Supplementary Materials for "Hidden population size estimation from respondent-driven sampling: a network approach"

Forrest W. Crawford[1], Jiacheng Wu[1], and Robert Heimer[2]

1. Department of Biostatistics

2. Department of Epidemiology of Microbial Diseases

Yale School of Public Health

1 Posterior distribution of N

We find the posterior predictive distribution of N by marginalizing over subgraphs $\widehat{G}_S \in \mathcal{C}(G_R, \mathbf{d})$, N, and p,

$$\begin{aligned}\Pr(N|\mathbf{Y}) &= \sum_{\widehat{G}_S \in \mathcal{C}(G_R,\mathbf{d})} \int_0^1 \int_0^\infty \Pr(N,p|\widehat{G}_S,\mathbf{Y}) \Pr(\widehat{G}_S,\lambda|\mathbf{Y}) \ \mathrm{d}\lambda \ \mathrm{d}p \\ &= \frac{\pi(N)}{\kappa(\mathbf{Y})} \sum_{\widehat{G}_S \in \mathcal{C}(G_R,\mathbf{d})} \frac{\pi(\widehat{G}_S)}{\kappa(\widehat{G}_S,\mathbf{Y})} \int_0^\infty L(\widehat{G}_S,\lambda;\mathbf{Y})\pi(\lambda) \ \mathrm{d}\lambda \\ &\qquad \times \int_0^1 L(N,p;\widehat{G}_S,\mathbf{Y})\pi(p) \ \mathrm{d}p. \end{aligned} \tag{1}$$

The integral over λ is

$$\begin{aligned}\int_0^\infty L(\widehat{G}_S,\lambda;\mathbf{Y}) \ \pi(\lambda) \ \mathrm{d}\lambda &= \int_0^\infty \left(\prod_{j \notin M} \mathbf{s}_j \right) \lambda^{n-m} \exp[-\lambda \mathbf{s}'\mathbf{w}] \ \frac{\xi^\eta \lambda^{\eta-1} e^{-\xi\lambda}}{\Gamma(\eta)} \ \mathrm{d}\lambda \\ &= \frac{\xi^\eta \Gamma(n-m+\eta) \prod_{j \notin M} \mathbf{s}_j}{\Gamma(\eta)(\mathbf{s}'\mathbf{w}+\xi)^{n-m+\eta}} \end{aligned} \tag{2}$$

and the integral over p is

$$\begin{aligned}\int_0^1 L(N,p|\widehat{G}_S,\mathbf{Y})\ \pi(p)\ \mathrm{d}p &= \left[\prod_{i=1}^n \binom{N-i}{d_i^u}\right]\int_0^1 p^{D^u}(1-p)^{nN-\binom{n+1}{2}-D^u}\ \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\ \mathrm{d}p \\ &= \frac{\left[\prod_{i=1}^n \binom{N-i}{d_i^u}\right]}{\mathrm{B}(\alpha,\beta)}\int_0^1 p^{D^u+\alpha-1}(1-p)^{nN-\binom{n+1}{2}-D^u+\beta-1}\ \mathrm{d}p \\ &= \left[\prod_{i=1}^n \binom{N-i}{d_i^u}\right]\frac{\mathrm{B}\left(D^u+\alpha, nN-\binom{n+1}{2}-D^u+\beta\right)}{\mathrm{B}(\alpha,\beta)}\end{aligned} \tag{3}$$

where the d_i^u's are computed from $\widehat{G}_S$ and $\mathbf{d}$, $D^u=\sum_{i=1}^n d_i^u$, and $\mathrm{B}(\cdot,\cdot)$ is the Beta function. The marginal posterior distribution of N is therefore

$$\begin{aligned}\Pr(N|\mathbf{Y}) = \frac{\pi(N)}{\kappa(\mathbf{Y})}\sum_{\widehat{G}_S}\frac{\pi(\widehat{G}_S)}{\kappa(\widehat{G}_S,\mathbf{Y})}\frac{\xi^\eta\Gamma(n-m+\eta)\prod_{j\notin M}\mathbf{s}_j}{\Gamma(\eta)(\mathbf{s}'\mathbf{w}+\xi)^{n-m+\eta}} \\ \times\left[\prod_{i=1}^n \binom{N-i}{d_i^u}\right]\frac{\mathrm{B}\left(D^u+\alpha, nN-\binom{n+1}{2}-D^u+\beta\right)}{\mathrm{B}(\alpha,\beta)}.\end{aligned} \tag{4}$$

2 Conditions for existence of moments of $\Pr(N|\mathbf{Y})$

The marginal posterior mass function of N is given by (4). We seek sufficient conditions for the posterior mass function to be proper and to have finite first and second moments when $\pi(N)\propto N^{-c}$. First, note that the sum over $\widehat{G}_S\in\mathcal{C}(G_R,\mathbf{d})$ is finite, so it suffices to consider only the conditional posterior for a particular G_S. Let d_i^u and $D^u=\sum_{i=1}^n d_i^u$ be defined from knowledge of G_S and $\mathbf{d}$. Then the posterior mass of N given G_S is

$$\Pr(N|G_S,\mathbf{Y})\propto\left[\prod_{i=1}^n\frac{(N-i)!}{(N-i-d_i^u)!}\right]\frac{\Gamma(nN-\binom{n+1}{2}-D^u+\beta)}{\Gamma(nN-\binom{n+1}{2}+\alpha+\beta)}N^{-c} \tag{5}$$

where we have used the definition of the Beta function as a ratio of Gamma functions. We first provide a bound for the product term, then the ratio of Gamma functions. Each term in the product obeys the bound

$$\begin{aligned}\frac{(N-i)!}{(N-i-d_i^u)!} &\le \frac{(N-i)^{N-i+1/2}e^{-(N-i)+1}}{\sqrt{2\pi}(N-i-d_i^u)^{N-i-d_i^u+1/2}e^{-(N-i-d_i^u)}} \\ &\le \frac{e^{-d_i^u+1}}{\sqrt{2\pi}}\left(\frac{N}{N-n-d_i^{\max}}\right)^{N-i+1/2}(N-n-d_i^{\max})^{d_i^u}\end{aligned} \tag{6}$$

(via Stirling's approximation) where $d_i^{\max}=\max_i d_i$. Then

$$\prod_{i=1}^n\frac{(N-i)!}{(N-i-d_i^u)!}\le\text{const}\times\left(\frac{N}{N-n-d_i^{\max}}\right)^{nN-\binom{n+1}{2}+n/2}(N-n-d_i^{\max})^{D^u}. \tag{7}$$

where the d_i^u's are computed from $\widehat{G}_S$ and $\mathbf{d}$ and $D^u = \sum_{i=1}^n d_i^u$. Second,

$$\begin{aligned}
\frac{\Gamma(nN - \binom{n+1}{2} - D^u + \beta)}{\Gamma(nN - \binom{n+1}{2} + \alpha + \beta)} &\leq \frac{(nN - \binom{n+1}{2} - D^u + \beta - 1)^{nN - \binom{n+1}{2} - D^u + \beta - 1/2} e^{-(nN - \binom{n+1}{2} - D^u + \beta - 1) + 1}}{\sqrt{2\pi}(nN - \binom{n+1}{2} + \alpha + \beta - 1)^{nN - \binom{n+1}{2} + \alpha + \beta - 1/2} e^{-(nN - \binom{n+1}{2} + \alpha + \beta - 1)}} \\
&= \left(\frac{nN - \binom{n+1}{2} - D^u + \beta - 1}{nN - \binom{n+1}{2} + \alpha + \beta - 1} \right)^{nN - \binom{n+1}{2} + \beta} \\
&\quad \times \frac{(nN - \binom{n+1}{2} - D^u + \beta - 1)^{-D^u}}{(nN - \binom{n+1}{2} + \alpha + \beta - 1)^{\alpha}} \frac{e^{D^u + \alpha + 1}}{\sqrt{2\pi}} \\
&\leq \left(nN - \binom{n+1}{2} + \beta - 1 \right)^{-D^u - \alpha} \frac{e^{D^u + \alpha + 1}}{\sqrt{2\pi}}.
\end{aligned} \tag{8}$$

Combining (7) and (8), we have

$$\begin{aligned}
\Pr(N|\mathbf{Y}) &\leq \text{const} \times \left(\frac{N}{N - n - d_i^{\max}} \right)^{nN - \binom{n+1}{2} + n/2} (N - n - d_i^{\max})^{D^u} \\
&\quad \times \left(nN - \binom{n+1}{2} + \beta - 1 \right)^{-D^u - \alpha} N^{-c} \\
&= \text{const} \times \left(\frac{N}{N - n - d_i^{\max}} \right)^{nN - \binom{n+1}{2} + n/2} \left(\frac{N - n - d_i^{\max}}{nN - \binom{n+1}{2} + \beta - 1} \right)^{D^u} \\
&\quad \times \left(nN - \binom{n+1}{2} + \beta - 1 \right)^{-\alpha} N^{-c}
\end{aligned} \tag{9}$$

The first term converges to one, the second to a constant that does not depend on N, while the last two terms dominate in the right-hand tail, and for large N we have

$$\begin{aligned}
\Pr(N|\mathbf{Y}) &\approx \left(nN - \binom{n+1}{2} + \beta - 1 \right)^{-\alpha} N^{-c} \\
&\propto N^{-(\alpha + c)}.
\end{aligned} \tag{10}$$

It follows that a sufficient condition for the posterior distribution to be proper is $\alpha + c > 1$. The condition $\alpha + c > 2$ ensures that the posterior mean exists, and $\alpha + c > 3$ ensures that the second moment exists, and hence the posterior variance.

3 Monte Carlo algorithm

3.1 Sampling G_S

Crawford (2016) describes a procedure for drawing a proposal subgraph $\widehat{G}_S$ uniformly from the set of compatible subgraphs $\mathcal{C}(G_R, \mathbf{d})$. Let $m = |M|$ be the number of seeds. The

posterior distribution of G_S is

$$\Pr(\widehat{G}_S|\mathbf{Y}) \propto \frac{\prod_{j\notin M} \mathbf{s}_j}{(\mathbf{s}'\mathbf{w}+\xi)^{n-m+\eta}}\pi(\widehat{G}_S). \tag{11}$$

Suppose $G_S = (V_S, E_S)$ is the current estimate of the recruitment-induced subgraph. We propose a new subgraph by adding or removing an edge from this graph. To draw a new sample from $\mathcal{C}(G_R, \mathbf{d})$, we select vertices i and j, with $i \neq j$ at random. Then if $\{i,j\} \notin E_S$, $\mathbf{u}_i > 0$, and $\mathbf{u}_j > 0$, we propose to add the edge $\{i,j\}$ to E_S. If $\{i,j\} \in E_S$ and $\{i,j\} \notin E_R$, we propose to remove the edge $\{i,j\}$ from E_S. Otherwise, we select a different $\{i,j\}$ and try again. The vector of the number of susceptible vertices just before each recruitment is $\mathbf{s} = \text{lowerTri}(\mathbf{AC})'\mathbf{1} + \mathbf{C}'\mathbf{u}$ using the current subgraph estimate G_S and let $\mathbf{s}^+$ and $\mathbf{s}^-$ be the corresponding vectors obtained by adding or removing an edge between i and j. It is not necessary to compute $\mathbf{s}$ via matrix multiplication. Instead, Crawford (2016) provides the update expressions

$$\begin{aligned} \mathbf{s}_k^+ &= \mathbf{s}_k - \mathbb{1}\{k > j\}C_{ik} - C_{jk} \\ \mathbf{s}_k^- &= \mathbf{s}_k + \mathbb{1}\{k > j\}C_{ik} + C_{jk}, \end{aligned} \tag{12}$$

for $k = 1, \ldots, n$. Now let t_i^* be the time at which vertex i used all its coupons or the end of the study, whichever came first. Then the change in total edge-time is given by

$$\begin{aligned} \mathbf{s}^{+'}\mathbf{w} &= \mathbf{s}'\mathbf{w} - (t_i^* - \min(t_j, t_i^*) + t_j^* - t_j) \\ \mathbf{s}^{-'}\mathbf{w} &= \mathbf{s}'\mathbf{w} + (t_i^* - \min(t_j, t_i^*) + t_j^* - t_j). \end{aligned} \tag{13}$$

Using these expressions, the ratio of posterior probabilities for N reduces to a simple form. To illustrate, suppose we wish to add the edge i, j to $G_S = (V_S, E_S)$, where $\{i,j\} \notin E_S$, $\mathbf{u}_i \geq 1$, and $\mathbf{u}_j \geq 1$. For a proposal $G_S^+ = (V_S, E_S^+)$ identical to G_S except that $\{i,j\} \in E_S^+$, $\mathbf{u}_i^+ = \mathbf{u}_i - 1$, and $\mathbf{u}_j^+ = \mathbf{u}_j - 1$, the ratio is

$$\frac{\Pr(G_S^+|\mathbf{Y})}{\Pr(G_S|\mathbf{Y})} = \left(\prod_{j\notin M} \frac{\mathbf{s}_j^+}{\mathbf{s}_j}\right)\left(\frac{\mathbf{s}'\mathbf{w}+\xi}{\mathbf{s}^{+'}\mathbf{w}+\xi}\right)^{n-m+\eta}\frac{\pi(G_S^+)}{\pi(G_S)}. \tag{14}$$

To illustrate the ratio for removing the edge i, j, suppose $G_S = (V_S, E_S)$ has $\{i,j\} \in E_S$ and $\{i,j\} \notin E_R$. For a proposal $G_S^- = (V_S, E_S^-)$ identical to G_S except that $\{i,j\} \notin E_S^-$, $\mathbf{u}_i^- = \mathbf{u}_i + 1$, and $\mathbf{u}_j^- = \mathbf{u}_j + 1$, the ratio is

$$\frac{\Pr(G_S^-|\mathbf{Y})}{\Pr(G_S|\mathbf{Y})} = \left(\prod_{j\notin M} \frac{\mathbf{s}_j^+}{\mathbf{s}_j}\right)\left(\frac{\mathbf{s}'\mathbf{w}+\xi}{\mathbf{s}^{-'}\mathbf{w}+\xi}\right)^{n-m+\eta}\frac{\pi(G_S^-)}{\pi(G_S)}. \tag{15}$$

Suppose G_S^* is the proposal graph and let $\Pr(G_S^*|G_S)$ be the probability of proposing G_S^* from G_S, with N fixed. To decide whether to accept G_S^*, we form the Metropolis-Hastings acceptance probability,

$$\rho = \min\left\{1,\ \frac{\Pr(G_S^*|\mathbf{Y})}{\Pr(G_S|\mathbf{Y})}\frac{\Pr(G_S|G_S^*)}{\Pr(G_S^*|G_S)}\right\}. \tag{16}$$

The form of $\Pr(G_S^*|G_S)$ is given by Crawford (2016).

3.2 Sampling N given G_S

The posterior distribution of N conditional on a given compatible subgraph G_S is

$$\Pr(N|G_S, \mathbf{Y}) \propto \left[\prod_{i=1}^{n} \binom{N-i}{d_i^u}\right] \mathrm{B}\left(D^u + \alpha, nN - \binom{n+1}{2} - D^u + \beta\right) \pi(N) \tag{17}$$

Although this conditional distribution does not have a standard form, we can derive a close approximation using the negative binomial distribution when $\Pr(N|G_S, \mathbf{Y})$ has a mode. Let $d_1^u, \ldots, d_n^u$ be the number of pendant edges emanating from each sampled vertex at the moment they are recruited, calculated from G_S. Suppose for now that N is continuous-valued. We can calculate analytic derivatives of $\ell(N) = \log \Pr(N|G_S, \mathbf{Y})$ as follows:

$$\begin{aligned}
\frac{\partial \ell}{\partial N} &= \left[\sum_{i=1}^{n} \psi(N-i+1) - \psi(N-i-d_i^u+1)\right] \\
&\quad + \left[\psi\left(nN - \binom{n+1}{2} - D^u + \beta\right) - \psi\left(nN - \binom{n+1}{2} + \alpha + \beta\right)\right] n - \frac{c}{N} \\
\frac{\partial^2 \ell}{\partial N^2} &= \left[\sum_{i=1}^{n} \psi^{(1)}(N-i+1) - \psi^{(1)}(N-i-d_i^u+1)\right] \\
&\quad + \left[\psi^{(1)}\left(nN - \binom{n+1}{2} - D^u + \beta\right) - \psi^{(1)}\left(nN - \binom{n+1}{2} + \alpha + \beta\right)\right] n^2 + \frac{c}{N^2}
\end{aligned} \tag{18}$$

where $\psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$ is the digamma function and $\psi^{(1)}(x) = \frac{\partial^2 \log \Gamma(x)}{\partial x^2}$ is the polygamma function. Let $\hat{N} = \mathrm{argmax}_N \ell(N)$ be the mode of $\Pr(N|G_S, \mathbf{Y})$ and let

$$v = \left(-\left.\frac{\partial^2 \ell}{\partial N^2}\right|_{N=\hat{N}}\right)^{-1} \tag{19}$$

be an approximation to the variance. To draw from $\Pr(N|G_S, \mathbf{Y})$ we employ a proposal distribution to generate a candidate N^* and use a Metropolis-Hastings correction to draw from the relevant conditional posterior. We will use $\hat{N}$ and v to construct a proposal distribution for N given G_S. Consider $N^* \sim \mathrm{NegBin}(\hat{N}, r)$, where we have parameterized the negative binomial distribution by its mean and size r. The variance of the proposal distribution under this parameterization is $N + N^2/r$, so to achieve a proposal variance of v, where $v > N$, set $r = N^2/(v-N)$. The proposal distribution is

$$\Pr(N^* = k|\hat{N}) = \left(\frac{\hat{N}}{r+\hat{N}}\right)^k \frac{\Gamma(r+k)}{k!} \Bigg/ \sum_{j=N_{\min}}^{\infty} \left(\frac{\hat{N}}{r+\hat{N}}\right)^j \frac{\Gamma(r+j)}{j!\Gamma(r)}, \tag{20}$$

where we have normalized by the probability that $N^* \geq N_{\min}$. Then the Metropolis-Hastings ratio for the proposal N^* conditional on G_S is

$$\rho = \min\left\{1, \frac{\Pr(N^*|G_S, \mathbf{Y})}{\Pr(N|G_S, \mathbf{Y})} \frac{\Pr(N|\hat{N})}{\Pr(N^*|\hat{N})}\right\}. \tag{21}$$

The infinite sum in the denominator of (20) cancels in the ratio (21).

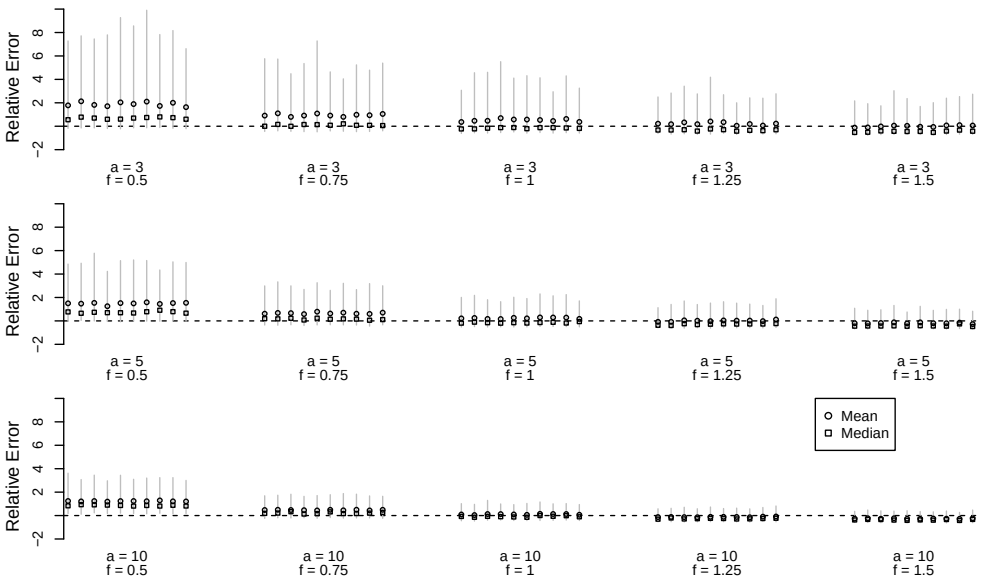

Figure 1: Posterior estimates of N when the prior mean of p is not equal to the true value p_{true}. We set the prior mean to $\mathbb{E}[p] = fp_{\text{true}}$ for $f \in \{0.5, 0.75, 1, 1.25, 1.5\}$ and evaluate estimates for different values of α.

4 Simulation results under mis-specification

4.1 Simulation results under mis-specification of the prior mean

First, we evaluate posterior estimates of $N = 10^5$ when the prior mean for p is not equal to the true value. To evaluate the sensitivity of estimates under mis-specification of the prior mean, we specify the Beta prior for p such that $\mathbb{E}_\pi[p] = fp_{\text{true}}$, where $f > 0$ is the fraction by which the prior mean of p is mis-specified. We investigate estimates of N with $f \in \{0.5, 0.75, 1, 1.25, 1.5\}$, shown in Figure 1, for different values of α. The middle column in the figure corresponds with $f = 1$, giving a match between the prior mean and the true value of p. Overall, specifying $\mathbb{E}[p] < p_{\text{true}}$ results in over-estimation of N, and specifying $\mathbb{E}[p] > p_{\text{true}}$ results in under-estimation of N. In most cases, the 95% posterior quantile intervals for N cover the true value of N in the simulation. The worst results are obtained when the prior mean is highly mis-specified, and large α gives high prior precision. Overall, the requirement that the first two moments of the posterior distribution for N exist necessitates a somewhat informative prior distribution, and gross misspecification of this prior (along with high prior precision) can skew estimates away from the true value of N.

4.2 Simulation results under mis-specification of the population graph model

Real-world social networks exhibit more complex structure than Erdős-Rényi networks, so it is of interest to understand the properties of the proposed estimation framework when the Erdős-Rényi assumption does not hold. Handcock et al (2014) assess estimates under

an graph model in which vertices are of two types, with different connection probabilities within and between types. We consider a simple undirected graph $G = (V, E)$ with $|V| = N$ in which there are two types of vertices $V = (V_0, V_1)$. Edges between type-0 vertices occur independently with constant probability p_{00}, edges between type-1 vertices occur with probability p_{11}, and edges between type-0 and type-1 vertices occur with probability $p_{01} = p_{10}$. This graph model is a 2-block case of the general stochastic blockmodel, in which connection probabilities within groups are homogeneous, and there are possibly different connection probabilities between groups. To evaluate estimates of $N = 10^5$ under mis-specification of the population graph model, we keep the average number of edges in the graph constant, and perturb the connection probabilities. Fix $p \in (0, 1)$ and fix the expected number of edges $\mathbb{E}[|E|] = \binom{N}{2}p$. For notational ease, let $n_0 = |V_0|$ and $n_1 = |V_1|$.

4.2.1 Balance in within-block connection probabilities

Let $h \geq 0$ and let

$$\begin{aligned} p_{00} &= p + h \\ p_{11} &= p + h \\ p_{01} &= p - h\frac{\binom{n_0}{2} + \binom{n_1}{2}}{n_0 n_1} \end{aligned}$$

be the within- and between-group connection probabilities. The expected number of edges is

$$\begin{aligned} \mathbb{E}[|E|] &= \binom{n_0}{2}p_{00} + \binom{n_1}{2}p_{11} + n_0 n_1 p_{01} \\ &= \binom{n_0}{2}(p + h) + \binom{n_1}{2}(p + h) + n_0 n_1 \left(p - h\frac{\binom{n_0}{2} + \binom{n_1}{2}}{n_0 n_1}\right) \\ &= \binom{N}{2}p \end{aligned}$$

Therefore by changing the value of h, we can scale continuously between a graph with Erdős-Rényi distribution, and a 2-block model, while keeping the expected number of edges constant.

However, not every positive value of h is possible. In order for the connection probabilities to reside in the $[0, 1]$ interval, we must have $0 \leq h \leq h_{\max}$, where

$$h_{\max} = \min\left\{1 - p,\ \frac{p n_0 n_1}{\binom{n_0}{2} + \binom{n_1}{2}}\right\}$$

Define $\epsilon \in [0, 1]$ let $h = \epsilon h_{\max}$. Then keeping $N = 10^5$, n_0, n_1, and p constant and varying ϵ from 0 to 1, we have a graph model whose average number of edges is preserved, that scales continuously from an Erdős-Rényi graph to a disconnected 2-block model with equal within-block connection probabilities. Choosing $\epsilon = 0$ yields the Erdős-Rényi graph with

density p. We use this procedure to generate population graphs G for given p and ϵ, and estimate N using the proposed methodology.

Figure 2 shows posterior estimates on the relative error scale. Let $q = n_0/N$ be the fraction of vertices in the smaller block. Estimates appear to exhibit small dependence on the value of ϵ, with some positive bias evident when the model is maximally mis-specified and $\epsilon = 1$. These results indicate relative robustness of inferences to minor deviations to the Erdős-Rényi assumption, in which block structure exists, but the blocks have relatively constant connection probabilities, $p_{00} = p_{11}$ with $\mathbb{E}[|E|] = \binom{N}{2}p$. As we show below, when these within-block connection probabilities differ substantially, estimates can exhibit appreciable bias.

4.2.2 Imbalance in within-block connection probabilities

Suppose $n_0 < n_1$, $g \in \mathbb{R}$ and let

$$\begin{aligned} p_{00} &= p + g \\ p_{11} &= p - g \\ p_{01} &= p + g\frac{\binom{n_1}{2} - \binom{n_0}{2}}{n_0 n_1} \end{aligned}$$

be the within- and between-group connection probabilities. As before, the expected number of edges is $\mathbb{E}[|E|] = \binom{N}{2}p$, but the within-group connection probabilities p_{00} and p_{11} can differ substantially. Constraints on the connection probabilities imply that $g_{\min} \leq g \leq g_{\max}$, where

$$g_{\min} = \max\left\{-p, p\frac{n_0 n_1}{\binom{n_0}{2} - \binom{n_1}{2}}\right\}$$

$$g_{\max} = \min\left\{p, 1-p, (1-p)\frac{n_0 n_1}{\binom{n_1}{2} - \binom{n_0}{2}}\right\}$$

Then letting $\delta \in [0, 1]$, setting $g = g_{\min} + \delta(g_{\max} - g_{\min})$ allows us to scale continuously between the extremes of g. For different values of δ, the model scales continuously from an Erdős-Rényi graph to a 2-block model in which one block of vertices has greater or lesser density than the other block. Choosing $\delta = -g_{\min}/(g_{\max} - g_{\min})$ so that $g = 0$ yields the Erdős-Rényi graph with density p.

This type of mis-specification under extreme imbalance can result in biased estimates. Figure 2 shows posterior estimates on the relative bias scale. Let $q = n_0/N$ be the fraction of vertices in the smaller block. We observe positive bias when δ is close to the extremes $\delta_{\min}$ and $\delta_{\max}$, with the largest bias occurring when q is small. Bias is smallest, as expected, when δ takes an intermediate value, and the underlying network is close to Erdős-Rényi in structure. Estimates are more accurate as α increases, and prior variance decreases.

5 An approximation for prior elicitation

Suppose we wish to find values of α and β that place the prior mean of N approximately equal to $\hat{N}$, a prior estimate of N. Recall that d_i^u follows the Beta-Binomial distribution,

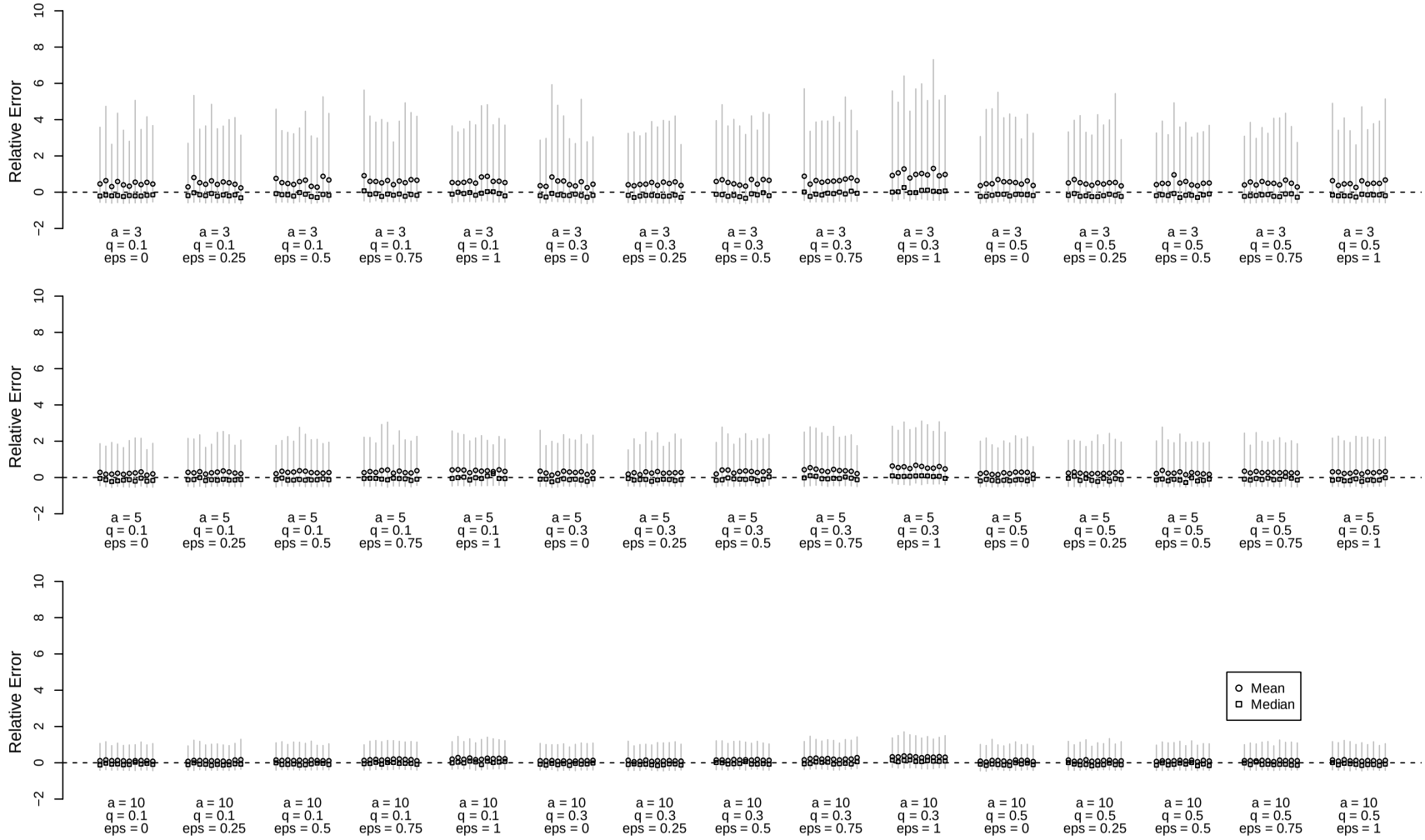

Figure 2: Estimates of $N = 10^5$ under mis-specification of the population graph model as described in Section 4.2.1. Setting ϵ ("eps" above) to zero yields the Erdős-Rényi graph with density p.

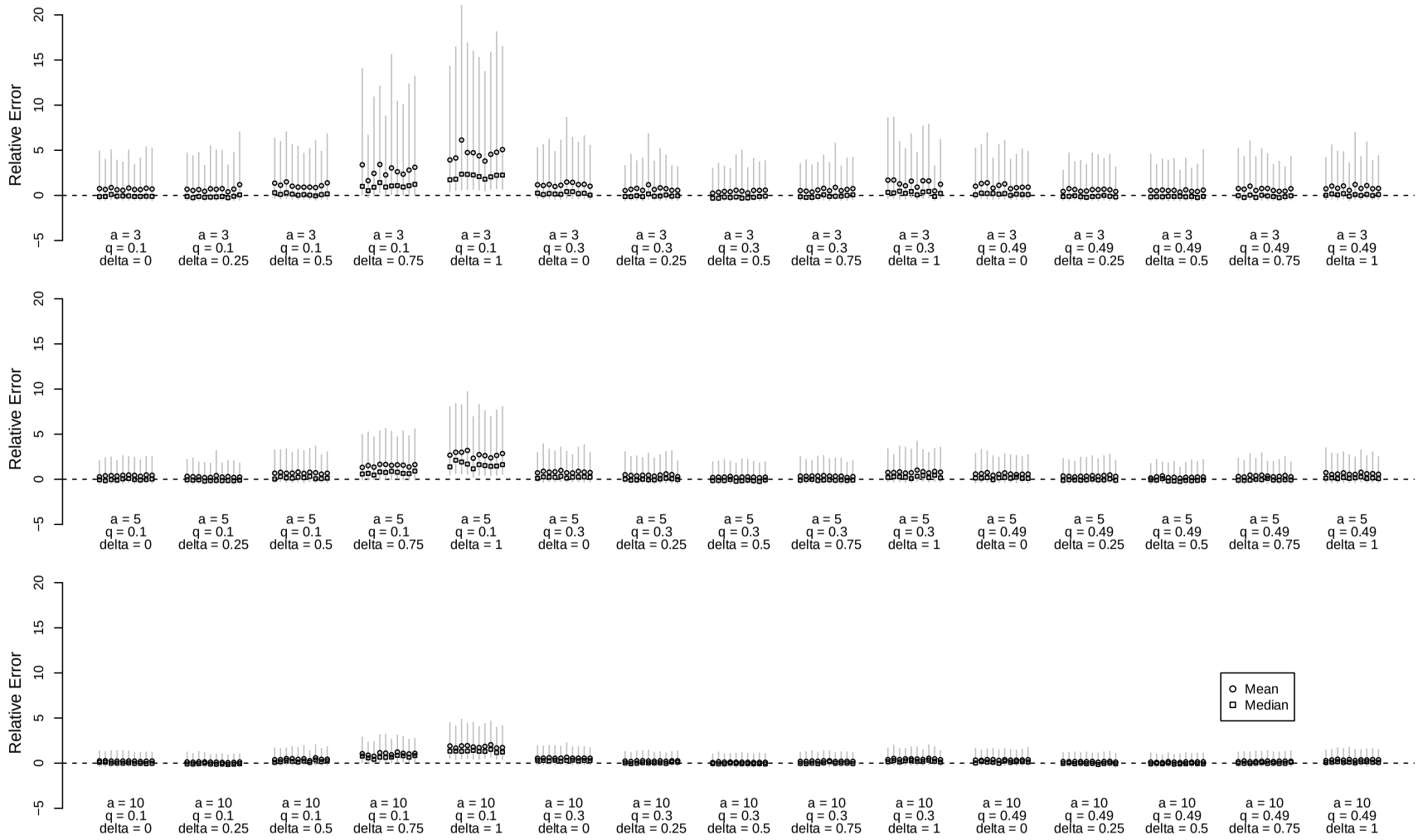

Figure 3: Estimates of $N = 10^5$ under mis-specification of the population graph model, as described in Section 4.2.2. Intermediate values of δ ("delta" above) correspond to graphs whose distribution is close to Erdős-Rényi with density p.

and let $\bar{p} = \alpha/(\alpha+\beta)$. Then $\mathbb{E}[d_i^u] = (N-i)\bar{p}$ and

$$\mathbb{E}\left[\sum_{i=1}^{n} d_i^u\right] = \bar{p}\left(nN - \binom{n+1}{2}\right). \tag{22}$$

Equating observed and expected values of d_i^u and rearranging, we have an estimator for N given $\bar{p}$,

$$\tilde{N} = \frac{n+1}{2} + \frac{1}{\bar{p}n}\sum_{i=1}^{n} d_i^u \tag{23}$$

or an estimator for $\bar{p}$ given N,

$$\tilde{p} = \frac{\sum_{i=1}^{n} d_i^u}{nN - \binom{n+1}{2}}. \tag{24}$$

Now let $N = \hat{N}$ in (24). Since G_S is not directly observed in an RDS study, the d_i^u's are not available. However, we can place a sharp lower bound on the numerator of (24) by conditioning on the observed degrees. Let r_i be the number of subjects recruited by subject i over the course of the study. The number of edges belonging to vertex i connecting to unrecruited vertices at the time of its recruitment cannot be smaller than r_i. But at most $i-1$ edges of i can connect to already-recruited vertices, so $\max\{r_i, d_i - (i-1)\}$ is a lower bound for d_i^u. Recall that M is the set of seeds. Then we have the lower bound

$$\max\{r_i, d_i - i + 1\} \leq d_i^u \tag{25}$$

This leads us to a lower bound for $\bar{p}$ that depends only on $\hat{N}$ and information contained in $\mathbf{d}$ and G_R:

$$\frac{\sum_{i=1}^{n} \max\{r_i, d_i - i + 1\}}{n\hat{N} - \binom{n+1}{2}} \leq \tilde{p} \tag{26}$$

Let p_{lo} denote this lower bound. One strategy for prior elicitation is to restrict the prior distribution of p so that $\Pr(p < p_{\text{lo}})$ is small. We therefore fix α and find β so that $\Pr(p > p_{\text{lo}}|\alpha, \beta) = 0.99$.

6 Results of SS-size method on the St. Petersburg dataset

Table 1 shows the estimated number of PWID in St. Petersburg using the SS-size method implemented in the "sspse" package (Handcock and Gile, 2015; Handcock et al, 2014, 2015). Table 2 shows the results of regression analyses to determine whether the reported degrees in the St. Petersburg data decrease over time as the sample accrues. We estimated the change in expected degree as a function of recruitment order, with and without an outlier who reported a degree of 200. Figure 4 shows the reported degrees.

Prior Parameters			Estimates			Implied Prevalence	
n/N	Max N	Size	Mean	2.5%	97.5%	20-45yrs	All
Beta($\gamma = 1$)	200000	raw	2735	2209	3206	0.18%	0.06%
Beta($\gamma = 5$)	200000	raw	2714	2209	3206	0.18%	0.06%
Beta($\gamma = 10$)	200000	raw	2731	2209	3405	0.18%	0.06%
Beta($\gamma = 1$)	500000	raw	2056	1812	2312	0.14%	0.04%
Beta($\gamma = 5$)	500000	raw	2059	1812	2312	0.14%	0.04%
Beta($\gamma = 10$)	500000	raw	2066	1812	2312	0.14%	0.04%
Beta($\gamma = 1$)	200000	imputed	43996	12976	99110	2.93%	0.96%
Beta($\gamma = 5$)	200000	imputed	38456	12577	78574	2.56%	0.84%
Beta($\gamma = 10$)	200000	imputed	38192	10982	81265	2.55%	0.83%
Beta($\gamma = 1$)	500000	imputed	25223	6809	73274	1.68%	0.55%
Beta($\gamma = 5$)	500000	imputed	23018	6809	48783	1.53%	0.50%
Beta($\gamma = 10$)	500000	imputed	26611	6809	63774	1.77%	0.58%
Flat($\gamma = 1$)	200000	raw	1436	1212	1611	0.10%	0.03%
Flat($\gamma = 5$)	200000	raw	1430	1212	1611	0.10%	0.03%
Flat($\gamma = 10$)	200000	raw	1427	1212	1611	0.10%	0.03%
Flat($\gamma = 1$)	500000	raw	1352	1313	1812	0.09%	0.03%
Flat($\gamma = 5$)	500000	raw	1351	1313	1812	0.09%	0.03%
Flat($\gamma = 10$)	500000	raw	1355	1313	1812	0.09%	0.03%
Flat($\gamma = 1$)	200000	imputed	30896	3804	92730	2.06%	0.67%
Flat($\gamma = 5$)	200000	imputed	31690	3206	91334	2.11%	0.69%
Flat($\gamma = 10$)	200000	imputed	19562	3006	48267	1.30%	0.43%
Flat($\gamma = 1$)	500000	imputed	25639	3311	74767	1.71%	0.56%
Flat($\gamma = 5$)	500000	imputed	29026	2812	101250	1.94%	0.63%
Flat($\gamma = 10$)	500000	imputed	34214	4311	103249	2.28%	0.74%

Table 1: Estimates from the "sspse" software of the number of people who inject drugs in St. Petersburg, Russia. We obtained posterior estimates under the flat (uniform) prior and Beta prior for the sample proportion n/N. The Conway-Maxwell-Poisson (CMP) distribution is the prior for the population degree distribution $f(d|\eta)$. We obtain results under two values for the maximum possible N: 200,000 and 500,000. We set the prior mean of N to 83118 and the prior standard deviation to $\gamma \times 5799$ where $\gamma \geq 1$, based on the estimate by (Heimer and White, 2010). By increasing γ to 5, 10, and 20, we obtain priors for N with greater variance. We set the mean, standard deviation, and maximum of the degree distribution equal to their sample counterparts.

	All degrees			Excluding $d = 200$		
Method	Slope	SE	p-value	Slope	SE	p-value
Linear	9.24×10^{-4}	1.27×10^{-3}	0.47	8.91×10^{-4}	7.92×10^{-4}	0.26
Poisson	9.00×10^{-5}	4.67×10^{-5}	0.54	8.88×10^{-5}	4.73×10^{-5}	0.06
M (Huber)	1.23×10^{-3}	6.68×10^{-4}		1.23×10^{-3}	6.69×10^{-4}	
M (Bisquare)	1.26×10^{-3}	6.73×10^{-4}		1.26×10^{-3}	6.74×10^{-4}	

Table 2: Regression results for the slope of the time-ordered sample of degrees in the St. Petersburg data. The SS method of Handcock et al (2014) and Handcock et al (2015) assumes that the average degree of recruited subjects decreases as the sample accrues. We fit linear, Poisson, and M estimates with Huber and bisquare weighting for the full set of degrees, and with one outlier ($d = 200$) removed. Estimated slope for the regression line is always positive, indicating that degrees appear to increase in this sample.

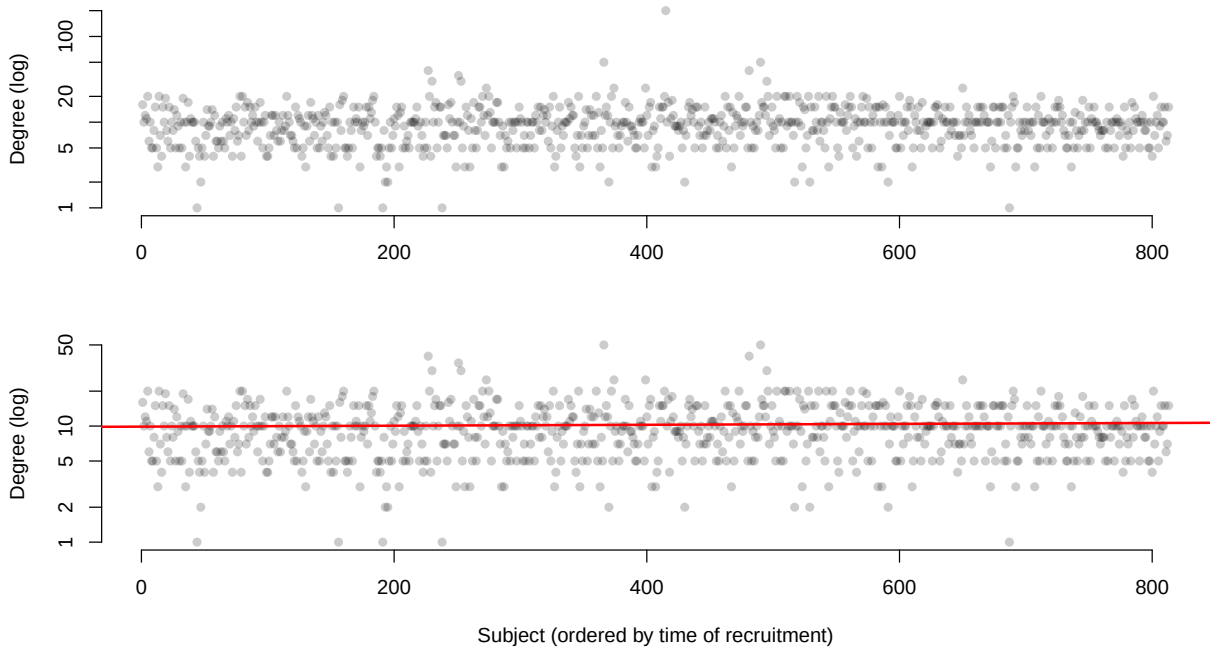

Figure 4: Degrees of recruited subjects in the St. Petersburg study of PWID. The mean reported degree is 10.26, with SD 8.5. One subject reported degree 200. The linear regression line, with slightly positive slope, is overlaid.

References

Crawford FW (2016) The graphical structure of respondent-driven sampling. Sociological Methodology 46:187–211

Handcock MS, Gile KJ (2015) sspse: Estimating Hidden Population Size using Respondent Driven Sampling Data. Los Angeles, CA, URL `http://CRAN.R-project.org/package=sspse`, r package version 0.5-1

Handcock MS, Gile KJ, Mar CM (2014) Estimating hidden population size using respondent-driven sampling data. Electronic Journal of Statistics 8(1):1491–1521

Handcock MS, Gile KJ, Mar CM (2015) Estimating the size of populations at high risk for hiv using respondent-driven sampling data. Biometrics 71(1):258–266

Heimer R, White E (2010) Estimation of the number of injection drug users in St. Petersburg, Russia. Drug and Alcohol Dependence 109(1):79–83