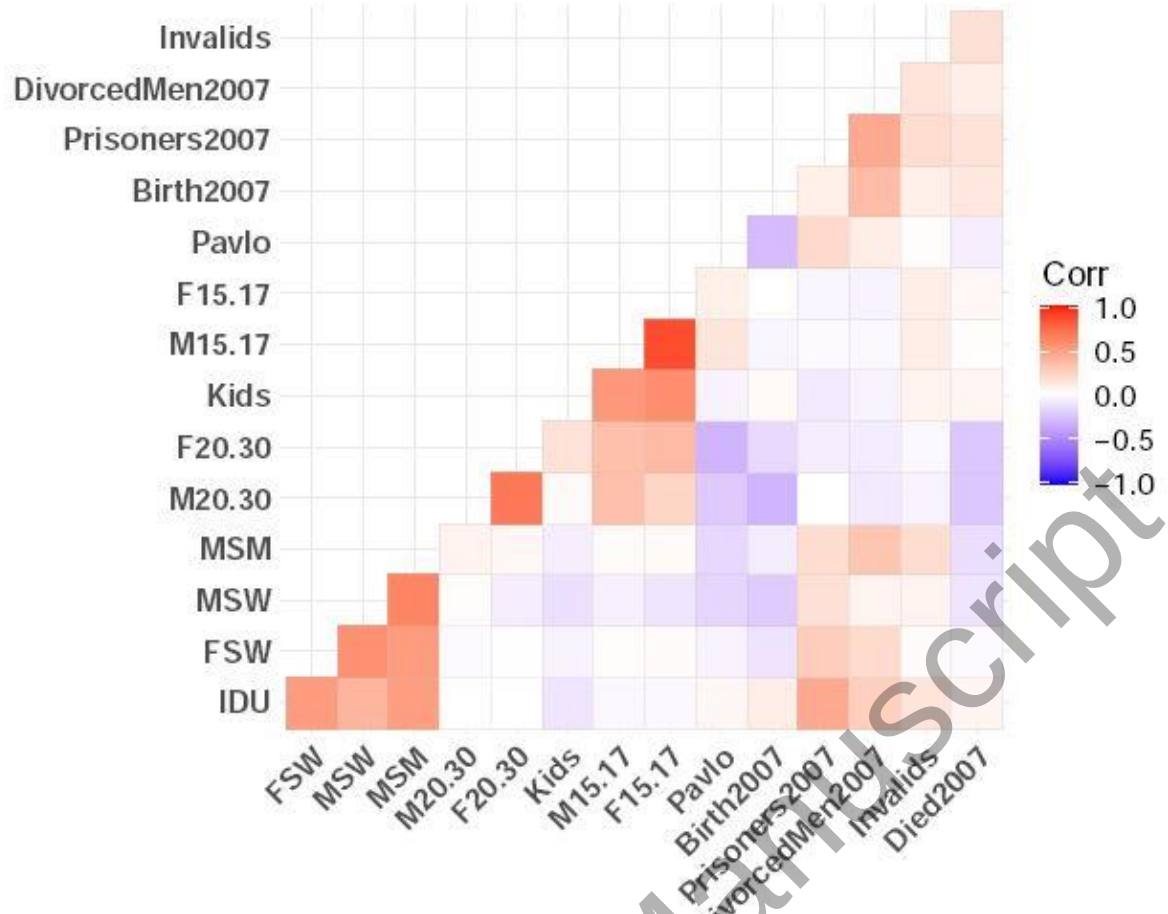

Fig. 5 Estimated correlation matrix for the Ukraine data, arranged by a hierarchical clustering algorithm.

Accepted Manuscript

Table 1 Estimated unknown subpopulation sizes and 95% credible intervals. Values are rounded to the nearest 100. “Correlated” indicates the results from our correlated model with the correlated scaling, “ Paniotto et al. (2009) Raw“ indicates the estimates reported in Paniotto et al. (2009) using the NSUM MLE, and “ Paniotto et al. (2009) Adjusted” indicates estimates reported using their method which adjusts for the level of respect.

Subpopulation	Correlated	Paniotto et al. (2009) Raw	Paniotto et al. (2009) Adjusted
FSW	85,200 (48,100 - 150,000)	34,000 (27,000 - 39,000)	81,000 (65,000 - 93,000)
MSW	6,190 (1,950 - 20,700)	2,400 (1,800 - 3,400)	3,700 (2,800 - 5,200)
MSM	12,300 (5,160 - 28,800)	7,200 (5,300 - 9,100)	14,000 (10,000 - 17,000)
IDU	401,000 (242,000 - 643,000)	103,000 (85,000 - 112,000)	358,000 (285,000 - 389,000)

Table 2 Table of selected group parameter estimates for the correlated NSUM model. Age is standardized with mean 43.7 and standard deviation 19.0. Significance at $\alpha = 0.05$ is denoted by *. The level of respect question was not asked for “people who died in 2007.”

Subpopulation	Age	Age^2	Level of Respect
Men 20-30	-0.38*	-0.17*	0.04*
Female 20-30	-0.37*	-0.14*	0.07*
Kids	-0.31*	0.00	0.12*
Prisoners 2007	-0.24*	-0.24*	0.21*
Divorced Men 2007	-0.22*	-0.32*	-0.01
Birth 2007	-0.15*	-0.12*	0.16*
FSW	-0.69*	-0.10	0.22*
MSW	-0.72*	0.20	0.02
MSM	-1.05*	-0.20	0.58*

Accepted Manuscript

IDU	-0.57*	-0.31*	0.04

Accepted Manuscript