Hidden Population Estimation with Auxiliary Information

Anonymous Author
Anonymous Institution

Abstract

Many populations defined by illegal or stigmatized behavior are difficult to sample using conventional survey methodology. Respondent Driven Sampling (RDS) is a participant referral process frequently employed in this context to collect information. This sampling methodology can be modeled as a stochastic process that explores the graph of a social network, generating a partially observed subgraph between study participants. The methods currently used to impute the missing edges in this subgraph exhibit biased downstream estimation. We leverage auxiliary participant information and concepts from indirect inference to ameliorate these issues and improve estimation of the hidden population size. These advances result in smaller bias and higher precision in the estimation of the study participant arrival rate, the sample subgraph, and the population size. Lastly, we use our method to estimate the number of People Who Inject Drugs (PWID) in the Kohtla-Jarve region of Estonia.

1 Introduction

Valid statistical inference tasks require understanding the data sampling mechanism (Heckathorn, 1997). Often this means identifying a sampling frame, e.g., an enumeration of units in the population of interest, and sampling from it with a known rule. However, many populations lack a conventional sampling frame because they are characterized by behaviors that are illegal (Frost et al., 2006; Johnston et al., 2010) or stigmatized (Hladik et al., 2012; Kerr et al., 2018). These "hidden" populations include intravenous drug users (Crawford, 2016), undocumented immigrants (Johnston et al., 2010), and other vulnerable groups. Respondent Driven Sampling (RDS) is a participant referral process frequently employed by researchers when a sampling frame is unavailable because it preserves the privacy and safety of at-risk populations (Heckathorn, 1997).

RDS begins with a small convenience sample of individuals, who are interviewed and asked to recruit other members of the target population with a limited number of incentivized coupons provided by the researchers. When individuals redeem their coupons, they receive an incentive, are enrolled in the study, and are asked to recruit as well. Both access and trust are achieved by incentivizing members of the hidden population to recruit along social connections, thereby verifying the safety of participation. Additionally, anonymity is preserved since only the researchers and a participant's recruiter know an individual's membership status.[1]

Although this sampling mechanism provides access to the hidden population of interest while accommodating privacy concerns, it creates unique inferential challenges (Heckathorn, 1997). The current literature has mainly focused on estimating prevalence of health-related characteristics in the hidden population, e.g., HIV (Montealegre et al., 2013) and syphilis (Frost et al., 2006). In order to conduct inference under this unique sampling design, researchers create simple approximate models for RDS recruitment, often treating the implicit social network as a nuisance parameter (Gile, 2011; Volz and Heckathorn, 2008). In recent years, focus has shifted to uncovering more about this underlying graph (Crawford et al., 2018a; Verdery et al., 2017) for its use in downstream estimation.

There are various approaches to estimating the overall size of a hidden population that do not account for the sampling mechanism, such as RDS, and hence may perform poorly. Simple capture-recapture meth-

Preliminary work. Under review by AISTATS 2024. Do not distribute.

[1] Various additional layers of protection are possible, such as "coupons" being digital and recruiter information remaining anonymous to recruits.

ods require random sampling and so ignore the mechanism altogether (White, 1982), and multiplier methods (Fearon et al., 2017) depend on every survey participant accurately reporting the hidden population membership status of their acquaintances, which is unrealistic in many sensitive contexts. Successive Sampling has been used to estimate population size from RDS samples (Johnston et al., 2010), however this method does not incorporate all the network information available. The key problem with these approaches is that they effectively ignore the underlying graph structure in the population. To address this, Crawford et al. (2018b) propose to first estimate the unobserved edges in a subgraph of the population in order to develop a model for the hidden population size. Estimating missing graph information requires working with a model over a complex combinatorial space, and we will illustrate that the proposed maximum likelihood and Bayesian estimators are necessarily biased or sensitive to the specification of the prior (Crawford et al., 2018b).

We make two improvements to existing estimators. First, we apply indirect inference, a strategy that helps debias canonical estimators via simulation, and then we incorporate auxiliary information collected during RDS into the estimation process. Section 2 introduces the structure of the RDS stochastic process model and its likelihood. In Section 3, we describe our indirect inference estimator (IIE) (Jiang and Turnbull, 2004) and show that this estimator is less biased than the MLE asymptotically. Section 4 reviews the population size model conditional on the complete subgraph and proposes a method to incorporate additional information into the population size estimation procedure. Section 5 and 6 demonstrate, through simulation studies and a case study respectively, the impact of indirect inference estimation and auxiliary information on population size estimation.

2 RDS Model Setup and Issues

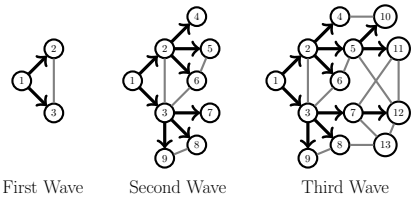


Figure 1: G_R is composed of coupon exchanges $\rightarrow$. G_S includes both observed $\rightarrow$ and unobserved connections —.

Throughout we consider a setting where our population is represented by a graph $G = (V, E)$, where V is the set of $|V| = N$ nodes in the graph and E is the set of all pairwise connections, or edges, between individuals. Respondent Driven Sampling (RDS) starts with a set of seeds (node 1 in Figure 1), and then proceeds by recruiting other participants (the middle and right panels of Figure 1) over the edges of the original graph G. This process continues until a stopping rule is reached (e.g., a predetermined number of recruited individuals or a budget constraint are met). At the end of this process, a researcher is in possession of a recruitment subgraph $G^R = (V^R, E^R) \subset G$ on $|V^R| = n \leq N$ individuals. The labels of the nodes in V^R denote the order in which they arrived at the study (and so participant i was interviewed before participant j if $i < j$). Importantly, this is *not* the vertex induced subgraph of G that would have been observed by projecting the original graph G onto the vertices V^R. We will call this induced subgraph $G^S = (V^S, E^S)$ and note that, while $V^S = V^R$, we only have that $E^R \subseteq E^S$. *If we had access to G^S then estimating the size of the graph G would be a simple task.*

There are two reasons that edges in G^S are missing in G^R. First, recruiters may run out of coupons before they recruit all of their neighbors (e.g., pariticipant 13 in Figure 1). Second, if participant i recruits participant k before participant j does, a connection $\{j, k\} \in E^S$ will not be observed because an individual cannot participate in the study multiple times (e.g., participant 6 is recruited by participant 2 before participant 3 can recruit them in Figure 1).

We describe the data collected in RDS studies that carry information about G^S.[2] Typical RDS studies ask participants how many hidden population members they know. For participant i, this is their degree in the larger graph G, $d_i = |\{\{i, j\} \in E : i \in V^R, j \in V, i \neq j\}|$. The vector of observed degrees, $\mathbf{d} = (d_1, d_2, \ldots, d_n)$, is ordered by arrival to the study. Additionally, we define a vector $\mathbf{w}$ such that w_i is the time between the arrival of participant $i-1$ and participant i. This makes the full data observed at the end of an RDS study $\mathbf{Y} = (G^R, \mathbf{d}, \mathbf{w})$, $\mathbf{Y} \in \mathcal{Y}$.

Our RDS arrival process model is described by wait times attached to edges in G between recruiters with unused coupons and unrecruited members of the hidden population, termed "susceptible edges" (Crawford et al., 2018b). When the wait time associated with edge $\{i, j\}$ expires, participant i recruits participant j (as long as j has not been previously recruited); d_j and w_j are then recorded and $\{i, j\}$ is added to G^R. We assume that edge times are independent and identically distributed according to an exponential distribution (Crawford, 2016).

[2]The notation of Crawford et al. (2018b) is used when possible for referential convenience.

Assumption 1 (Exponential Wait Times) *Upon entering the study, a participant immediately becomes active, and their susceptible edges are assigned a wait time that is drawn independently from an exponential distribution with common parameter $\lambda \in \mathbb{R}^+$. (This combines assumptions 4 and 6 in Crawford et al. (2018b).)*

Let $A^S \in \{0,1\}^{n\times n}$ be the adjacency matrix associated with graph G^S, where $A^S_{i,j} = 1$ if $\{i,j\} \in E^S$ and 0 if not; let $u_i \in \mathbf{u} = (u_1, u_2, \ldots, u_n)$ be the number of connections study participant i has to unrecruited hidden population members, $u_i = |\{\{i,j\} \in E : j \notin V^R\}|$; and let M be the seed set. Additionally, let $\mathrm{lt} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ be the lower-triangular function, i.e., for any $A \in \mathbb{R}^{n\times n}$, we have $\{\mathrm{lt}(A)\}_{i,j} = A_{i,j}1(i \le j)$. The joint likelihood for parameters A^S and λ is

$$\mathcal{L}_n(\mathbf{Y}|A^S, \lambda) = \left(\prod_{j\notin M} \lambda s_j\right) \exp\left(-\lambda \mathbf{s}^\top \mathbf{w}\right), \qquad (1)$$

where $\mathbf{s} = \mathrm{lt}(A^S C)^\top 1 + C^\top \mathbf{u}$, and $C \in \mathbb{R}^{n\times n}$ is the coupon matrix defined by $C_{ij} = 1$ if participant i has at least one coupon before the j^{th} participant is recruited, and zero otherwise (Definition 4 from Crawford et al. (2018b)). We note that A^S only enters the likelihood through the susceptible edge vector, $\mathbf{s}$. Let $\mathcal{A} = \{0,1\}^{n\times n}$, then the maximum likelihood estimator (MLE) corresponding to Equation (1) is

$$\left\{\widehat{A}^S_n, \widehat{\lambda}_n\right\} = \arg \max_{A^S\in\mathcal{A}, \lambda\in\mathbb{R}^+} \mathcal{L}_n(\mathbf{Y}|A^S, \lambda). \qquad (2)$$

Both G^R and $\mathbf{d}$ function as graphical constraints ensuring that the estimated adjacency matrix, $\widehat{A}^S_n$, is *compatible* with the observed data.

Definition 1 (Compatibility) *An estimated subgraph $\widehat{G}^S = (\widehat{V}^S, \widehat{E}^S)$ represented by the estimated adjacency matrix $\widehat{A}^S_n$ is compatible with the observed data, $\mathbf{Y}$, if the following three conditions hold: 1. $V^R = \widehat{V}^S$; 2. $E^R \subseteq \widehat{E}^S$; 3. The degree of each $i \in \widehat{V}^S$ does not exceed d_i. (This is Definition 5 from Crawford et al. (2018b).)*

2.1 Issues with Maximum Likelihood Estimation

Beyond computational difficulties associated with maximizing functions over graph space, the MLE in Equation (2) can exhibit severe bias even for moderately large sample sizes. We start by noting that *if* A^S were known, Equation (1) reduces to the likelihood of exponentially distributed data. It is well known that the MLE for the rate parameter of an exponential, λ, has a bias that diminishes as the sample size, n, increases: $|\mathbb{E}(\widehat{\lambda}) - \lambda| = \lambda/(n-1)$. However, in RDS, A^S is not known, *and* the magnitude of the bias is related to the rate of increase of both n and N (the unobserved population size). Specifically, when A^S is unknown, Equation (1) has $n+1$ unknown parameters that are meant to be estimated based on n observations and the graphical constraints imposed by A^S — while the parameters remain identifiable due to these constraints, it does not mean that high quality estimation is possible. This is especially true for RDS, as the constraints are often loose in this context ($n \ll N$ and so $n/N \nrightarrow 1$).

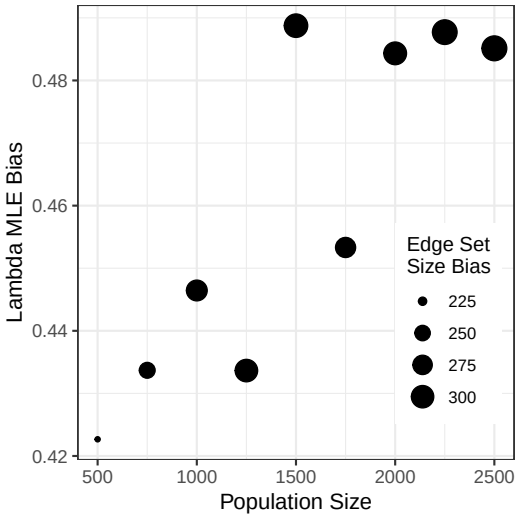


Figure 2: This figure depicts the bias of $\widehat{\lambda}_n$ and $\left|\widehat{A}^S_n\right|$. We can see that the bias of λ and the edge set size are positively correlated and increase as the sample proportion decreases.

In Figure 2, we plot the observed bias in $\widehat{A}^S_n$ (summarized by the bias in the total number of edges: $|\mathbb{E}(|\widehat{A}^S_n|) - |A^S||$) and $\widehat{\lambda}_n$ following an RDS simulated according to the generative model in Equation (1) with $\lambda = 1$, a single seed participant, five coupons per participant, and $n = 100$. The population graph, G, is simulated from an Erdos-Renyi model with edge probability p (details of this model choice are provided in Section 4). On the x-axis, we vary the *total* population size, N. We see that as n/N decreases and the constraints loosen, the bias increases. The intuition behind this is as follows. For a given λ and $i \in \{1, 2, \ldots, n\}$, the MLE of s_i *without* graphical constraints is $1/(\lambda w_i)$, which has expectation $\mathbb{E}\{1/(\lambda w_i)\} = \infty$. This suggests that if $n/N \nrightarrow 1$ as $n \to \infty$ and $N \to \infty$, then the MLE of $s_i\lambda$ will have positive bias. RDS is used in settings where $n << N$ (and so the constraints on $\mathbf{s}$ are minimal), so an alter-

native to the MLE is needed for high quality inference. *We aim to resolve these biases using an alternative estimator motivated by concepts from indirect inference.*

3 Indirect Inference Estimator

We define the indirect inference estimator, derive its theoretical properties, and demonstrate its improvement in estimating RDS model parameters empirically.

3.1 Indirect Inference

The indirect inference estimator (IIE) relies on specifying a calibration statistic. The choice of this statistic is not unique, but often there is a natural option in a given problem domain (Jiang and Turnbull, 2004); we use the MLE for λ as our calibration statistic. The IIE is constructed by finding parameter settings under which the expected value of the calibration statistic matches its observed value.

To formalize the indirect inference estimator (IIE) in our setting, we require a few definitions. Let $\lambda^\dagger : \mathcal{Y} \to \mathbb{R}$ and $A^S_\dagger : \mathcal{Y} \to \{0,1\}^{n\times n}$ be functions that map the data, $\mathbf{Y}$, to the solutions of Equation (2). Additionally, define $A^S_\lambda : \mathcal{Y}\times\mathbb{R}^+ \to \{0,1\}^n$ so that for observed data, $\mathbf{Y}$, and value $\lambda' > 0$, $A^S_\lambda(\mathbf{Y},\lambda')$ is the solution to Equation (2) holding λ fixed at λ'.

We propose the following estimation procedure for our model parameters. Let $\widetilde{\lambda}_n$ solve

$$\mathbb{E}_{\mathbf{Z}\sim P_{A^S_\lambda(\mathbf{Y},\widetilde{\lambda}_n),\widetilde{\lambda}_n}}\left\{\lambda^\dagger(\mathbf{Z})\right\} = \lambda^\dagger(\mathbf{Y}), \tag{3}$$

and $\widetilde{A}^S_n = A^S_\lambda(\mathbf{Y},\widetilde{\lambda}_n)$, then the IIE is the pair $(\widetilde{\lambda}_n, \widetilde{A}^S_n)$. The expectation in Equation (3) is taken over simulated data $\mathbf{Z} = (G^R, \mathbf{d}, \mathbf{w}^*) \in \mathcal{Y}$, where $\mathbf{w}^* \sim P_{A^S,\lambda}$ and $P_{A^S,\lambda}$ is the generative model described in Equation (1). The procedure for calculating the IIE is summarized in Algorithm 1.

To understand *why* an IIE can reduce bias, we first discuss the IIE for exponentially distributed data, which we observed in Section 2.1 are closely related to the data generated by RDS. The important benefit of this setting is that we are able to derive the analytic form of the IIE.

Suppose $\mathbf{X} = (X_1, ..., X_n)$ comprises n independent draws from an exponential distribution indexed by $\lambda \in \mathbb{R}^+$. The likelihood of $\mathbf{X}$ is

$$\mathcal{L}(\lambda|\mathbf{X}) = \prod_{i=1}^n \lambda\exp(-\lambda X_i) = \lambda^n\exp(-\lambda\sum_{i=1}^n X_i).$$

The MLE is $\widehat{\lambda}_n = n/\left(\sum_{i=1}^n X_i\right)$, which is distributed according to an Inverse-Gamma distribution with shape and scale parameters $(n, n\lambda)$. The absolute bias of the MLE is

$$\left|\mathbb{E}\left(\widehat{\lambda}_n - \lambda\right)\right| = \frac{\lambda}{n-1},$$

which is linear in λ. Again choosing the MLE as the calibration statistic in the IIE procedure, we see that the IIE is $\widetilde{\lambda} = (n-1)/(\sum_i X_i)$, which is unbiased.

We also compare the mean squared error (MSE) of the two estimators,

$$\begin{aligned}
&\mathrm{MSE}(\widehat{\lambda}_n) = \left(\frac{\lambda}{n-1}\right)^2 + \frac{n^2\lambda^2}{(n-1)^2(n-2)},\\
&\mathrm{MSE}(\widetilde{\lambda}_n) = \frac{\lambda^2}{(n-2)},\\
&\mathrm{MSE}(\widehat{\lambda}_n) - \mathrm{MSE}(\widetilde{\lambda}_n) =\\
&\qquad\left(\frac{\lambda}{n-1}\right)^2 + \frac{(2n-1)\lambda^2}{(n-1)^2(n-2)} > 0.
\end{aligned}$$

Not only is $\widetilde{\lambda}_n$ unbiased, it is also more accurate.

In general, the IIE is unbiased for the parameter λ if the bias of the MLE is linear in the parameter. The exponential likelihood example above suggests that this is possible in our setting. We formally describe the asymptotic behavior of the bias of the IIE as compared to the MLE in the next subsection.

Algorithm 1: The Indirect Inference Estimator

Goal: Find the estimator,

$$\widetilde{\lambda}_n \in \arg\min_\lambda \left|\mathbb{E}_{\mathbf{Z}\sim P_{A^S_\lambda(\mathbf{Y},\lambda),\lambda}}\left\{\lambda^\dagger(\mathbf{Z})\right\} - \lambda^\dagger(\mathbf{Y})\right|$$

Generate a grid of λ^k values, $k \in \{1, 2, .., K\}$
for *k in* $\{1, 2, \ldots, K\}$ **do**
  for *j in* $\{1, 2, \ldots, J\}$ **do**
    Find $\widehat{A}^S_{n,k} = A^S_\lambda(\mathbf{Y}, \lambda^k)$
    Simulate wait time vector $w^{k,j}$ from the model defined by parameters $\widehat{A}^S_{n,k}, \lambda^k$
    Find $\widehat{\lambda}^{k,j}_n$ by maximizing Equation (2) with generated data $\mathbf{Z}^{k,j} = (G^R, \mathbf{d}, w^{k,j})$
  end
  Save set $\left\{\lambda^k, \widehat{A}^S_{n,k}, \widehat{\lambda}^k_n = \left(\sum_{j=1}^J \widehat{\lambda}^{k,j}_n\right)/J\right\}$
end
Calculate $k^* = \arg\min_{k\in\{1,2,\ldots,K\}} \left|\widehat{\lambda}^k_n - \lambda^\dagger(\mathbf{Y})\right|$
Output estimators $\left\{\widetilde{\lambda}_n, \widetilde{A}^S_n\right\} = \left\{\lambda^{k^*}, A^S_{k^*}\right\}$

3.2 Asymptotics

To characterize the asymptotic behavior of the IIE we assume that the MLE admits an Edgeworth expansion (Hall, 2013).

Assumption 2 *As $n \to \infty$,*

$$\widehat{\lambda}_n = \lambda + \frac{A(V,\lambda)}{\sqrt{n}} + \frac{B(V,\lambda)}{n} + \frac{C(V,\lambda)}{n^{3/2}} + o_p(n^{-3/2}),$$

where V has a distribution that does not depend on λ and $A(V,\lambda)$ $B(V,\lambda)$ and $C(V,\lambda)$ are random vectors that only depend on λ and V.

Such an expansion holds for the MLE under general conditions; see Section 2.4 of Hall (2013) for details. Under this expansion, it can be seen that the bias is of order $n^{-1/2}$.

Proposition 1 *Given Assumption 2, as $n \to \infty$,*

$$\mathbb{E}\left(\widetilde{\lambda}_n\right) = \lambda + \frac{\mathbb{E}\left\{C^*(V,\lambda)\right\}}{n^{3/2}} + o_p(n^{-3/2}),$$

where V is a random variable with a distribution that does not depend on λ and $C^*(V,\lambda)$ is a random vector that only depends on λ and V.

Proposition 1 follows from Assumption 2 and Corollary 2.1 in Gouriéroux et al. (2000). It shows that the IIE does not have bias terms of order $n^{-1/2}$ and n^{-1}, while the MLE does.

3.3 Empirical Performance: Study Participant Arrival Rate and Subgraph Accuracy Improvements

In this section, we empirically evaluate the finite sample behavior of our proposed IIE estimator for the two model parameters in the likelihood of Equation (1). We simulate RDS trajectories of size 100 over various graph sizes, with an average wait time of $\lambda = 1$ and each recruit having 5 coupons. The hidden population graph, G, is simulated from an Erdos-Renyi model with edge probability p (details of this model choice are provided in Section 4). In our simulations, we vary $N \in \{1000, 5000, 10000\}$ and $Np \in \{5, 10, 15\}$. Algorithm 1 is used to construct the IIE, and we compare it to the MLE.

Table 1 demonstrates that the IIE, $\widetilde{A}_n^S$, has a higher true positive rate than the MLE in all simulation settings. Importantly, Table 3 in Appendix A shows that these improvements do not come at the expense of the true negative rate.

The rate parameter λ is of independent interest for assessing coupon uptake speed and the time necessary for recruiting a target sample size. Table 4 in Appendix A indicates that over a range of population sizes and graph densities, the IIE, $\widetilde{\lambda}_n$, outperforms the MLE in terms of MSE.

Remark 1 *Consistent with Figure 2 and the intuition developed in Section 2.1, the advantage of both $\widetilde{\lambda}_n$ and $\widetilde{A}_n^S$ over $\widehat{\lambda}_n$ and $\widehat{A}_n^S$ respectively is slightly greater in high average degree and low sample proportion settings generally.*

Table 1: Graph True Positive Rate (%)

		MLE		IIE	
Pop.	Deg.	Average	Std.	Average	Std.
1000	5	56.66	0.85	67.61	1.47
1000	10	36.52	0.82	50.48	2.00
1000	15	29.49	0.79	47.71	2.38
5000	5	58.76	0.96	69.93	1.58
5000	10	37.00	0.91	51.73	1.92
5000	15	30.57	1.08	49.48	2.48
10000	5	59.25	0.93	72.18	1.52
10000	10	37.52	0.93	54.15	2.08
10000	15	30.50	0.84	51.30	2.22

These are the true positive rates of the estimated subgraphs for a series of population sizes (Pop.) and average degrees (Deg.). The standard deviations reported quantify the Monte Carlo error associated with these estimates based on 100 simulations.

4 Hidden Population Size Estimation

One of the primary goals of sampling hard-to-reach populations is to estimate their total size, N. Imagine that the population graph $G = (V, E)$ is a sample from an Erdos-Renyi graph model with parameters N and p (that is, there are N individuals in the graph and the probability of a connection between any two of them is p). While this is a very simple model, it has demonstrated practical utility when estimating hidden population size, forming the basis for methods such as the snowball sampling estimator (Frank and Snijders, 1994) and the network scale-up estimator (Killworth et al., 1998). Under an Erdos-Renyi model, the degree of each individual in G is distributed as

$$d_i \sim \text{Binomial}(N-1, p).$$

If we had access to a simple random sample of individuals, then we could directly estimate N based on this likelihood.

Unfortunately, RDS does not yield a simple random sample from the population (e.g., an individual's probability of being sampled depends on their degree (Heckathorn, 1997; Gile, 2011)). Conditional on the (unobserved) A^S, it is possible to write down the distribution for the number of edges individual $i \in \{1, 2, \ldots, n\}$ shares with unsampled members of the hidden population at the time of individual i's recruitment. Let $d_i^u = d_i - \sum_{j=1}^{i-1} 1\left(\{i,j\} \in E^S\right)$, and

note that, unlike d_i, this quantity is independently and identically distributed from a Binomial distribution, $d_i^u \sim \text{Binomial}(N-i, p)$.

4.1 Revising Current Approaches

Based on this population size model, Crawford et al. (2018b) propose an approximate Bayesian MCMC sampling scheme with strong priors on p and A^S to conduct inference on N. Unfortunately, they find that informative priors on p are necessary for ensuring finite first and second moments of the posterior distribution for N. For example, the most diffuse prior on p they use in their simulations has a variance of about 5×10^{-6}. Moreover, they require an informative prior on the graph space $\pi(A^S) \propto \exp(-\gamma|E^S|)$, where $\gamma = -\log(p/(1-p))$ ranges from about 5 to 9, imposing heavy penalties on graphs with large edge sets. These priors inflate the posterior mean of N, resulting in significant upward bias.

Prior selection is non-trivial in our problem. Choosing a non-informative prior risks an improper posterior (Kahn, 1987), but, given the nature of the populations we aim to study, it is unlikely that strong informative priors are scientifically justifiable. Moreover, full posterior inference for N is not possible due to computational constraints, requiring multiple approximations (Hunter and Handcock, 2006; Crawford et al., 2018b). We avoid these issues by reformulating the problem as regularized maximum likelihood estimation, which incorporates information on edge prevalence, p, via a regularization term. Given regularization function $R(p) = \log \text{Beta}(p; a, b)$ for $a, b \in \mathbb{R}^+$, we define the regularized MLE estimates of N, p conditional on $\widehat{A}_n^S$ and $\widetilde{A}_n^S$ respectively,

$$\left\{\widehat{p}, \widehat{N}\right\} = \arg\max_{p,N} \ \log \mathcal{L}(N, p|\widehat{A}_n^S) + R(p),$$
$$\left\{\widetilde{p}, \widetilde{N}\right\} = \arg\max_{p,N} \ \log \mathcal{L}(N, p|\widetilde{A}_n^S) + R(p).$$

4.2 Improving Estimation Using Auxiliary Information

The RDS data collection process commonly includes a large survey that can be used to improve population size estimation. In particular, it is common to track how information accumulates over the RDS process, and this measurement necessarily carries information about the underlying network. For example, an RDS interview may begin with a quiz about local free resources, important public health issues, or beneficial health practices (e.g., for People Who Inject Drugs this might include drug therapy options or needle exchange sites). The interview ends with the interviewer revealing the answers to the quiz so that each study participant leaves the study with the same amount of information. The performance of a study participant on this quiz is the graph dependent outcome, $\mathbf{Q}$. Below we propose a model for $\mathbf{Q}$ that, when combined with the IIE approach of Section 3, provides substantial improvements over the population size estimates of the previous section.

Remark 2 *Other graph-dependent outcomes are certainly possible: measurements may depend on participant interactions with their friends or require participants to quantify some characteristic of their referral chain. These different types of* $\mathbf{Q}$ *would simply require different models from the ones we study below, but would otherwise be easily incorporated into the analysis.*

Define monotonically increasing functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, $\mathbf{1}_n = (1, 1, \ldots, 1) \in \mathbb{R}^n$, and an n dimensional distribution F_n. If we assume that there is communication over the network, then the performance of an interviewee on the quiz should be proportional to their connections to previously recruited study participants,

$$\mathbf{Q} = f\left\{\alpha + \gamma g\left(\mathbf{m}\right)\right\} + \epsilon_n, \ \epsilon_n \sim F_n, \tag{4}$$

where $\mathbf{m} = \left\{A^S \cdot \text{lt}(\mathbf{1}_n \mathbf{1}_n^\top)\right\} \mathbf{1}_n$. In Equation (4), α represents an average hidden population member's knowledge of the quiz subject without outside intervention and γ is the intensity of communication flow. Adding information about the outcome $\mathbf{Q}$ to our analysis will improve our estimation of $\mathbf{m}$, which will improve estimators of A^S and N.

We now augment our IIE procedure with the auxiliary information contained in $\mathbf{Q}$. We expand $\mathbf{Y}$ to include the regression information, $\mathbf{Y}^r = (\mathbf{Q}, G^R, d, \mathbf{w}) \in \mathbf{Y}^r$. Define $\lambda_r^\dagger : \mathcal{Y}^r \to \mathbb{R}$ as the function that maps the data, $\mathbf{Y}^r$, to the MLE for λ, and define $A_\lambda^{S,r} : \mathcal{Y}^r \times \mathbb{R} \to \{0,1\}^{n \times n}$ so that for value $\lambda' > 0$, $A_\lambda^{S,r}(\mathbf{Y}, \lambda')$ is the MLE estimator of A^S holding λ fixed at λ'. Let $\widetilde{\lambda}_n^r$ solve

$$\mathbb{E}_{\mathbf{Z}^r \sim P_{A_\lambda^{S,r}(\mathbf{Y}^r, \widetilde{\lambda}_n^r), \widetilde{\lambda}_n^r}} \left\{\lambda^\dagger(\mathbf{Z}^r)\right\} = \lambda^\dagger(\mathbf{Y}^r), \tag{5}$$

and $\widetilde{A}_n^{S,r} = A_\lambda^{S,r}(\mathbf{Y}^r, \widetilde{\lambda}_n^r)$, then the IIE is now the pair $(\widetilde{\lambda}_n^r, \widetilde{A}_n^{S,r})$. The expectation in Equation (5) is over simulated data $\mathbf{Z}^r = (\mathbf{Q}, G^R, \mathbf{d}, \mathbf{w}^*) \in \mathcal{Y}$, where $\mathbf{w}^* \sim P_{A^S,\lambda}$ and $P_{A^S,\lambda}$ is the generative model described in Equation (1). Algorithm 2 in Appendix D builds on Algorithm 1 and provides the complete description for this computation. The regularized MLE estimator of population size conditional on the IIE with auxiliary information is

$$\left\{\widetilde{p}^r, \widetilde{N}^r\right\} = \arg\max_{p,N} \ \log \mathcal{L}(N, p|\widetilde{A}_n^{S,r}) + R(p).$$

5 Population Size Estimation Simulations

In this section, we empirically evaluate the IIEs of hidden population size with and without auxiliary information. The first simulation study compares our estimators to state-of-the-art competitors on a variety of population sizes and graph densities. The second simulation showcases the robustness of our estimators to the misspecification of the graph model in Section 4.

Simulation 1 For each simulation, we draw a hidden population graph from an Erdos-Renyi model, $G \sim \text{ER}(N, p)$, varying $N \in \{1000, 5000, 10000\}$ and $Np \in \{5, 10, 15\}$. We then simulate an RDS study of size $n = 100$ over this graph, starting from 3 random seeds. The RDS follows the generative model specified in Equation (1) with $\lambda = 1$ and 5 coupons. Letting I_n be the n-dimensional identity matrix and $\mathbf{m} = \left\{A^S \cdot \text{lt}(\mathbf{1}_n \mathbf{1}_n^\top)\right\} \mathbf{1}_n$, we observe a vector of study participant attributes, $\mathbf{Q}$, drawn according to

$$\mathbf{Q} = \alpha + \gamma \mathbf{m} + \epsilon_n, \quad \epsilon_n \sim N(0, I_n \sigma^2),$$

which is within the class of models outlined in Equation (4). We set $\alpha = 0$, $\gamma = 1$ and $\sigma^2 = 1$ and experiment with regularization information on p to explore the utility of social network edge density information when estimating population size. This procedure is repeated 200 times for each simulation setting.

We compare our estimators to the MLE derived in Crawford et al. (2018b) as well as to several estimators proposed in Handcock et al. (2014) that use the successive sampling (SS) method. Under a uniform prior, the SS estimators, which are posterior summaries, require the researcher to specify the maximum that the population size can attain, $N_{\max}$. For a given N, we use values $N_{\max} \in \{3N, 5N, 8N\}$.

Figure 3 reports the results across all nine simulation setups. First, we note that the IIEs with and without auxiliary information have lower maximum absolute deviation (MAD) than the MLE over a range of hidden population graph sizes and densities. The weak regularization information setting is defined by $R(p) = \log \text{Beta}(p; a, b)$, where $\text{Beta}(p; a, b)$ is centered at p with $a = 0.1$. Consistent with Remark 1, the improvements of the IIE without auxiliary information over the MLE are greater in high average degree settings. The improvements of the IIE with auxiliary information over the IIE without auxiliary information follow the same pattern. When comparing to the SS approach, we note that the estimators based on this procedure are very sensitive to the prior specification. In fact, the MAD for the SS Mean estimator (the posterior mean) where $N_{\max} = kN$ for $k \in \mathbb{R}^+$ is almost exactly $|(k-2)N - n|/2$, which is the absolute difference between the prior mean and N.

In Appendix C, we explore the role of regularization in our estimator. Figure 4 in Appendix C shows that in the strong regularization setting, where $R(p) = \log \text{Beta}(p; a, b)$ and $a = 10$, the improvements of $\widetilde{N}$ and $\widetilde{N}^r$ over $\widehat{N}$ are higher in larger populations.

Simulation 2: Graph model misspecification We assess the sensitivity of our population estimate results to the Erdos-Renyi model assumption. Following Crawford et al. (2018b) and Gile et al. (2018), we divide the hidden population into two groups, $V_A \subseteq V$ and $V_B = V \setminus V_A$. The probability of an edge between members of the same group is p_{in}, and the probability of a connection between members of different groups is p_{out}. For constant $c \in [0, 1]$, we set $p_{out} = cp_{in}$. We let $p^* = \mathbb{P}(E_{ij} \in E)$, where nodes i and j are drawn uniformly at random from V.

For our simulations, we set $N = 5000$ and $p^* = 0.002$ (implying an average degree of 10). Additionally, we let $R(p) = \log \text{Beta}(p; a, b)$, where $a = 100$ and $\text{Beta}(p; a, b)$ is centered at p_{in}, mimicking an ignorance of the two block structure. This simulation setting tests the sensitivity of our results to the misspecification of the graph model and (very strong) incorrect regularization information. As expected, Table 5 in Appendix B indicates that the MAD of $\widehat{N}$, $\widetilde{N}$, and $\widetilde{N}^r$ is higher when the blocks are evenly split and the difference between p_{in} and p_{out} is large. For example, when $c = 0.3$ and the groups are evenly split, the estimators demonstrate a $150\% - 300\%$ increase in MAD over the estimators in the correctly specified setting, while when $c = 0.9$ and $N_A/N = 0.75$, the increase is only $19\% - 34\%$. Encouragingly, the IIE with and without auxiliary information still perform better than the MLE in this misspecified setting, however the benefits are smaller.

6 Application: How many people inject drugs in the Kohtla-Jarve region of Estonia?

According to the European Drug Report 2023, from 2015-2021 Estonia had the highest per capita prevalence of People Who Inject Drugs (PWID) in Europe. There is also evidence of high HIV (Degenhardt et al., 2017) and drug overdose death (related to the introduction of Fentanyl) rates among PWID in Estonia during this time period (Uusküla et al., 2020). Estimating the number of PWID is imperative for understanding the magnitude of this public health crisis and the necessary scope of potential policy solutions. Specifically, syringe exchange programs were launched

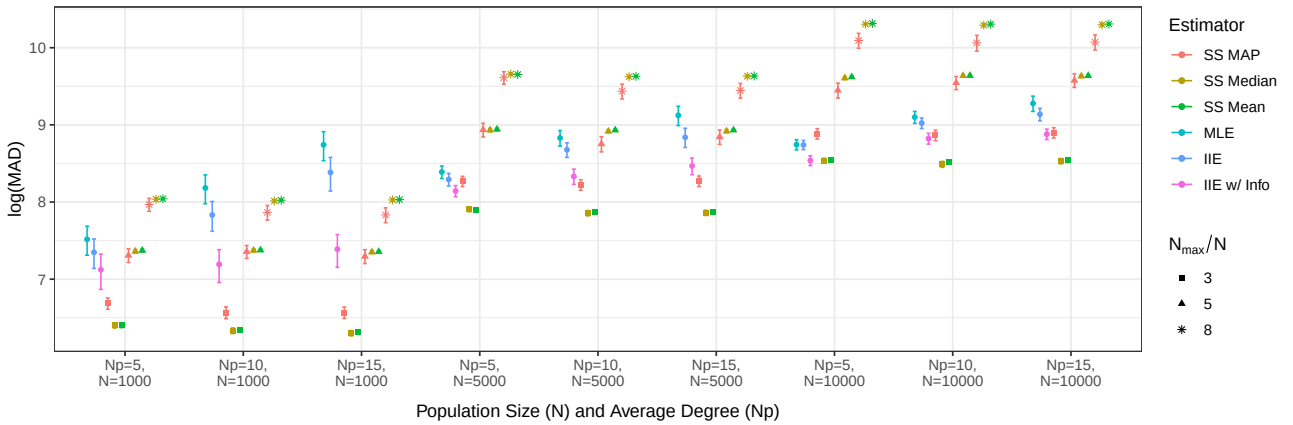


Figure 3: This figure compare the performance of $\widehat{N}, \widetilde{N}$, and $\widetilde{N}^r$ under weak regularization information over a series of population sizes, N, and average degrees, Np, with 90% Monte Carlo confidence intervals.

in 1997 to lower the prevalence of HIV among PWID in Estonia. Without estimates of the PWID population size, it is difficult to confirm that this program's current resources are sufficient (Wu et al., 2017).

Wu et al. (2017) use data from an RDS sample conducted in 2012 to estimate the number of PWID in the Kohtla-Jarve region of Estonia. They compare a series of models including the standard multiplier method (Fearon et al., 2017), successive sampling (Johnston et al., 2010) and the network-based approach (Crawford et al., 2018b). This RDS sample began with 6 seeds and includes 600 participants from the Kohtla-Jarve region. The data on each member of the study includes their arrival time, degree, recruiter identity, and allotted coupons. We use the IIE approach of Section 4, estimating the population size to be $\widetilde{N} = 795$. This is contained within the intervals implied by previous estimates (Wu et al., 2017).

These data further include an indicator of whether the participant is using antiretroviral therapy (ART) for HIV. We use this covariate and the RDS sample to construct a data-realistic simulation study to showcase how a hypothetical network-based covariate could assist in estimating population size. A simple change to the study would have asked each person to share their ART status with their social connections in the PWID population (to hopefully increase screening for HIV and uptake of ART). The auxiliary information to be collected from each RDS participant is then a measurement of how many people have shared their ART status with them since the beginning of the study. Letting $\mathbf{x}_{ART} \in \{0,1\}^n$ be the indicator of ART status, the responses to this question, $\mathbf{Q} = (q_1, q_2, \ldots, q_n)$, could follow a Poisson model similar to the one described in Section 5, $\mathbf{Q} = \text{Poisson}\left[\left\{A^S \cdot \text{lt}(\mathbf{1}_n\mathbf{1}_n^\top)\right\}\mathbf{x}_{ART}\right]$. For this simulation we choose a subgraph A^S that is compatible with the RDS observed and set $\lambda = \widetilde{\lambda}_n = 0.23$ (estimated from the original data without auxiliary information). Based on these two values, $N = 1105$ in this simulation. We incorporate the auxiliary information in $\mathbf{Q}$ to improve our estimation of N as outlined in Section 4.2. Table 2 compares $\widetilde{N}^r$ and $\widehat{N}$, and we see that when such auxiliary information is available, leveraging it improves (by approximately 20%) population size estimation.

Table 2: Population Estimation MAD

Algorithm	MAD	Std.
MLE	219.1	9.3
IIE w/ Info	181.3	6.8

This table displays the MAD from the population size conditional on simulation parameters in Section 6.

7 Conclusion

RDS provides access to populations often excluded from scientific discourse. Although this sampling process presents a variety of inferential problems, it also contains valuable information on the social network connecting study participants. This paper expands on the existing literature with new mechanisms for improving estimation of the study participant arrival rate, complete subgraph, and population size. The first accounts for the the bias of the MLE using concepts from indirect inference, and the second proposes a mechanism for including auxiliary information. Both methods combine to achieve cutting edge performance.

References

Crawford, F. W. (2016). The graphical structure of respondent-driven sampling. *Sociological methodology*, 46(1):187–211.

Crawford, F. W., Aronow, P. M., Zeng, L., and Li, J. (2018a). Identification of homophily and preferential recruitment in respondent-driven sampling. *American journal of epidemiology*, 187(1):153–160.

Crawford, F. W., Wu, J., and Heimer, R. (2018b). Hidden population size estimation from respondent-driven ampling: a network approach. *Journal of the American Statistical Association*, 113(522):755–766.

Degenhardt, L., Peacock, A., Colledge, S., Leung, J., Grebely, J., Vickerman, P., Stone, J., Cunningham, E. B., Trickey, A., Dumchev, K., et al. (2017). Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of hiv, hbv, and hcv in people who inject drugs: a multistage systematic review. *The Lancet Global Health*, 5(12):e1192–e1207.

Fearon, E., Chabata, S. T., Thompson, J. A., Cowan, F. M., and Hargreaves, J. R. (2017). Sample size calculations for population size estimation studies using multiplier methods with respondent-driven sampling surveys. *JMIR public health and surveillance*, 3(3):e7909.

Frank, O. and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, 10:53–53.

Frost, S. D., Brouwer, K. C., Firestone Cruz, M. A., Ramos, R., Ramos, M. E., Lozada, R. M., Magis-Rodriguez, C., and Strathdee, S. A. (2006). Respondent-driven sampling of injection drug users in two us–mexico border cities: recruitment dynamics and impact on estimates of hiv and syphilis prevalence. *Journal of Urban Health*, 83(1):83–97.

Gile, K. J. (2011). Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *Journal of the American Statistical Association*, 106(493):135–146.

Gile, K. J., Beaudry, I. S., Handcock, M. S., and Ott, M. Q. (2018). Methods for inference from respondent-driven sampling data. *Annual Review of Statistics and Its Application*, 5:65–93.

Gouriéroux, C., Renault, E., and Touzi, N. (2000). 13 calibration by simulation for small sample bias correction. *Simulation-based inference in econometrics: Methods and applications*, page 328.

Hall, P. (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.

Handcock, M. S., Gile, K. J., Kim, B. J., and McLaughlin, K. R. (2023). *sspse: Estimating Hidden Population Size using Respondent Driven Sampling Data*. University of California, Los Angeles (urlhttps://github.com/HPMRG/sspse), Los Angeles, CA. R package version 1.1.0.

Handcock, M. S., Gile, K. J., and Mar, C. M. (2014). Estimating hidden population size using respondent-driven sampling data. *Electronic journal of statistics*, 8(1):1491.

Heckathorn, D. D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems*, 44(2):174–199.

Hladik, W., Barker, J., Ssenkusu, J. M., Opio, A., Tappero, J. W., Hakim, A., Serwadda, D., and Group, C. S. (2012). Hiv infection among men who have sex with men in kampala, uganda–a respondent driven sampling survey. *PloS one*, 7(5):e38143.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

Hunter, D. R. and Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of computational and graphical statistics*, 15(3):565–583.

Jiang, W. and Turnbull, B. (2004). The indirect method: inference based on intermediate statistics—a synthesis and examples. *Statistical Science*, 19(2):239–263.

Johnston, L. G., Thurman, T. R., Mock, N., Nano, L., and Carcani, V. (2010). Respondent-driven sampling: A new method for studying street children with findings from albania. *Vulnerable Children and Youth Studies*, 5(1):1–11.

Kahn, W. D. (1987). A cautionary note for bayesian estimation of the binomial parameter n. *The American Statistician*, 41(1):38–40.

Kerr, L., Kendall, C., Guimarães, M. D. C., Mota, R. S., Veras, M. A., Dourado, I., de Brito, A. M., Merchan-Hamann, E., Pontes, A. K., Leal, A. F., et al. (2018). Hiv prevalence among men who have sex with men in brazil: results of the 2nd national survey using respondent-driven sampling. *Medicine*, 97(1 Suppl).

Khabbazian, M., Hanlon, B., Russek, Z., and Rohe, K. (2017). Novel sampling design for respondent-driven sampling. *Electronic Journal of Statistics*, 11(2):4769–4812.

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.

Lee, C. and Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50.

Montealegre, J. R., Risser, J. M., Selwyn, B. J., McCurdy, S. A., and Sabin, K. (2013). Effectiveness of respondent driven sampling to recruit undocumented central american immigrant women in houston, texas for an hiv behavioral survey. *AIDS and Behavior*, 17(2):719–727.

Uusküla, A., Talu, A., Vorobjov, S., Salekešin, M., Rannap, J., Lemsalu, L., and Des Jarlais, D. (2020). The fentanyl epidemic in estonia: factors in its evolution and opportunities for a comprehensive public health response, a scoping review. *International Journal of Drug Policy*, 81:102757.

Verdery, A. M., Fisher, J. C., Siripong, N., Abdesselam, K., and Bauldry, S. (2017). New survey questions and estimators for network clustering with respondent-driven sampling data. *Sociological methodology*, 47(1):274–306.

Volz, E. and Heckathorn, D. D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of official statistics*, 24(1):79.

White, G. C. (1982). *Capture-recapture and removal methods for sampling closed populations*. Los Alamos National Laboratory.

Wu, J., Crawford, F. W., Raag, M., Heimer, R., and Uusküla, A. (2017). Using data from respondent-driven sampling studies to estimate the number of people who inject drugs: application to the kohtla-järve region of estonia. *Plos one*, 12(11):e0185711.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Additional Simulation Results for Section 3.3

This section contains simulation results that are referenced in Section 3.3 of the main text. It continues the empirical evaluation of the IIE for the two model parameters in Equation (1): A^S, the subgraph between study participants, and λ, the study paricipant arrival rate. Table 1 in Section 3.3 of the main text shows that $\widetilde{A}_n^S$, the IIE of A^S, has a higher true positive rate than $\widehat{A}_n^S$, the MLE of A^S, across all simulation settings.

We first evaluate the error rates of $\widetilde{A}_n^S$ and $\widehat{A}_n^S$ in more detail. Table 3 reports the true negative rates (TNR) of $\widehat{A}_n^S$ and $\widetilde{A}_n^S$ over a range of graph densities and population sizes. It shows that there is no discernible difference between the TNR of the IIE and MLE in these settings. Therefore, the higher true positive rates of $\widetilde{A}_n^S$ depicted in Table 1 do not come at the expense of overall accuracy.

We also compare the performance of the IIE and the MLE for λ in terms of MSE. Table 4 shows that the IIE, $\widetilde{\lambda}_n$, is considerably more accurate than the MLE, $\widehat{\lambda}_n$, over a range of graph sizes and densities. We observe that $\widetilde{\lambda}_n$ has an MSE that is less than 50% of the MSE of $\widehat{\lambda}_n$ across all settings. Additionally, the difference in MSE between $\widehat{\lambda}_n$ and $\widetilde{\lambda}_n$ is slightly higher with larger population sizes, which correspond to lower sample proportions (since the sample size is held fixed at $n = 100$), and higher average degrees.

Table 3: True Negative Rates of $\widehat{A}_n^S$ and $\widetilde{A}_n^S$ (%)

		MLE		IIE	
Pop.	Deg	Average	Std.	Average	Std.
1000	5	99.61	0.01	99.61	0.01
1000	10	99.08	0.01	99.08	0.01
1000	15	98.61	0.02	98.61	0.02
5000	5	99.92	0.00	99.92	0.00
5000	10	99.83	0.01	99.82	0.01
5000	15	99.72	0.01	99.72	0.01
10000	5	99.96	0.00	99.96	0.00
10000	10	99.90	0.00	99.90	0.00
10000	15	99.87	0.01	99.87	0.00

These are the true negative rates of $\widehat{A}_n^S$ and $\widetilde{A}_n^S$ for a series of population sizes (Pop.) and average degrees (Deg.). The standard deviations reported quantify the Monte Carlo error associated with these estimates based on 100 simulations.

Table 4: MSE of $\widehat{\lambda}_n$ and $\widetilde{\lambda}_n$

		IIE		MLE	
Pop.	Deg.	Mean	Sd	Mean	Sd
1000	5	0.09	0.02	0.21	0.02
1000	10	0.11	0.02	0.28	0.03
1000	15	0.09	0.02	0.24	0.03
5000	5	0.11	0.02	0.25	0.02
5000	10	0.13	0.03	0.36	0.04
5000	15	0.10	0.02	0.27	0.03
10000	5	0.10	0.02	0.28	0.03
10000	10	0.12	0.03	0.32	0.04
10000	15	0.09	0.02	0.29	0.03

These are the MSEs of the λ estimators for a series of population sizes (Pop.) and average degrees (Deg.). The standard deviations reported quantify the Monte Carlo error associated with these estimates over 100 simulations.

B Stochastic Block Model Sensitivity Analysis

This section presents additional results for Simulation 2 from Section 5 in the main text. In this simulation, we test the robustness of our population size estimators to misspecification in the graph model. The Erdos-Renyi model we employ assumes that edges between members of the population form with the same probability, p. However, individuals may be more likely to form connections with one group of people than another. Consider the following generative model for the population graph, $G = (V, E)$. The hidden population is divided into two groups, $V_A \subseteq V$ and $V_B = V \setminus V_A$ with sizes $N_A = |V_A|$ and $N_B = |V_B|$. The probability of an edge between members of the same group is p_{in}, and the probability of a connection between members of different groups is p_{out}. For constant $c \in [0, 1]$, we set $cp_{\text{out}} = p_{\text{in}}$ so that $p_{\text{in}} \geq p_{\text{out}}$. This is an example of a stochastic block model, which is used throughout network analysis (Holland et al., 1983; Lee and Wilkinson, 2019; Khabbazian et al., 2017). We let $p^* = \mathbb{P}(E_{ij} \in E)$, where nodes i and j are drawn uniformly at random from V. Defining E_{out} and E_{in} as the set of edges between and within groups respectively, we derive an expression for p^* in terms of p_{out} and c,

$$\begin{aligned} p^* &= \mathbb{P}(E_{ij} \in E) \\ &= \mathbb{P}(E_{ij} \in E_{\text{out}}) * p_{\text{out}} + \mathbb{P}(E_{ij} \in E_{\text{in}}) * p_{\text{in}} \\ &= \frac{2N_A N_B}{(N_A + N_B)(N_A + N_B - 1)} p_{\text{out}} \\ &\quad + \frac{N_A(N_A - 1) + N_B(N_B - 1)}{(N_A + N_B)(N_A + N_B - 1)} p_{\text{in}}. \end{aligned}$$

Since $cp_{\text{out}} = p_{\text{in}}$,

$$p^* = \frac{2N_AN_B + c(N_A(N_A-1) + N_B(N_B-1))}{(N_A+N_B)(N_A+N_B-1)} p_{\text{out}}.$$

We use this expression to set the overall edge prevalence in the simulations summarized by Table 5, making $N = 5000$ and $Np^* = 10$.

To assess the sensitivity of our population size estimators to the Erdos-Renyi model assumption, we vary N_A/N and c. As $N_A/N \to 1$ (or 0) or $c \to 1$, the Erdos-Renyi model becomes a better approximation of the truth. As $N_A/N \to 0.5$ and $c \to 0$, there is more heterogeneity in the graph edge probabilities, and the approximation becomes worse. We can see this pattern in Table 5. The first line of the table, $N_A/N = 1$ and $c = 1$, shows the MAD of our population size estimators under the Erdos-Renyi model for comparison. When $c = 0.3$ and $N_A/N = 0.5$, the error of the estimators is highest, and when $c = 0.9$ and $N_A/N = 0.75$, it is lowest. The rest of Table 5 illustrates a continuous spectrum between these two extremes. Lastly, as mentioned in Section 5 of the main text, the IIEs with and without auxiliary information still perform better than the MLE in this misspecified setting.

Table 5: MAD with Incorrect Strong Regularization Information for the Stochastic Block Model

N_A/N	c	MLE	IIE	IIE w/ Info
1.0	1.0	861.6	560.9	459.4
0.50	0.3	2318.7	2071.8	2021.4
0.50	0.6	1711.1	1425.1	1344.8
0.50	0.9	1116.1	755.6	638.7
0.75	0.3	1784.1	1519.8	1443.4
0.75	0.6	1502.3	1178.5	1079.9
0.75	0.9	1039.5	672.6	614.2

This table displays the Mean Absolute Deviation (MAD) of the population estimators over a series of N_A/N and c values. We use very strong regularization information centered at p_{in} to mimic ignorance of the two block structure. These results are averaged over simulations with Monte Carlo standard deviation error below 25.

C Simulation Results under Strong Regularization for Section 5

In this section, we present the results of a simulation under strong regularization. As described in Section 4.1 of the main text, we use a regularized MLE approach to estimate population size to avoid specifying informative priors that are difficult to justify scientifically. The regularization function, $R(p) = \log \text{Beta}(p; a, b)$, incorporates information on edge prevalence, p — where $\text{Beta}(p; a, b)$ is a Beta distribution that is centered at p with a variance that is inversely proportional to a. Here we use the same setup as Simulation 1 but vary the hyperparameters in the regularizer.

In Figure 3 of Section 5 in the main text, we compare the MAD of $\widehat{N}$ (MLE), $\widetilde{N}$ (IIE), and $\widetilde{N}^r$ (IIE with auxiliary information) with $a = 0.1$, and a series of Successive Sampling (SS) estimators. We observe that $\widetilde{N}^r$ improves on $\widetilde{N}$, and both are more accurate than $\widehat{N}$. Additionally, the performances of the SS estimators are highly dependent on their prior. In Figures 4a and 4b, we show the log(MAD) of $\widehat{N}, \widetilde{N}$, and $\widetilde{N}^r$ with $a = 1$ and $a = 10$ respectively. The relationships between estimators $\widehat{N}, \widetilde{N}$ and $\widetilde{N}^r$ mirror Figure 3. Encouragingly, the MAD of our population size estimators decreases significantly as a increases, and, with strong regularization information, $\widehat{N}, \widetilde{N}$ and $\widetilde{N}^r$ are consistently more accurate than the SS estimators.

D IIE and Successive Sampling Algorithm Details

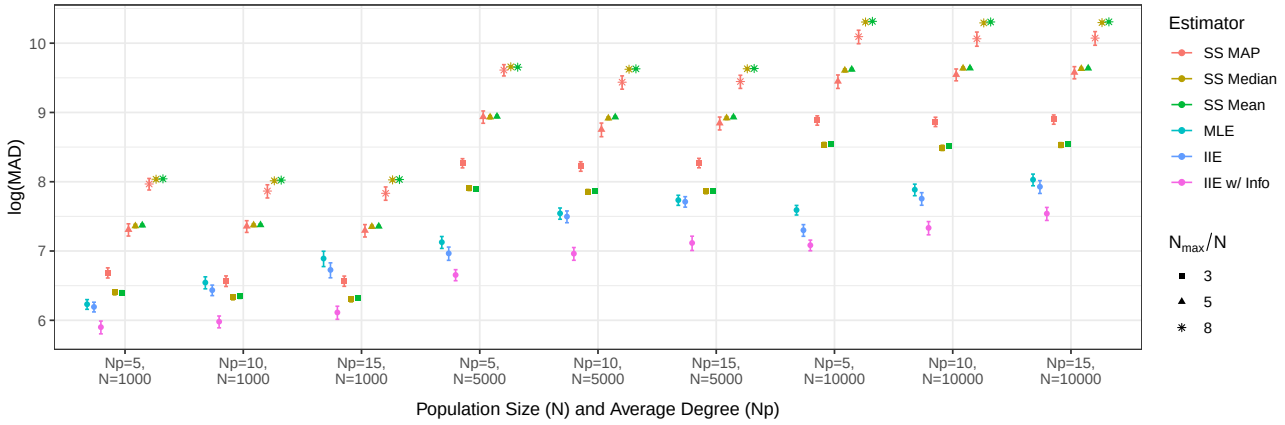
In this section, we present the details of Algorithms 1 and 2 (Algorithm 1 is introduced in Section 3 of the main text).

Both algorithms construct the IIE by finding the parameters under which the expected value of a calibration statistic is equal to the observed value, where we set the calibration statistic equal to the MLE of λ. In the simulations of Section 5 in the main text, we use $K = 9$ grid values centered at $\widehat{\lambda}_n$, the MLE for the observed data. Specifically, we set $\lambda^k = \widehat{\lambda}_n - (k-4) \times 0.1$ for $k \in \{1, 2, \ldots, 9\}$. For each set of candidate parameters, $\left\{\lambda^k, A^S_\lambda(\mathbf{Y}, \lambda^k)\right\}$ and $\left\{\lambda^k, A^{S,r}_\lambda(\mathbf{Y}^r, \lambda^k)\right\}$ for Algorithms 1 and 2 respectively, we approximate the expected value (with $J = 25$) of the MLE of λ, labeling this quantity $\widehat{\lambda}^k_n$. The IIE are the parameters under which $\widehat{\lambda}^k_n$ is closest to $\widehat{\lambda}_n$.

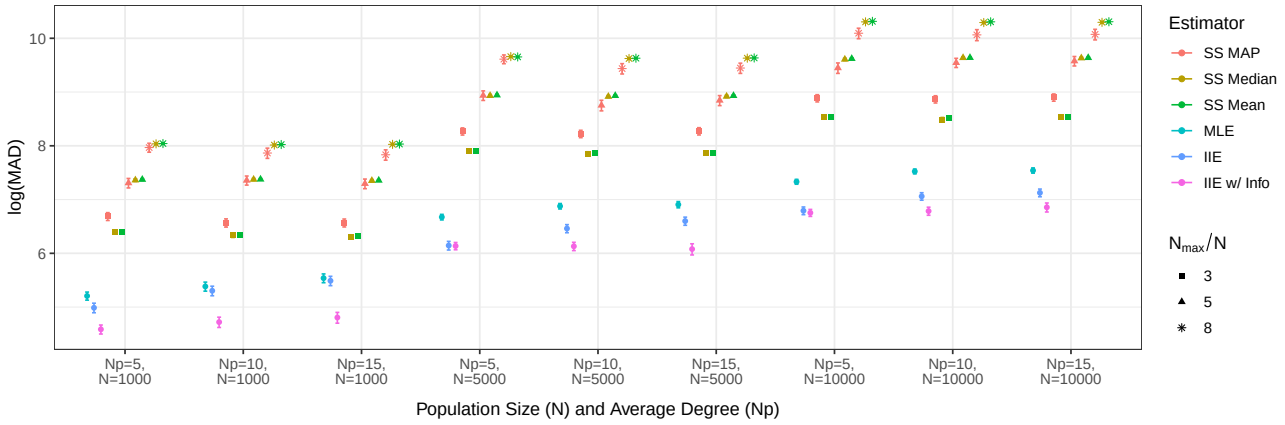
As described in Section 4.2 of the main text, Algorithm 2 augments Algorithm 1 with auxiliary information. We note that this implies the MLE is taken with respect to different likelihoods in Algorithms 1 and 2. Defining $\beta \in \mathbb{R}^p$ for $p \in \mathbb{N}$ as the parameter that indexes the distribution of $\mathbf{Q}$, the MLE referenced in Algorithm 2 is

$$\left\{\widehat{A}^S_n, \widehat{\lambda}_n, \widehat{\beta}_n\right\} = \arg \max_{A^S \in \mathcal{A}, \lambda \in \mathbb{R}^+, \beta \in \mathbb{R}^p} \mathcal{L}_n(\mathbf{Y}, \mathbf{Q} | A^S, \lambda, \beta).$$

Algorithms 1 and 2 take about 24 hours to run with a sample size of 100 implemented by the code included



(a) $R(p) = \log \text{Beta}(a, b)$ where $a = 1$



(b) $R(p) = \log \text{Beta}(a, b)$ where $a = 10$

Figure 4: This figure compares the performance of $\widehat{N}, \widetilde{N}$, and $\widetilde{N}^r$ under strong regularization information over a series of population sizes, N, and average degrees, Np, with 90% Monte Carlo confidence intervals.

in the Supplementary Material.

Lastly, we use the SSPSE package (Handcock et al., 2023) under a "flat" prior setting to construct the SS estimators analyzed in Figures 3 and 4.

Algorithm 2: The Indirect Inference Estimator with Auxiliary Information

We want to find the estimator,

$$\widetilde{\lambda}_n^r \in \arg\min_{\lambda \in \mathbb{R}^+} \left| \mathbb{E}_{\mathbf{Z} \sim P_{A_\lambda^{S,r}(\mathbf{Y}^r,\lambda),\lambda}} \left\{ \lambda^\dagger(\mathbf{Z}^r) \right\} - \lambda^\dagger(\mathbf{Y}^r) \right|$$

Define β as the parameter that indexes the distribution of $\mathbf{Q}$;

Generate a grid of λ^k values, $k \in \{1, 2, .., K\}$;

for *k in* $\{1, 2, .., K\}$ **do**

  for *j in* $\{1, 2, .., J\}$ **do**

    Find $A_{k,j}^S = \max_{A^S,\beta} \mathcal{L}(A^S, \beta | \lambda^k, \mathbf{Y}^r)$;

    Simulate wait time vector $w^{k,j}$ from the model defined by parameters $A_{k,j}^S, \lambda^k$;

    Find $\widehat{\lambda}_n^{k,j}, \widehat{A}_n^{S,k,j}$ by maximizing the likelihood conditional on the generated data $\mathbf{Z}_{k,j}^r = (w^{k,j}, G^R, \mathbf{d}, \mathbf{Q})$;

  end

  Save vector $(\lambda^k, A_k^S, \beta^k, \widehat{\lambda}_n^k = \frac{\sum_{j=1}^J \widehat{\lambda}_n^{k,j}}{n})$;

end

Calculate $k^* = \arg\min_k \left| \widehat{\lambda}_n^k - \lambda_r^\dagger(\mathbf{Y}^r) \right|$;

Our estimator is then

$$\left(\widetilde{\lambda}_n^r, \widetilde{A}_n^{S,r}, \widetilde{\beta}^r \right) = \left(\lambda^{k^*}, A_{k^*}^S, \beta^{k^*} \right)$$