New methods for inference from Respondent-Driven Sampling Data

Krista J. Gile

University of Massachusetts, Amherst
Joint work with Mark S. Handcock
UCLA*

June 22, 2012

*Research supported by NICHD grant 7R29HD034957 and NIDA grant 7R01DA012831, UW Networks Project (Martina Morris, PI) and by NSF grant DMS-0354131, UW NSF VIGRE, Postdoctoral Prize Research Fellowship, Nuffield College.
Hard-to-Reach Population Methods Research Group (HPMRG) (and Collaborators)

- Ian Fellows, UCLA
- Krista J. Gile, UMass, Amherst
- Mark S. Handcock, UCLA
- Lisa G. Johnston, Tulane University, UCSF
- Corinne M. Mar, University of Washington
- Miles Ott, Brown University
- Miruna Petrescu-Prahova, University of Washington
- Matt Salganik, Princeton University
- Amber Tomas, Mathematica Policy Research
Outline of Presentation

1. Link-Tracing Hidden Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Network Model-Assisted Estimator
7. Sensitivity Analysis
8. Application
9. Discussion
Outline of Presentation

1. Link-Tracing Hidden Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Network Model-Assisted Estimator
7. Sensitivity Analysis
8. Application
9. Discussion
Sampling Hard-to-Reach Populations

- Motivation: UNAIDS
  - Requires HIV prevalence estimates for all countries
  - Most countries: concentrated in high-risk populations:
    Injecting drug users, men who have sex with men, and sex workers
  - Hard-to-reach networked populations.
- Other applications: Unregulated workers, jazz musicians

Traditional Survey Sampling:

- Probability sample (e.g. simple random sampling, stratified random sampling)
- Analyze data using sampling weights

Hidden populations: No practical conventional sampling frame.
Suppose:

- Each population joined by informal social network of relationships.
- Researchers can access some members of the population.

Then:

- Begin with a reachable convenience sample (the *seeds*).
- Expand sample by following social network ties.

This is Link-tracing Network Sampling.
Outline of Presentation

1. Link-Tracing Hidden Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Network Model-Assisted Estimator
7. Sensitivity Analysis
8. Application
9. Discussion
Respondent-Driven Sampling - Link-tracing variant:

- Seed Dependence: Follow only a few links from each sampled
- Confidentiality: Respondents distribute uniquely identified coupons. No names. (respondent-driven)
- Estimation based on Network positions: Several approaches

- Effective at obtaining large varied samples in many populations.
- Widely used: over 100 studies, in over 30 countries. Often HIV-risk populations.


Stylized population
Start with seeds . . .
Seeds recruit the first wave . . .
First wave recruit the second wave . . .
and so on . . .
(and with un-sampled)
degree of node $i = \#$ of ties of node $i$
$homophily = \frac{\text{Percent realized infected to infected ties}}{\text{Percent realized uninfected to infected tie}}$
Outline of Presentation

1. Link-Tracing Hidden Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Network Model-Assisted Estimator
7. Sensitivity Analysis
8. Application
9. Discussion
Link-Tracing Sampling:

- Challenges
  - Sampling depends on (typically) partially-observed network data
  - Convenience mechanism for initial sample leads to non-probability sample
  - Unknown population size = unknown sampling frame
- Sampling designs have much in common, but no consensus on inferential approach

Respondent-Driven Sampling subject to all of these
Classic Design-Based Inference:  
Generalized Horvitz-Thompson Estimator

• Goal: Estimate proportion “infected”:  

\[ \mu = \frac{1}{N} \sum_{i=1}^{N} z_i \]

where population labeled 1, 2, ..., \( N \),

\[ z_i = \begin{cases} 
1 & \text{i infected} \\
0 & \text{i uninfected.} 
\end{cases} \]

• Generalized Horvitz-Thompson Estimator:

\[ \hat{\mu} = \frac{\sum_i S_i \frac{z_i}{\pi_i}}{\sum_i S_i \frac{1}{\pi_i}} \]

where

\[ S_i = \begin{cases} 
1 & \text{i sampled} \\
0 & \text{i not sampled} 
\end{cases} \quad \pi_i = P(S_i = 1). \]

Key Point: Requires \( \pi_i \forall i : S_i = 1 \).
Simulation Study

Simulate Population
- 1000, 835, 715, 625, 555, or 525 nodes
- 20% “Infected”

Simulate Social Network (from ERGM, using statnet)
- Mean degree 7
- Homophily on Infection: \( R = \frac{P(\text{infected to infected tie})}{P(\text{uninfected to infected tie})} = 5 \) (or other)
- Differential Activity: \( w = \frac{\text{mean degree infected}}{\text{mean degree uninfected}} = 1 \) (or other)

Simulate Respondent-Driven Sample
- 500 total samples
- 10 seeds, chosen proportional to degree
- 2 coupons each
- Coupons at random to relations
- Sample without replacement

Repeat 1000 times!

Blue parameters varied in study.
Outline of Presentation

1. Link-Tracing Hidden Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Network Model-Assisted Estimator
7. Sensitivity Analysis
8. Application
9. Discussion
One Approach: Random walk approximation

Consider:

- Connected undirected network
- Random walk on network
- A Markov chain on nodes
- Then stationary distribution proportional to nodal degree.
One Approach: Random walk approximation

Respondent-driven Sampling:

- Approximate link-tracing process by this Markov chain
- Assume sample can be treated as from stationary distribution
- Then sampling probabilities proportional to degree.


**Volz-Heckathorn Estimator (VH):** inverse probability weighted by degrees

\[ \hat{\mu} = \frac{\sum_i S_i \frac{z_i}{d_i}}{\sum_i S_i \frac{1}{d_i}} \]

where \( d_i \) = degree of node \( i \), \( S_i \) sample indicator, \( z_i \) quantity of interest.
Volz-Heckathorn, $w=1$

Estimated Proportion Infected

Sample %: 50% 60% 70% 80% 90% 95%
Varying Sample Percentage, $w=1.4$

Estimated Proportion Infected

Sample %: 50% 60% 70% 80% 90% 95%
Outline of Presentation

1. Link-Tracing Hidden Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Network Model-Assisted Estimator
7. Sensitivity Analysis
8. Application
9. Discussion
Finite Population Correction

Consider:

- A distribution uniform over all networks with given nodal degrees
- Marginalizing over this distribution of networks, transition probabilities of random walk proportional to degree

Furthermore, consider:

- A without-replacement random walk, over the same distribution of networks
- Then transition probabilities equivalent to *successive sampling*

Successive Sampling (aka PPSWOR):

- Select the first unit (node) with probability proportional to size (degree).
- Select each additional unit with probability proportional to size

*from the remaining unsampled units*
New Estimator based on Successive Sampling

Estimate sampling probabilities based on successive sampling

These probabilities:

- Depend on population size
- Depend on sizes of all units
- Are not available in closed form

Approach:

- Assume population size known (sensitivity analysis)
- Novel iterative algorithm

Successive Sampling (SS) Estimator: Algorithm

- Goal: Estimate sampling probabilities ($\pi_k$) by degree $k$.
- A function of population degree distribution $\mathbb{N}$, $\pi_k(\mathbb{N})$.

1. Initial: $\pi_k(\mathbb{N}^0) \propto k$.
2. For $i = 1 \ldots r$:
   
   (a) Estimate degree distribution $\mathbb{N}^i$ by Generalized Horvitz-Thompson Estimator
   
   (b) Compute $\pi_k(\mathbb{N}^i)$ by simulation:
      
      i. Simulate $M$ SS samples from $\mathbb{N}^i$
      
      ii. $\pi_k(\mathbb{N}^i) = \mathbb{E}[V_k; \mathbb{N}^i] \approx \frac{U_k + 1}{M \cdot \mathbb{N}_k^i + 1}$,

      where $V_k$ is the number of sample units of degree $k$, and $U_k$ is the number sampled in the $M$ simulations.

3. Use $\hat{\pi} = \pi(\mathbb{N}^r)$ to estimate $\mu$:

   $\hat{\mu}_{SS} = \frac{\sum_i S_i \frac{z_i}{\hat{\pi}_d_i}}{\sum_i S_i \frac{1}{\hat{\pi}_d_i}}$. 
Standard Error Estimation:

Population Bootstrap:

- Simulate Population
  - Estimate $z$ by $d$ distribution
  - Estimate infection mixing matrix by $z$
- Simulate without-replacement sampling
  - Choose recruit $z$ according to mixing matrix
  - Choose recruit $d$ by successive sampling
  - Update available population and mixing matrix
- Compute SS Estimates
- Results:
  - Performs well across differential activity ($w$) and sample fraction
  - Performs well with homophily
  - Unreliable when seeds biased.
Volz-Heckathorn, $w=1$

Sample %: 50% 60% 70% 80% 90% 95%
Volz-Heckathorn, $w=1.4$

Estimated Proportion Infected

Sample %: 50% 60% 70% 80% 90% 95%
SS, $w=1.4$

Estimated Proportion Infected

Sample %: 50% 60% 70% 80% 90% 95%
All Infected Seeds, varying Homophily, 50%

Expected Prevalence Estimate (Truth = 0.20)
All Infected Seeds, varying number of seeds, 50%
Outline of Presentation

1. Link-Tracing Hidden Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Network Model-Assisted Estimator
7. Sensitivity Analysis
8. Application
9. Discussion
Seed Bias

- Depends on network structure (homophily)
- Depends on branching structure (waves)
- Also, need finite population correction.

Mathematically a random walk that is:

- Branching

- Without-Replacement

- on a Non-regular graph
Seed Bias

- Depends on network structure (homophily)
- Depends on branching structure (waves)
- Also, need finite population correction.

Mathematically a random walk that is:

- **Branching**  
  in an infinite space
- **Without-Replacement**  
  on a regular graph (lattice)
- on a **Non-regular graph**  
  with replacement, non-branching

Joint treatment analytically elusive.
Network Model-Assisted Estimator

- Interested in sampling probabilities $\pi_i = \mathbb{E}(S_i)$.
- Should reflect:
  - Nodal degree $d_i$
  - Sample fraction
  - Seed selection
  - Homophily and Branching Structure
Approach

Idealizations:

1. For known network $y$, seeds $s$, compute $\pi_i = \mathbb{E}(S_i | y, s)$.
2. For known network model, $\eta$, $\pi_i = \sum_{y \in \mathcal{Y}} P(y | \eta) \mathbb{E}(S_i | y, s)$

We do not know $y$ or $\eta$. So we estimate $\eta$. 
Exponential Random Graph Model

Exponential-family model for network $Y$, conditional on infection status $z$ and nodal degrees $d$.

$$P(Y = y) = \frac{\exp[\eta \cdot m(y, z, d)]}{c(\eta)}$$

$y \in \mathcal{Y}$, the space $\mathcal{Y}$ consists of all binary undirected networks consistent with $d$ and $z$, and

$$c(\eta) = \sum_{u \in \mathcal{Y}} \exp[\eta \cdot m(u, z, d)]$$

A restriction of the common exponential-family random graph model (ERGM).

Here,

$$m(y, z, d) = \sum_{i,j} y_{ij} z_i (1 - z_j)$$

Require:

- $\mathbb{N}$ (degree-infection distribution of population)
- Sufficient statistic: $m(y, z, d)$ (number of cross-ties)
Fitting the Model

Problem: Requires (unknown) population proportions and sufficient statistic.

Solution: Use design-based estimators

\[
\hat{N}_{kl} = \frac{1}{N} \sum_{i=1}^{N} \frac{S_i \mathbb{I}(d_i = k, z_i = l)}{\hat{\pi}_i}
\]

\[
\hat{m}(\eta) = \sum_{i=1}^{N} \frac{S_i (x_i(1 - z_i) + (d_i - x_i)z_i)}{2\hat{\pi}_i}
\]

where \( x_i = \sum_j z_j y_{ij} \) requires the observation of \( x_i \forall i : S_i = 1 \).

For sampling \( S_i \), degree \( d_i \), infection \( z_i \)

Problem: This, in turn, requires sampling probabilities.

Solution: Novel iterative algorithm to find self-consistent solution.
Model-Assisted Estimator: Algorithm

• Goal: Estimate sampling probabilities ($\pi_i$) by degree $d_i$ and infection $z_i$.

• A function of homophily ($\eta$), and population of degrees and infection $N$.

• Estimate $\hat{\pi}_i$ proportional to degree $d_i$.

• Iterate the following steps:
  – Estimate $N$ and $m(\eta)$ using $\hat{\pi}_i$.
  – Find corresponding model parameter $\eta$ (statnet R package)
  – Simulate $M$ networks, and samples from networks. Estimate $\hat{\pi}_i$ by simulation.

• Use the resulting estimated probabilities, $\hat{\pi}_i$, to form weighted estimator.

$$\hat{\mu}_{MA} = \frac{\sum_i S_i \frac{z_i}{\hat{\pi}_i}}{\sum_i S_i \frac{1}{\hat{\pi}_i}}.$$
Standard Error Estimation

Population Bootstrap:

- Simulate M populations
  - Estimate $z$ by $d$ distribution
  - Estimate $\eta$
  - Simulate networks according to $\eta$
- Simulate RDS samples
  - Fix seed distribution
  - Sample without replacement
- Compute MA estimates. Average estimates over M populations
- Results:
  - Performs well across differential activity ($w$), sample fraction, seed bias
  - Computationally expensive
Estimated Sampling Probabilities

- Infected
- Uninfected

Estimated Nodal Inclusion Probability vs. Nodal Degree

- 5
- 10
- 15

- 0.0
- 0.2
- 0.4
- 0.6
- 0.8
- 1.0
Simulation Study

Critical Questions:

• Does Model-Assisted estimator perform as well as SS estimator for \( w \neq 1 \) and large sample fraction?
• Does Model-Assisted estimator correct for seed bias?
• How well does parametric bootstrap perform?
• What about unknown population size and network structure?

Comparison of Estimators:

• Mean: Naive Sample Mean
• SH: Salganik-Heckathorn: based on MME of number of cross-relations
• VH: Existing Volz-Heckathorn Estimator
• SS: New SS Estimator
• MA: New Network Model-Assisted Estimator
50% Sample, \( w = 1, R = 1 \), Random Seeds
70% Sample, \( w = 1.8, \ R = 1, \) Random Seeds

Estimate of Proportion Infected, Truth=0.20

\[ \begin{array}{c}
\text{Mean} \\
\text{SH} \\
\text{VH} \\
\text{SS} \\
\text{MA}
\end{array} \]
50% Sample, \( w = 1, R = 5 \), Infected Seeds
70% Sample, \( w = 1.8, R = 5 \), Infected Seeds
All Infected Seeds, varying Homophily

Mean
VH
SS

Expected Prevalence Estimate (Truth = 0.20)
0.10 0.15 0.20 0.25 ... 1 3 5
All Infected Seeds, varying Homophily

Expected Prevalence Estimate (Truth = 0.20)

- Mean
- SH
- VH
- SS
All Infected Seeds, varying Homophily

Expected Prevalence Estimate (Truth = 0.20)

- Mean
- SH
- VH
- SS
- MA
All Infected Seeds, varying number of seeds (waves)
All Infected Seeds, varying number of seeds (waves)

Expected Prevalence Estimate (Truth = 0.20)

Max Wave Seeds

4 20

5 10

6 6

Mean

SH

VH

SS

Legend:

- Mean
- SH
- VH
- SS
## All Infected Seeds, varying number of seeds (waves)

<table>
<thead>
<tr>
<th>Max Wave</th>
<th>Seeds</th>
<th>Expected Prevalence Estimate (Truth = 0.20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>20</td>
<td>●</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>●●●●</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>●●●●●</td>
</tr>
</tbody>
</table>

- **SH**
- **VH**
- **SS**
- **MA**

**Mean**

- **Black**
- **SH**
- **VH**
- **SS**
- **MA**
## Parametric Bootstrap

<table>
<thead>
<tr>
<th>% homoph. sample</th>
<th>R</th>
<th>( w )</th>
<th>sample bias</th>
<th>SE observed</th>
<th>SE bootstrap</th>
<th>coverage 95%</th>
<th>coverage 90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>1</td>
<td>1</td>
<td>No</td>
<td>0.0140</td>
<td>0.0137</td>
<td>94.1%</td>
<td>88.8%</td>
</tr>
<tr>
<td>70%</td>
<td>1</td>
<td>1.8</td>
<td>No</td>
<td>0.0073</td>
<td>0.0075</td>
<td>94.9%</td>
<td>90.4%</td>
</tr>
<tr>
<td>50%</td>
<td>5</td>
<td>1</td>
<td>Initial</td>
<td>0.0188</td>
<td>0.0175</td>
<td>93.7%</td>
<td>87.9%</td>
</tr>
<tr>
<td>50%</td>
<td>5</td>
<td>1.8</td>
<td>Initial</td>
<td>0.0079</td>
<td>0.0080</td>
<td>95.0%</td>
<td>87.3%</td>
</tr>
<tr>
<td>50%</td>
<td>5</td>
<td>1</td>
<td>Referral</td>
<td>0.0216</td>
<td>0.0225</td>
<td>91.7%</td>
<td>84.7%</td>
</tr>
</tbody>
</table>
Outline of Presentation

1. Link-Tracing Hidden Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Network Model-Assisted Estimator
7. Sensitivity Analysis
8. Application
9. Discussion
Sensitivity Analysis

- **Unknown Population Size**
  - Repeat simulations with inaccurate population estimate.

- **Unknown Network Structure**
  - Repeat simulations with more complex network model.
$N = 1000$, 50% **Sample**, $w = 1$, $R = 1$, Random Seeds
$N = 715, \, 70\% \text{ Sample}, \, w = 1.8, \, R = 1, \text{ Random Seeds}$
$N = 1000$, $50\%$ Sample, $w = 1$, $R = 5$, Infected Seeds
$N = 715$, 70\% Sample, $w = 1.8$, $R = 5$, Infected Seeds
Increased Triangles (4 × edges with shared partner)
Increased Geometric Function of Edge-Triangles (10 ×)

Estimate of Proportion Infected, Truth = 0.20

- Mean
- SH
- VH
- SS
- MA

Graph showing the distribution of estimates for different categories (SH, VH, SS, MA) with a mean of 0.20.
Outline of Presentation

1. Link-Tracing Hidden Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Network Model-Assisted Estimator
7. Sensitivity Analysis
8. Application
9. Discussion
HIV Prevalence among MSM in a Caribbean City

HIV of MSM

Estimated HIV Prevalence

VH

SS (± 1 se)

Mean

Population Proportion MSM

0.002 0.005 0.010 0.020

0.05 0.06 0.07 0.08 0.09 0.10 0.11
HIV Prevalence among IDU in an Eastern European City

Estimated Prevalence

<table>
<thead>
<tr>
<th></th>
<th>0.76</th>
<th>0.78</th>
<th>0.80</th>
<th>0.82</th>
<th>0.84</th>
<th>0.86</th>
<th>0.88</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Correct</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seeds Seeds Offspring</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean
SH
VH
SS
MA (+/− 1se)
HIV Prevalence among IDU in an Eastern European City

Estimated Prevalence

- Random Correct Seeds & Seeds Offspring

Mean

- SH
- VH
- SS
- MA (+/− 1se)
<table>
<thead>
<tr>
<th>Wave</th>
<th>Uninfected Recruiter</th>
<th>Avg</th>
<th>Infected Recruiter</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>7</td>
<td>0</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>8 2 1 3</td>
<td>0.93</td>
<td>17 11 5</td>
<td>0.75</td>
</tr>
<tr>
<td>8</td>
<td>4 2 2</td>
<td>1.25</td>
<td>15 21 8</td>
<td>1.08</td>
</tr>
<tr>
<td>7</td>
<td>2 1 1 3</td>
<td>1.71</td>
<td>10 2 4 4</td>
<td>1.1</td>
</tr>
<tr>
<td>6</td>
<td>4 1 1 1</td>
<td>0.86</td>
<td>9 5 2 4</td>
<td>1.05</td>
</tr>
<tr>
<td>5</td>
<td>1 2 1</td>
<td>1</td>
<td>9 1 2 6</td>
<td>1.28</td>
</tr>
<tr>
<td>4</td>
<td>2 2</td>
<td>0.5</td>
<td>11 4 2 4</td>
<td>0.95</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>0</td>
<td>6 2 7</td>
<td>1.67</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>8 11 4</td>
<td>1.07</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>7 1 4</td>
<td>1.15</td>
</tr>
<tr>
<td>0</td>
<td>-</td>
<td>0</td>
<td>1 2 3</td>
<td>2.33</td>
</tr>
</tbody>
</table>

Total 30 8 6 9 0.89

Legend: Number of Recruits: 0 1 2 3
HIV Prevalence among IDU in an Eastern European City

<table>
<thead>
<tr>
<th>Estimated Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.76</td>
</tr>
<tr>
<td>0.78</td>
</tr>
<tr>
<td>0.80</td>
</tr>
<tr>
<td>0.82</td>
</tr>
<tr>
<td>0.84</td>
</tr>
<tr>
<td>0.86</td>
</tr>
<tr>
<td>0.88</td>
</tr>
</tbody>
</table>

Random Correct Seeds &
Seeds Seeds Offspring

- Mean
- SH
- VH
- SS
- MA (+/- 1se)
HIV Prevalence among IDU in an Eastern European City

Estimated Prevalence

0.76 0.78 0.80 0.82 0.84 0.86 0.88

Random Correct Seeds & Seeds Offspring

Mean

SH

VH

SS

MA (+/- 1se)
Outline of Presentation

1. Link-Tracing Hidden Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Network Model-Assisted Estimator
7. Sensitivity Analysis
8. Application
9. Discussion
Discussion: New Estimators

Prevalence Estimate (Truth = 0.20)

Samp %  50%  70%  50%  70%
w 1  1.8  1  1.8
R 1  1  5  5
Seeds Random Random Infected Infected

Mean
SH
VH
SS
MA
### Discussion: Respondent-Driven Sampling - Assumptions

<table>
<thead>
<tr>
<th>Network Structure Assumptions</th>
<th>Sampling Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Walk Model</td>
<td>Network size large ( N &gt;&gt; n )</td>
</tr>
<tr>
<td>Remove Seed Dependence</td>
<td>Homophily weak enough Connected graph</td>
</tr>
<tr>
<td>To Estimate Probabilities</td>
<td>All ties reciprocated</td>
</tr>
<tr>
<td>Additional Assumptions of SS</td>
<td>Known network size ( N )</td>
</tr>
<tr>
<td>Additional Assumptions of MA</td>
<td>Non-random mixing observable</td>
</tr>
<tr>
<td></td>
<td>Network model form</td>
</tr>
</tbody>
</table>

Sampling with replacement
Single non-branching chain
Sufficiently many sample waves
Degree accurately measured
Random referral
No seed bias

**Assumptions of Volz-Heckathorn Estimator**
### Discussion: Respondent-Driven Sampling - Assumptions

<table>
<thead>
<tr>
<th>Network Structure Assumptions</th>
<th>Sampling Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Walk Model</td>
<td><strong>Network size large</strong> ($N \gg n$)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Remove Seed Dependence</td>
<td><strong>Homophily weak enough</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Connected graph</strong></td>
</tr>
<tr>
<td>To Estimate Probabilities</td>
<td><strong>All ties reciprocated</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional Assumptions of SS</td>
<td><strong>Known network size</strong> $N$</td>
</tr>
<tr>
<td>Additional Assumptions of MA</td>
<td><strong>Non-random mixing observable</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Network model form</strong></td>
</tr>
</tbody>
</table>

*Assumptions of Successive Sampling Estimator*
## Discussion: Respondent-Driven Sampling - Assumptions

<table>
<thead>
<tr>
<th>Network Structure Assumptions</th>
<th>Sampling Assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Walk Model</td>
<td></td>
</tr>
<tr>
<td>Network size large $(N \gg n)$</td>
<td>Sampling with replacement</td>
</tr>
<tr>
<td></td>
<td>Single non-branching chain</td>
</tr>
<tr>
<td>Remove Seed Dependence</td>
<td>Homophily weak enough</td>
</tr>
<tr>
<td></td>
<td>Connected graph</td>
</tr>
<tr>
<td>To Estimate Probabilities</td>
<td>All ties reciprocated</td>
</tr>
<tr>
<td></td>
<td>Degree accurately measured</td>
</tr>
<tr>
<td></td>
<td>Random referral</td>
</tr>
<tr>
<td>Additional Assumptions of SS</td>
<td>Known network size $N$</td>
</tr>
<tr>
<td></td>
<td>No seed bias</td>
</tr>
<tr>
<td>Additional Assumptions of MA</td>
<td>Non-random mixing observable</td>
</tr>
<tr>
<td></td>
<td>Network model form</td>
</tr>
<tr>
<td></td>
<td>Sampling model form</td>
</tr>
</tbody>
</table>

Assumptions of Model-Assisted Estimator
Discussion: Model-Assisted Estimator

- Sampling probabilities based on degrees, finite population effects, seeds, homophily
- Natural framework for bootstrap standard error estimation
- Extensions:
  - Measurable aspects of Network (neighborhoods, perhaps clustering)
  - Measurable aspects of Sampling Process (differential recruitment, biased referral)
  - Inference for other features of simulated population
- Improved computational efficiency.
Discussion: Hidden Population Sampling

Hidden Population Sampling

- Still many assumptions, high variance.
- Typically, RDS not advisable if alternatives available.
- RDS used in varied populations:
  - recent immigrants, unregulated workers, Nigerian rioters.

Network Sampling (link-tracing)

- Two main challenges: non-random seeds, unknown population size.

Social Network Analysis

- Here, network used for sampling, nuisance for estimation.
  - Often, it is of independent interest.
- First fitting of network model to data with initial convenience sample.
References:


Thank You!