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.

Link-Tracing Sampling

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:

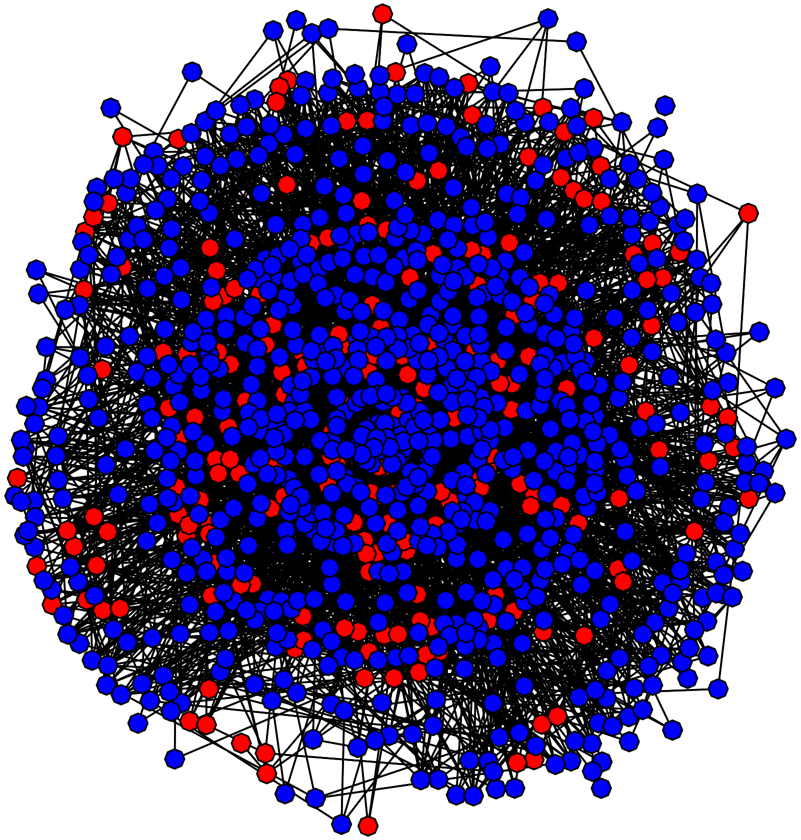
- Seed Dependence: Follow only a few links from each sampled
- Confidentiality: Respondents distribute uniquely identified coupons. No names. (*respondent-driven*)
- Estimation based on Network positions: Several approaches

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

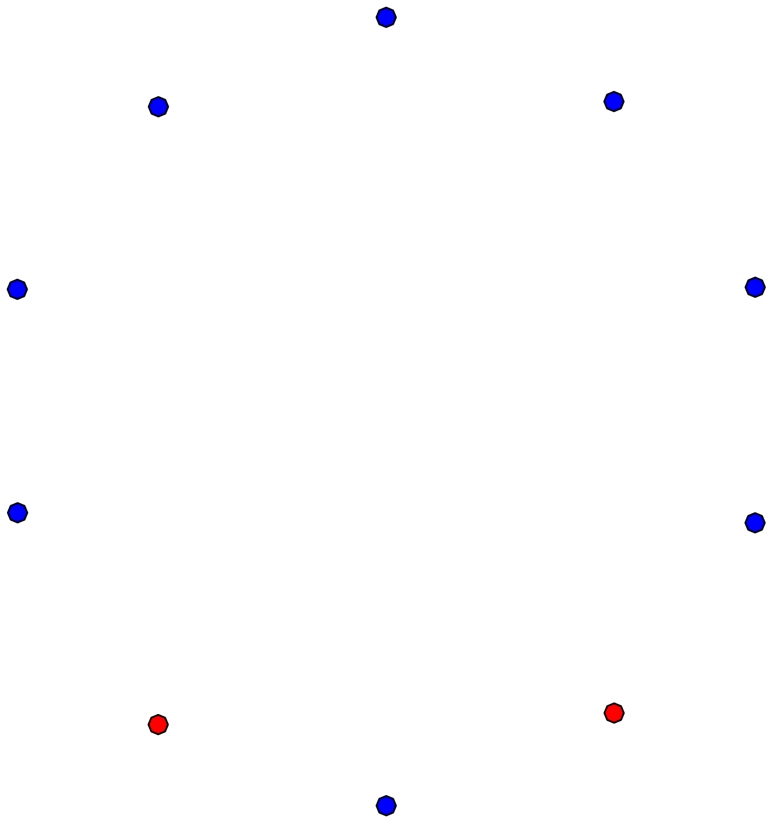
Heckathorn, D.D., "Respondent-driven sampling: A new approach to the study of hidden populations." *Social Problems*, 1997.

Salganik, M.J. and D.D. Heckathorn, "Sampling and estimation in hidden populations using respondent-driven sampling." *Sociological Methodology,* 2004.

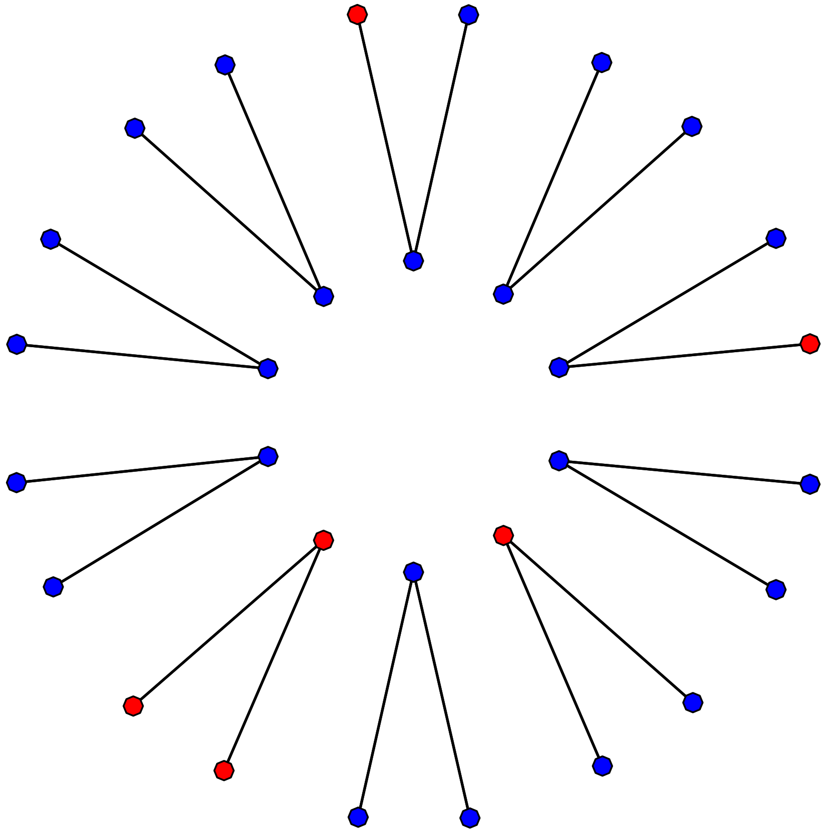
Stylized population



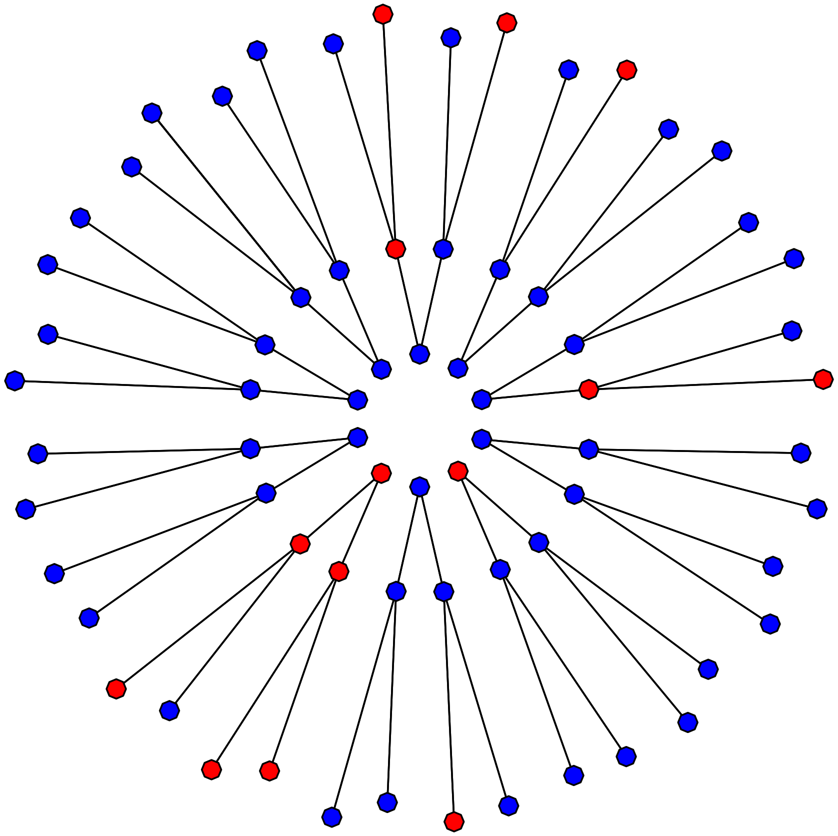
Start with seeds . . .



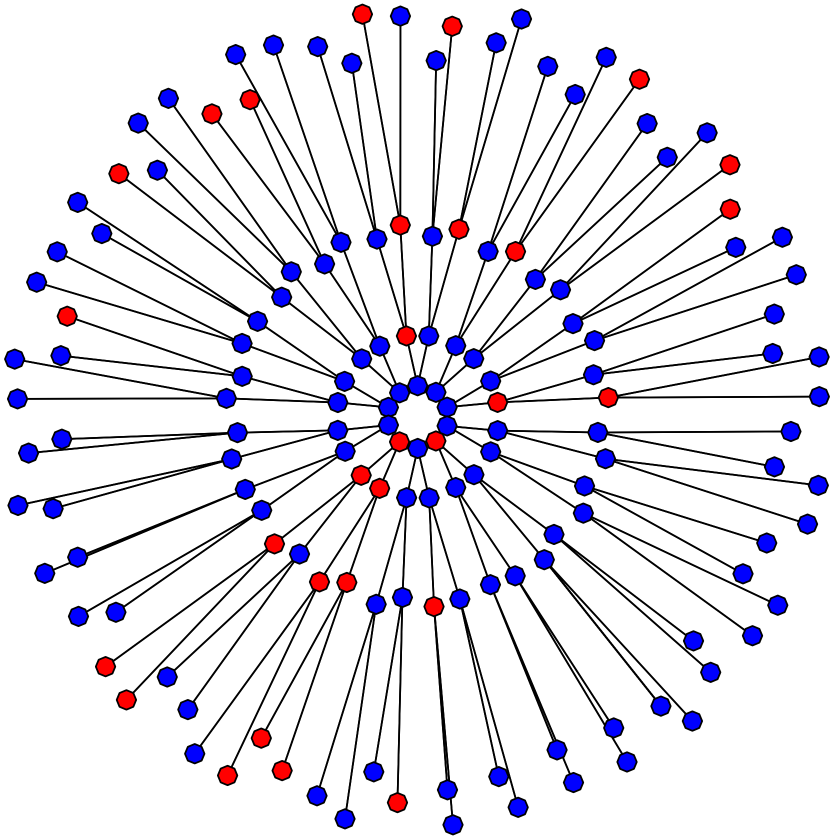
Seeds recruit the first wave . . .

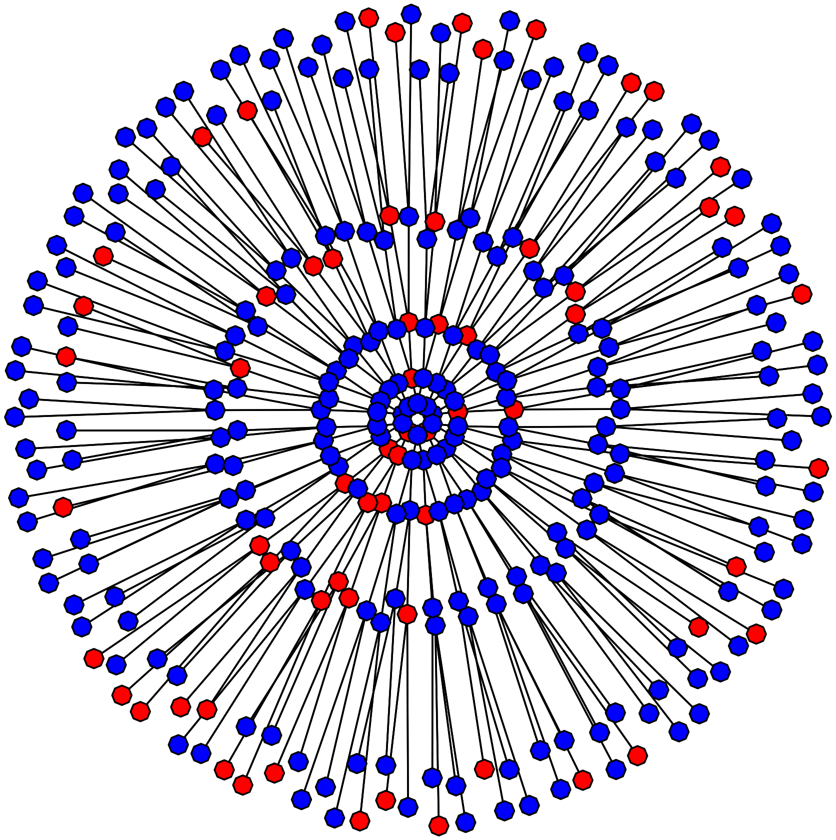


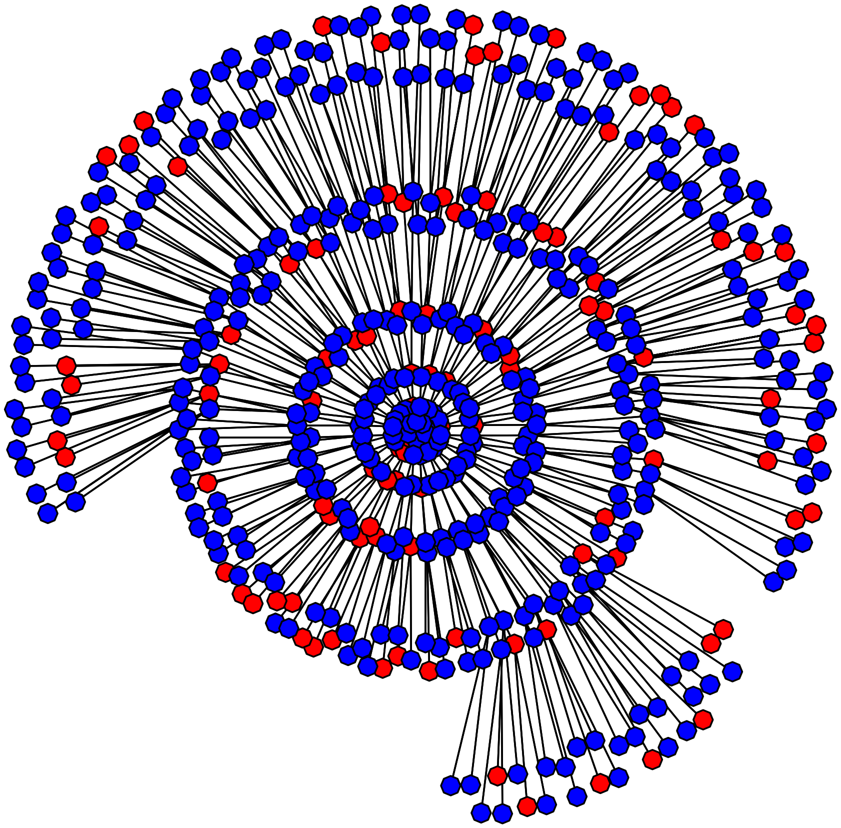
First wave recruit the second wave . . .



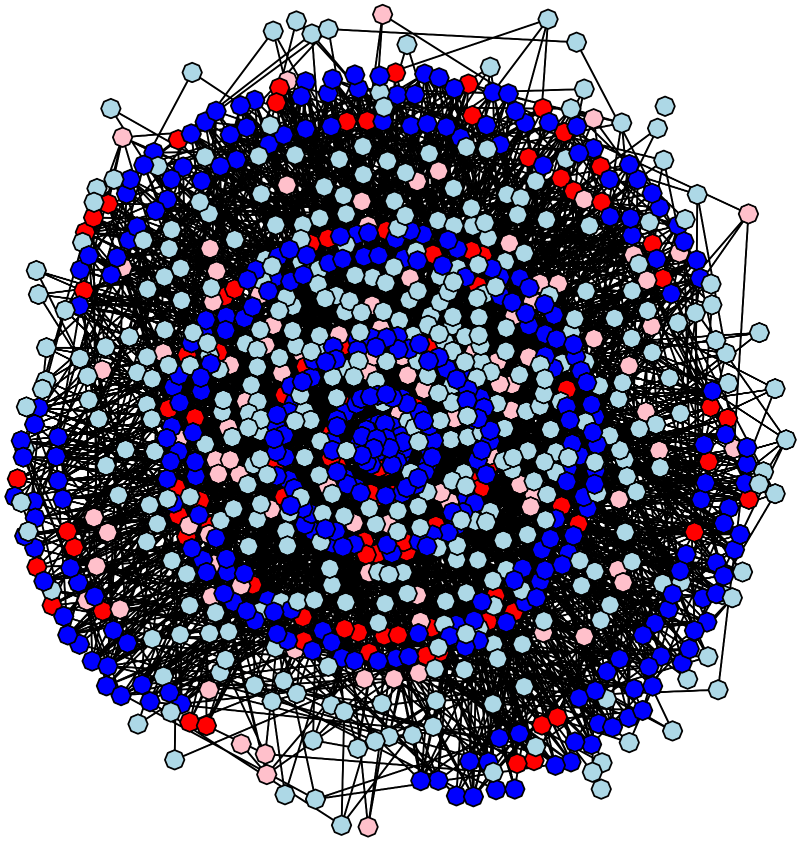
and so on . . .

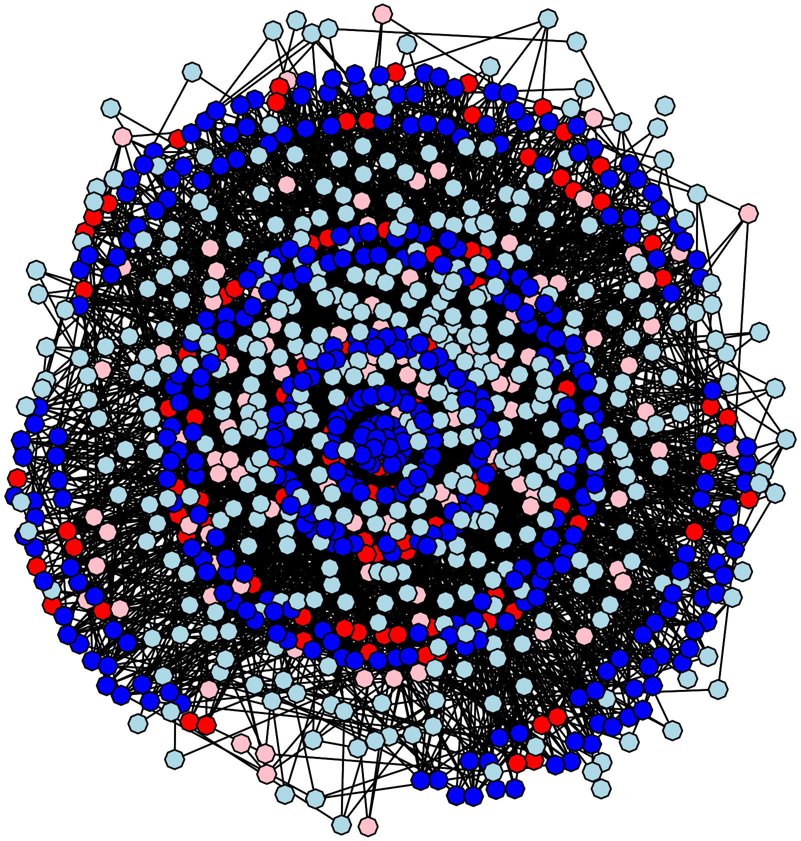




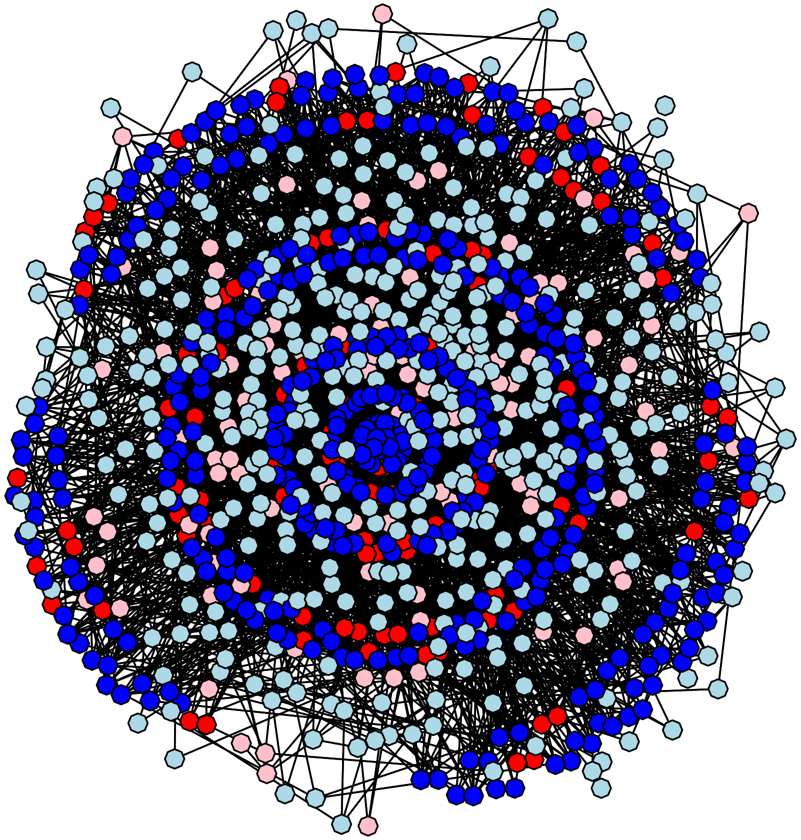


(and with un-sampled)





degree of node i = # of ties of node i



$$\textit{homophily} = \frac{\text{Percent realized infected to infected ties}}{\text{Percent realized uninfected to infected tie}}$$

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 $1, 2, \ldots N$,

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

- 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

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

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.

Salganik, M.J., and D.D. Heckathorn, "Sampling and estimation in hidden populations using respondent-driven sampling." *Sociological Methodology,* 2004.

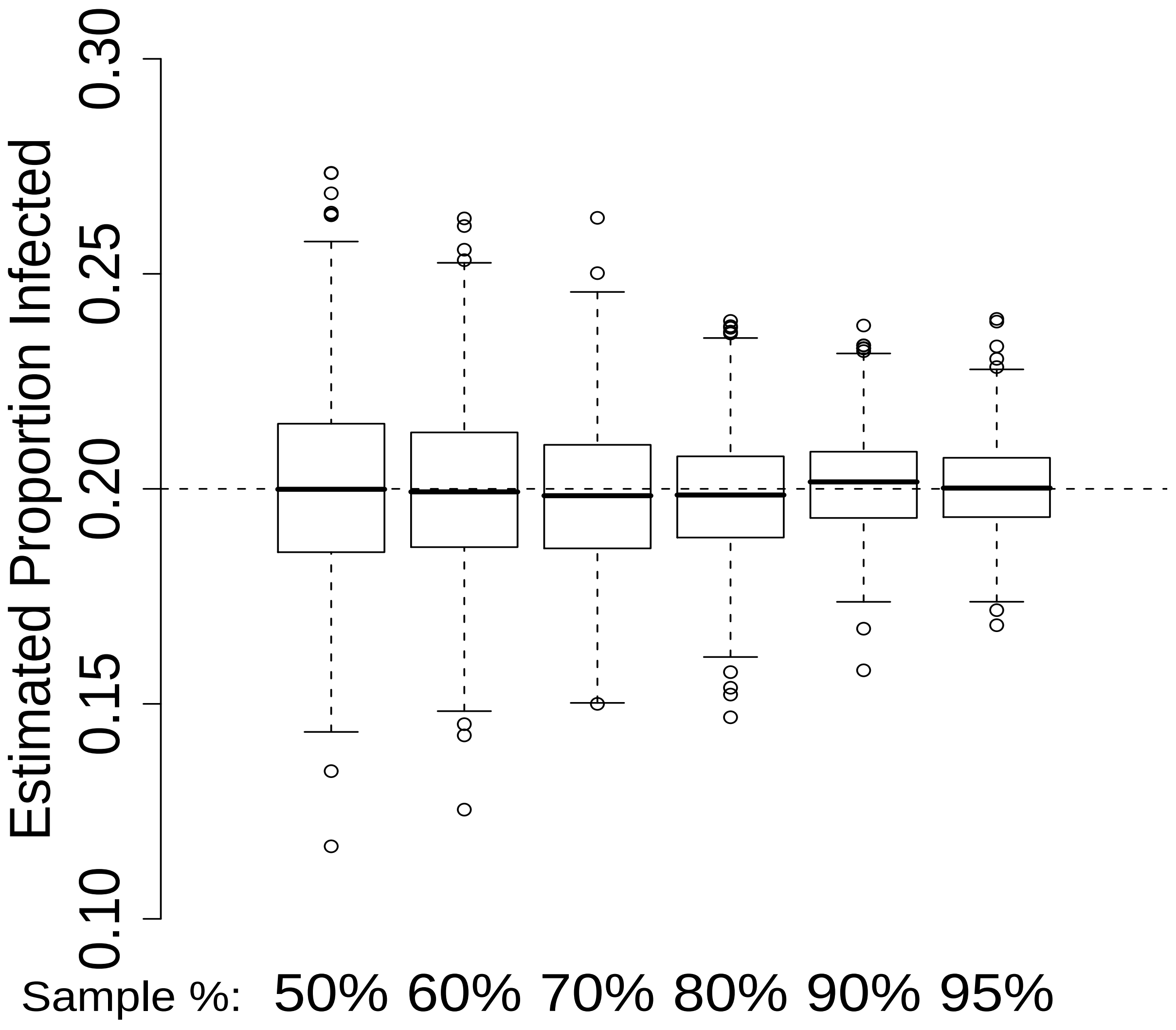
Volz, E., and D.D. Heckathorn, "Probability Estimation Theory for Respondent Driven Sampling," *Journal of Official Statistics,* 2008.

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

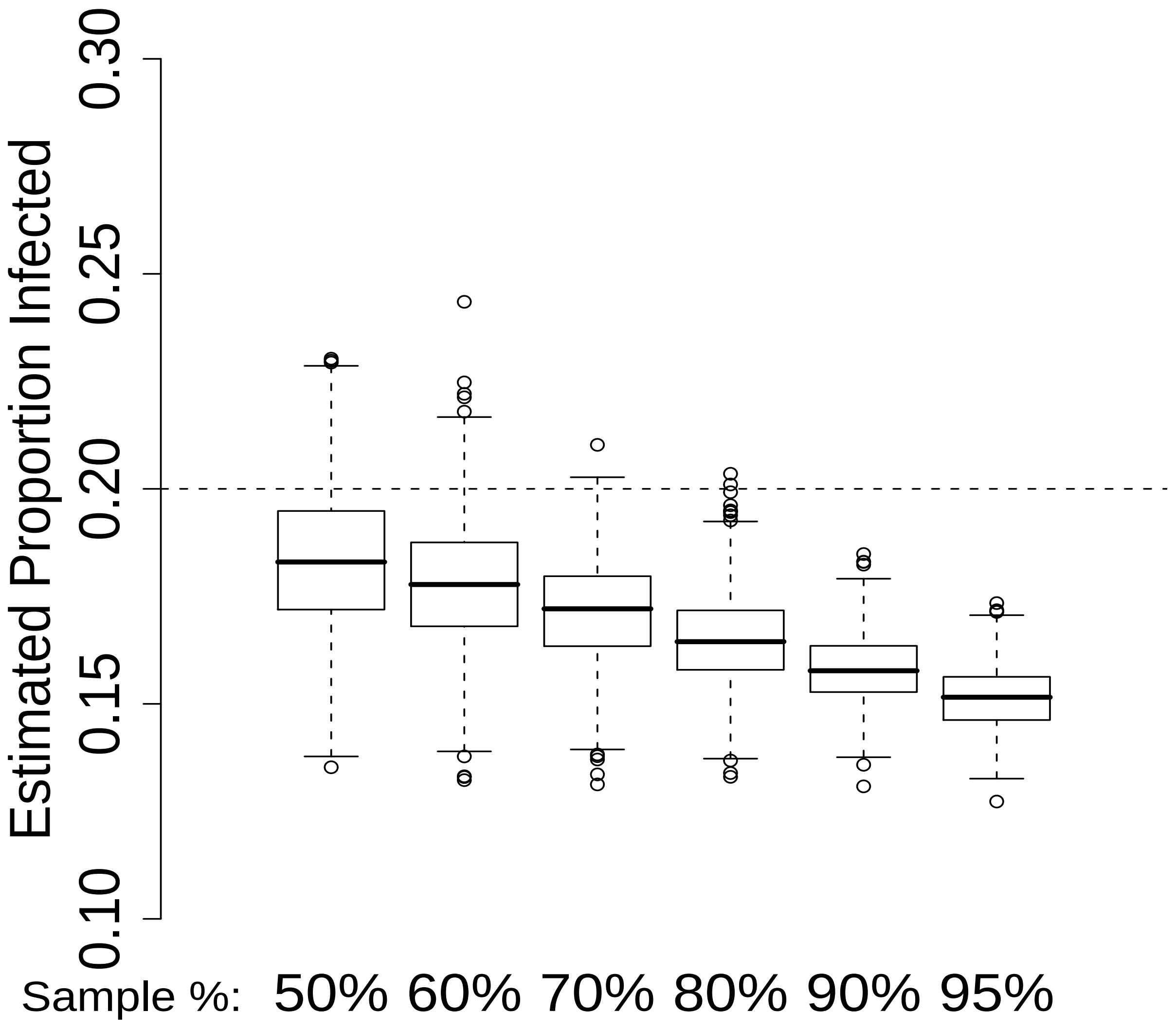
$$\hat{\mu} = \frac{\sum_i S_i \frac{z_i}{d_i}}{\sum_i S_i \frac{1}{d_i}}$$

where $d_i =$ degree of node i, S_i sample indicator, z_i quantity of interest.

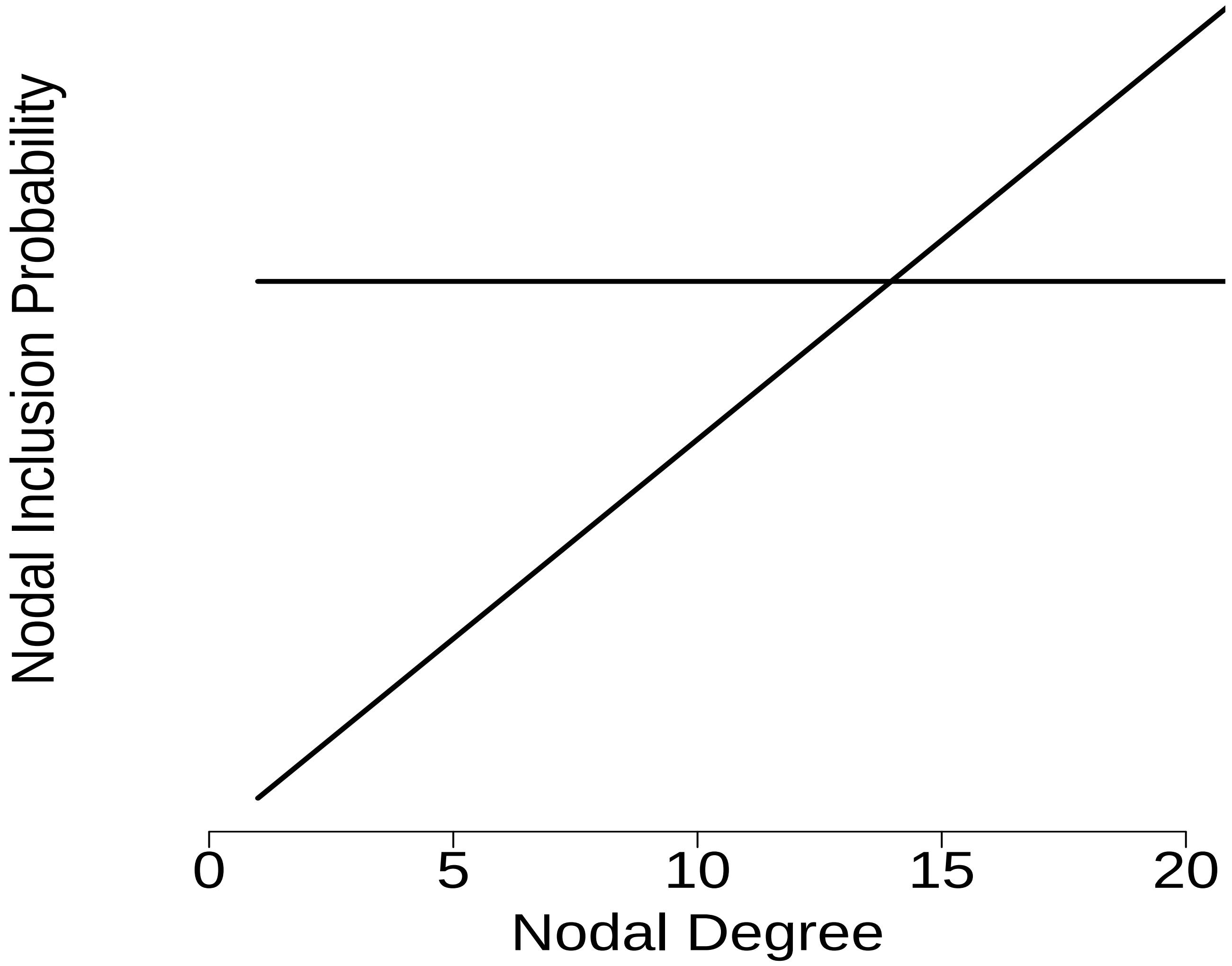
Volz-Heckathorn, w=1



Varying Sample Percentage, w=1.4



Finite Population Bias



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

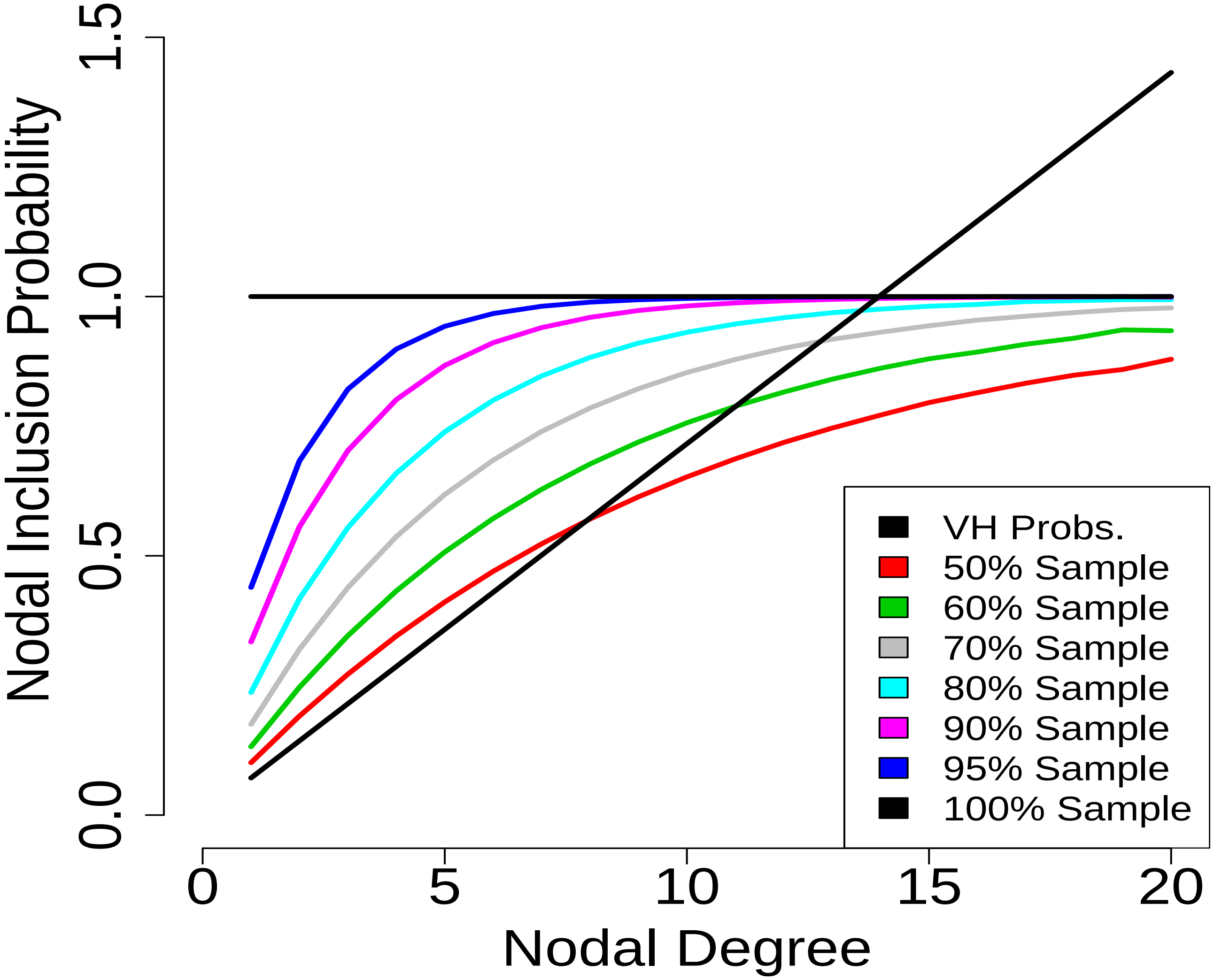
Furthermore, consider:

- 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*

Successive Sampling Mapping



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

Gile, K.J. “Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation,” *Journal of the American Statistical Association,* 2011.

Successive Sampling (SS) Estimator: Algorithm

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

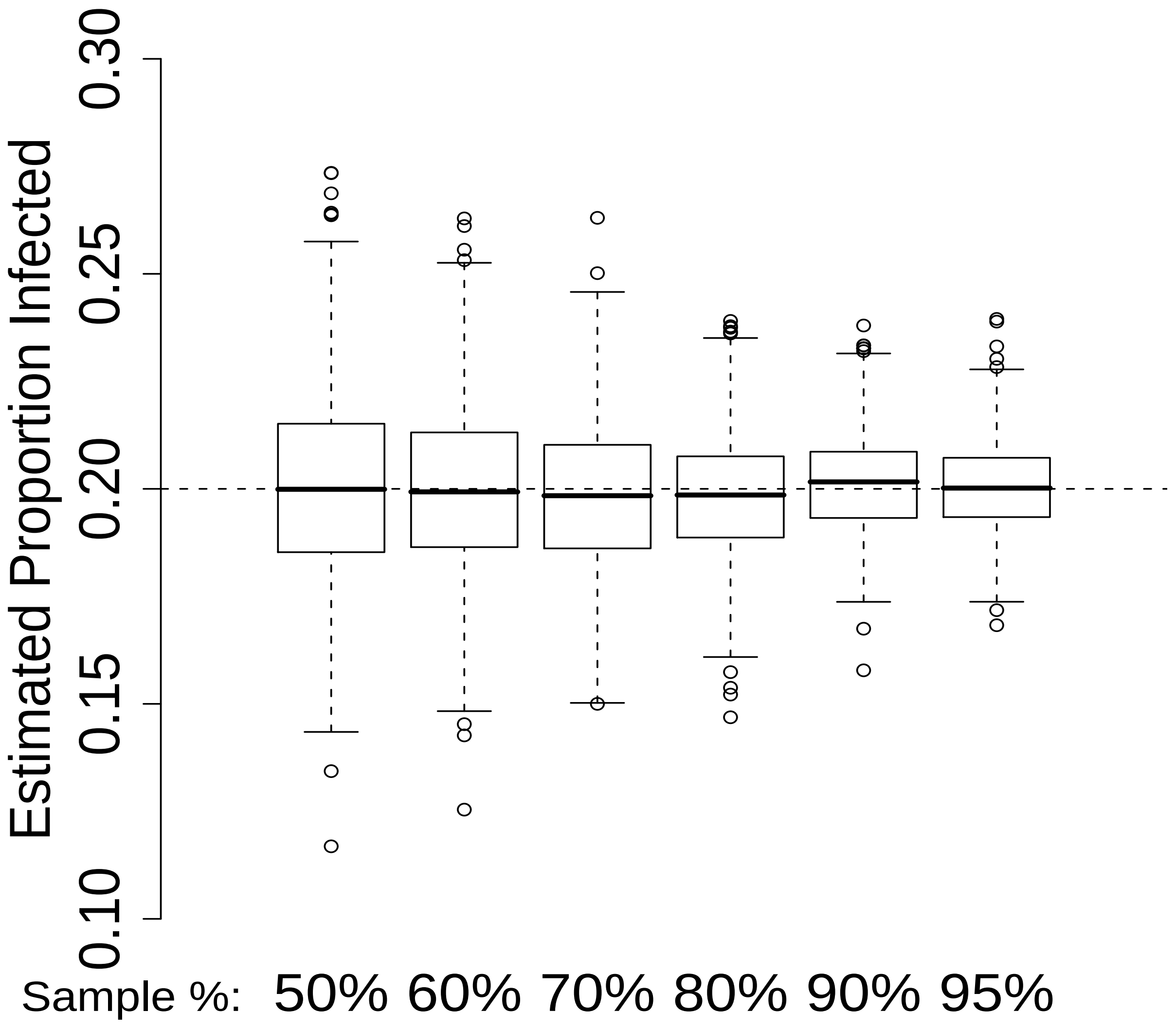
1. Initial: $\pi_k(\mathbb{N}^0) \propto k$.
2. For $i = 1 \ldots r$:
   (a) Estimate degree distribution $\mathbb{N}^i$ by Generalized Horvitz-Thompson Estimator
   (b) Compute $\pi_k(\mathbb{N}^i)$ by simulation:
      i. Simulate M SS samples from $\mathbb{N}^i$
      ii.
$$\pi_k(\mathbb{N}^i) = \frac{\mathbb{E}[V_k;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

Standard Error Estimation:

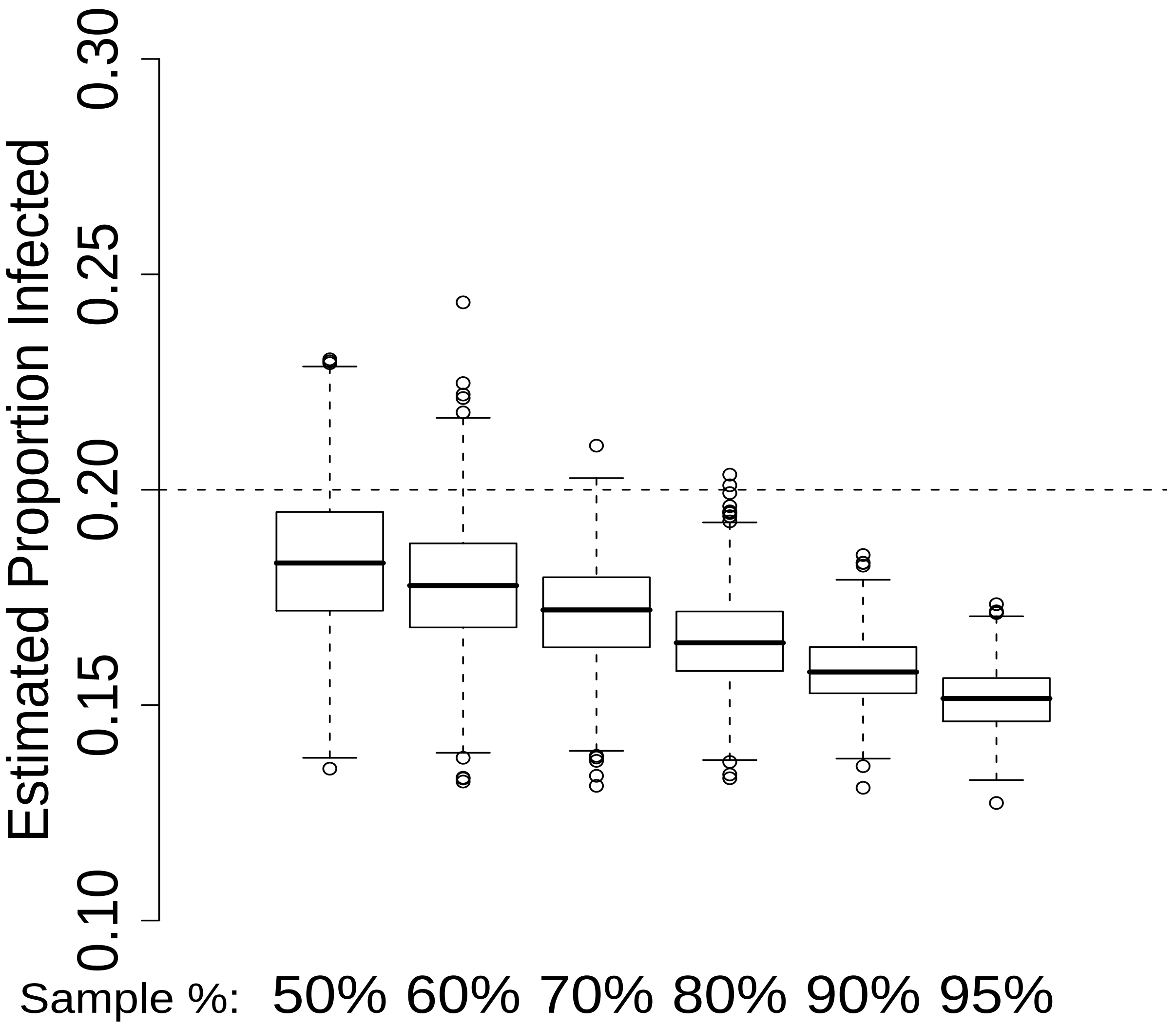
Population Bootstrap:

- 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.

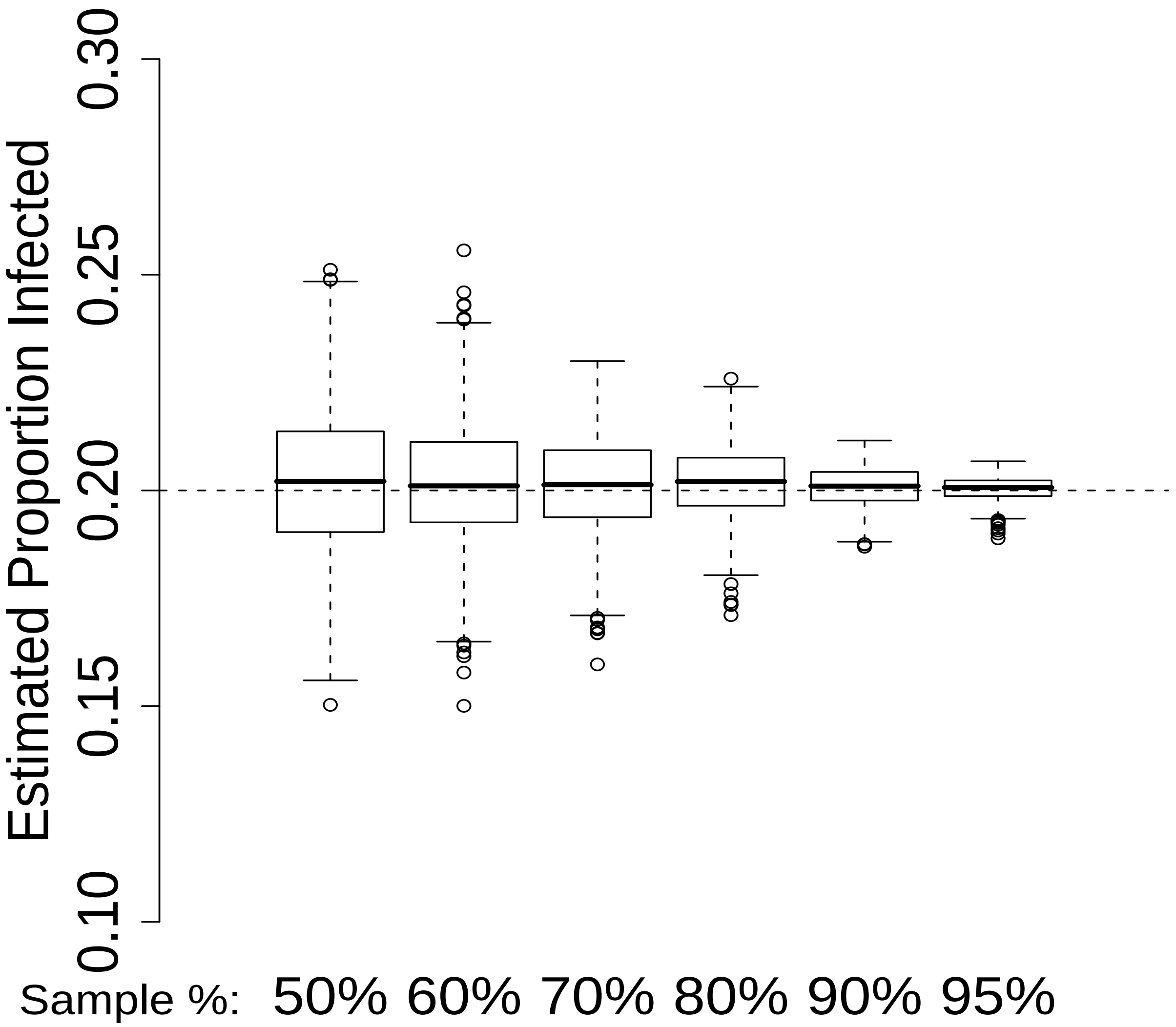
Volz-Heckathorn, w=1



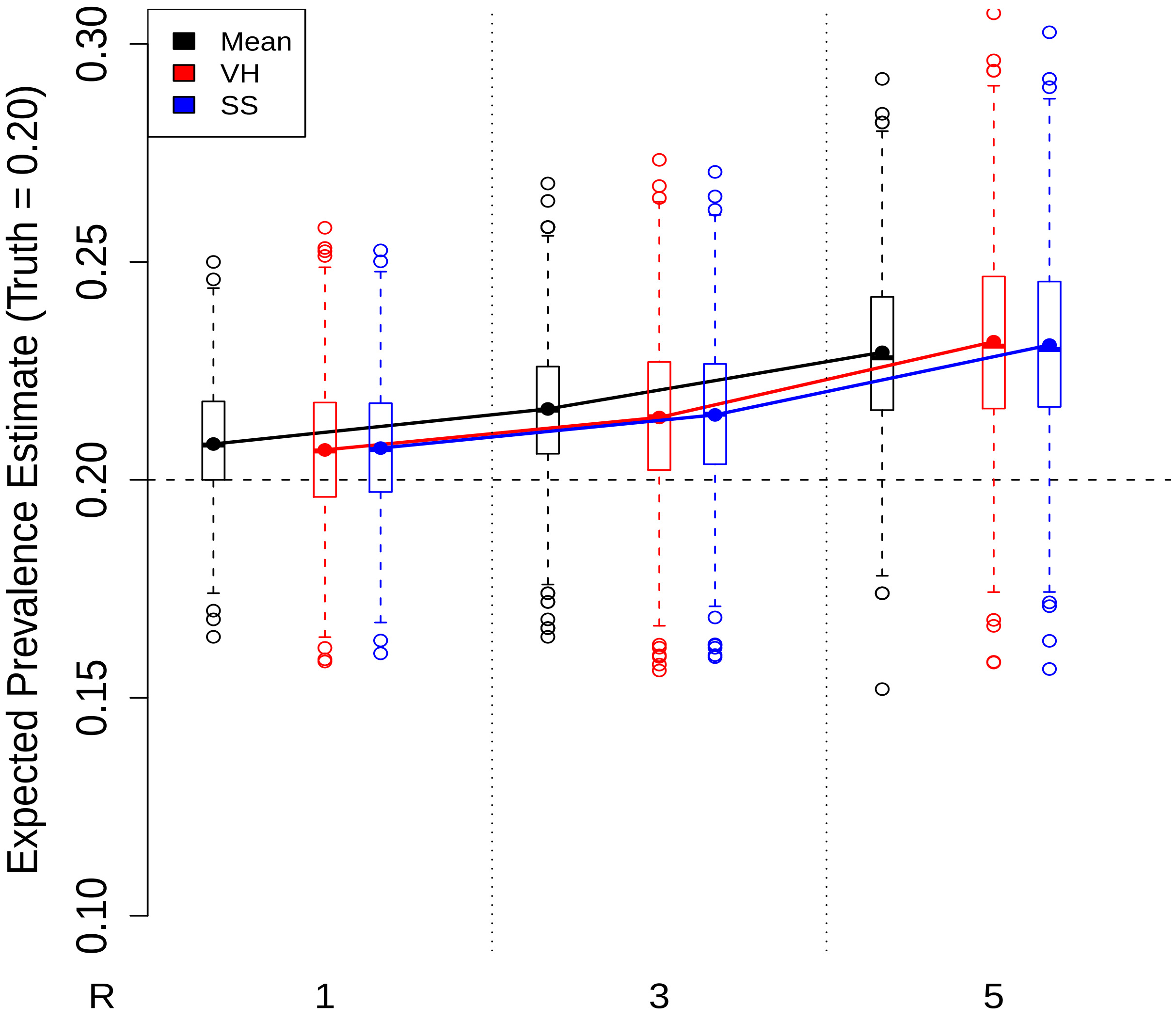
Volz-Heckathorn, w=1.4



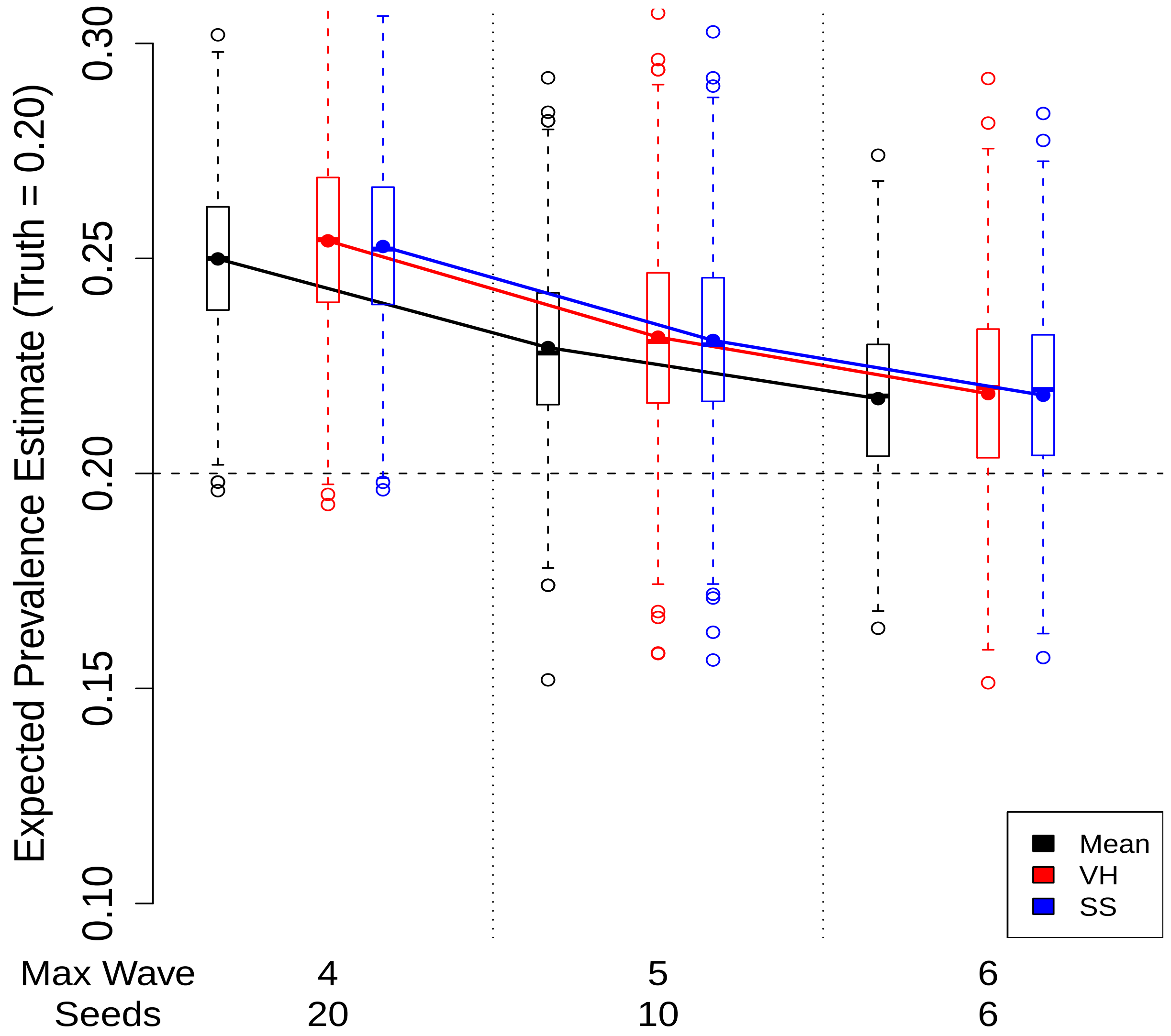
SS, w=1.4


Estimated Proportion Infected
0.10
0.15
0.20
0.25
0.30
Sample %:
50%
60%
70%
80%
90%
95%

All Infected Seeds, varying Homophily, 50%



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, η, $\pi_i = \sum_{y \in \mathscr{Y}} P(y|\eta)\mathbb{E}(S_i|y, s)$

We do not know y or η. So we estimate η.

Exponential Random Graph Model

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

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

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

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

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

Here,

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

Require:

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

Fitting the Model

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

Solution: Use design-based estimators

$$\hat{\mathbb{N}}_{kl} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{S}_i \mathbb{I}(\mathbf{d}_i = k, z_i = l)}{\hat{\pi}_i}$$

$$\hat{m}(\eta) = \sum_{i=1}^{N} \frac{\mathbf{S}_i \left(\mathbf{x}_i(1 - z_i) + (\mathbf{d}_i - \mathbf{x}_i)z_i\right)}{2\hat{\pi}_i}$$

where $\mathbf{x}_i = \sum_j z_j y_{ij}$ requires the observation of $\mathbf{x}_i \ \forall \ i : \mathbf{S}_i = 1$.
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 (π_i) by degree d_i and infection z_i.
- A function of homophily (η), and population of degrees and infection $\mathbb{N}$.

- Estimate $\hat{\pi}_i$ proportional to degree d_i.
- Iterate the following steps:
  - Estimate $\mathbb{N}$ and $m(\eta)$ using $\hat{\pi}_i$.
  - Find corresponding model parameter η (`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.

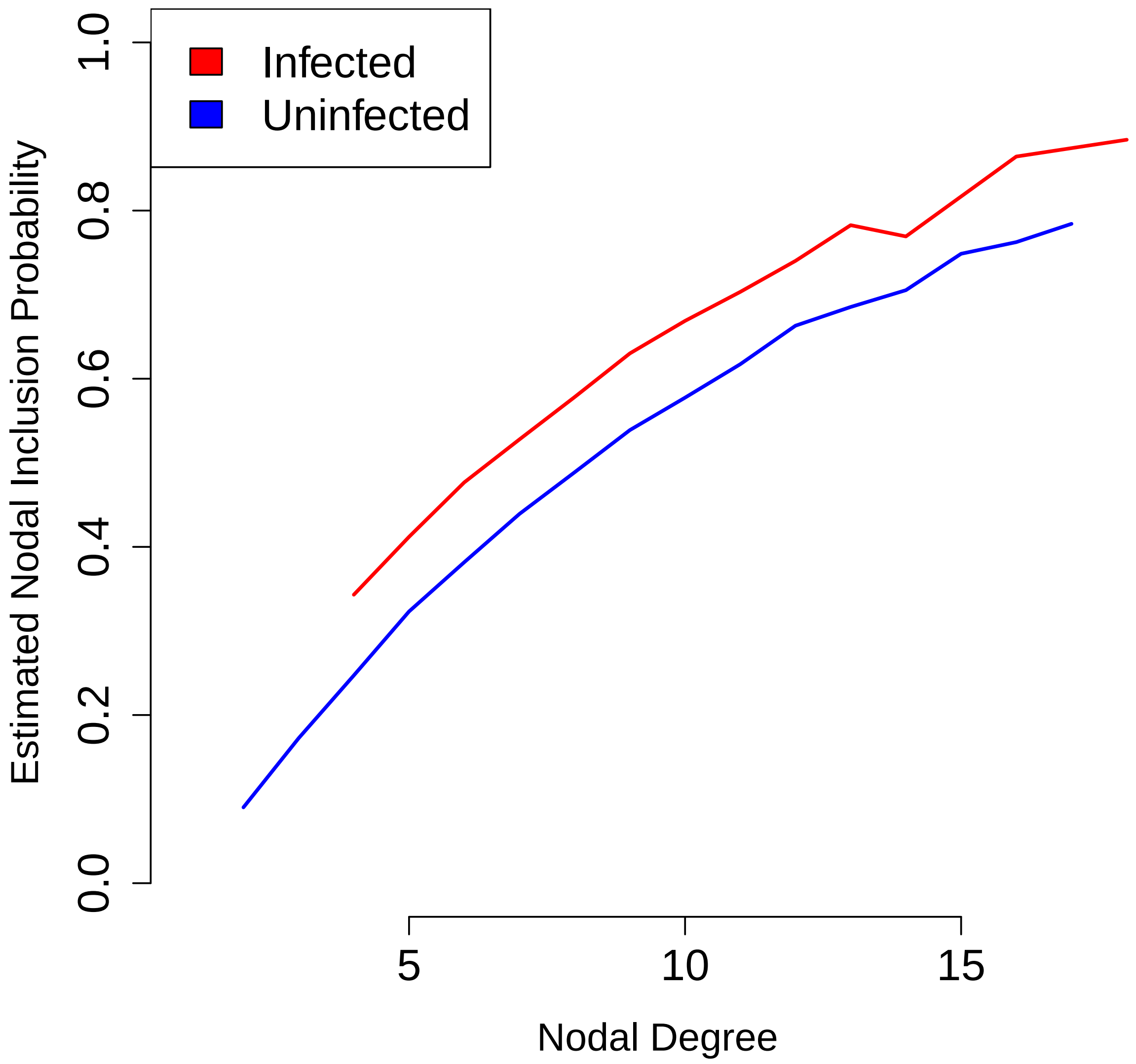
$$\hat{\mu}_{MA} = \frac{\sum_i S_i \frac{z_i}{\hat{\pi}_i}}{\sum_i S_i \frac{1}{\hat{\pi}_i}}.$$

Standard Error Estimation

Population Bootstrap:

- Simulate M populations
  - Estimate z by d distribution
  - Estimate η
  - Simulate networks according to η
- 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



Simulation Study

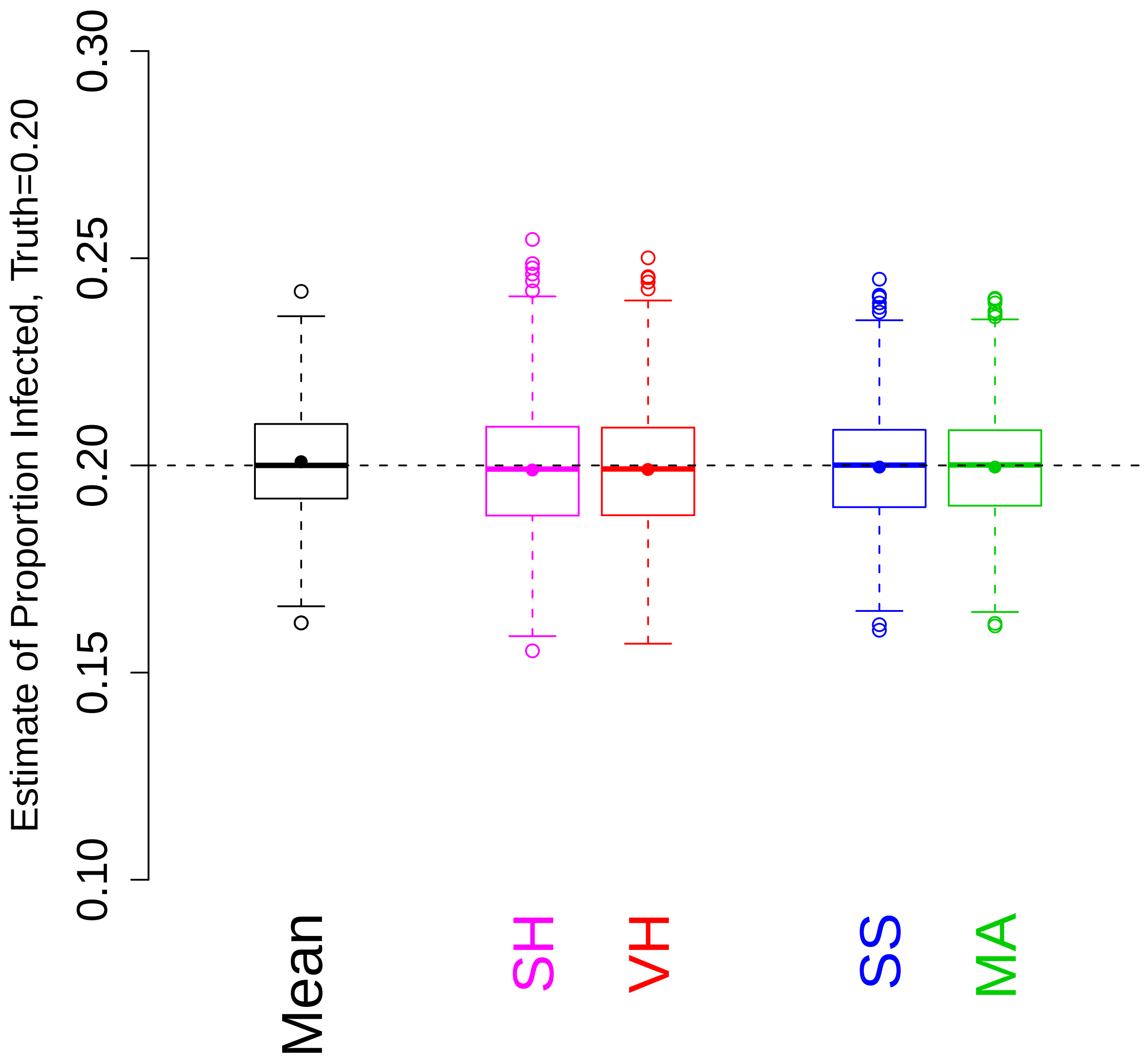
Critical Questions:

- 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?

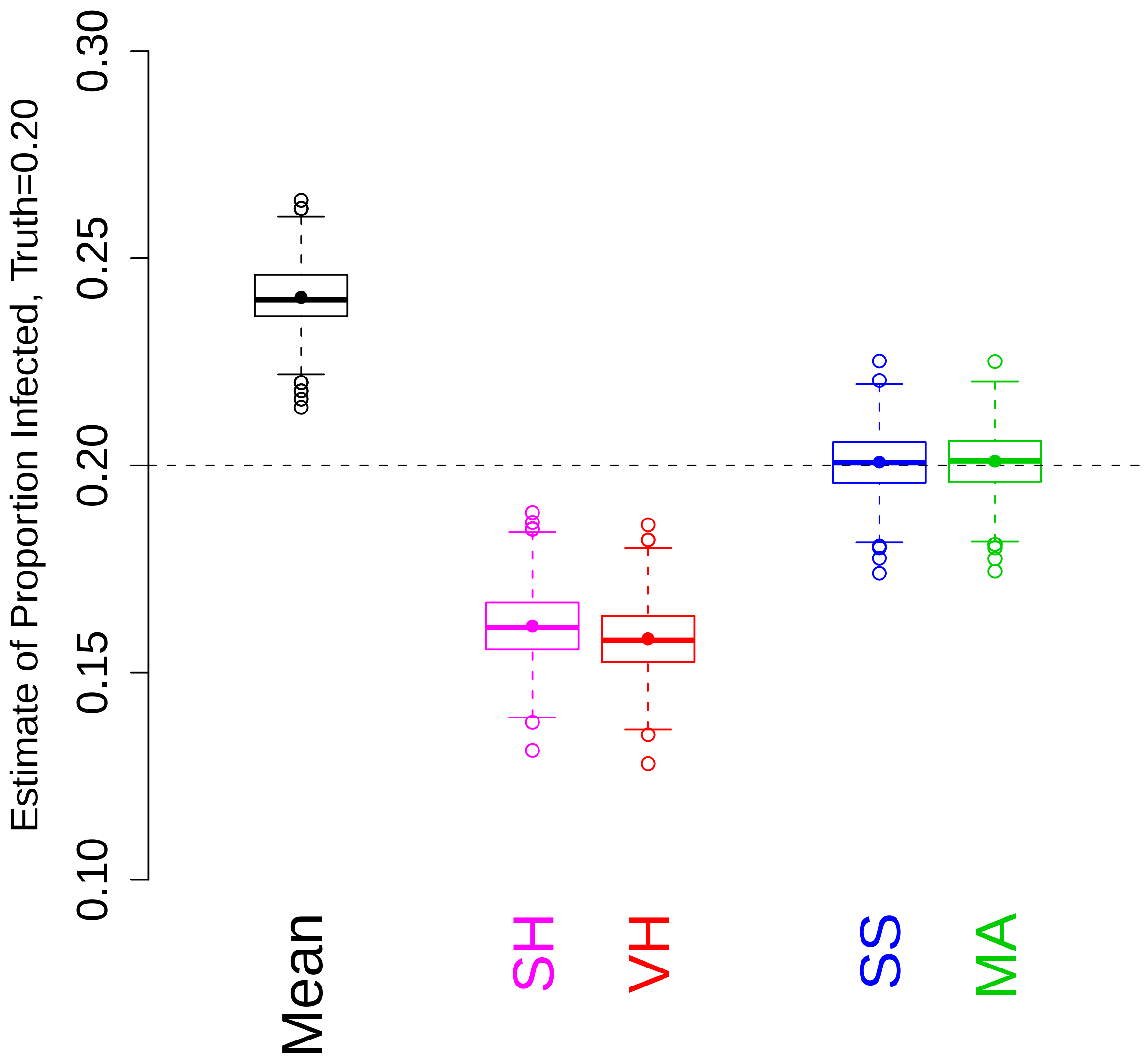
Comparison of Estimators:

- 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

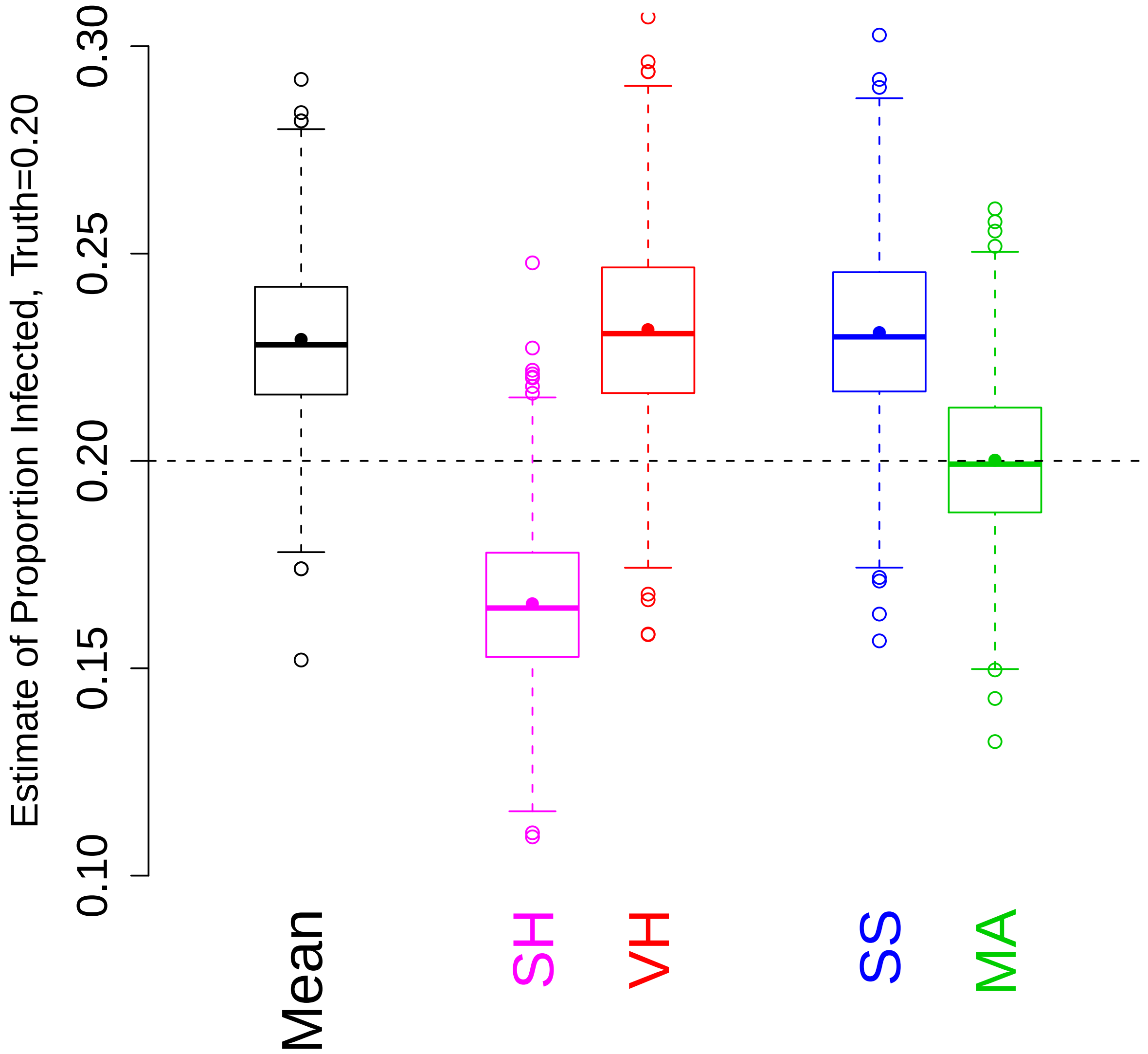
50% **Sample,** $w = 1$**,** $R = 1$**, Random Seeds**



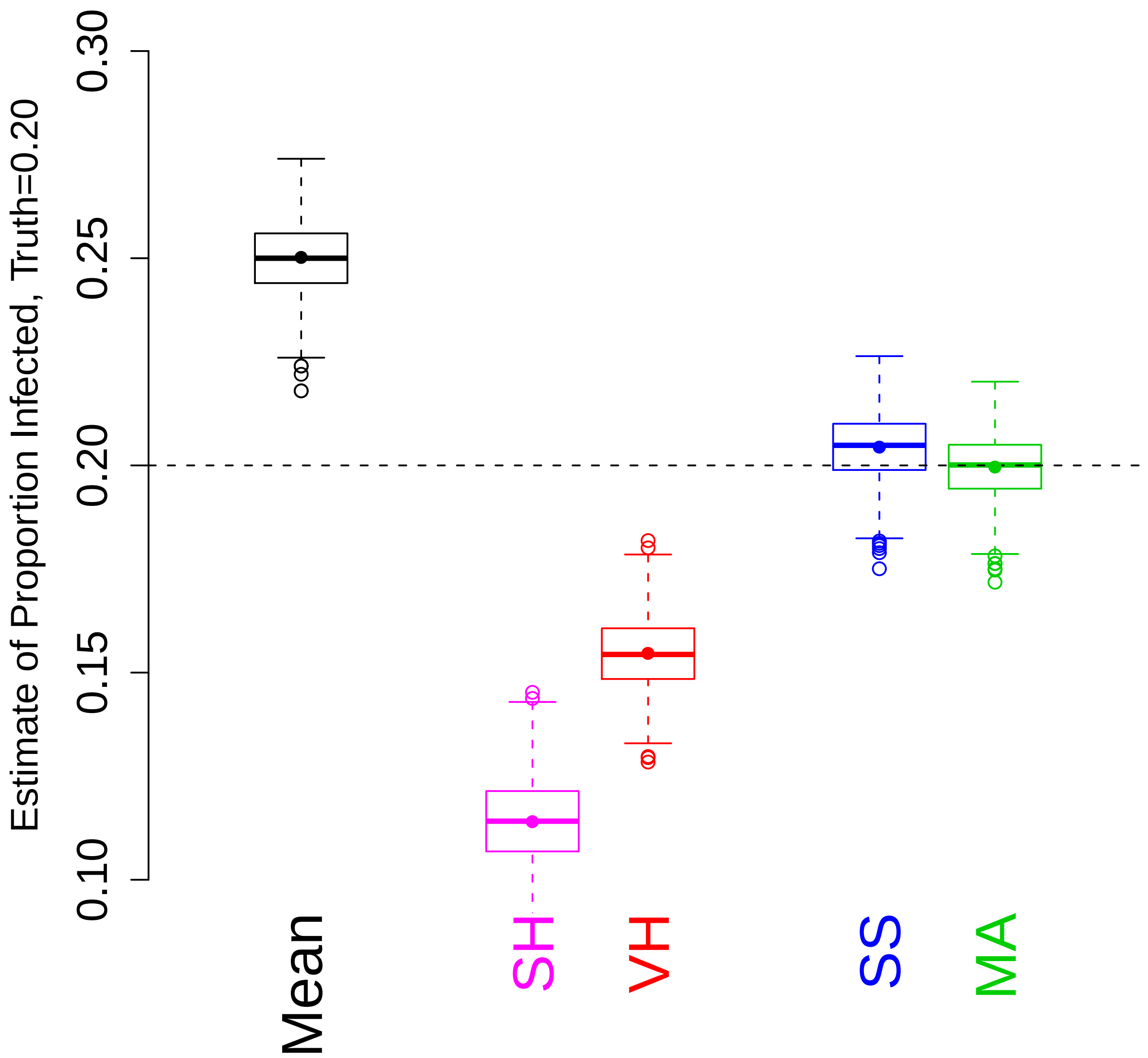
70% Sample, $w = 1.8$, $R = 1$, Random Seeds



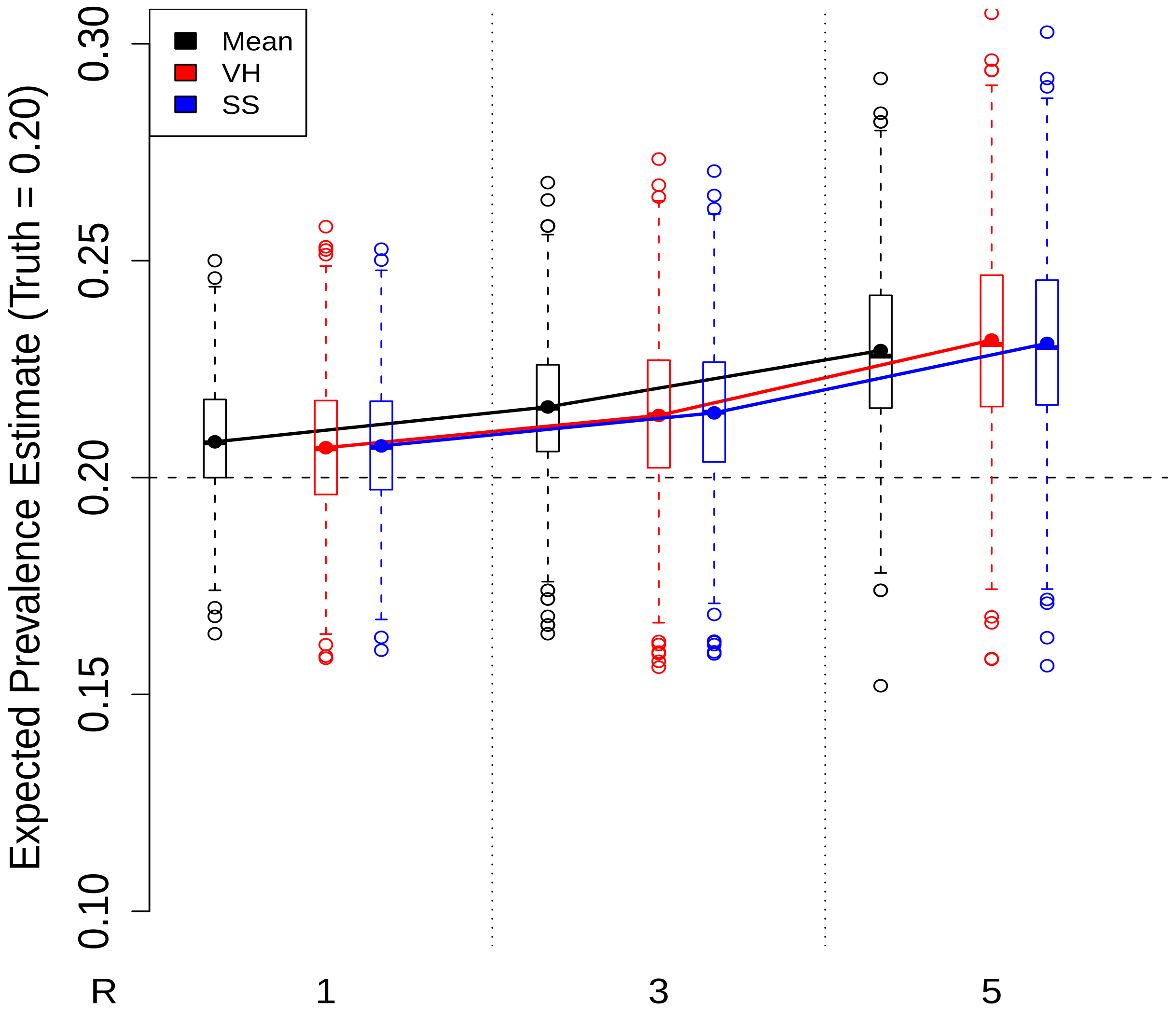
50% Sample, $w = 1$, $R = 5$, Infected Seeds



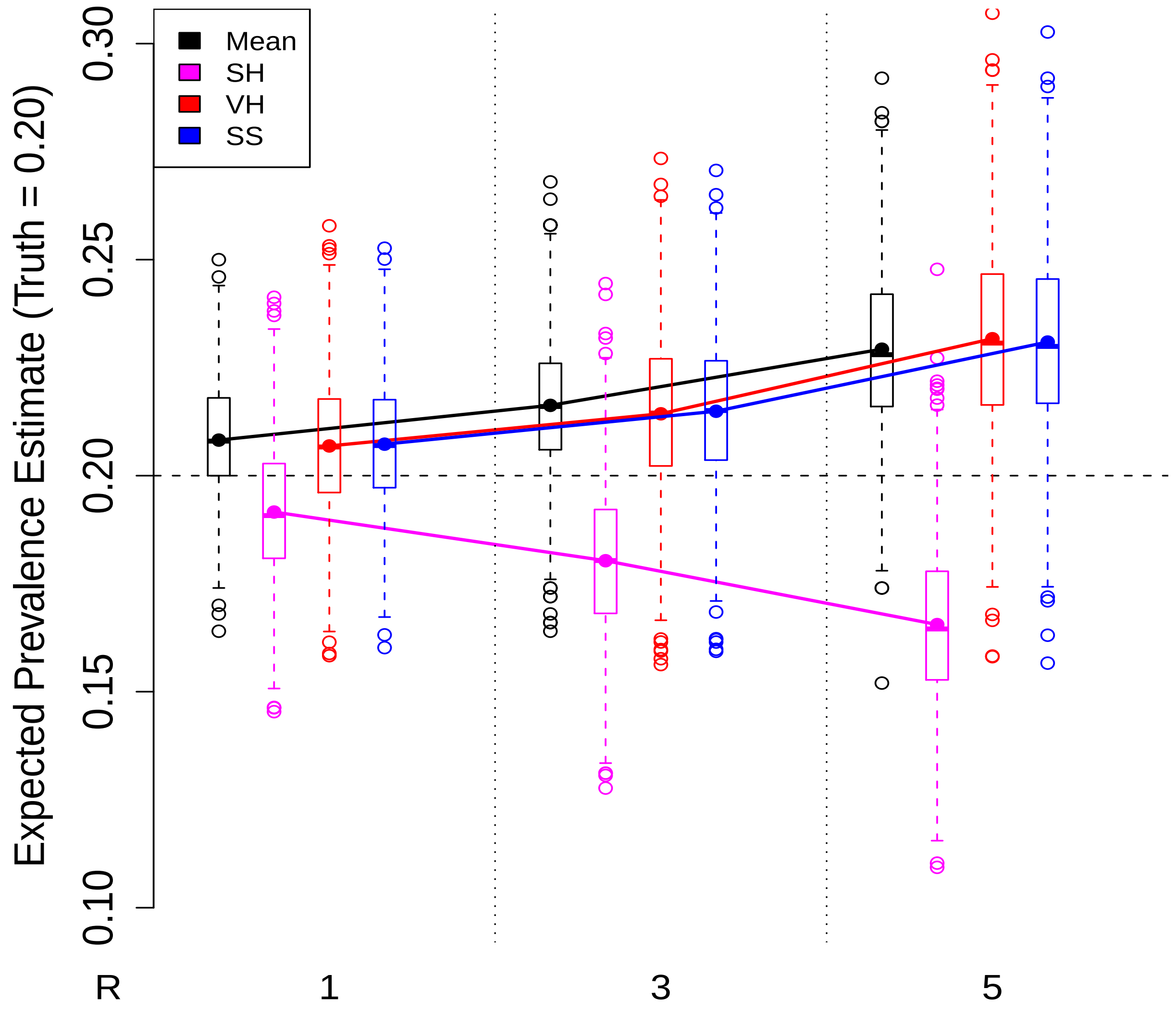
70% Sample, $w = 1.8$, $R = 5$, Infected Seeds



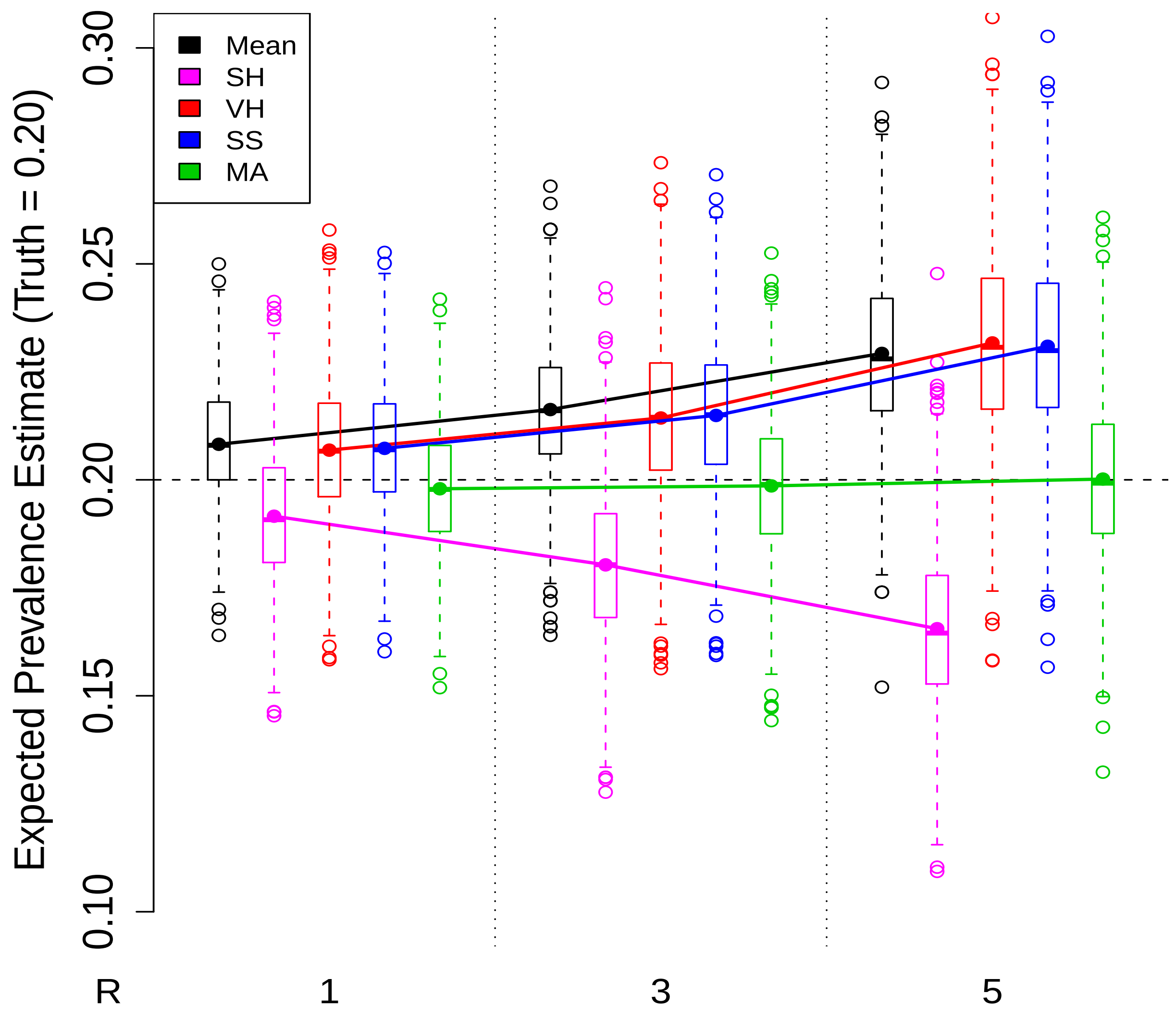
All Infected Seeds, varying Homophily



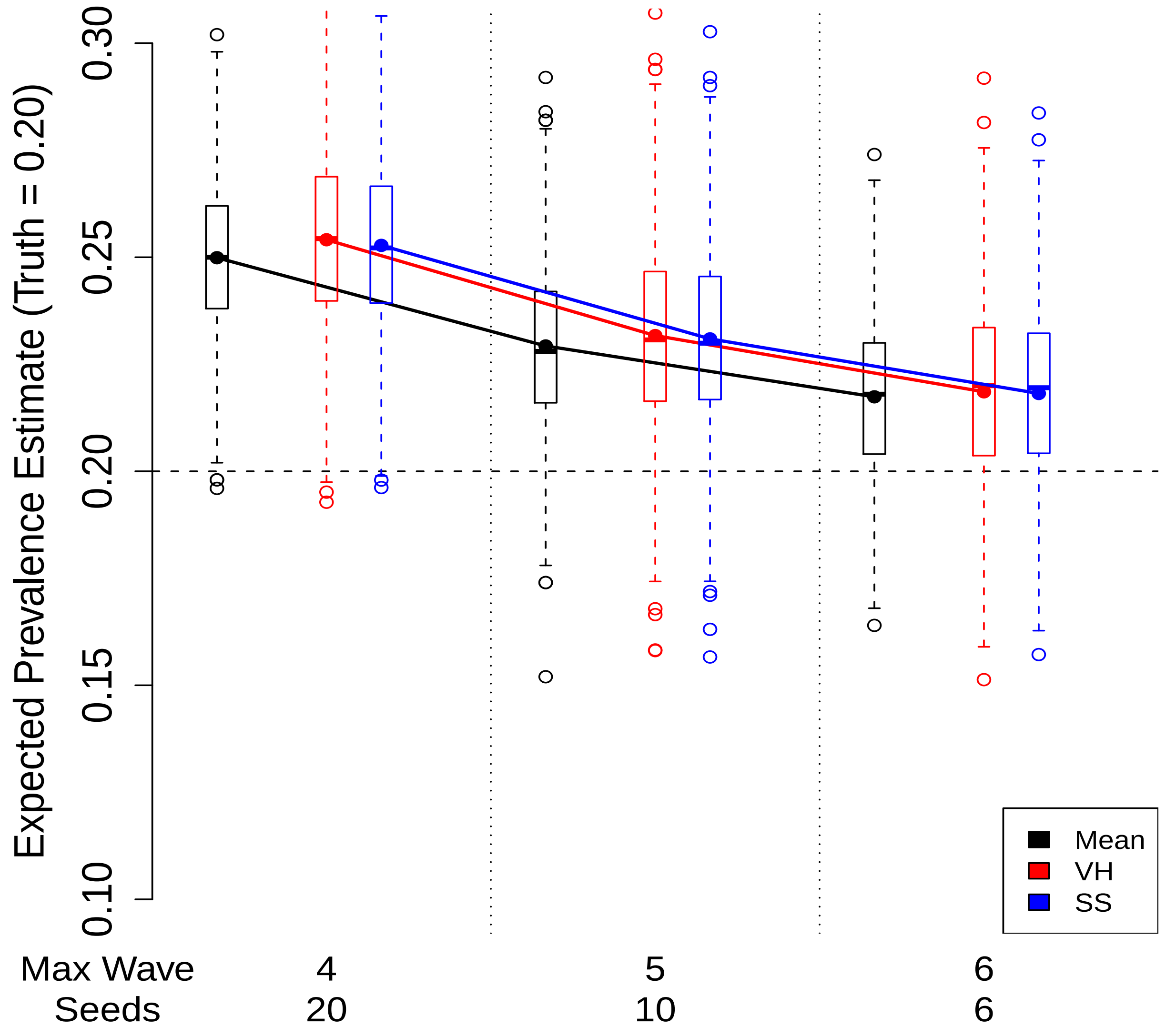
All Infected Seeds, varying Homophily



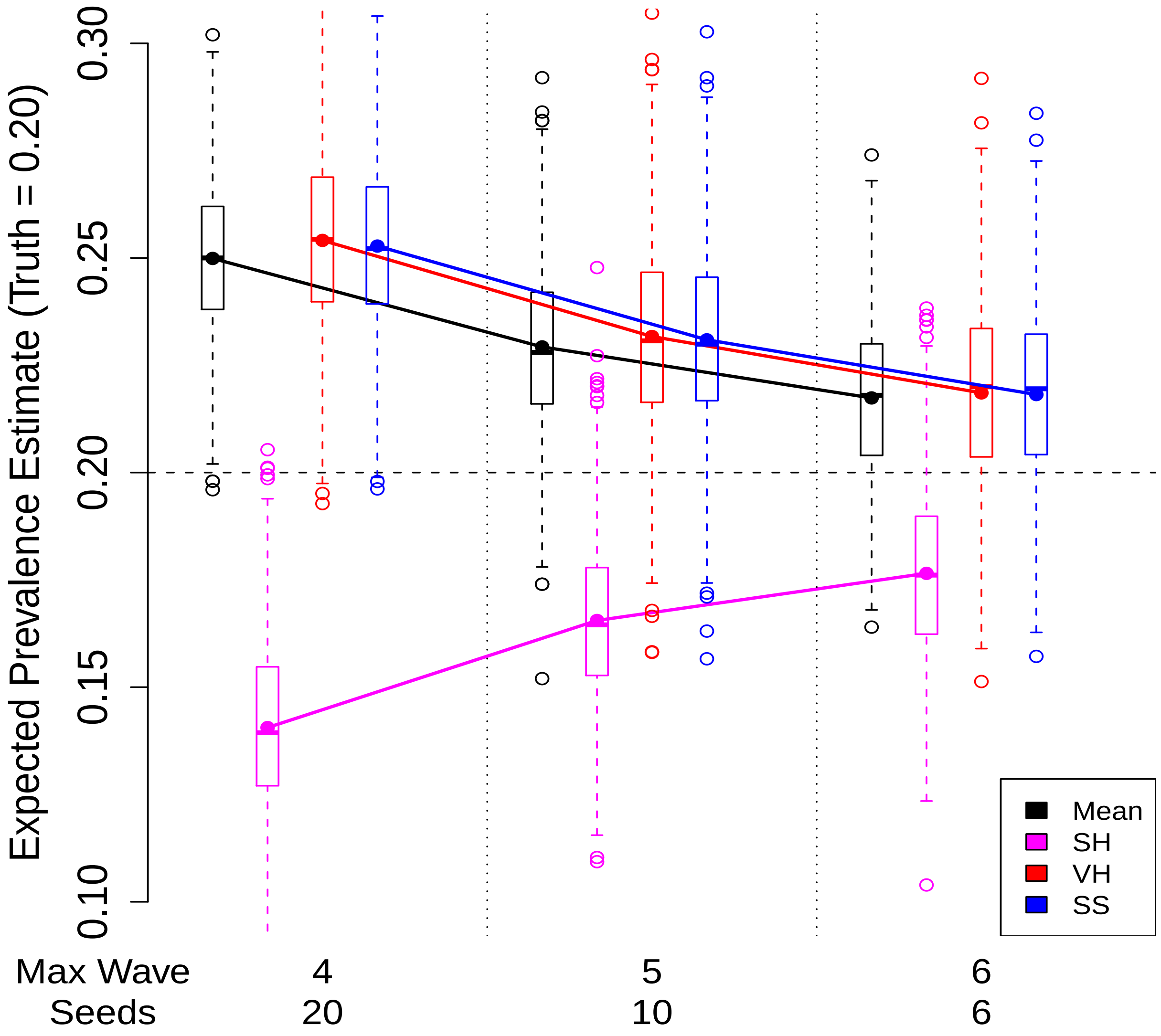
All Infected Seeds, varying Homophily



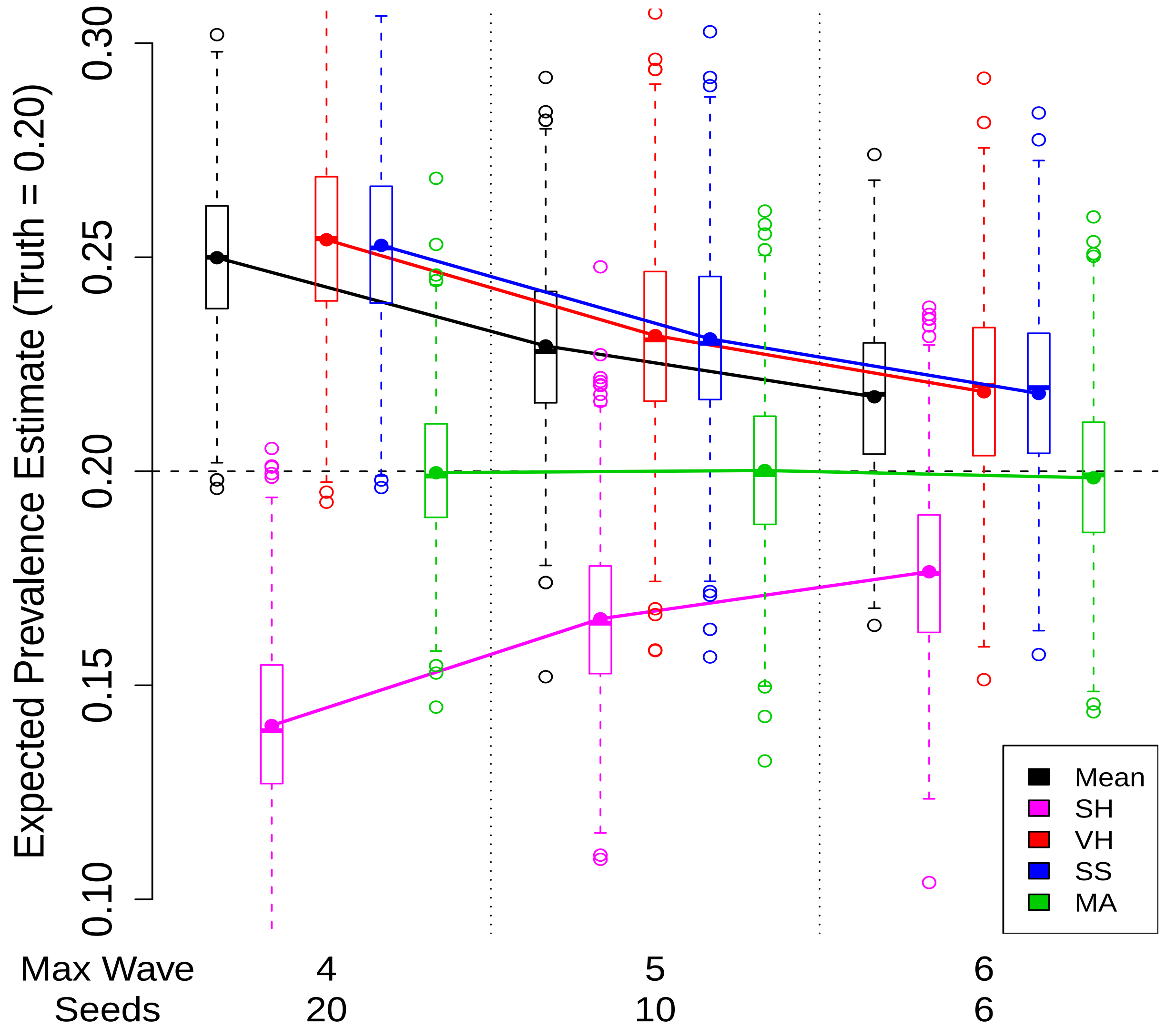
All Infected Seeds, varying number of seeds (waves)



All Infected Seeds, varying number of seeds (waves)



All Infected Seeds, varying number of seeds (waves)



Parametric Bootstrap

% sample	homoph. R	w	sample bias	SE observed	SE bootstrap	coverage 95%	coverage 90%
50%	1	1	No	0.0140	0.0137	94.1%	88.8%
70%	1	1.8	No	0.0073	0.0075	94.9%	90.4%
50%	5	1	Initial	0.0188	0.0175	93.7%	87.9%
50%	5	1.8	Initial	0.0079	0.0080	95.0%	87.3%
50%	5	1	Referral	0.0216	0.0225	91.7%	84.7%

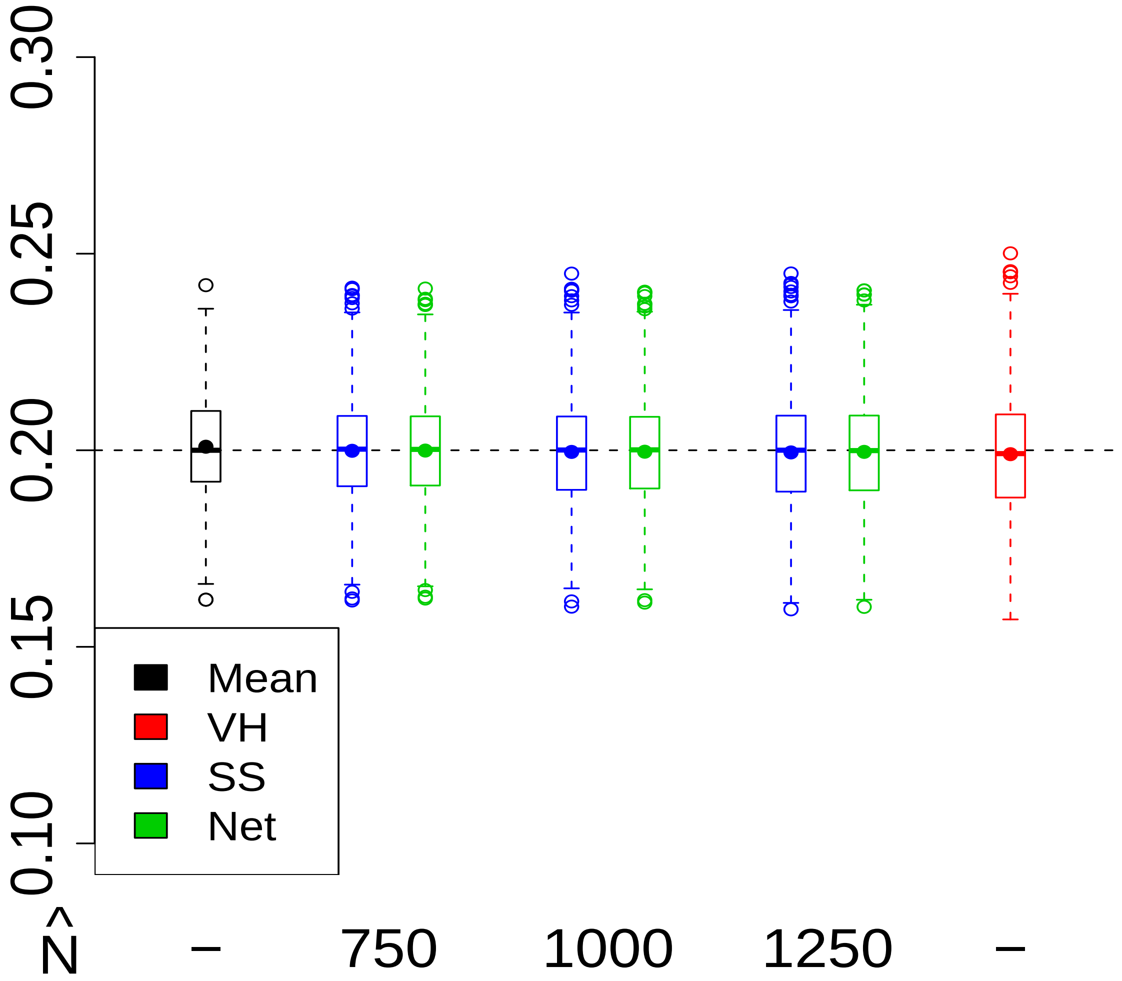
Outline of Presentation

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

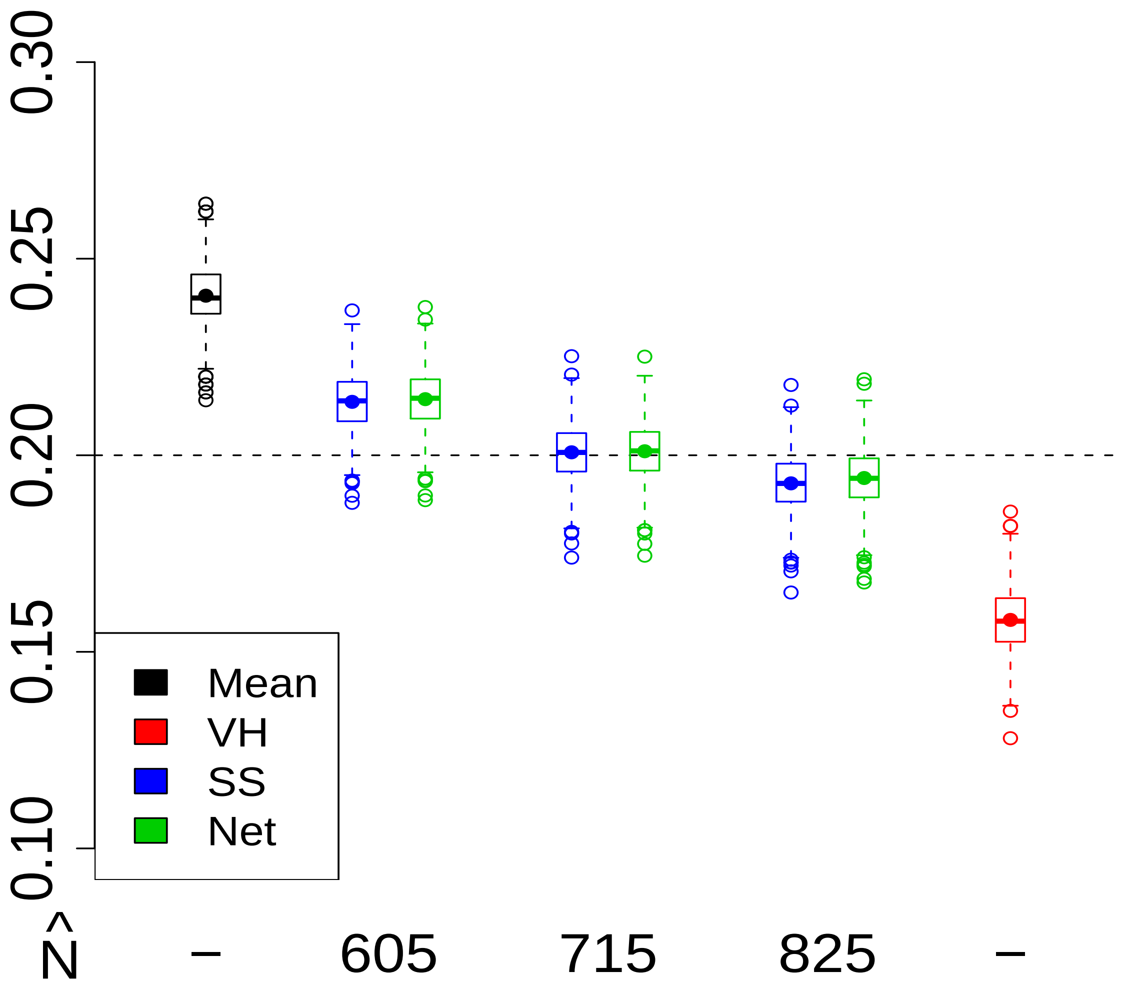
Sensitivity Analysis

- Unknown Population Size
  - Repeat simulations with inaccurate population estimate.
- Unknown Network Structure
  - Repeat simulations with more complex network model.

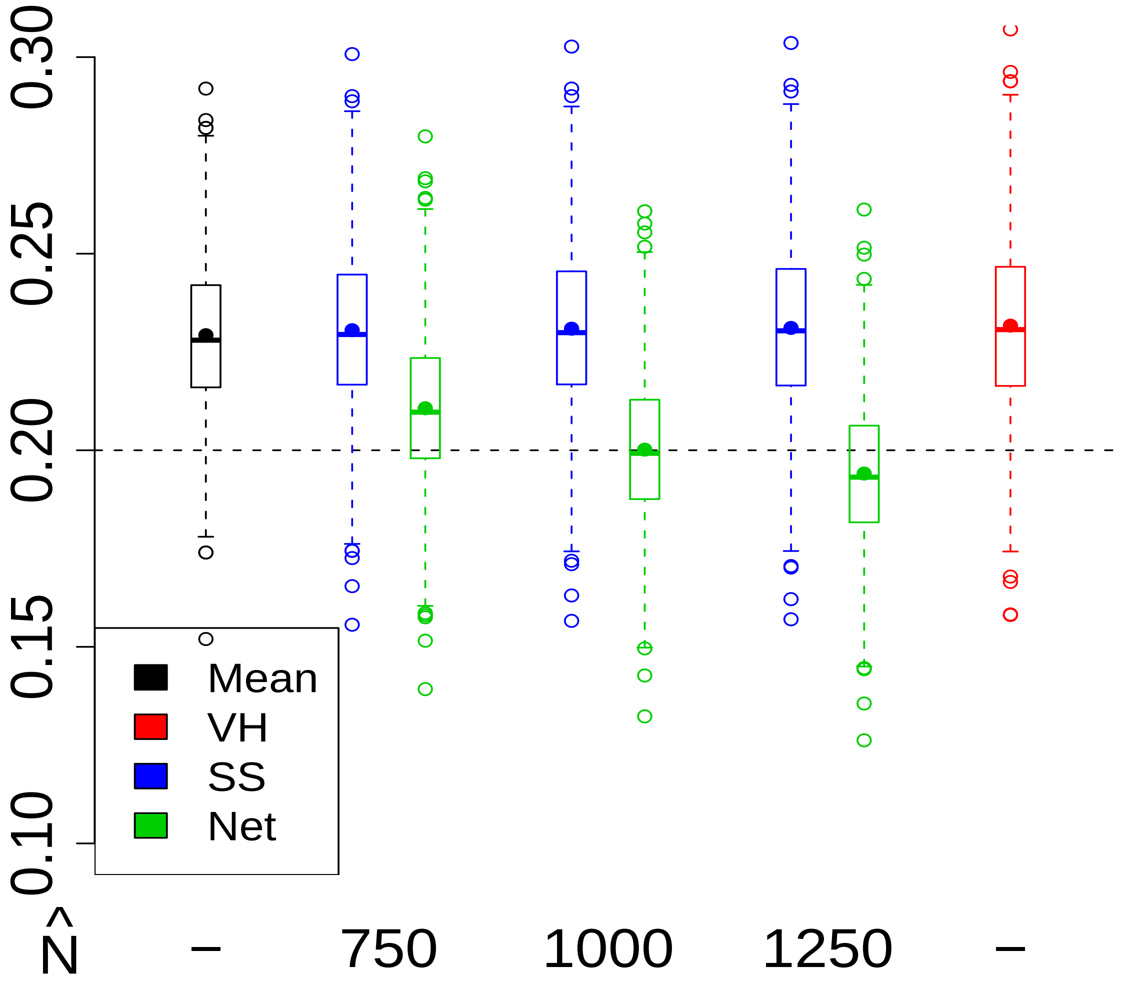
$N = 1000$, 50% Sample, $w = 1$, $R = 1$, Random Seeds



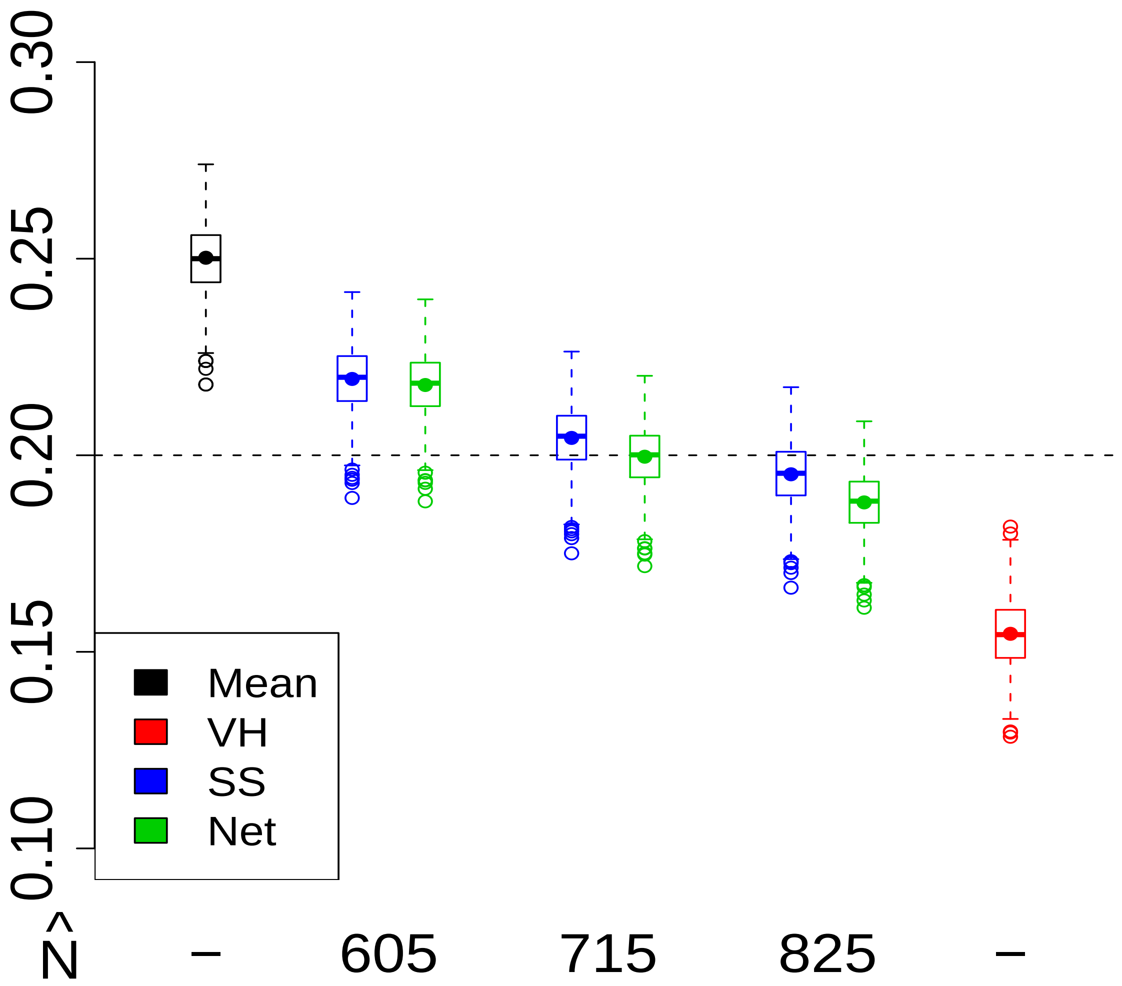
$N = 715$, 70% **Sample,** $w = 1.8$, $R = 1$, **Random Seeds**



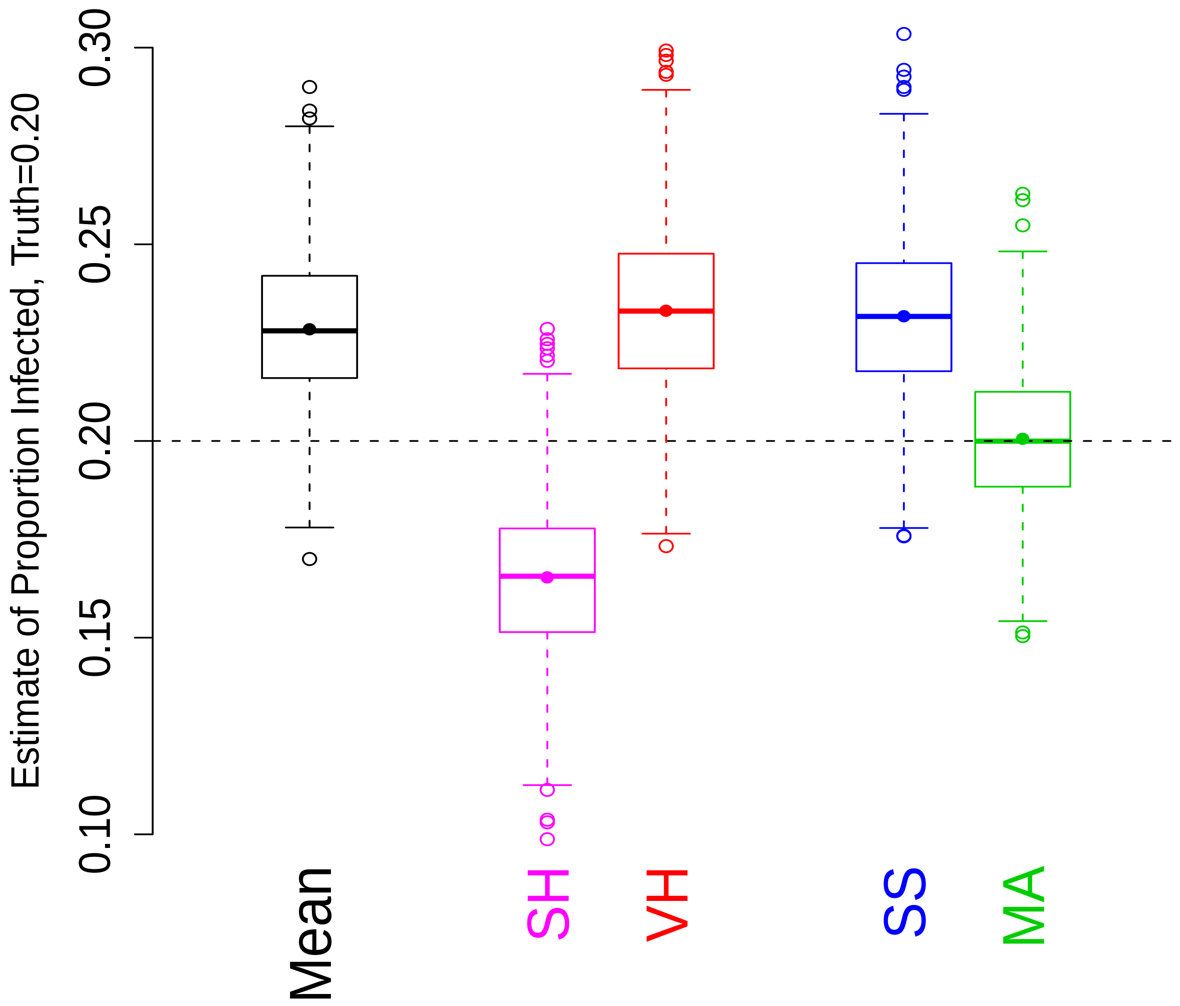
$N = 1000$, 50% Sample, $w = 1$, $R = 5$, Infected Seeds



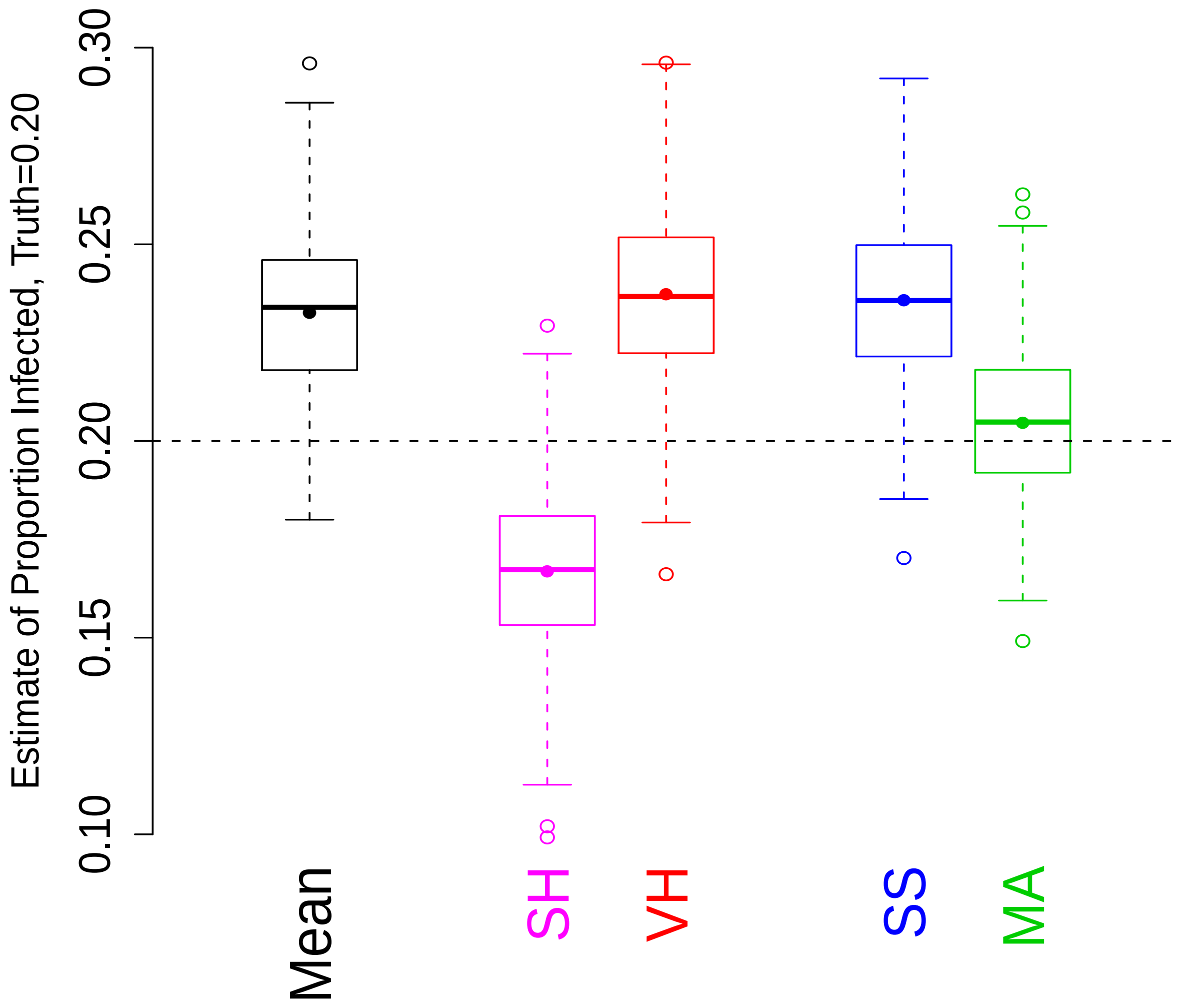
$N = 715$, 70% Sample, $w = 1.8$, $R = 5$, Infected Seeds



Increased Triangles (4 × edges with shared partner)



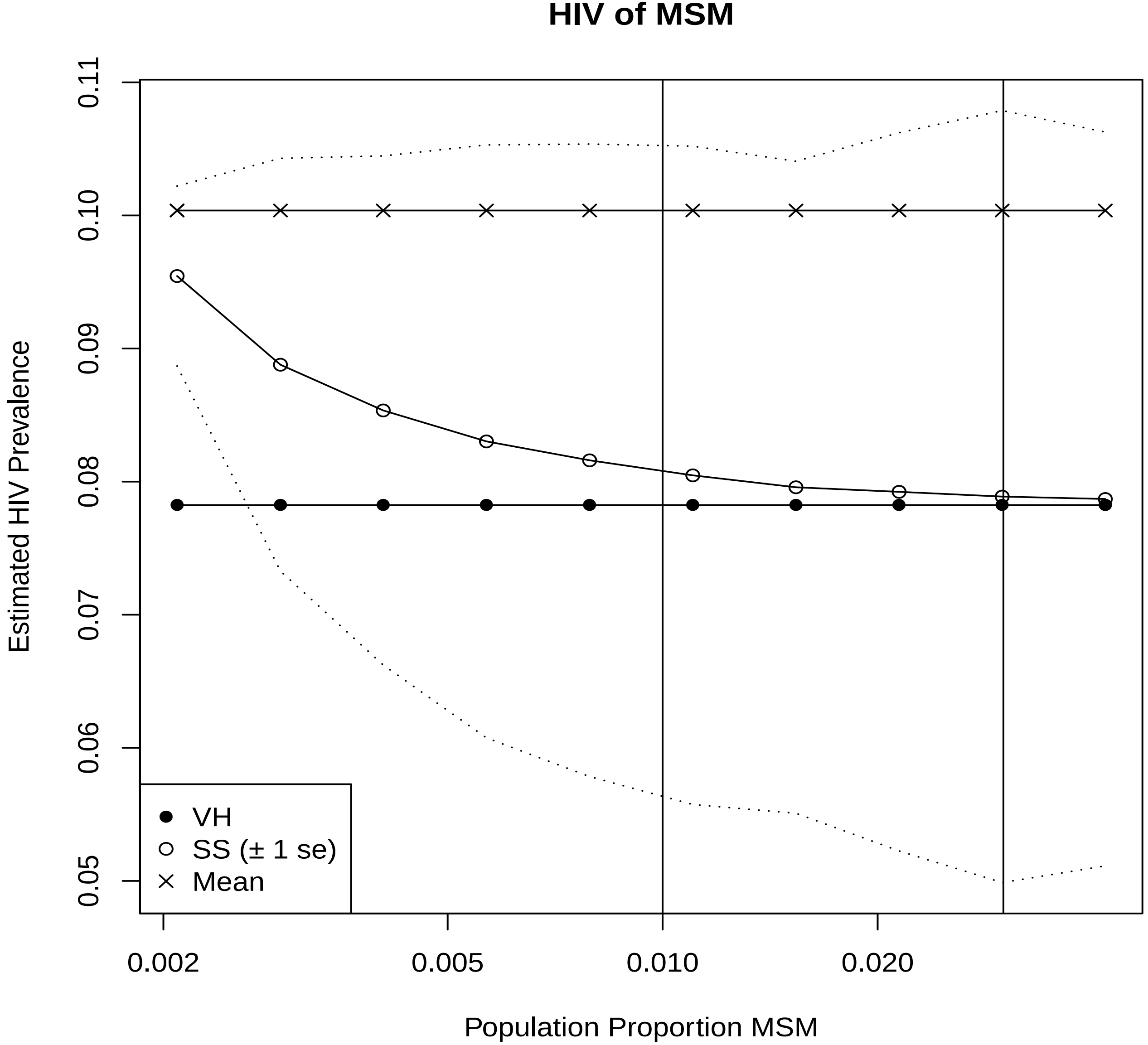
Increased Geometric Function of Edge-Triangles (10 ×)



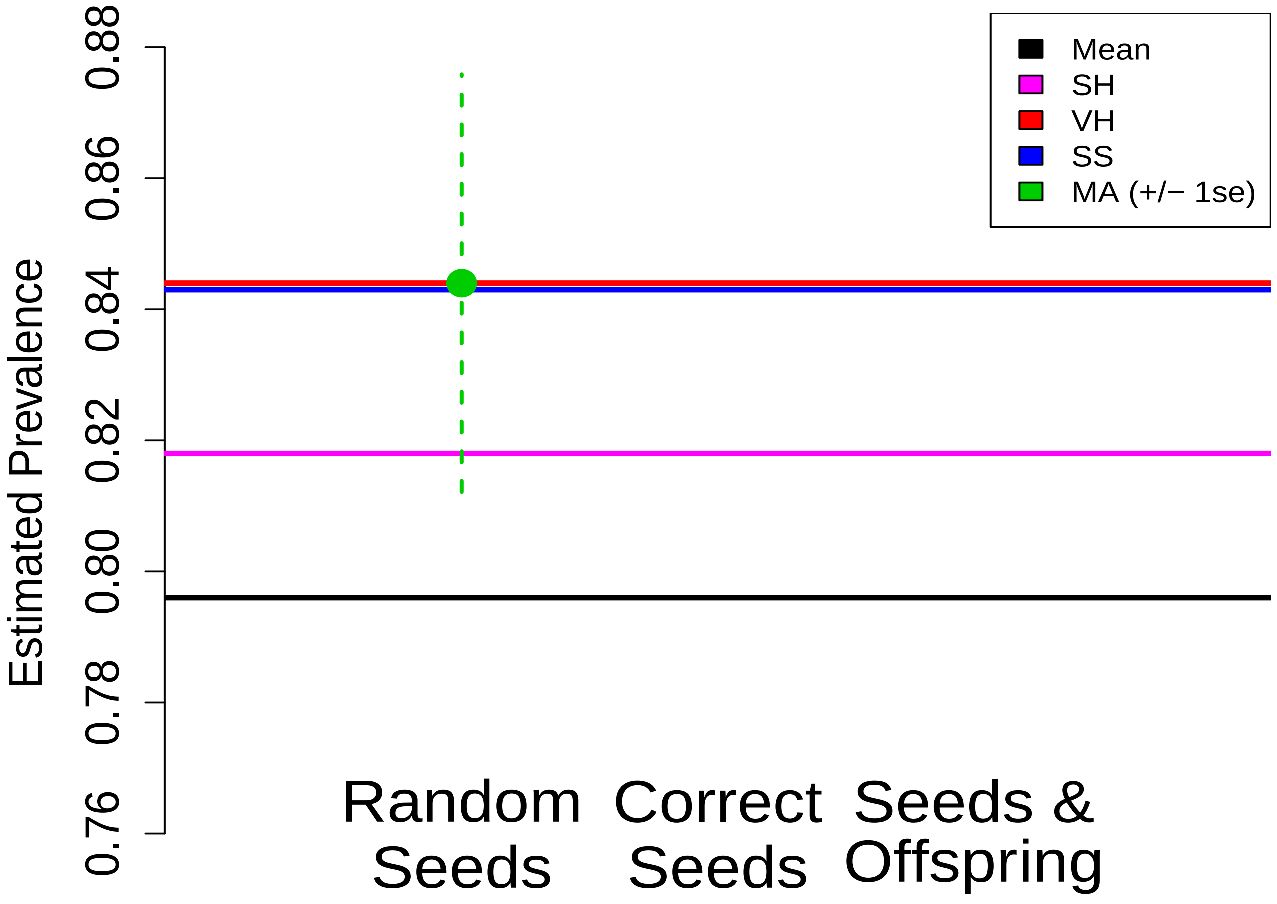
Outline of Presentation

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

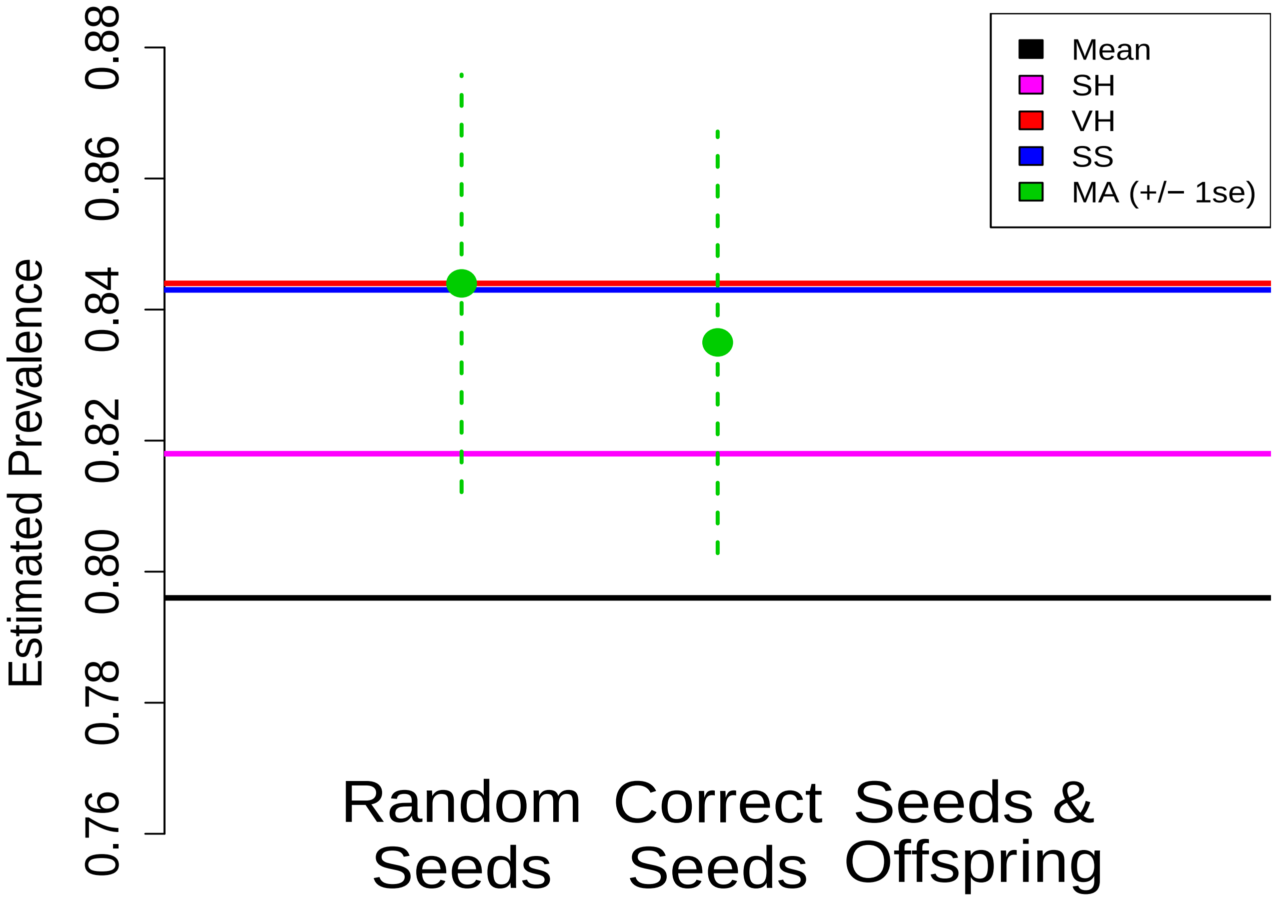
HIV Prevalence among MSM in a Caribbean City



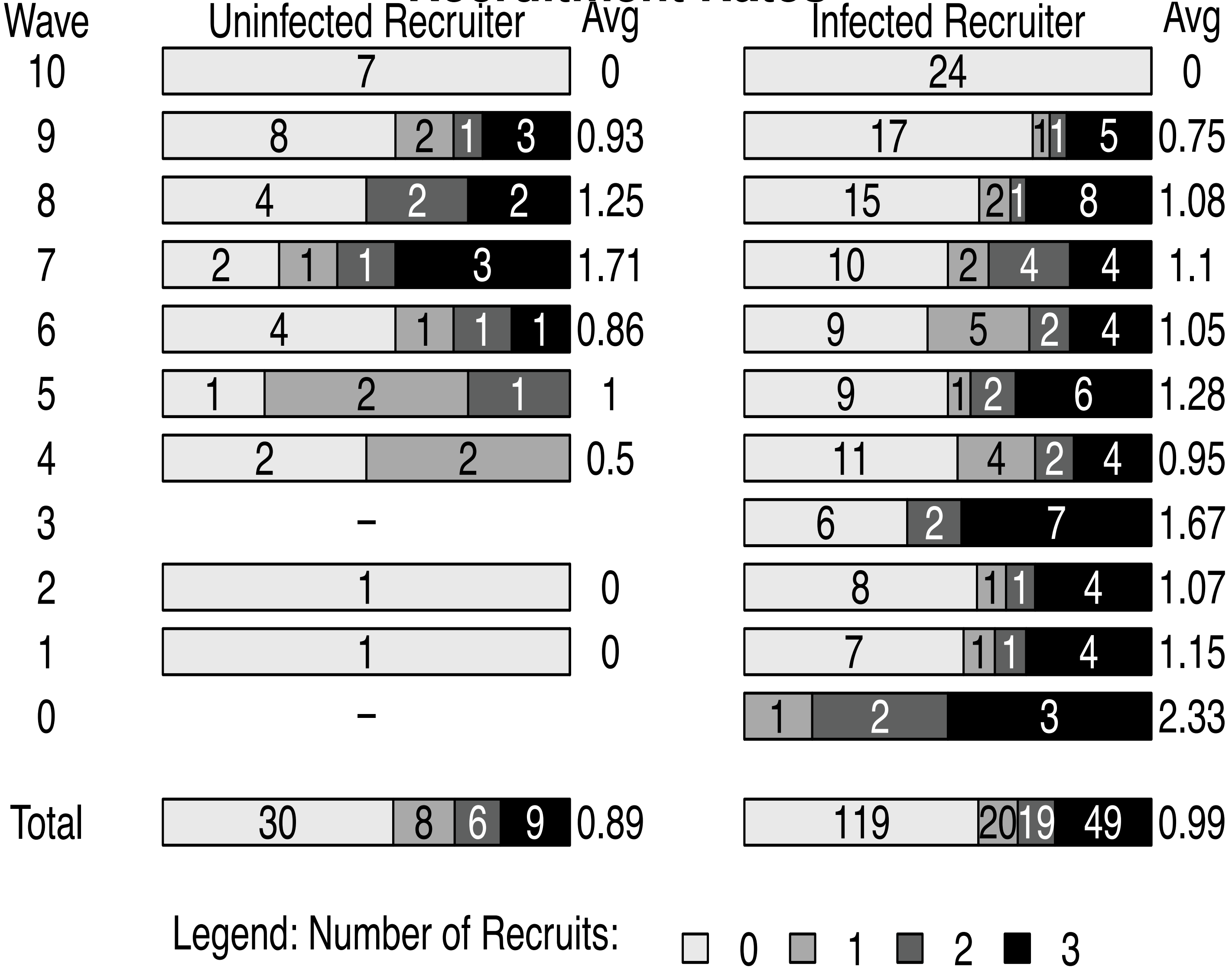
HIV Prevalence among IDU in an Eastern European City



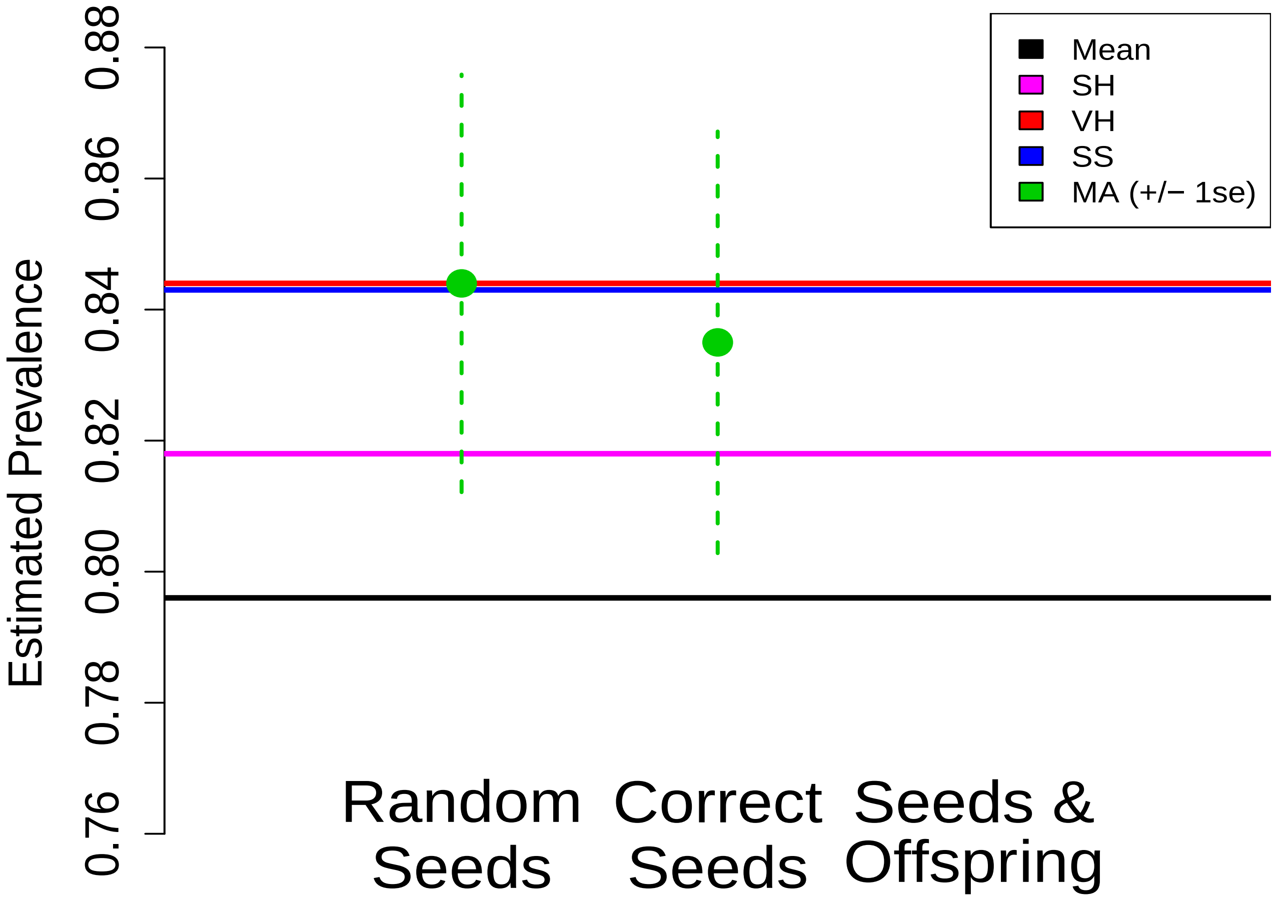
HIV Prevalence among IDU in an Eastern European City



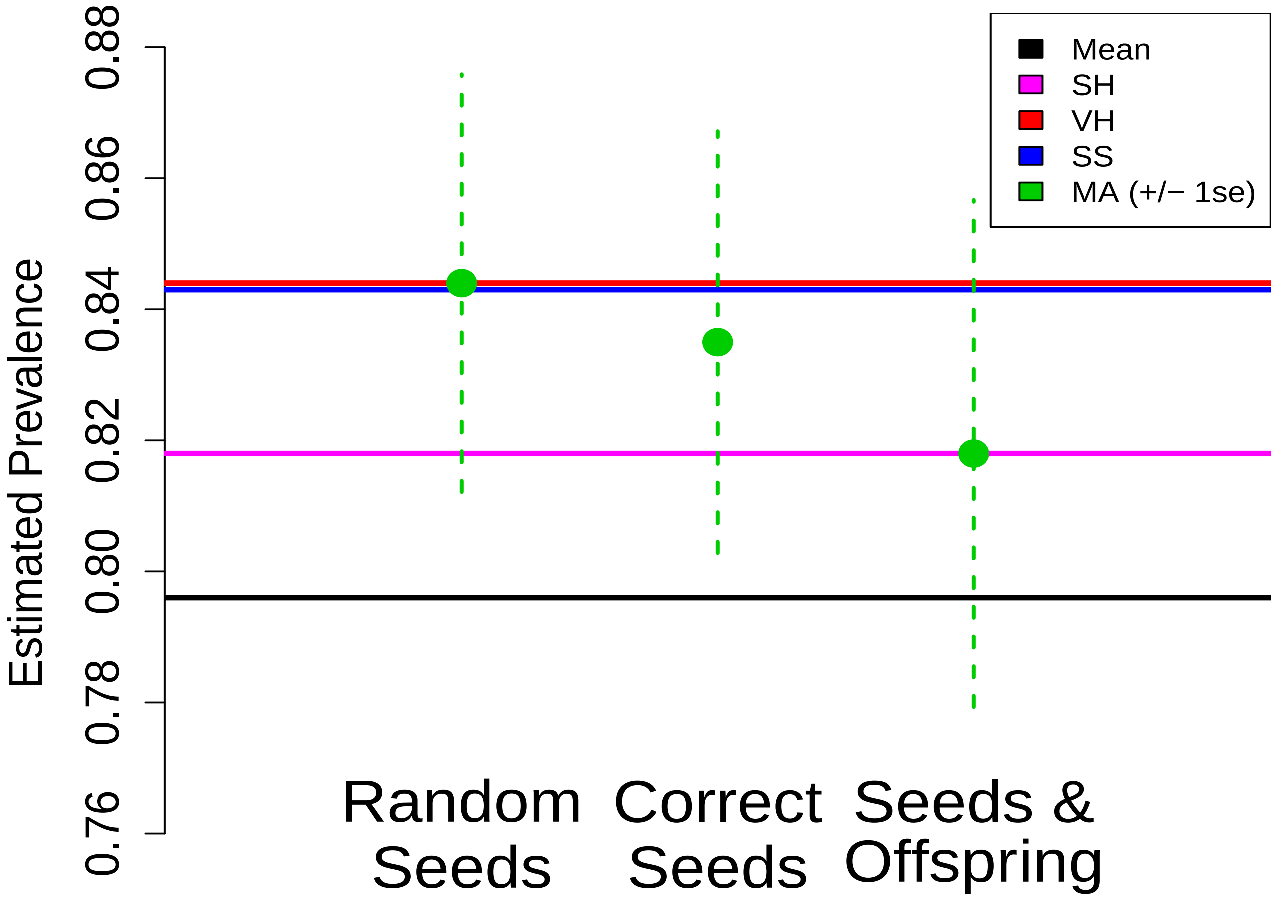
Recruitment Rates



HIV Prevalence among IDU in an Eastern European City



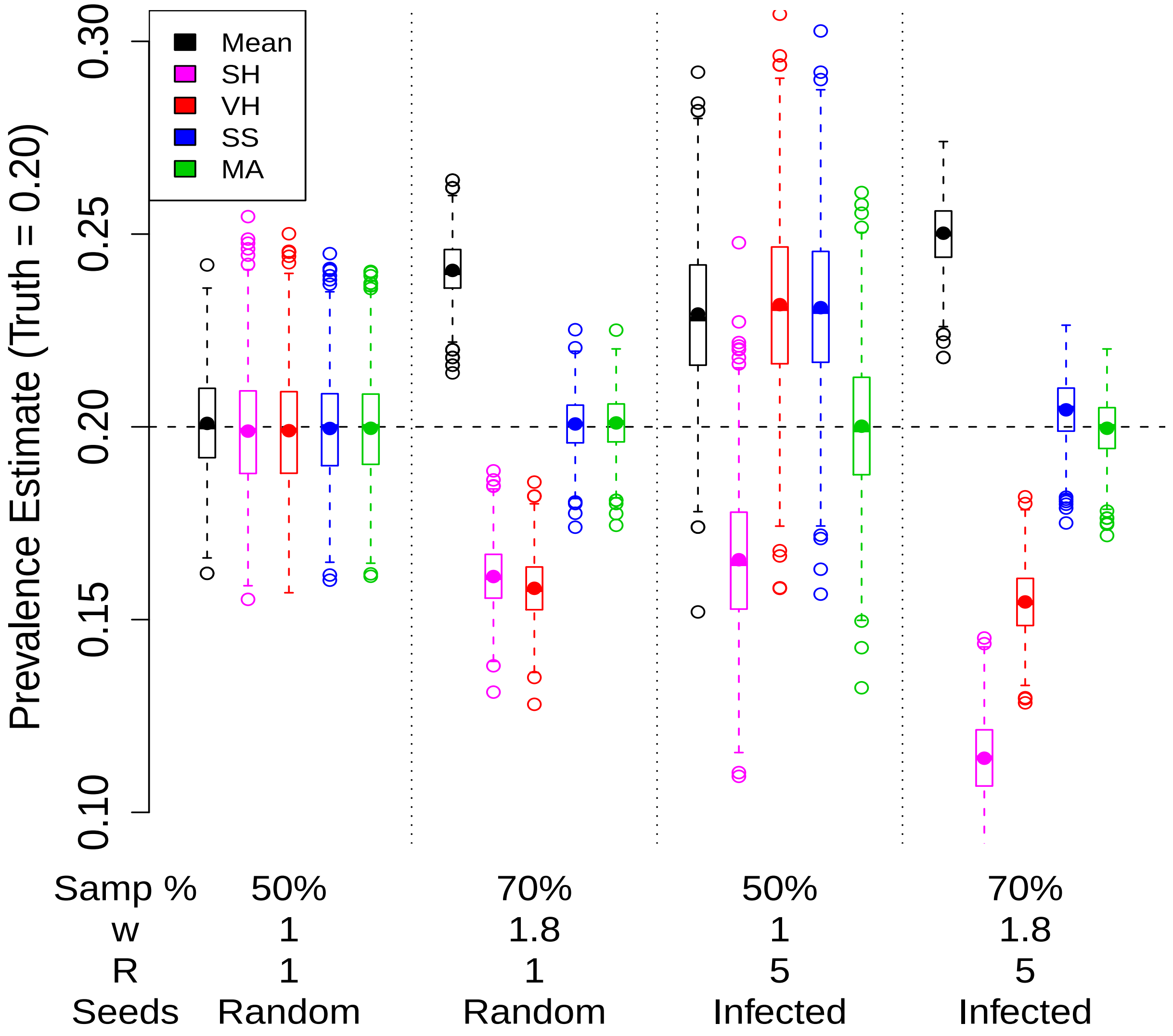
HIV Prevalence among IDU in an Eastern European City



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



Discussion: Respondent-Driven Sampling - Assumptions

	Network Structure Assumptions	Sampling Assumptions
Random Walk Model	Network size large ($N >> n$)	Sampling with replacement Single non-branching chain
Remove Seed Dependence	Homophily weak enough Connected graph	Sufficiently many sample waves
To Estimate Probabilities	All ties reciprocated	Degree accurately measured Random referral
Additional Assumptions of SS	~~Known network size N~~	~~No seed bias~~
Additional Assumptions of MA	~~Non-random mixing observable~~ ~~Network model form~~	~~Sampling model form~~

Assumptions of Volz-Heckathorn Estimator

Discussion: Respondent-Driven Sampling - Assumptions

	Network Structure Assumptions	Sampling Assumptions
Random Walk Model	~~Network size large ($N >> n$)~~	~~Sampling with replacement~~ ~~Single non-branching chain~~
Remove Seed Dependence	Homophily weak enough Connected graph	Sufficiently many sample waves
To Estimate Probabilities	All ties reciprocated	Degree accurately measured Random referral
Additional Assumptions of SS	Known network size N	No seed bias
Additional Assumptions of MA	~~Non-random mixing observable~~ ~~Network model form~~	~~Sampling model form~~

Assumptions of Successive Sampling Estimator

Discussion: Respondent-Driven Sampling - Assumptions

	Network Structure Assumptions	Sampling Assumptions
Random Walk Model	~~Network size large ($N >> n$)~~	~~Sampling with replacement~~ ~~Single non-branching chain~~
Remove Seed Dependence	~~Homophily weak enough~~ Connected graph	~~Sufficiently many sample waves~~
To Estimate Probabilities	All ties reciprocated	Degree accurately measured Random referral
Additional Assumptions of SS	Known network size N	~~No seed bias~~
Additional Assumptions of MA	Non-random mixing observable Network model form	Sampling model form

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:

- Krista J. Gile, *Inference from Partially-Observed Network Data*, Ph.D. Dissertation, Department of Statistics, University of Washington, 2008.

- Krista J. Gile and Mark S. Handcock, "Respondent-Driven Sampling: An Assessment of Current Methodology," *Sociological Methodology,* 2010, available on arXiv.

- Krista J. Gile, "Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation," *Journal of the American Statistical Association*, 2011, available on arXiv.

- Krista J. Gile and Mark S. Handcock, "Network Model-Assisted Inference from Respondent-Driven Sampling Data," under revision, available on arXiv.

Thank You!