Hidden Population Estimation with Indirect Inference and Auxiliary Information

Justin Weltz, Eric Laber, and Alexander Volfovsky

Duke University

Introduction

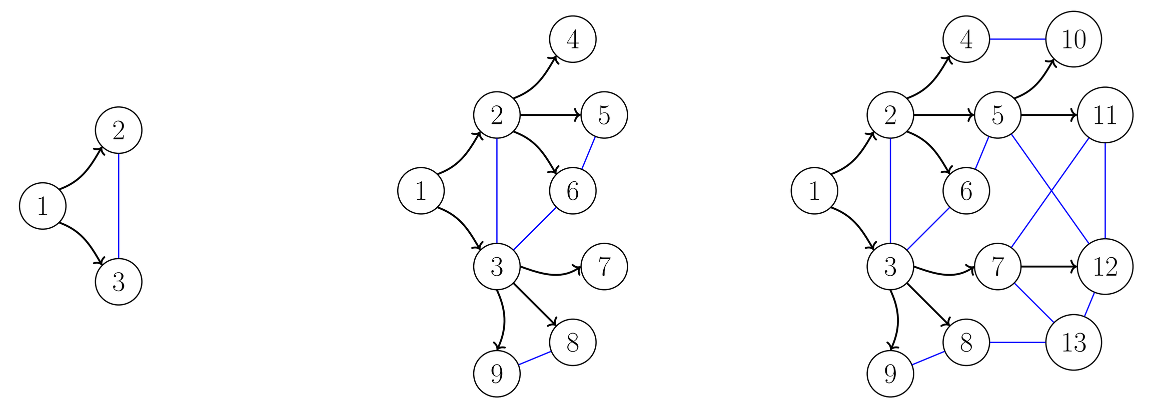
How do we sample a population that is hard-to-reach to perform statistical inference?

Examples
- People who are unhoused
- People who are undocumented
- People who inject drugs

Sampling Difficulties
- Identification
- Trust
- Anonymity

Respondent-Driven Sampling (RDS)

Interview seed participants → Give current study participants coupons to recruit others → New study participants redeem coupons and are interviewed → Give current study participants coupons to recruit others

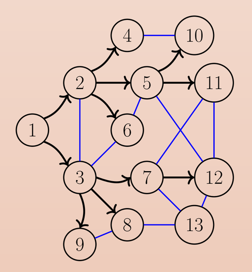


RDS Data and Inference

We consider social network $G = (V, E)$, where V is the set of N individuals and E is the set of pairwise connections, or edges, between individuals.

Data:
- The recruitment subgraph, $G^R = (V^R, E^R) \subset G$
- Participants' degrees
- Participants' arrival times

Inference Goals:
- ➔ The complete sample subgraph, $G^S = (V^S, E^S)$
- ➔ The populations size, N



- G^R is composed of coupon exchanges →
- G^S includes both observed → and unobserved connections —

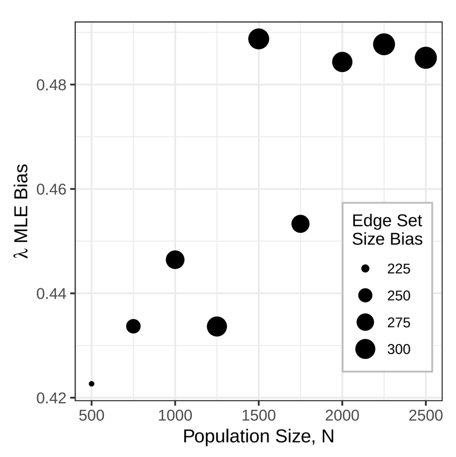
Problems

Assuming the wait times associated with the recruitment process are i.i.d. exponential with mean $1/\lambda$, the likelihood is

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

- A^S is the adjacency matrix representation of G^S
- $\boldsymbol{s}$ contains the number of active coupons before each recruitment time
- $\boldsymbol{w}$ contains the time periods between recruitments
- M is the set of original study participants

Problem 1: Biased maximum likelihood estimation (MLE)



This figure depicts the bias of the MLEs for λ and $|E^S|$. *We can see that the biases of these estimators are positively correlated and increase as the sample proportion decreases.*

Problem 2: Ignoring useful RDS survey information

The RDS data collection process commonly includes a large survey that can be used to improve estimation.

Solutions

Solution 1: Indirect Inference Estimation (IIE)

The IIE is constructed by finding parameter settings under which the expected value of a calibration statistic matches its observed value.

For each λ' in a grid of λ values,

Simulate from the RDS model consistent with λ' → Calculate the MLEs for these datasets → Calculate the average MLE over datasets

Repeat this process for each grid value. *The IIE is the RDS model whose average MLE is closest to the observed MLE.*

Solution 2: Auxiliary Information

In the RDS survey, it is possible to track how information accumulates over the RDS process, **and this measurement necessarily carries information about the underlying network.**

Population Estimation

To estimate the population size, we must specify a model for the graph G. For example, assume the population graph is a sample from an Erdos-Renyi graph model with parameters N and p.
- N individuals in the graph
- p is the probability of a connection between any two population members

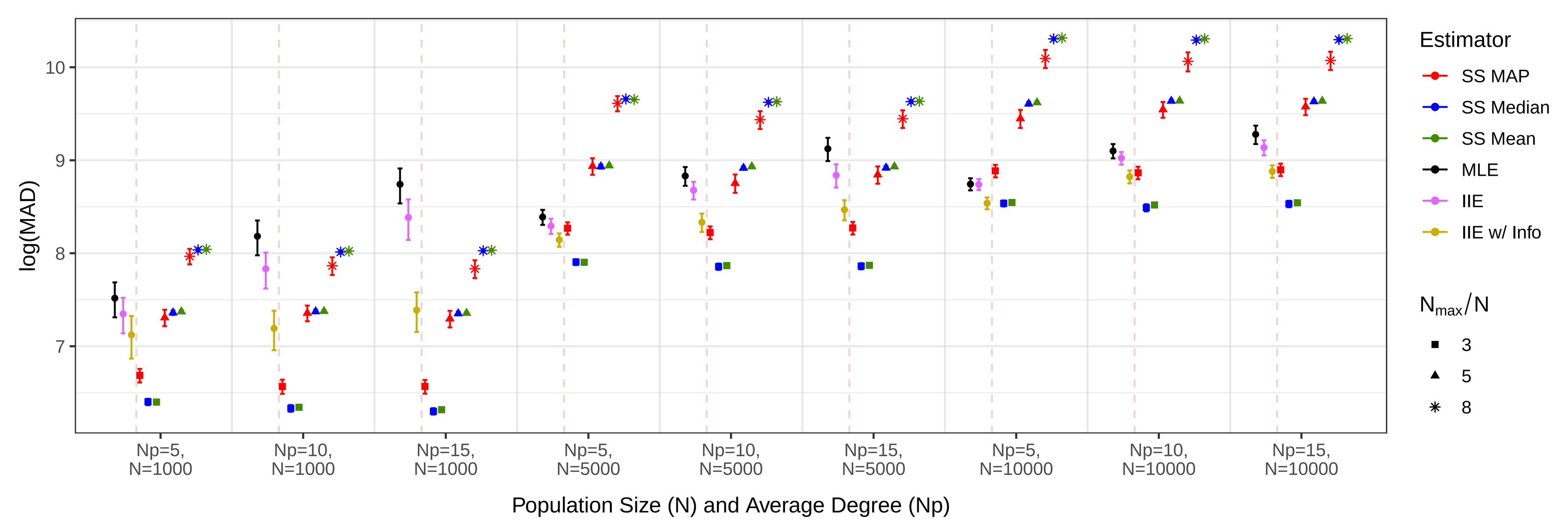
Label the number of edges individual i shares **with unsampled members of the hidden population** at the time of their recruitment as d_i^u. This quantity **depends on** A^S and $d_i^u \overset{\text{iid}}{\sim} \text{Bin}(N - i, p)$.

Estimating Population Size

RDS Data $\xrightarrow{\text{IIE}}$ $\widehat{A^S}, \hat{\lambda}$ $\xrightarrow{\text{Regularized MLE}}$ $\hat{N}, \hat{p}$

Simulation Results

We compare the maximum absolute deviation (MAD) of successive sampling estimators [Gile, 2011], the MLE, the IIE, and the IIE with auxiliary information over a series of population sizes, N, and average degrees, Np, with 90% Monte Carlo confidence intervals.



Case Study

RDS was conducted in Estonia among people who inject drugs (PWID).
- From 2015-2021, Estonia had the highest per capita prevalence of PWID in Europe
- To lower the prevalence of HIV among PWID in Estonia, syringe exchange programs were launched in 1997 [Wu et al., 2017]
- **Estimating the size of the PWID population sheds light on the magnitude of this public health crisis and the necessary scope of potential policy solutions**

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