

Overview of RDS from a Statistical Perspective

by the

Hard-to-Reach Population Methods Research Group*

Information available at

<http://www.hpmsg.org/>

<http://www.hpmsg.org/workshop>

*The project has been supported by the Presidents Emergency Plan for AIDS Relief (PEFPAR) through the US Centers for Disease Control and Prevention (CDC) under the terms of Cooperative Agreement U2GPS001468-5.

Hard-to-Reach Population Methods Research Group

- Ian E. Fellows, Fellows Statistics
- Lisa G. Johnston, Tulane University, UCSF
- Krista J. Gile, University of Massachusetts - Amherst
- Corinne M. Mar, University of Washington
- Mark S. Handcock, UCLA

Outline of Presentation

1. [Link-Tracing Hard-to-Reach Population Sampling](#)
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Discussion

Standard Survey Sampling

Stylized description

- Choose a *population* of interest and a population characteristic of interest μ
- Determine the *sampling frame*: $i = 1, \dots, N$ sample units.
- Choose variables to measure on them:
outcome $z_i, i = 1, \dots, N$, *control variables* $x_i, i = 1, \dots, N$,
- Choose a *sampling design*:
e.g., simple random sampling, stratified sampling on x , stratified sampling on z
- Choose a sample of units $i = 1, \dots, n$ and collect data on the sampled units
- Estimate the population characteristics of interest based on the sample

Estimation

- Goal: Estimate the population mean of z :

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

where

$$z_i = \begin{cases} 1 & i \text{ has the characteristic} \\ 0 & i \text{ does not have the characteristic.} \end{cases}$$

- Sample indicators

$$S_i = \begin{cases} 1 & i \text{ sampled} \\ 0 & i \text{ not sampled} \end{cases}$$

- Inclusion probabilities

$$\pi_i = P(S_i = 1) \quad i = 1, \dots, N$$

e.g. simple random sampling

$$\pi_i = n/N \quad i = 1, \dots, N$$

Classic Design-Based Inference:

- Goal: Estimate proportion “infected” :

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

where

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

- Horvitz-Thompson Estimator:

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

where

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

Classic Design-Based Inference

- Goal: Estimate proportion “infected” :

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

where

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

- Hajek Estimator:

$$\hat{\mu} = \frac{\sum_i \frac{S_i}{\pi_i} z_i}{\sum_i \frac{S_i}{\pi_i}}$$

where

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

Hajek Estimator

- The Hajek is useful when the population size N is not known
- The Hajek is better when z is weakly or negatively correlated with π_i .
- The key point: Each estimator requires $\pi_i = P(S_i = 1) \quad \forall i : S_i = 1$
- We often need to model the sampling process to estimate these inclusion probabilities

Hard-to-Reach Population Sampling: Motivating Questions

- What proportion of Injecting Drug Users in Hanoi are HIV Positive?
- What proportion of unregulated workers in New York City experience workplace violations of code?
- What proportion of sex workers in rural China belong to ethnic minorities?

Limitation: No practical conventional sampling frame.

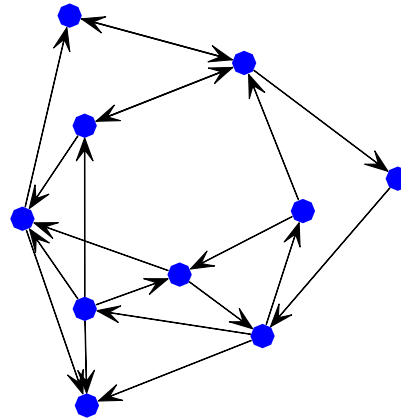
Hard-to-Reach Population Sampling:

Suppose:

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

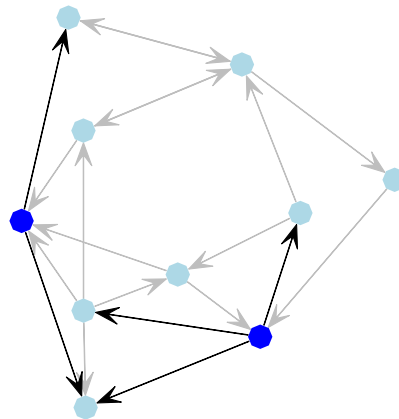
Surveying Networked Populations

- **Sampling Design:** Choose which part to observe:
“Ask 10% of IDUs about who they share needles with”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole community, but someone can not be contacted”
- **Boundary Specification Problem:**
“Should a visitor be considered a part of the community?”



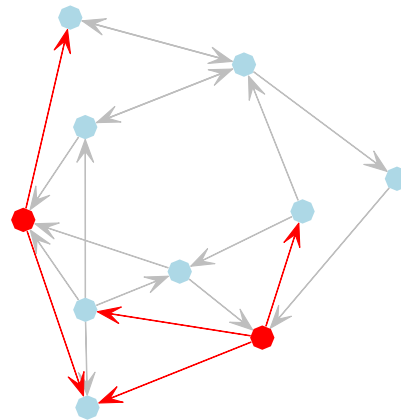
Surveying Networked Populations

- **Sampling Design:** Choose which part to observe:
“Ask 10% of IDUs about who they share needles with”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole community, but someone can not be contacted”
- **Boundary Specification Problem:**
“Should a visitor be considered a part of the community?”



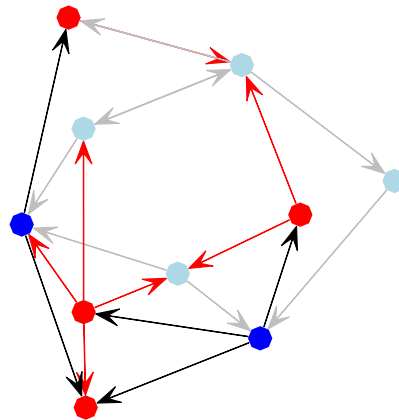
Surveying Networked Populations

- **Sampling Design:** Choose which part to observe:
“Ask 10% of IDUs about who they share needles with”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole community, but someone can not be contacted”
- **Boundary Specification Problem:**
“Should a visitor be considered a part of the community?”



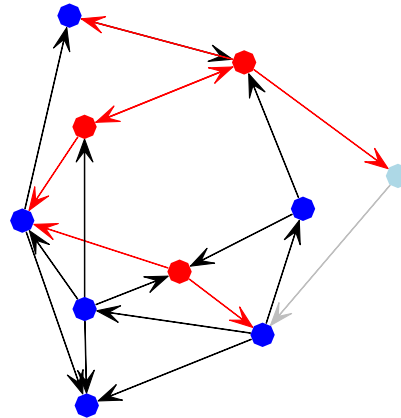
Surveying Networked Populations

- **Sampling Design:** Choose which part to observe:
“Ask 10% of IDUs about who they share needles with”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole community, but someone can not be contacted”
- **Boundary Specification Problem:**
“Should a visitor be considered a part of the community?”



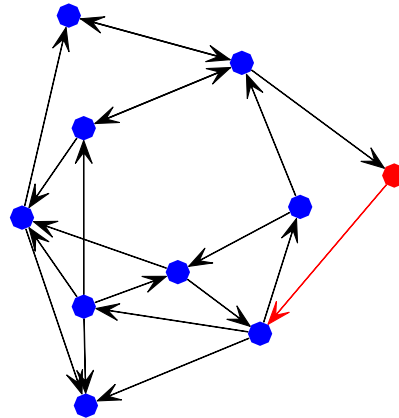
Surveying Networked Populations

- **Sampling Design:** Choose which part to observe:
“Ask 10% of IDUs about who they share needles with”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole community, but someone can not be contacted”
- **Boundary Specification Problem:**
“Should a visitor be considered a part of the community?”



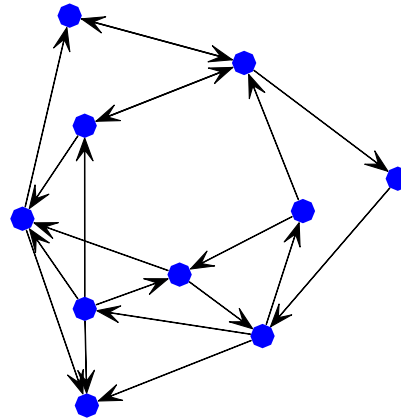
Surveying Networked Populations

- **Sampling Design:** Choose which part to observe:
“Ask 10% of IDUs about who they share needles with”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole community, but someone can not be contacted”
- **Boundary Specification Problem:**
“Should a visitor be considered a part of the community?”



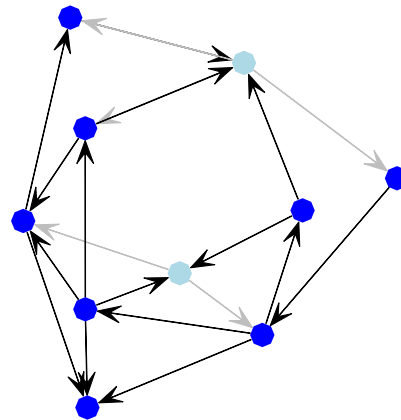
Surveying Networked Populations

- **Sampling Design:** Choose which part to observe:
“Ask 10% of IDUs about who they share needles with”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole community, but someone can not be contacted”
- **Boundary Specification Problem:**
“Should a visitor be considered a part of the community?”



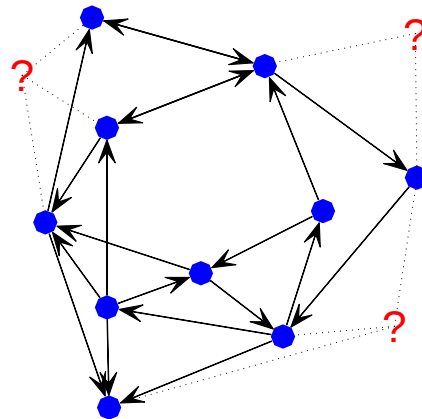
Surveying Networked Populations

- **Sampling Design:** Choose which part to observe:
“Ask 10% of IDUs about who they share needles with”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole community, but someone can not be contacted”
- **Boundary Specification Problem:**
“Should a visitor be considered a part of the community?”



Surveying Networked Populations

- **Sampling Design:** Choose which part to observe:
“Ask 10% of IDUs about who they share needles with”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole community, but someone can not be contacted”
- **Boundary Specification Problem:**
“Should a visitor be considered a part of the community?”



Link-Tracing Network Sampling

Suppose:

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

Sampling design:

- Begin with a reachable convenience sample (the *seeds*)
- Expand sample by the researchers sampling those tied to those already in the sample.

Often informally called “snowball” sampling.

Concerns:

- Seed Dependence: final sample depends on convenience sample of seeds
- Confidentiality: some populations prefer to stay 'hidden'
- Estimation: The sample depends on the unknown network

Design-based Inference for Describing Structure

- Approach:
 - Make probability statements about the outcomes and relations in the full network based on the observed part of the network
 - Base inference on the sampling design mechanism
 - Typically, weigh each observation by the inverse of probability of it being sampled
- Advantages:
 - Requires no assumptions about network structure
- Disadvantages:
 - Requires full knowledge of sampling mechanism, and sampling probabilities
 - Difficult to conduct complex analysis

Observable sampling probabilities under various sampling schemes

Sampling Scheme	Nodal Probabilities π_i		Dyadic Probabilities π_{ij}	
	Undirected	Directed	Undirected	Directed
Ego-centric	X	X	X	X
One-Wave	X			
k -Wave, $1 < k < \infty$				
Saturated	X			

“X” indicates observable sampling probabilities

An important case: Respondent-Driven Sampling

- Sampling design: Require respondents to choose from among their social circle rather than the researcher chooses.
 - Seed Dependence: follow only a few links from each sampled
 - Confidentiality: *respondent-driven*: respondents distribute uniquely identified coupons. no names.
 - Estimation: 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.

Respondent-Driven Sampling (RDS): Overview

Example:

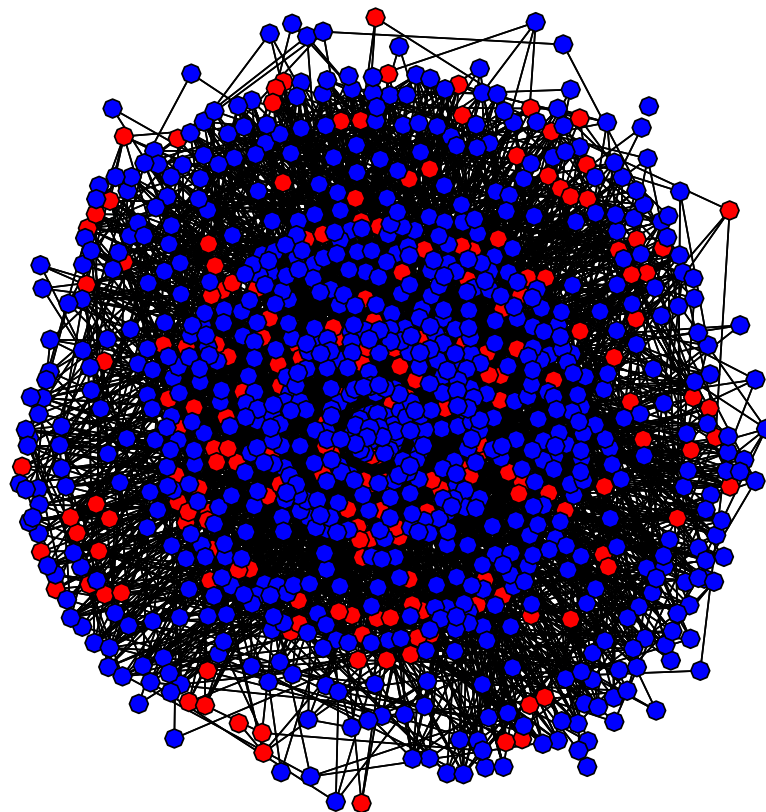
What proportion of Injecting Drug Users in Hanoi are HIV positive?

Hard-to-reach population

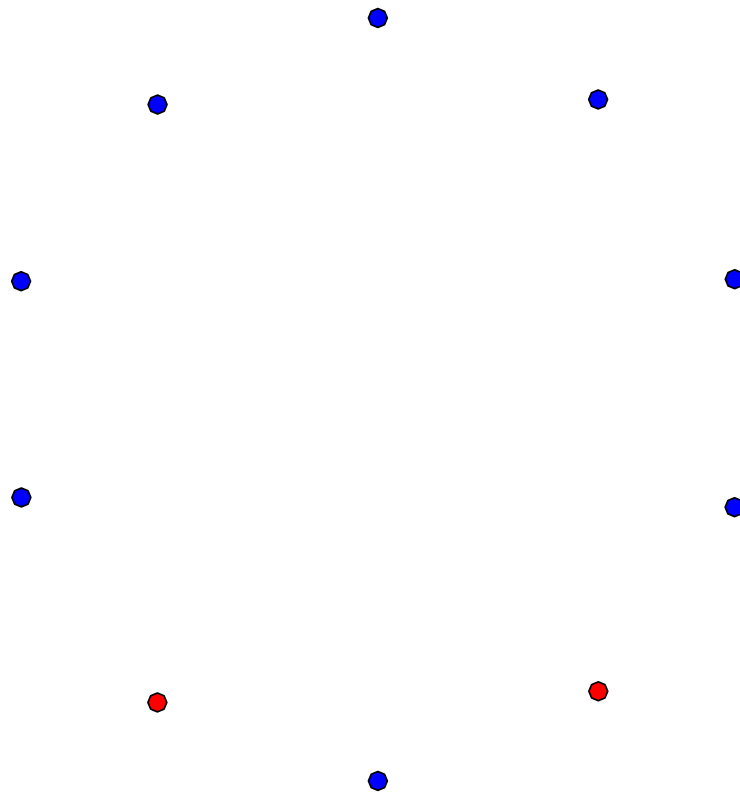
- Other Approaches:
 - Clinic-based sample
(convenience sample - not a probability sample)
 - Street-based sampling
(time-location sample - not probability sample of individuals)
 - Reported drug use in general population survey
(Sample from larger existing sampling frame -too expensive)
- RDS: “Something like” probability sample

M.J. Salganik and D.D. Heckathorn, “Sampling and estimation in hidden populations using respondent-driven sampling.” Sociological Methodology, 34, 193-239, 2004.

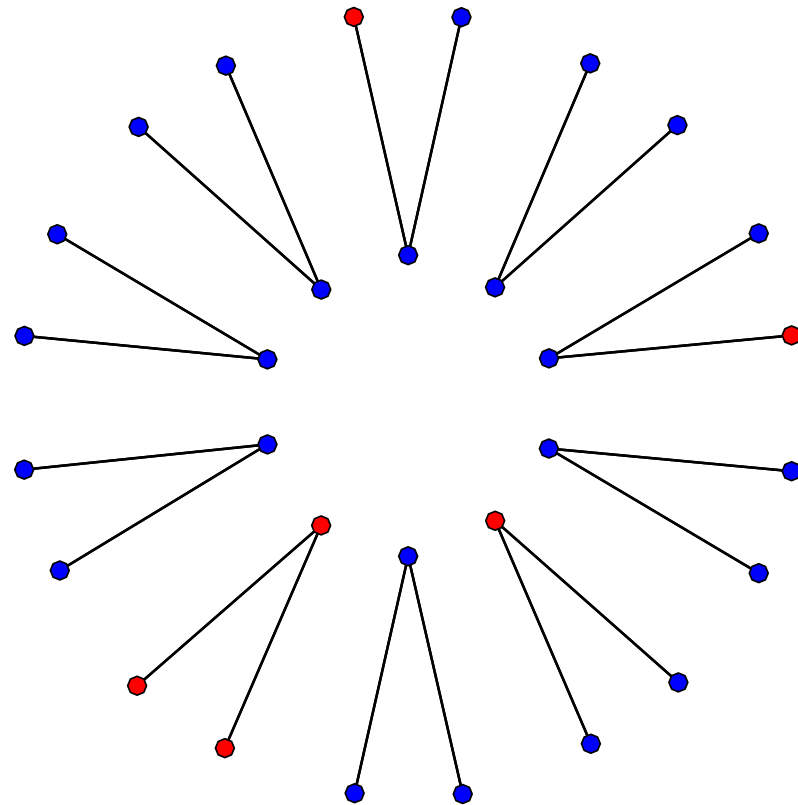
Stylized population



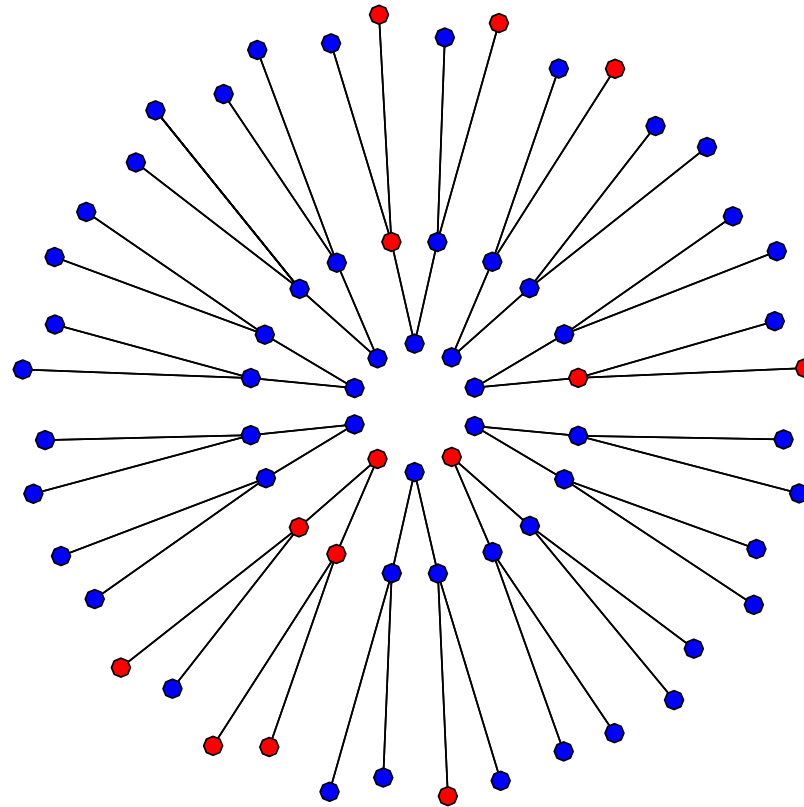
Start with seeds ...



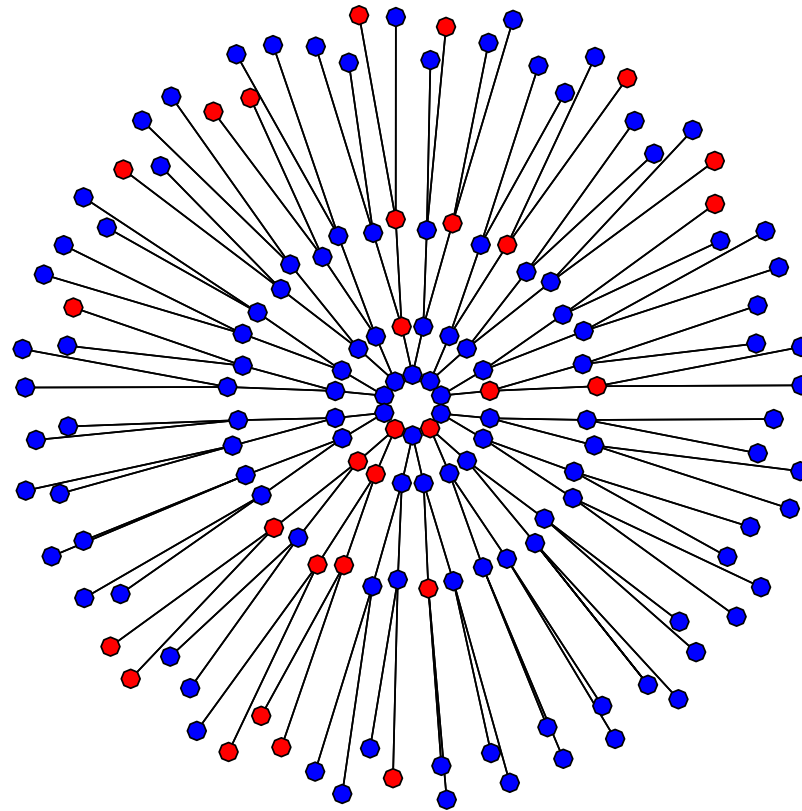
Seeds recruit the first wave ...

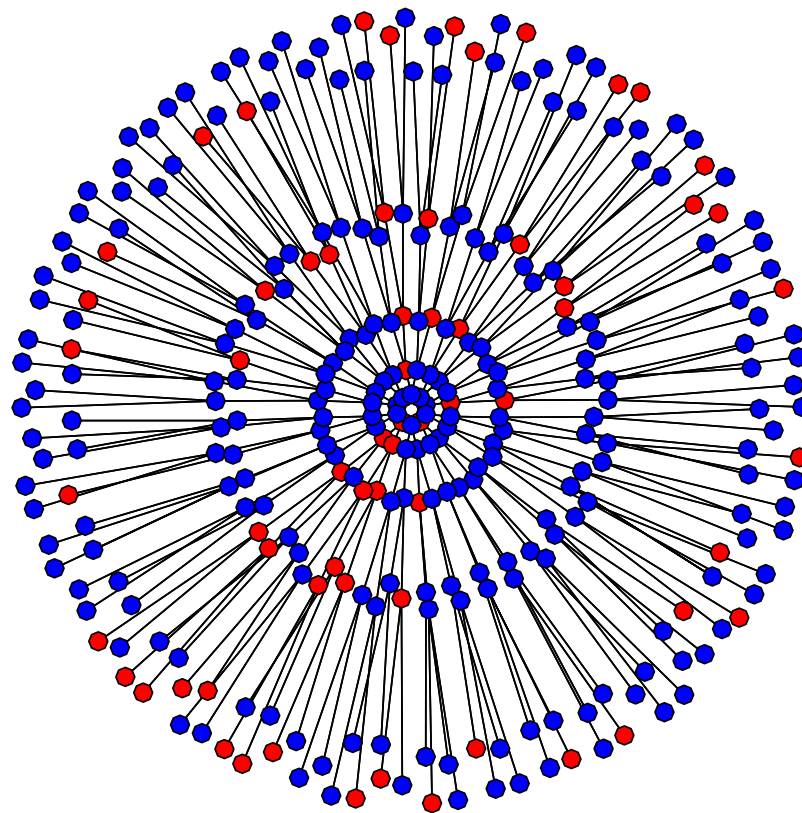


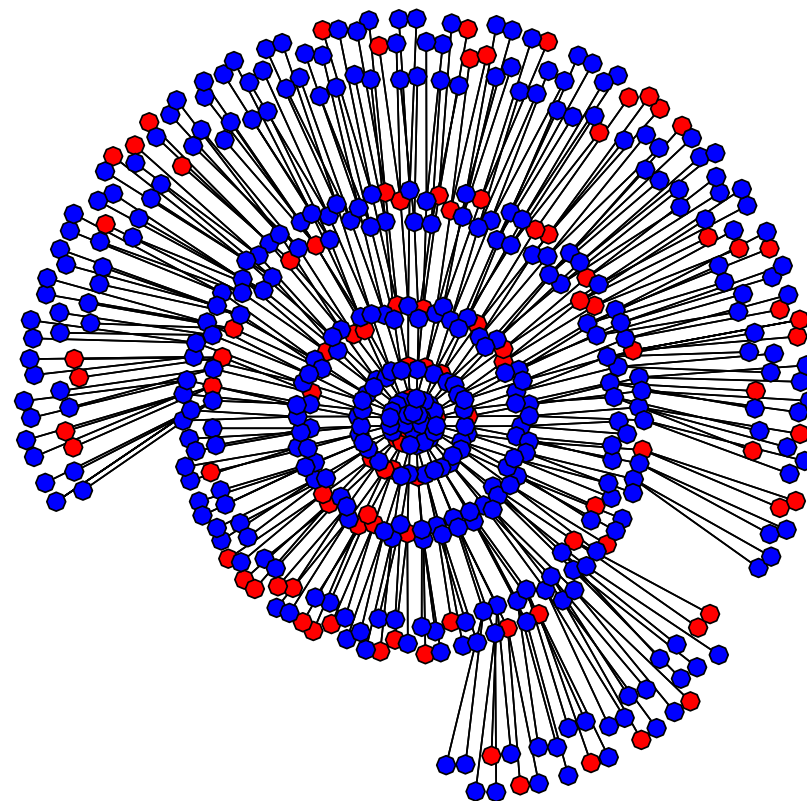
The first wave recruit the second wave ...



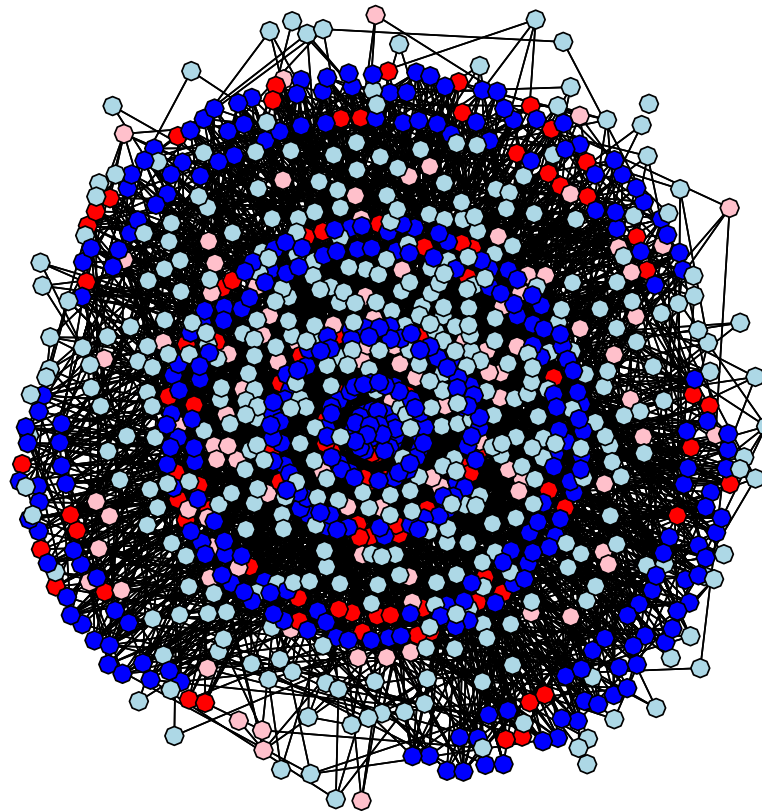
and so on ...

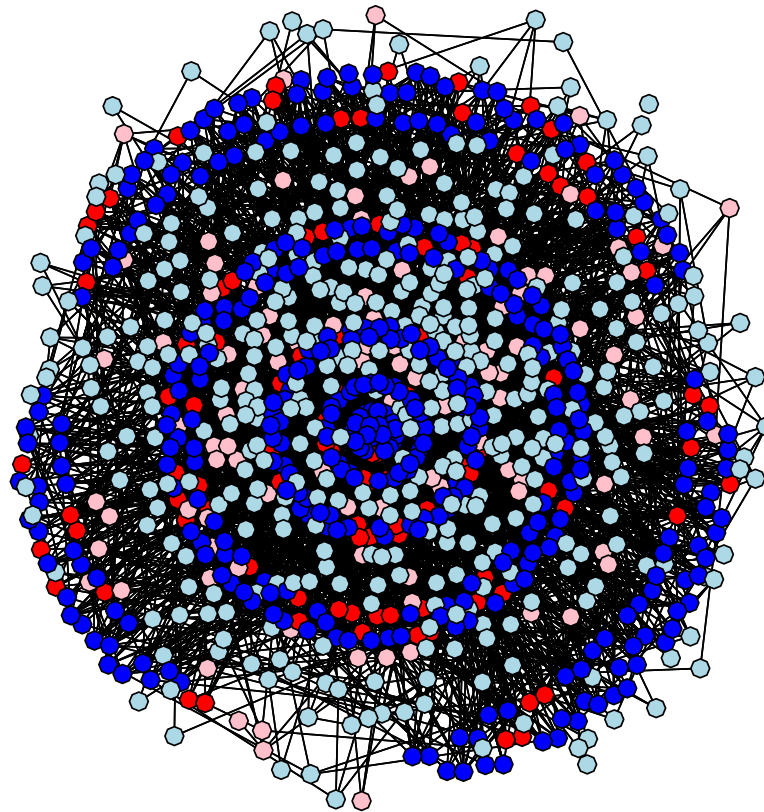




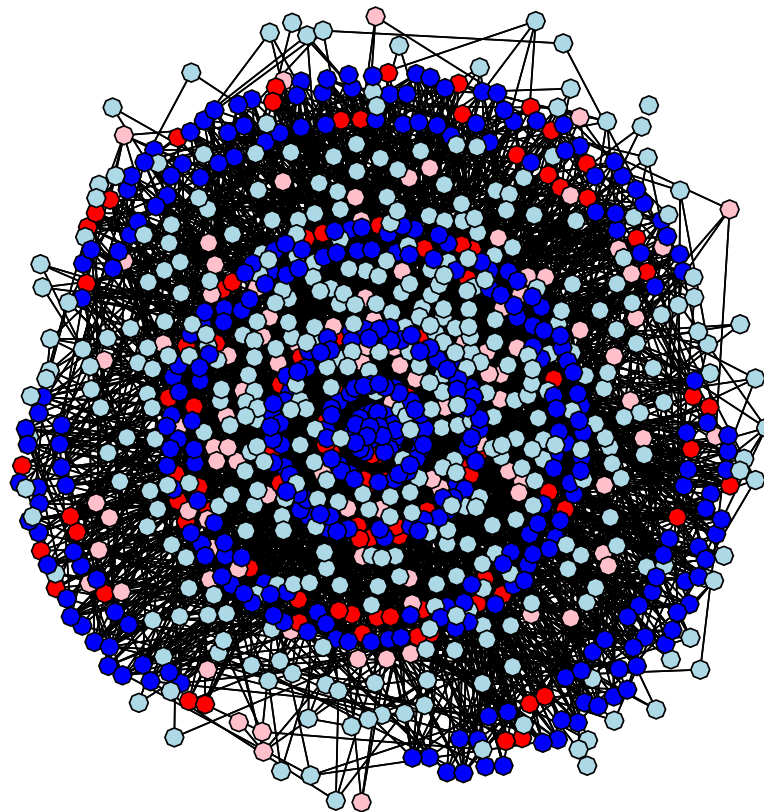


At the end (the unsampled are shaded)





degree of node i = # of ties of node i



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

Design-Based Inference:

- Goal: Estimate proportion “infected” :

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

where

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

- Hajek Estimator:

$$\hat{\mu} = \frac{\sum_i \frac{S_i}{\pi_i} z_i}{\sum_i \frac{S_i}{\pi_i}}$$

where

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

How do we we determine the sampling probabilities?

- The key point: The estimators require $\pi_i = P(S_i = 1) \quad \forall i : S_i = 1$
- We need to model the sampling process to estimate these inclusion probabilities

Outline of Presentation

1. Link-Tracing Hard-to-Reach Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. [Random Walk Approximation](#)
5. Successive Sampling Approximation
6. Discussion

One Approach: Random walk approximation

Respondent-driven Sampling:

- Approximate link-tracing process by a Markov chain representation
- 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.

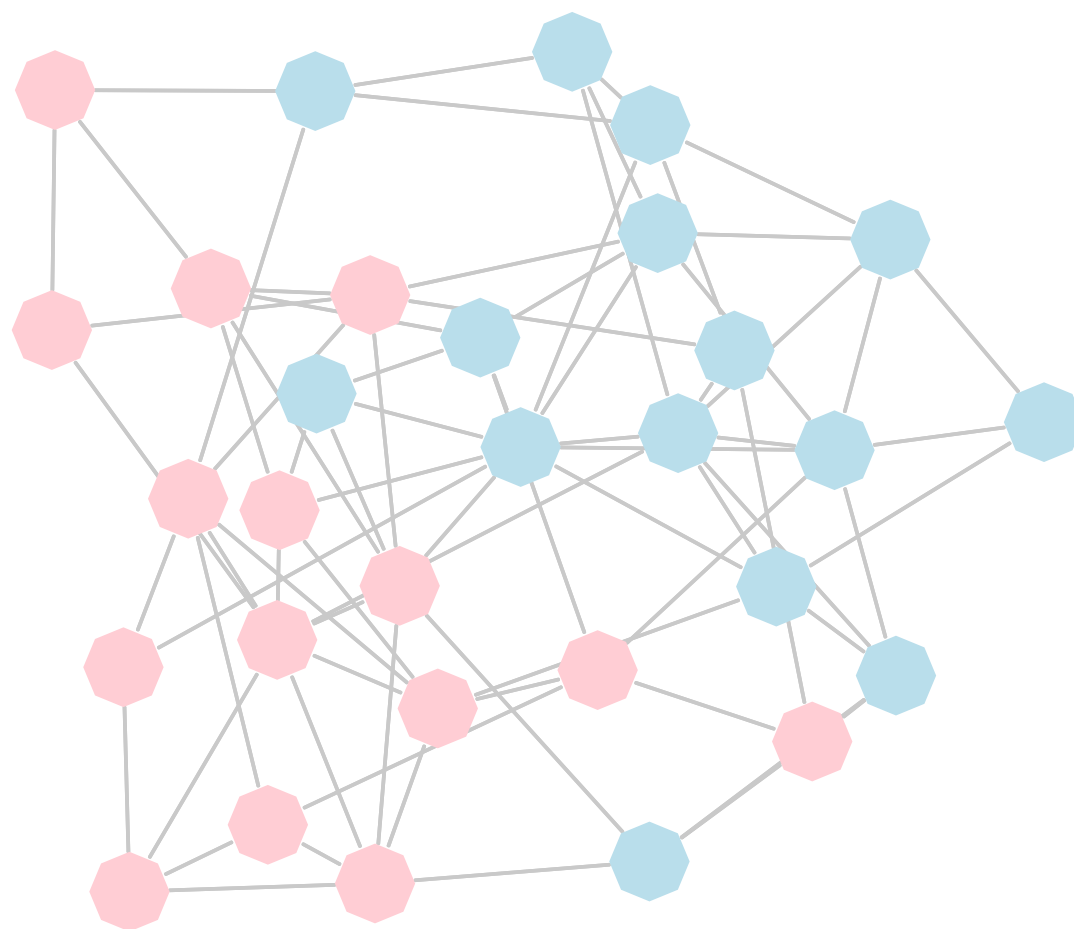
Volz, E., and D.D. Heckathorn, “Probability Estimation Theory for Respondent Driven Sampling,” *Journal of Official Statistics*, 2008.

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

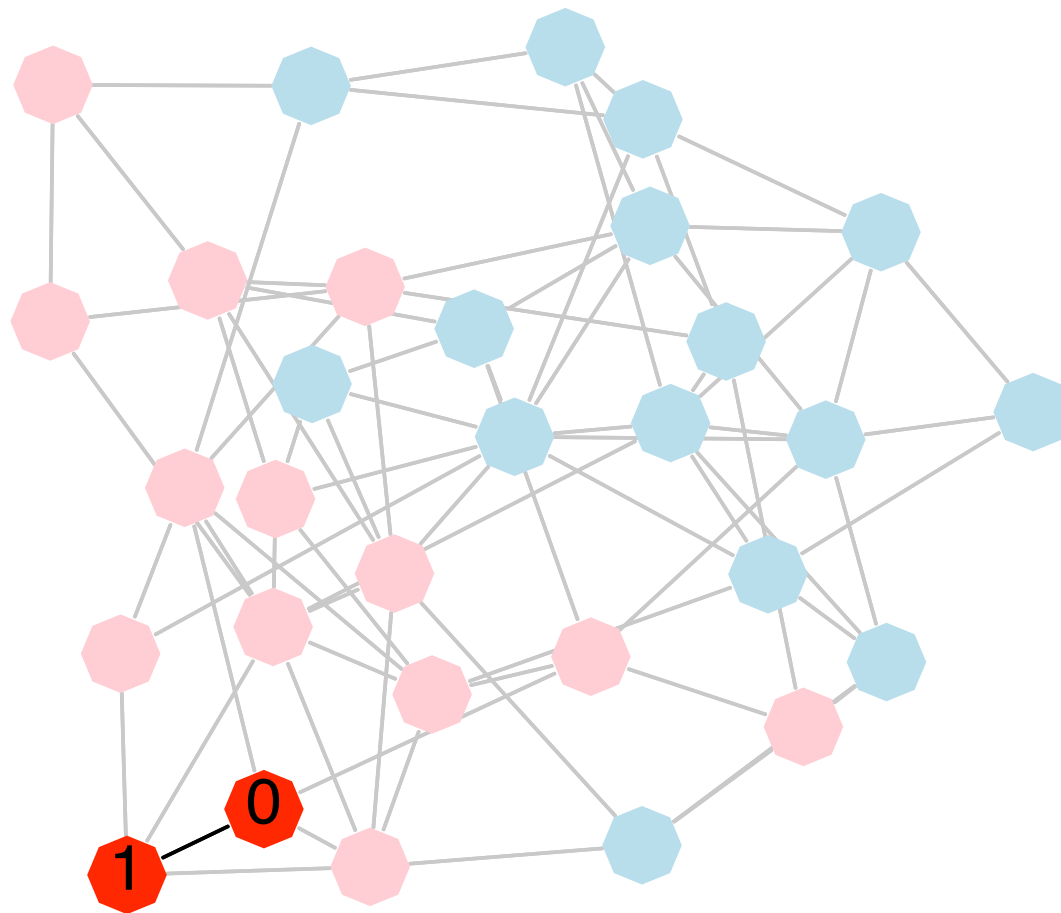
$$\hat{\mu} = \frac{\sum_i S_i \frac{z_i}{d_i}}{\sum_i S_i \frac{1}{d_i}}$$

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

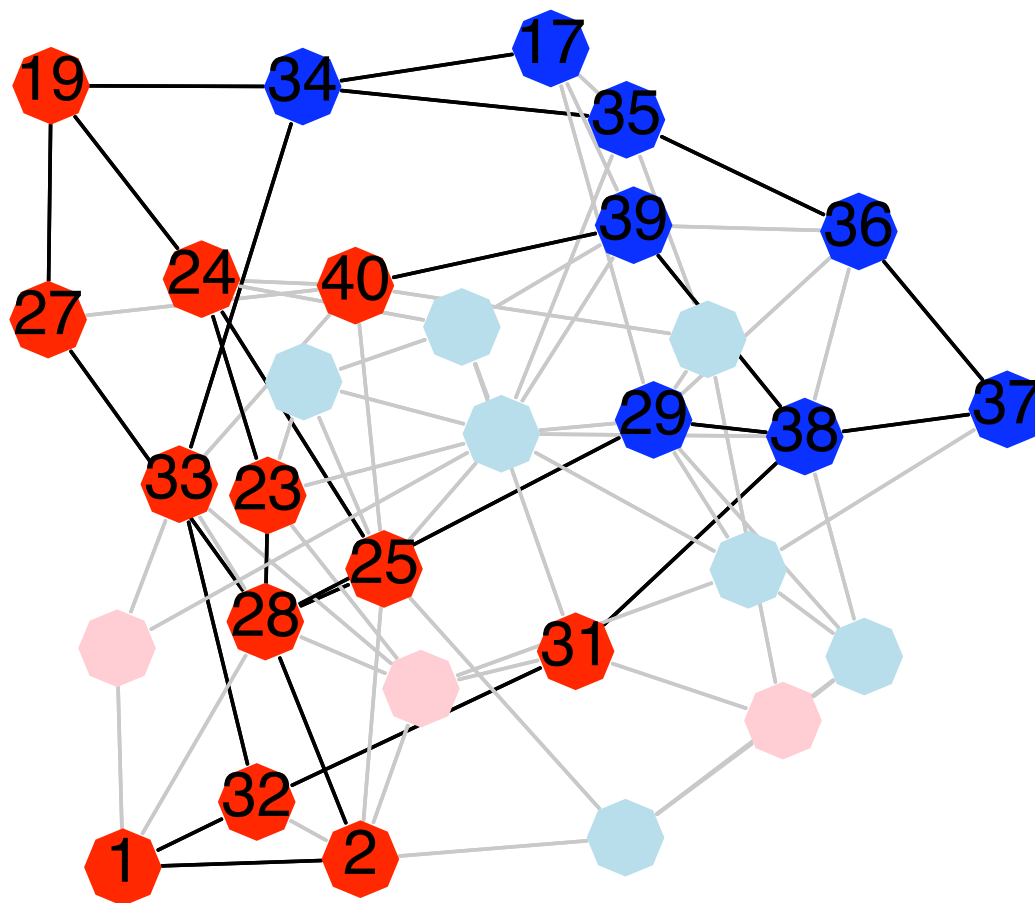
Hard-to-Reach Population



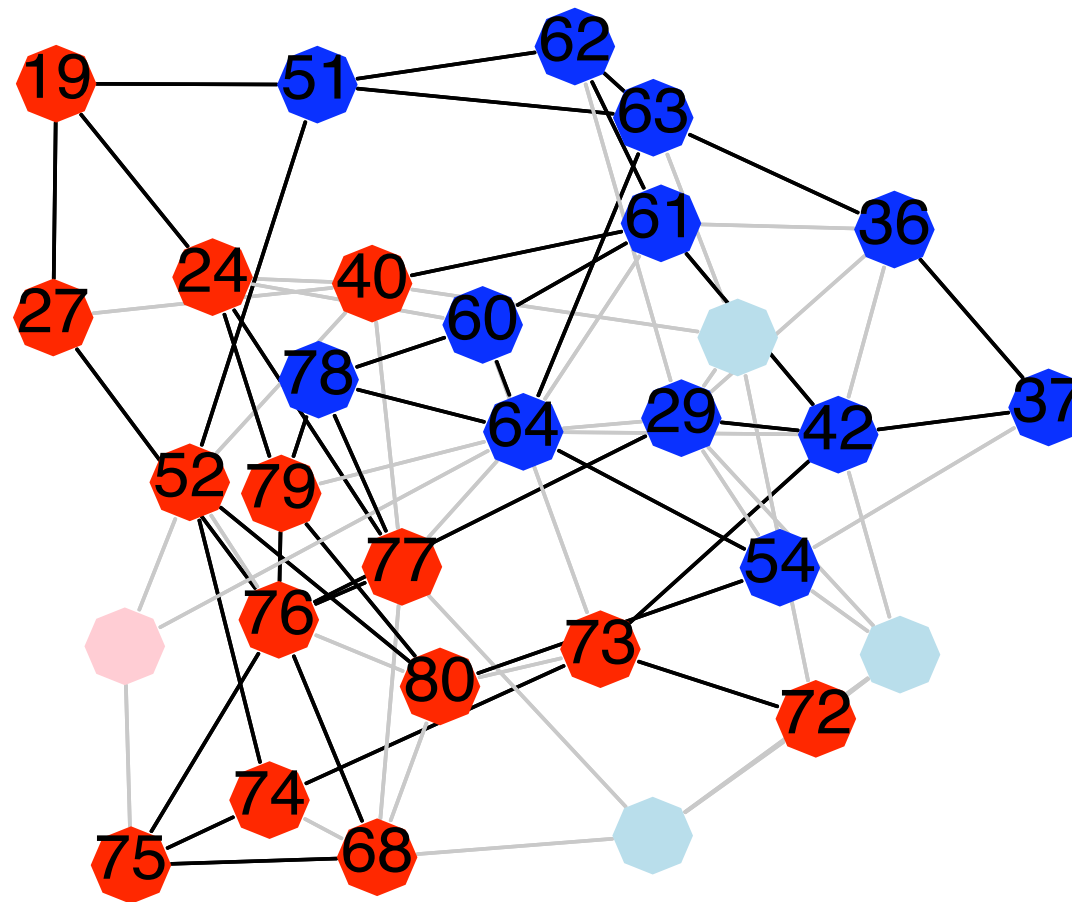
1-Step Random Walk



40-Step Random Walk (20 distinct nodes observed)



80-Step Random Walk (26 distinct nodes observed)



Some fundamental characteristics

- **degree** of a person: A measure of the *activity* or sl popularity of a person. Specifically, the number of social ties the person has to other people
- **Differential activity**: A measure of the relative popularity of the infected to the uninfected. Specifically, the ratio of the mean degree of the infected to the mean degree of the uninfected.
- Population **homophily**: A measure of like-with-like social ties. Specifically, the ratio of the number of uninfected to infected social ties *if there was no homophily* to the actual number
- **Recruitment homophily**: A measure of like-with-like recruitment. Specifically, the ratio of the number of recruits that have the same infection status as their recruiter to the number we would expect if there was no homophily.

Simulation Study to look at the performance of this approximation

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.

RDS-I Heckathorn et al. Estimators

Look at who recruits who

- Estimate the proportion infected people directly
- Idea: Compare the numbers of times that:
 - infected recruit uninfected, and
 - uninfected recruit infected
- The higher proportion infected, the more of the former
- So solve the balance equations for the proportion infected people
- Also adjust for differential activity
- Assumes the social ties are sampled at random

RDS-I Heckathorn et al. Estimators

- Works well for:
 - Differential activity different from unity
 - High homophily
 - Small sample fraction
 - No seed bias
- Works poorly for:
 - Large sample fraction
 - Seed bias

M.J. Salganik and D.D. Heckathorn, “Sampling and estimation in hidden populations using respondent-driven sampling.” Sociological Methodology, 34, 193-239, 2004.

Gile, K.J., and M.S. Handcock, “Respondent-Driven Sampling: An Assessment of Current Methodology,” Sociological Methodology, 40, 2010, available on arXiv.

RDS-II: Heckathorn et al. Estimators

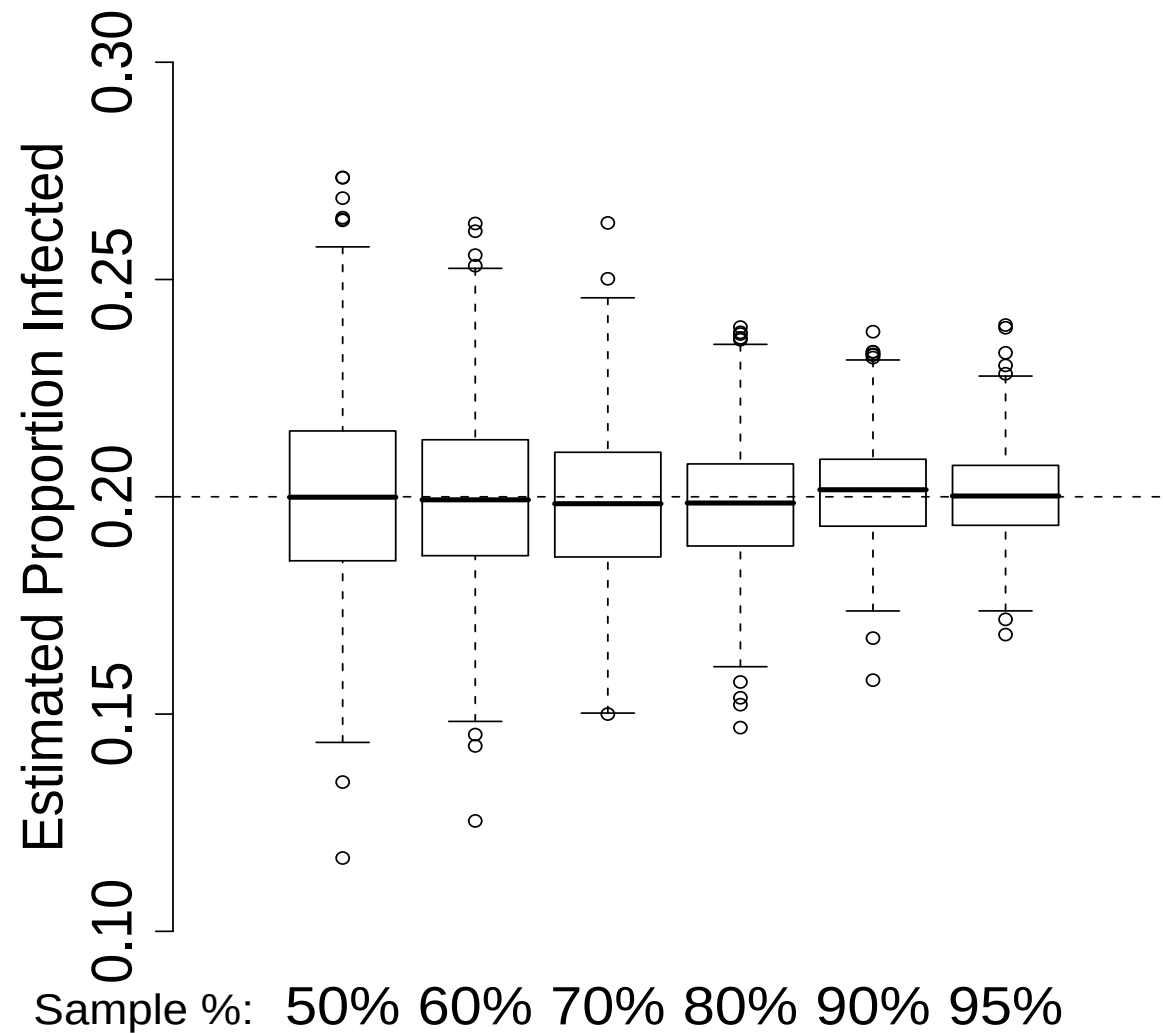
Use the nodal degree

- Based on infinite population approximation (and a few others)
- Assume sampling probability proportional to nodal degree.
- Works well for:
 - Small sample fraction
 - No seed bias
- Works poorly for:
 - Different mean degrees for infected and uninfected ($w \neq 1$)
 - Large sample fraction
 - Seed bias

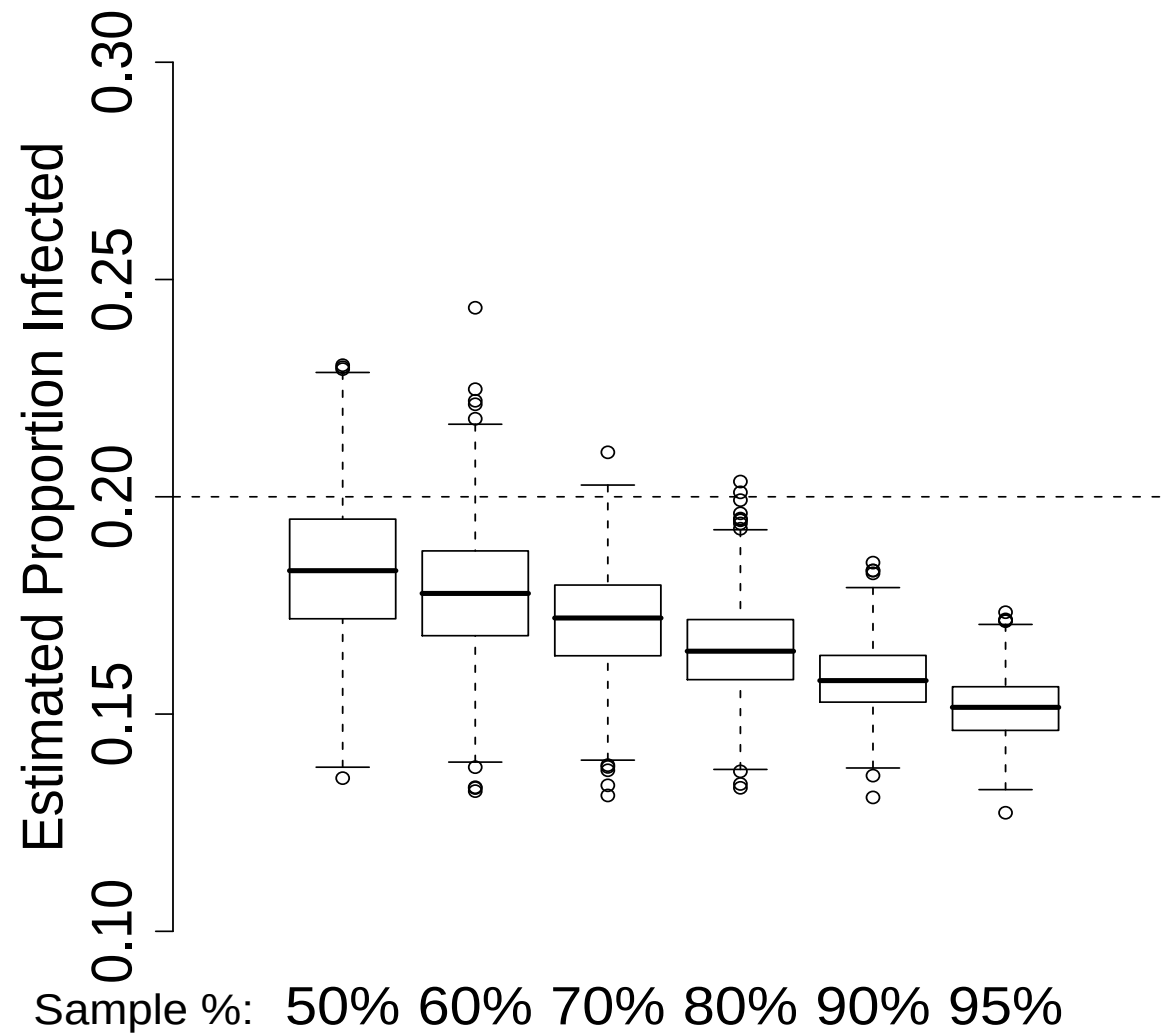
Erik Volz and Douglas D. Heckathorn, “Probability Estimation Theory for Respondent Driven Sampling,” Journal of Official Statistics, 24:1, 2008.

Gile, K.J., and M.S. Handcock, “Respondent-Driven Sampling: An Assessment of Current Methodology,” Sociological Methodology, 40, 2010, available on arXiv.

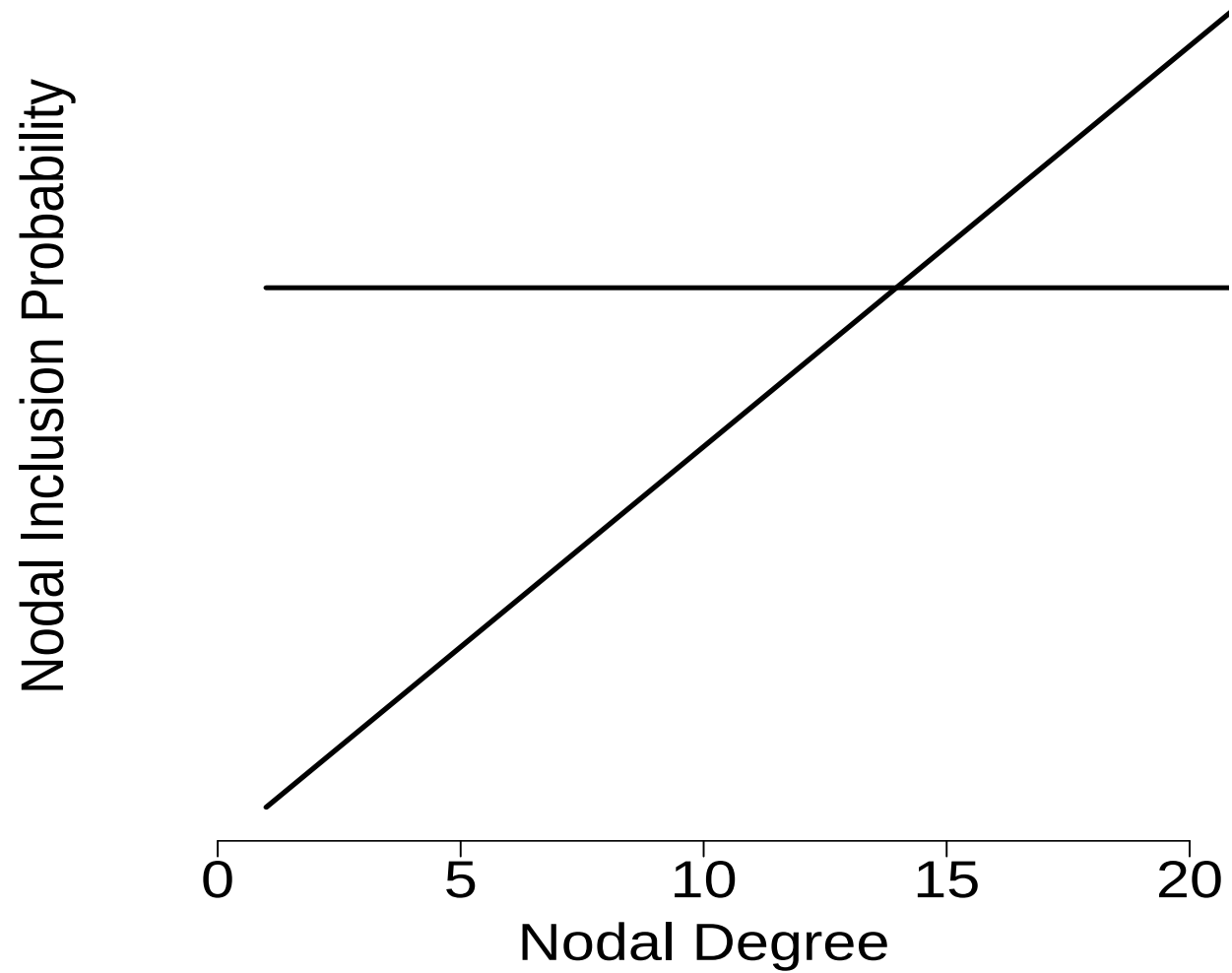
Volz-Heckathorn, No differential activity ($w=1$)



Varying Sample Percentage, Infected 40% more active ($w=1.4$)



Finite Population Bias



Outline of Presentation

1. Link-Tracing Hard-to-Reach Population Sampling
2. Respondent-Driven Sampling (RDS)
3. Inference for Respondent-Driven Sampling Data
4. Random Walk Approximation
5. Successive Sampling Approximation
6. Discussion

Finite Population Correction

Consider:

- A distribution uniform over all networks with given nodal degrees
- Then marginalizing over this distribution of networks, the transition probabilities of the random walk are very nearly proportional to degree

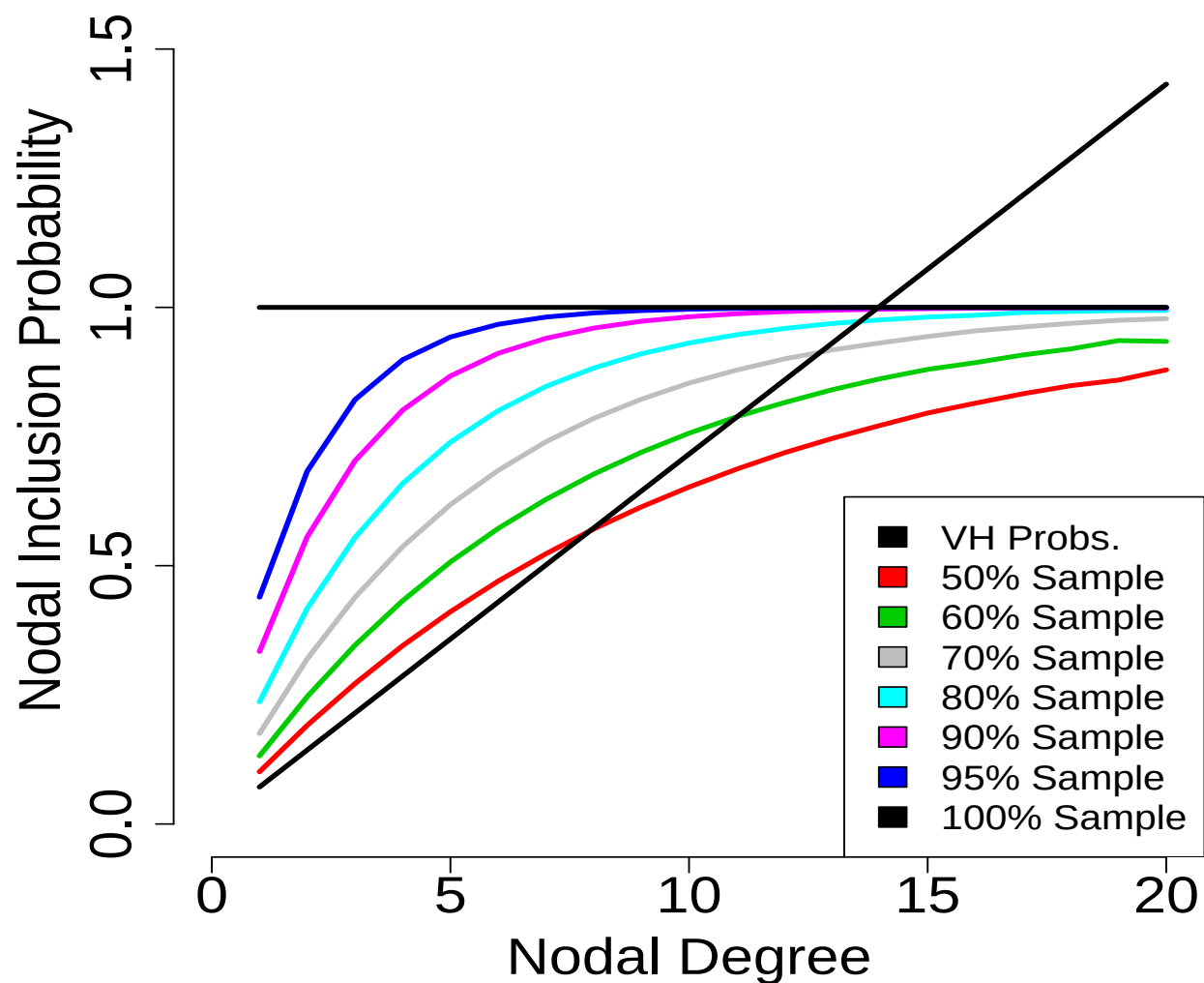
Furthermore, consider:

- 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 according to 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*, 106 (493), 135-146.

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 $i = 1 \dots r$:
 - (a) Estimate degree distribution $\mathbb{N}^i$ by Hajek Estimator
 - (b) Compute $\pi_k(\mathbb{N}^i)$ by simulation:
 - i. Simulate M SS samples from $\mathbb{N}^i$
 - ii.

$$\pi_k(\mathbb{N}^i) = \frac{\mathbb{E}[V_k; \mathbb{N}^i]}{\mathbb{N}_k^i} \approx \frac{U_k + 1}{M \cdot \mathbb{N}_k^i + 1},$$

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 μ :

$$\hat{\mu}_{SS} = \frac{\sum_i S_i \frac{z_i}{\hat{\pi}_{d_i}}}{\sum_i S_i \frac{1}{\hat{\pi}_{d_i}}}.$$

Modification in the presence of differential activity

- If z is highly correlated with d then Wu and Rao (2006) suggest using a maximum pseudo empirical likelihood estimator with the constraint:

$$\sum_i S_i P(z_i) \hat{\pi}_{d_i} = \frac{n}{N}$$

- We can estimate differential activity using the SS method.
- The Wu and Rao (2006) variant of $\hat{\mu}_{SS}$ performs slightly better when the infected are over 50% more active ($w > 1.5$).

A Super-population Framework for SS Estimator

- Assume nodal values are i.i.d. from an (unknown) distribution.
- Specifically, a non-parametric distribution over z and d .
- The observed data likelihood is then:

$$\mathcal{L}(\theta; \mathbf{d}, z) \equiv \sum_{\mathbb{N}} P_{\theta}(\mathbb{N}, \mathbf{d}, z) = \sum_{\mathbb{N}} P_{\theta}(\mathbb{N}) P(\mathbf{d}, z | \mathbb{N}) \quad (1)$$

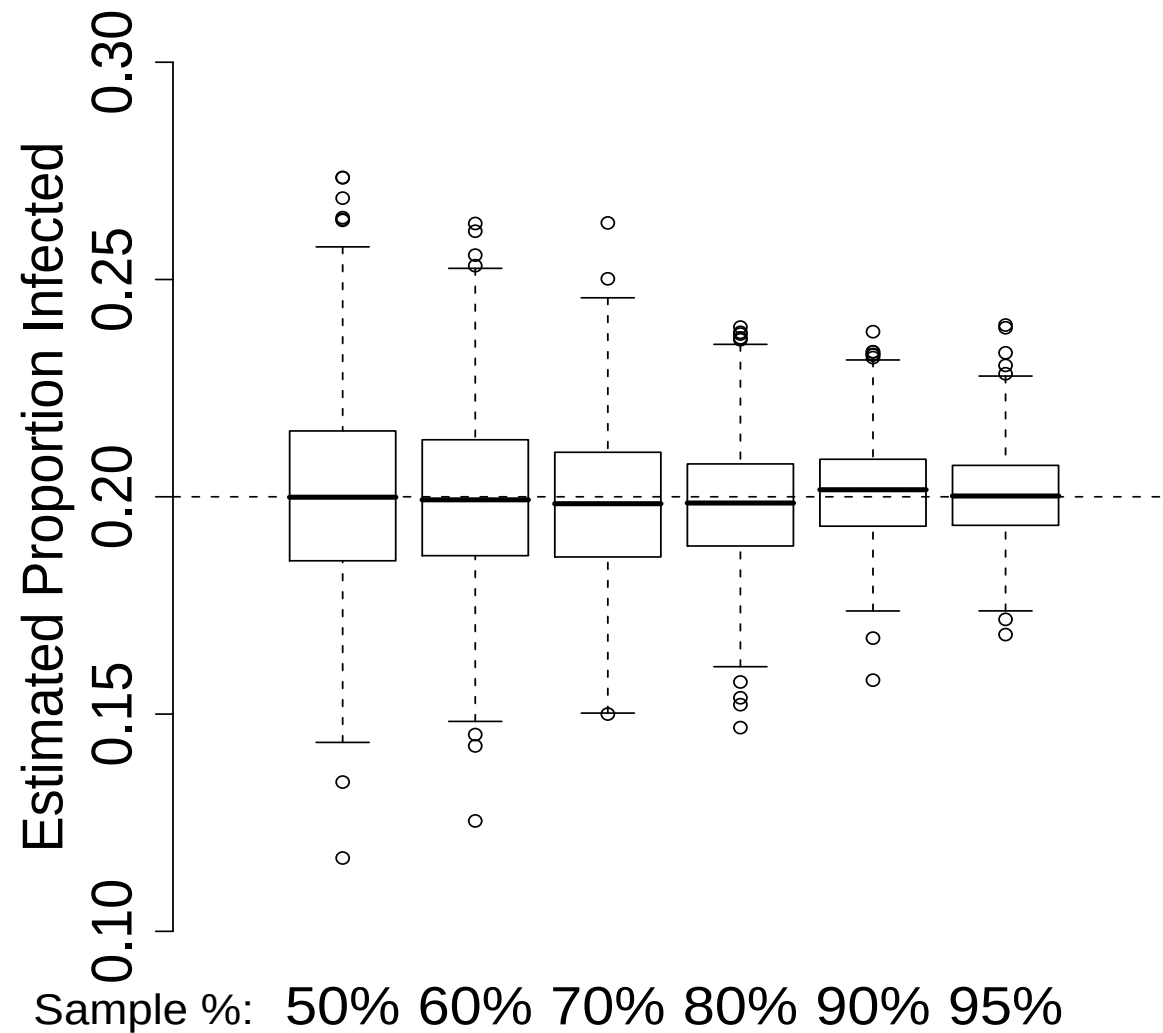
- The SS algorithm can then be viewed as an EM algorithm for the exponential-family (Sundberg 1976).

$$\mathbb{E}_{\theta^{(i)}} [\mathbb{N}] = \mathbb{E}_{\theta^{(i-1)}} [\mathbb{N} | \mathbf{d}, z] \quad (2)$$

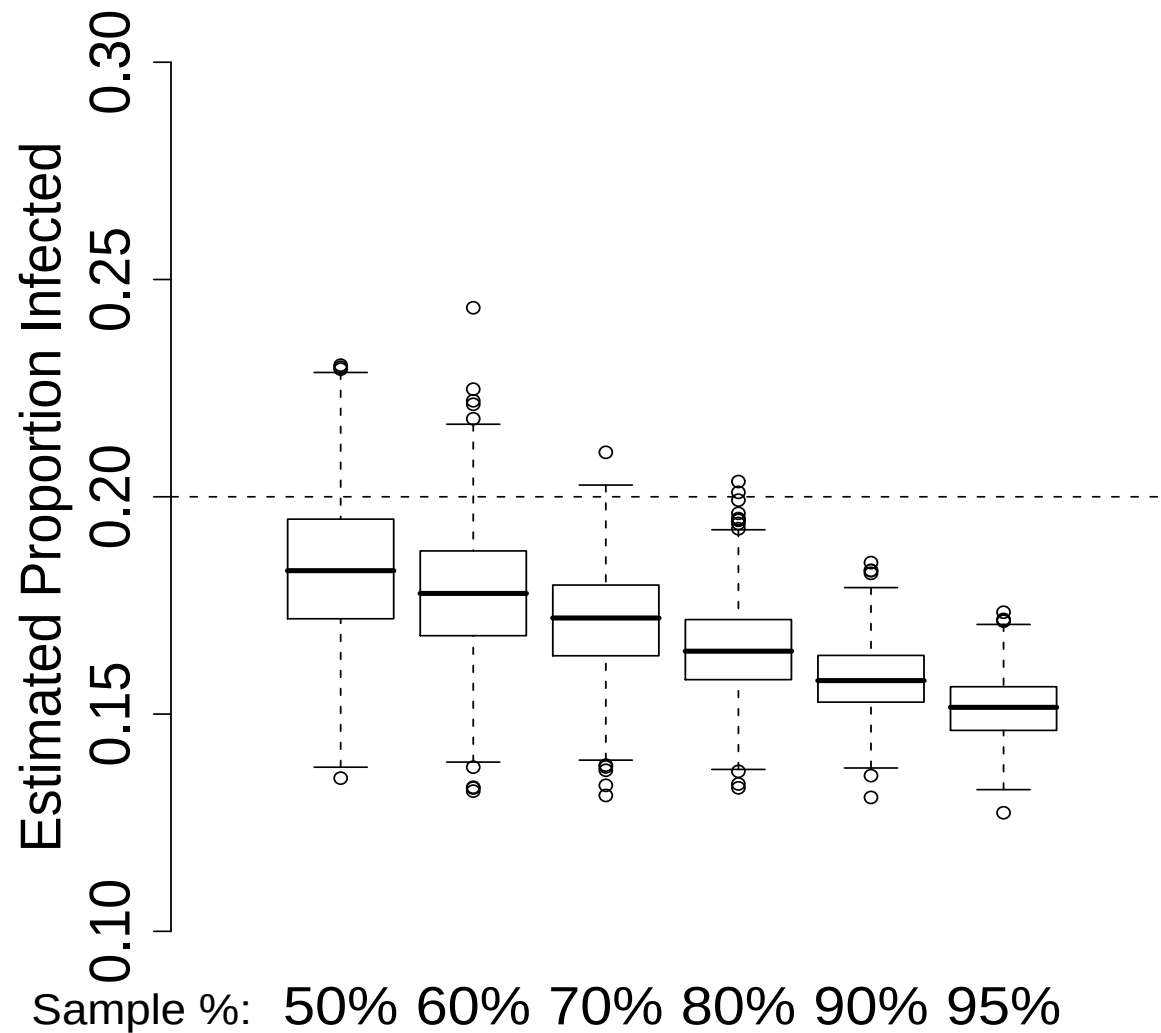
The SS algorithm makes the computational approximation:

$$\mathbb{E}_{\theta^{(i-1)}} [\mathbb{N}_k | \mathbf{d}, z] \approx \frac{N \frac{\mathbf{v}_k}{\hat{\pi}_k(\mathbb{N}^{(i-1)})}}{\sum_{l=1}^K \frac{\mathbf{v}_l}{\hat{\pi}_l(\mathbb{N}^{(i-1)})}}$$

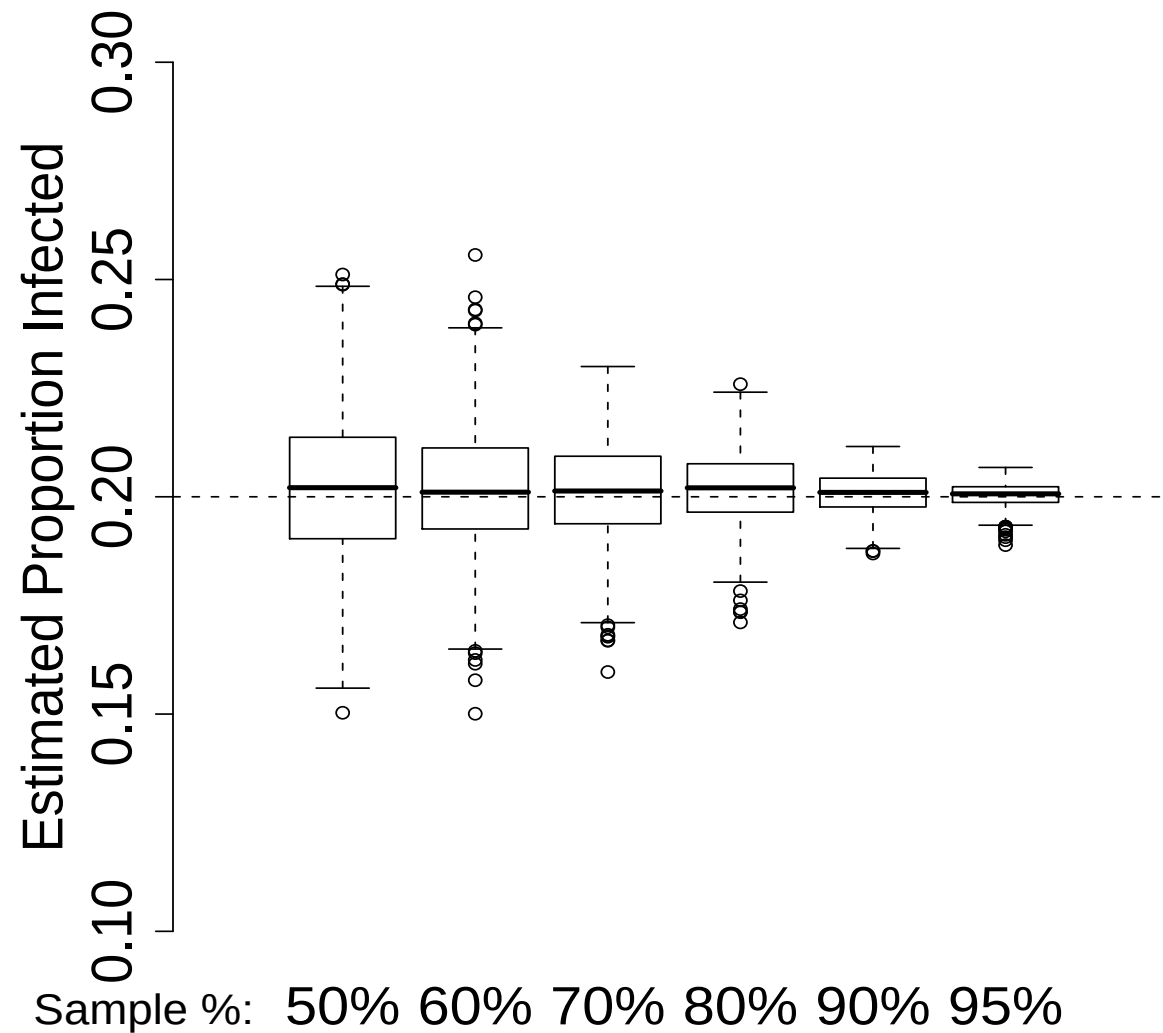
Volz-Heckathorn, No differential activity ($w=1$)



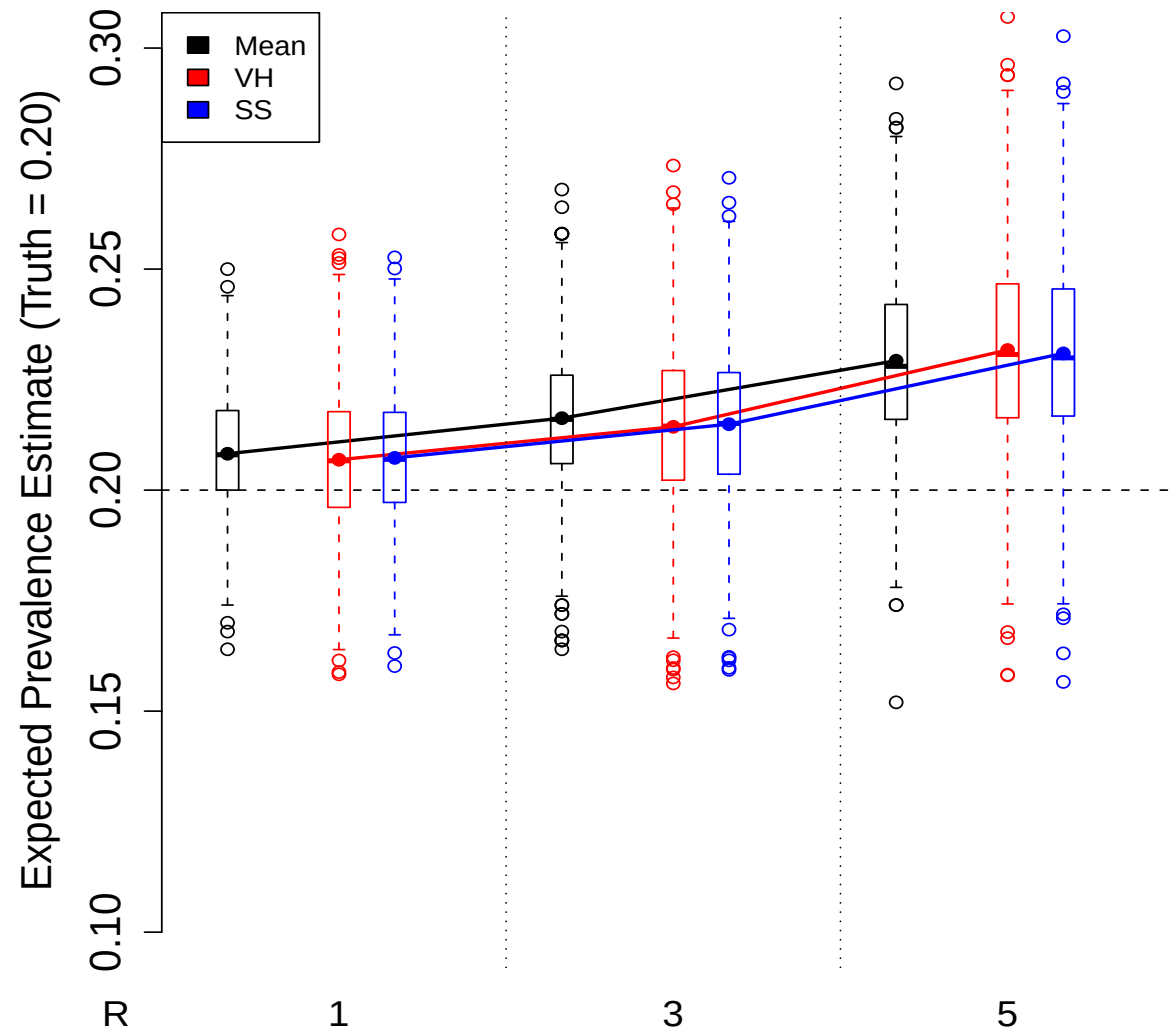
Volz-Heckathorn, Infected 40% more active ($w=1.4$)



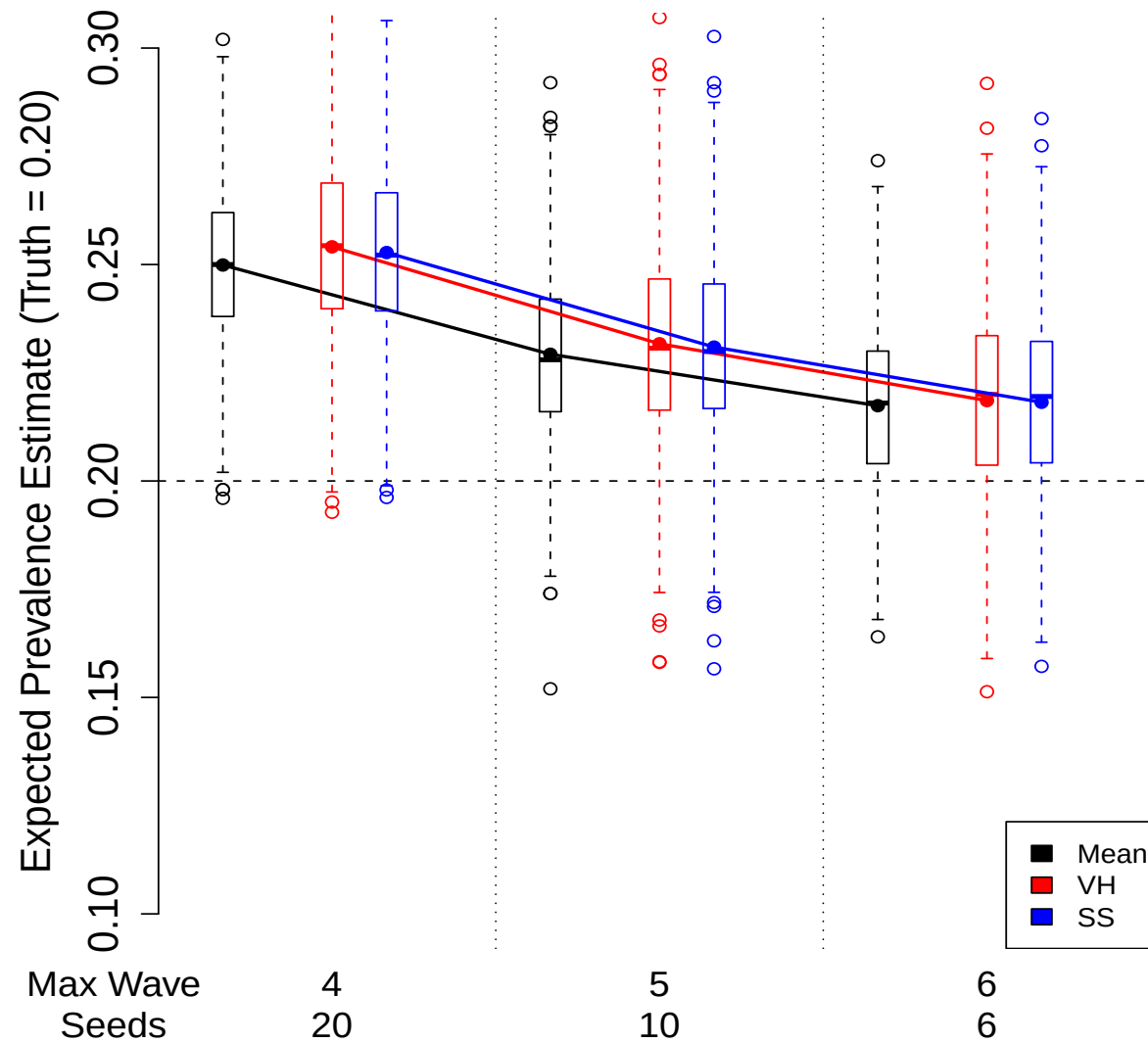
SS, Infected 40% more active ($w=1.4$)



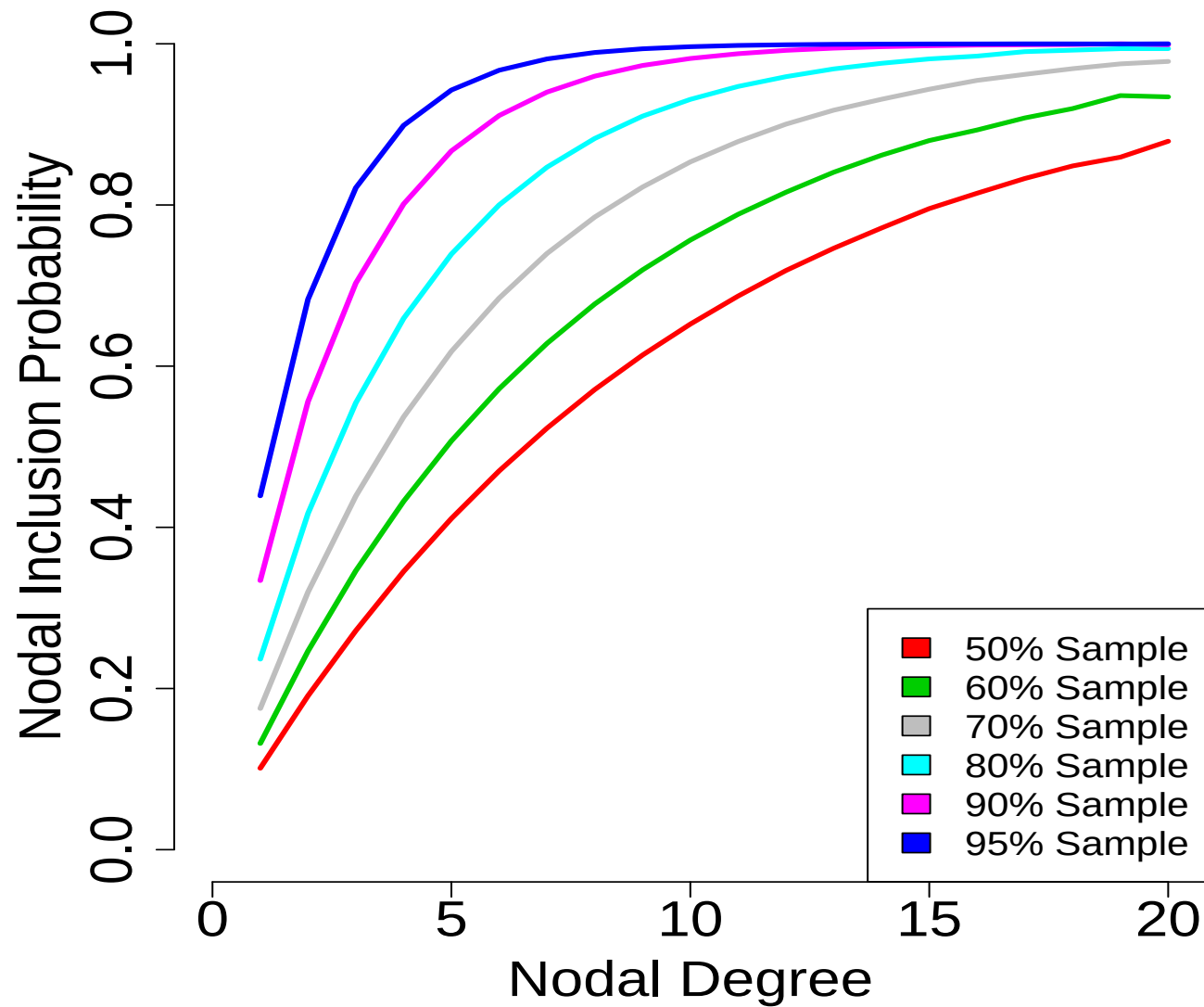
All Infected Seeds, varying Homophily, 50% sampled



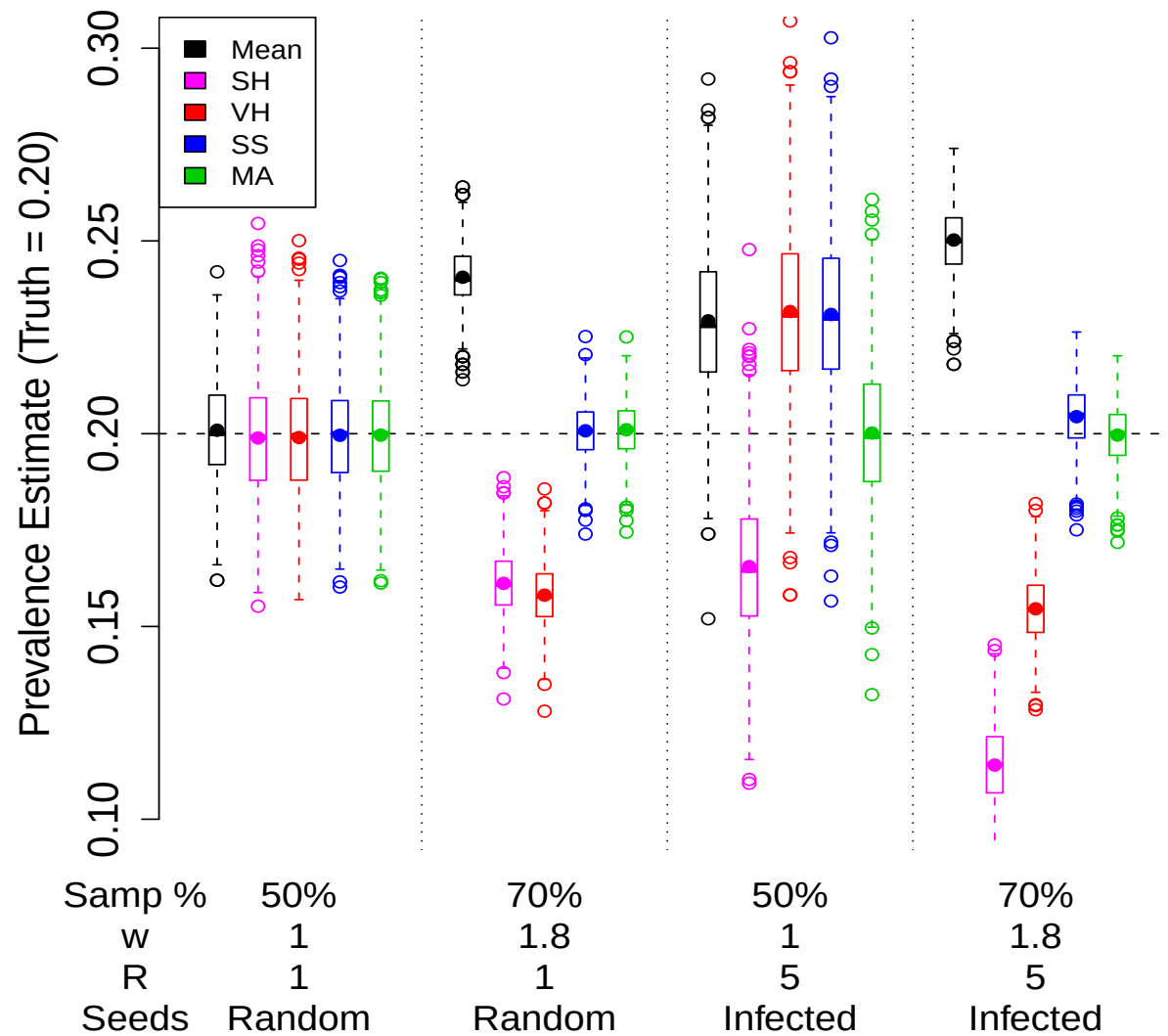
All Infected Seeds, varying number of seeds, 50% sampled



Mapping $Q_k(\mathbb{N}) : k \rightarrow \pi$



Discussion: New Estimators



Discussion: Respondent-Driven Sampling - Assumptions

	Network Structure Assumptions	Sampling Assumptions
Random Walk Model	Network size large ($N \gg n$)	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 ($N \gg n$)	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: Hard-to-Reach Population Sampling

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.

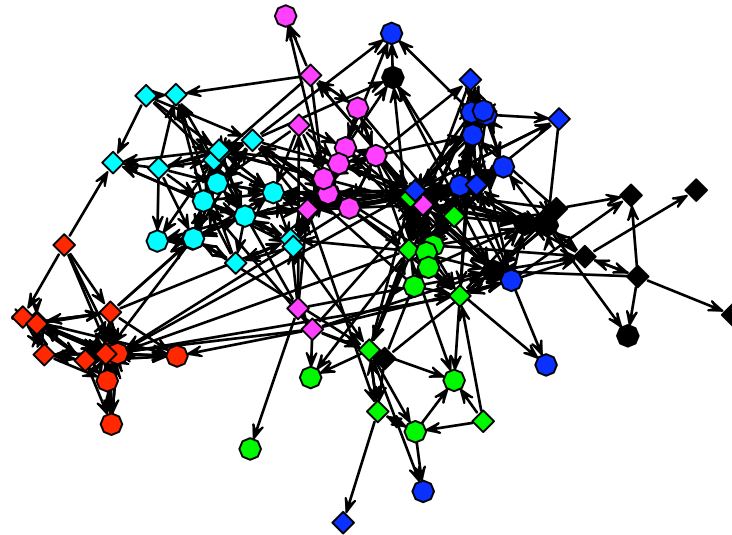
Discussion

- Challenges of Estimation from RDS:
 - Higher degree nodes more likely sampled
 - Degree-probability mapping depends on sample fraction
 - Seed characteristics may influence sample characteristics
 - Mediated by homophily and branching
- Successive Sampling Estimator:
 - Finite population corrections
 - Does not address homophily or Seed Bias
- Network Model-Assisted Estimator:
 - Addresses these concerns.
 - Unlike other Link-tracing methods, does not require initial probability sample
 - Still subject to many assumptions:
 - * Self-reported infected and uninfected contacts
 - * Known population size
 - * Adequate working network structure and sampling structure
 - * Measurement Error

References

- Bernhardt, A., D. Heckathorn, R. Milkman, and N. Theodore, “Documenting Unregulated Work: A Survey of Workplace Violations in New York City,” New York University, 2008.
- Frank, O. and T.A.B. Snijders, “Estimating the size of hidden populations using snowball sampling,” *Journal of Official Statistics*, 10, 1994.
- Gile, K. J. “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., and M.S. Handcock, “Respondent-Driven Sampling: An Assessment of Current Methodology,” *Sociological Methodology*, 40, 2010.
- Gile, K.J., and M.S. Handcock, “Network Model-Assisted Inference from Respondent-Driven Sampling Data.” *R&R Journal of the American Statistical Association*, 2011
- Handcock, M.S., D.R. Hunter, C.T. Butts, S.M. Goodreau, and M. Morris, “*statnet*: An R package for the Statistical Modeling of Social Networks.” , 2003.
- Heckathorn, D.D., and J. Jeffri, “Finding the Beat: Using Respondent-Driven Sampling to Study Jazz Musicians,” *Poetics*, 2000.
- Little, R.J. and D.B. Rubin, *Statistical Analysis with Missing Data.*, 2002.
- Salganik, M.J. and D.D. Heckathorn, “Sampling and estimation in hidden populations using respondent-driven sampling.” *Sociological Methodology*, 34, 193-239, 2004.
- Sarndal, C.E., B. Swensson, and J. Wretman, *Model Assisted Survey Sampling.*, 1992.
- Sugden, R. and T. Smith, “Ignorable and informative designs in survey sampling inference,” *Biometrika*, 71, 1984.
- Volz, E. and D. D. Heckathorn, “Probability Estimation Theory for Respondent Driven Sampling,” *Journal of Official Statistics*, 24:1, 2008.

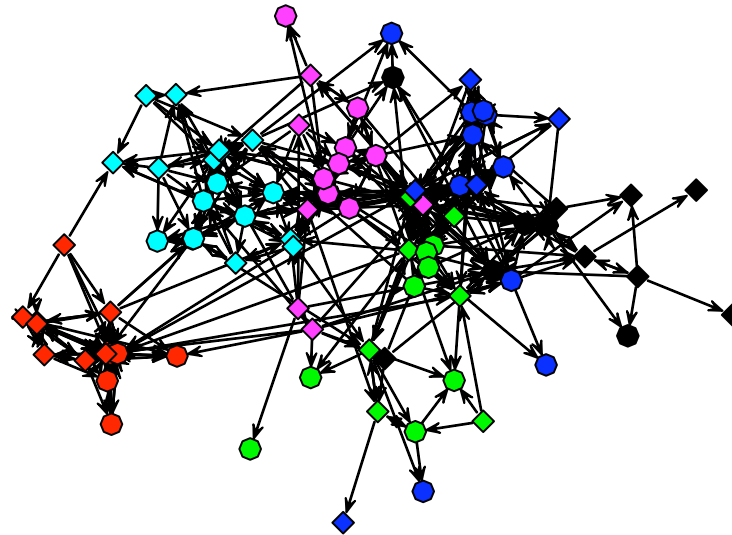
Link-tracing is (statistically) interesting



Sample following relations incident to nodes

- Sampling depends on Network
 - Sampling implicitly defined (adaptive). Network may be unknown
 - One sampled node may imply many sampled dyads
 - One sampled node may imply many more sampled nodes
 - Sample may depend heavily on initial sample
- Initial sample may be by unknown mechanism
- Population size may be unknown

Link-tracing is (statistically) interesting



Sample following relations incident to nodes

- Sampling depends on Network
 - Sampling implicitly defined (adaptive). Network may be unknown
 - One sampled node may imply many sampled dyads
 - One sampled node may imply many more sampled nodes
 - Sample may depend heavily on initial sample
- Initial sample may be by unknown mechanism
- Population size may be unknown

Comparison: Stratified Random Sample

From: Thompson, S.K., 1992 Sampling, New York: Wiley

- Sampling frame, strata
- Design sample within each stratum
- Known sampling probabilities used for inference

Stratified Random Sample: Design-Based Inference

Want to estimate

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

for population size N . Then sampling probability

$$\pi_i = \frac{n_{k_i}}{N_{k_i}}$$

N_{k_i} and n_{k_i} population and sample of strata k , to which i belongs. Then Horvitz-Thompson estimator:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^n \frac{x_i}{\pi_i}$$

is unbiased for μ .

Requires π_i for all sampled units.

Stratified Random Sample: Likelihood Inference

Observed Data: $X_i, i \in 1 \dots n$, AND $S_i, i \in 1 \dots N$, where $S_i = 1$ if unit i sampled.

Assume a model:

$$X_i = \mu + \beta_{k_i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

for k_i the strata of unit i . Parameter $\theta = \{\mu, \beta, \sigma\}$.

Inference based on:

$$\begin{aligned} L(\theta|S, X) &\propto P(S, X|\theta) &&= P(S|X, \theta)P(X|\theta) \\ &&&= \frac{1}{\prod_{j=1}^K \binom{N_j}{n_j}} P(X|\theta) \\ &\propto P(X|\theta) \end{aligned}$$

This requires *missing at random* (MAR), or *amenable* pattern, such that:

$$P(S|X, \theta) = P(S|X_{obs}).$$

Three Challenges of Link-Tracing

	Stratified Random Sampling	Link-Tracing Sampling
Sampling Design	Fully specified in advance	Depends on Network
Initial Sample Mechanism	Fully specified in advance	Often convenience sample
Target population	Known, well-defined	Often unknown size

Challenge 1: Sampling depends on network

- Design-based challenge: how to get sampling probabilities
- Likelihood challenge: is $P(S|X, \theta) = P(S|X_{obs})$?

Sampling depends on network: design-based

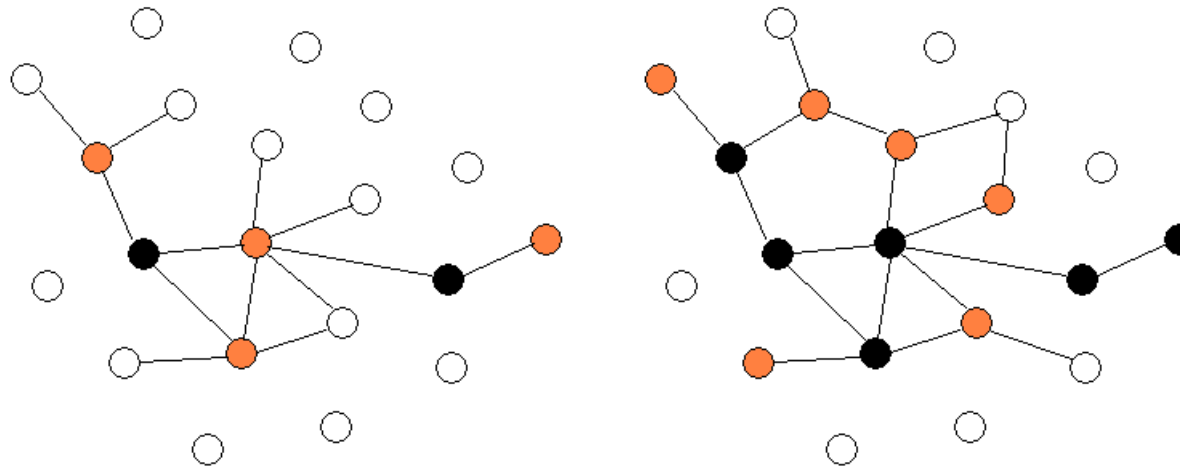
Simple Random Initial Sample. Observe all incident edges to sampled units:

$$\pi_i = 1 - \frac{\binom{N-m_i}{n}}{\binom{N}{n}}$$

Where m_i is the number possible initial units that would have resulted in i in the sample. Let y_{ij} indicate a tie between i and j . Then:

1-Wave: $m_i = 1 + \sum_{j \neq i} y_{ij}$ observed!

2-Waves: $m_i = 1 + \sum_{j \neq i} y_{ij} + \sum_{k \neq i} \sum_{j \neq i} y_{ik}(1 - y_{ij})y_{jk}$ not observed!



Sampling depends on network: design-based

Observable sampling probabilities:

Sampling Scheme	Nodal Probabilities π_i		Dyadic Probabilities π_{ij}	
	Undirected	Directed	Undirected	Directed
Simple Random	X	X	X	X
One-Wave	X			
k -Wave, $1 < k < \infty$				
Saturated	X			

- “X” indicates observable

Sampling Probabilities Unobserved for Many Simple Sampling Strategies

Snijders, T.A.B., 1992, “Estimation on the basis of snowball samples: how to weight.” Bulletin Methodologie Sociologique, 36, 59-70.

Handcock, M.S. and K.J. Gile, 2010, “Modeling social networks from sampled data.” , Annals of Applied Statistics, 4, Number 1, 5-25.

Sampling depends on network: likelihood

- Two types of data: Observed relations (Y_{obs}), and indicators of units sampled (S).

$$P(Y_{obs}, S|\theta) = \sum_{Unobserved} P(Y, S|\theta) = \sum_{Unobserved} P(S|Y, \theta)P(Y|\theta)$$

- θ is the model parameter

If $P(S|Y, \theta) = P(S|Y_{obs})$ (MAR), then $L(\theta|X, S) \propto \sum_{Unobserved} P(Y|\theta)$.

$$P(S|Y, \theta) = P(S|Y) = P(S_0|Y)P(S \setminus S_0|S_0, Y).$$

If all links followed to specified wave, $P(S \setminus S_0|S_0, Y) = \mathbb{I}\{S = s\}$.

Then require $P(S_0|Y) = P(S_0|Y_{obs})$. Any standard probability sampling method.

For many standard link-tracing designs, design *amenable* for likelihood inference.

Thompson, S.K. and O. Frank, 2000, “Model-based estimation with link-tracking sampling designs.” , Survey Methodology 26, 87-98.

Sampling depends on network: Approaches

1. Focus on Cases where probabilities observable (design-based)
2. Approximate sampling probabilities (design-based)
3. Treat amenable sample in likelihood frame

Sampling depends on network: Approaches

1. Focus on Cases where probabilities observable (design-based)
2. Approximate sampling probabilities (design-based)
3. Treat amenable sample in likelihood frame

Goodman, L.A., 1961 “*Snowball sampling.*,” *Annals of Mathematical Statistics*, 32, 148-70.

- Interested in counting reciprocals, triangles, cycles
- Trace links to get desired sample size
- More efficient than egocentric sample for the same number of nodes.

Sampling depends on network: Approaches

1. Focus on Cases where probabilities observable (design-based)
2. Approximate sampling probabilities (design-based)
3. Treat amenable sample in likelihood frame

Frank, O., 1971 The Statistical Analysis of Networks, London: Chapman and Hall

Frank, O., 2005 “Network sampling and model fitting” , in Carrigan, J.S.P., and Wasserman, S.S. (Eds.), Models and Methods in Social Network Analysis, Cambridge: Cambridge University Press.

- Given a network sampled through link-tracing, how to estimate features of network and population
- Sampling probabilities from special cases: one-wave, saturated, known network...

Sampling depends on network: Approaches

1. Focus on Cases where probabilities observable (design-based)
2. Approximate sampling probabilities (design-based)
3. Treat amenable sample in likelihood frame

*Thompson, S.K., 1990, “[Adaptive Cluster Sampling](#),” *Journal of the American Statistical Association*, 85, 1050-9.*

Thompson, S.K., 1992 Sampling, New York: Wiley

Thompson, S.K., and G.A.F. Seber, 1996 Adaptive sampling, New York: Wiley

- Saturated sampling
- Increase efficiency of estimation over simple random sampling

Sampling depends on network: Approaches

1. Focus on Cases where probabilities observable (design-based)
2. Approximate sampling probabilities (design-based)
3. Treat amenable sample in likelihood frame

Thompson, S.K., 1990, “Adaptive Cluster Sampling,” Journal of the American Statistical Association, 85, 1050-9.

Sampling depends on network: Approaches

1. Focus on Cases where probabilities observable (design-based)
2. [Approximate sampling probabilities \(design-based\)](#)
3. Treat amenable sample in likelihood frame

Salganik, M.J. and D.D. Heckathorn, 2004, “Sampling and estimation in hidden populations using [respondent-driven sampling](#).” Sociological Methodology, 34, 193-239.

Volz, E. and D. D. Heckathorn, 2008, “Probability Estimation Theory for [Respondent Driven Sampling](#),” Journal of Official Statistics, 24, 79-97.

- Treat sampling process as random walk on nodes.
- Stationary distribution probabilities proportional to degree.

Sampling depends on network: Approaches

1. Focus on Cases where probabilities observable (design-based)
2. [Approximate sampling probabilities](#) (design-based)
3. Treat amenable sample in likelihood frame

*Gile, K. J., 2010, “Improved Inference for [Respondent-Driven Sampling](#) Data with Application to HIV Prevalence Estimation,” *Journal of the American Statistical Association*, 106 (493), 135-146.*

- *Treat sampling process as successive sampling (PPSWOR) with sizes given by degrees.*
- *Estimate corresponding sampling probabilities.*

Sampling depends on network: Approaches

1. *Focus on Cases where probabilities observable (design-based)*
2. *Approximate sampling probabilities (design-based)*
3. *Treat amenable sample in likelihood frame*

Thompson, S.K. and O. Frank, 2000, “*Model-based estimation with link-tracking sampling designs.*” , *Survey Methodology* 26, 87-98.

Chow, M., and S.K. Thompson, 2003, “*Estimation with link-tracing sampling designs - a Bayesian approach*” , *Survey Methodology* 20, 197-205.

Handcock, M.S. and K.J. Gile, 2010, “*Modeling social networks from sampled data.*” , *Annals of Applied Statistics*, 4, Number 1, 5-25.

Thompson, S.K., 2006, “*Adaptive Web Sampling,*” *Biometrics*, 62, 1224-34.

Challenge 2: Unknown initial sample mechanism

Consider a hard-to-reach population, e.g. injecting drug users, or pages on the internet

- *Design-based challenge: how to get sampling probabilities*
- *Likelihood challenge: is $P(S|X, \theta) = P(S|X_{obs})$?*

Unknown initial sample: design-based

For initial sample S_0 , such that $S_{0j} = 1 \iff j$ in initial sample, define

$$M_{ij} = \begin{cases} 1 & S_{0j} = 1 \implies S_i = 1 \\ 0 & \text{else,} \end{cases}$$

determined by the network and sampling design.

Then

$$\pi_i = P(S_i > 0) = P\left(\sum_{j=1}^N M_{ij} S_{0j} > 0\right).$$

So π_i depends on the distribution of S_0 .

Unknown initial sample: likelihood

- *Two types of data: Observed relations (Y_{obs}), and indicators of units sampled (S).*

$$P(Y_{obs}, S|\theta) = \sum_{Unobserved} P(Y, S|\theta) = \sum_{Unobserved} P(S|Y, \theta)P(Y|\theta)$$

- *θ is the model parameter*

If $P(S|Y, \theta) = P(S|Y_{obs})$ (MAR), then $L(\theta|X, S) \propto \sum_{Unobserved} P(Y|\theta)$.

$$P(S|Y, \theta) = P(S|Y) = P(S_0|Y)P(S \setminus S_0|S_0, Y).$$

If all links followed to specified wave, $P(S \setminus S_0|S_0, Y) = \mathbb{I}\{S = s\}$.

Then require $P(S_0|Y) = P(S_0|Y_{obs})$. Any standard probability sampling method.

If $P(S_0|Y) \neq P(S_0|Y_{obs})$ (or unknown), not amenable for likelihood inference.

Unknown initial sample: Approaches

1. *Assume initial sample well-behaved*
2. *Assume initial sample design partially known*
3. *Assume many waves of sampling decrease dependence on initial sample*
4. *Condition on initial sample*

Unknown initial sample: Approaches

1. *Assume initial sample well-behaved*
2. *Assume initial sample design partially known*
3. *Assume many waves of sampling decrease dependence on initial sample*
4. *Condition on initial sample*

Most common. Won't dwell on.

Unknown initial sample: Approaches

1. *Assume initial sample well-behaved*
2. *Assume initial sample design partially known*
3. *Assume many waves of sampling decrease dependence on initial sample*
4. *Condition on initial sample*

Felix-Medina, M.H. and S.K. Thompson, 2004, “*Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations.*,” *Journal of Official Statistics*, 20, 19-38.

Felix-Medina, M.H. and P.E. Monjardin, 2006, “*Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations: A Bayesian-assisted approach.*,” *Survey Methodology*, 32, 187-95.

- *If part of the population is covered by a sampling frame, can still estimate population size.*
- *Requires sampling frame of venues*
- *Ignore ties within venue, assume cross-venue ties independent*

Unknown initial sample: Approaches

1. *Assume initial sample well-behaved*
2. *Assume initial sample design partially known*
3. *Assume many waves of sampling decrease dependence on initial sample*
4. *Condition on initial sample*

From: Felix-Medina, M.H. and S.K. Thompson, 2004, “Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations.,” Journal of Official Statistics, 20, 19-38.

Unknown initial sample: Approaches

1. *Assume initial sample well-behaved*
2. *Assume initial sample design partially known*
3. *Assume many waves of sampling decrease dependence on initial sample*
4. *Condition on initial sample*

Salganik, M.J. and D.D. Heckathorn, 2004, “Sampling and estimation in hidden populations using *respondent-driven sampling*.” *Sociological Methodology*, 34, 193-239.

Volz, E. and D. D. Heckathorn, 2008, “Probability Estimation Theory for *Respondent Driven Sampling*,” *Journal of Official Statistics*, 24, 79-97.

- *Treat sampling process as random walk on nodes.*
- *Stationary distribution independent of initial sample.*

Unknown initial sample: Approaches

1. *Assume initial sample well-behaved*
2. *Assume initial sample design partially known*
3. *Assume many waves of sampling decrease dependence on initial sample*
4. *Condition on initial sample*

Thompson, S.K., 2006, “*Targeted Random Walk Designs*,” *Survey Methodology*, 32, 11-24

- *Sample via random walk on nodes.*
- *Stationary distribution independent of initial sample.*
- *Control stationary distribution by transition design.*

Unknown initial sample: Approaches

1. *Assume initial sample well-behaved*
2. *Assume initial sample design partially known*
3. *Assume many waves of sampling decrease dependence on initial sample*
4. *Condition on initial sample*

Gile, K.J., and M.S. Handcock, 2010, “Network Model-Assisted Inference from Respondent-Driven Sampling Data” , manuscript

- *Condition on initial non-probability sample*
- *Fit network model*
- *Find self-consistent sampling probabilities and population characteristics given sample.*

Challenge 3: Unknown population

Consider a hard-to-reach population, e.g. injecting drug users, or pages on the internet

- *Design-based challenge: how to get sampling probabilities*
- *Likelihood challenge: is $P(S|X, \theta) = P(S|X_{obs})$? Can we fit model?*

Unknown population: design-based

$$\sum_{i=1}^N \pi_i = \sum_{i=1}^N S_i \pi_i + \sum_{i=1}^N (1 - S_i) \pi_i = n$$

- *Standard estimates require $\pi_i \forall i : S_i = 1$*
- *Knowing this implies $\sum_{i=1}^N (1 - S_i) \pi_i$ known*
- *Rarely this is known but $N - n$ unknown*

Typically, N unknown $\implies \pi_i$ unknown for many i .

Unknown population: likelihood

- *Two types of data: Observed relations (Y_{obs}), and indicators of units sampled (S).*

$$P(Y_{obs}, S|\theta) = \sum_{Unobserved} P(Y, S|\theta) = \sum_{Unobserved} P(S|Y, \theta)P(Y|\theta)$$

- *θ is the model parameter*

Suppose $P(S|Y, \theta) = P(S|Y_{obs})$, then

$$L(\theta|X, S) \propto \sum_{Unobserved} P(Y|\theta).$$

- *Many (most) network models defined for full network (e.g. Bernoulli model)*
- *$\sum_{Unobserved}$ difficult if N unknown (need N to marginalize).*

Network models hard to fit without N .

Unknown population: Approaches

1. *Assume N known*
2. *Estimate N*
3. *Ratio Estimator (design-based)*
4. *Condition on part of sample*

Unknown population: Approaches

1. *Assume N known*
2. *Estimate N*
3. *Ratio Estimator (design-based)*
4. *Condition on part of sample*

Most common. Won't dwell on.

Unknown population: Approaches

1. *Assume N known*
2. *Estimate N*
3. *Ratio Estimator (design-based)*
4. *Condition on part of sample*

*Frank, O. and T.A.B. Snijders, 1994, “Estimating the [size](#) of hidden populations using [snowball sampling](#),” *Journal of Official Statistics*, 10, 53-67.*

- *Repeated sampling through link-tracing gives information on population size*
- *Initial probability sample*
- *Treat distributions of numbers of re-sampled nodes (capture-recapture)*

Unknown population: Approaches

1. *Assume N known*
2. *Estimate N*
3. *Ratio Estimator (design-based)*
4. *Condition on part of sample*

Felix-Medina, M.H. and S.K. Thompson, 2004, “*Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations.*,” *Journal of Official Statistics*, 20, 19-38.

Felix-Medina, M.H. and P.E. Monjardin, 2006, “*Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations: A Bayesian-assisted approach.*,” *Survey Methodology*, 32, 187-95.

- *If part of the population is covered by a sampling frame, can still estimate population size.*
- *Requires sampling frame of venues*
- *Ignore ties within venue, assume cross-venue ties independent*

Unknown population: Approaches

1. *Assume N known*
2. *Estimate N*
3. *Ratio Estimator (design-based)*
4. *Condition on part of sample*

From: Felix-Medina, M.H. and S.K. Thompson, 2004, “Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations.,” Journal of Official Statistics, 20, 19-38.

Unknown population: Approaches

1. *Assume N known*
2. *Estimate N*
3. *Ratio Estimator (design-based)*
4. *Condition on part of sample*

Handcock, M.S., K.J. Gile, and C.M. Mar, 2010, “Estimating Hard-to-Reach Population Size using Respondent-Driven Sampling Data” , manuscript

- *Leverage assumed successive sampling approximation to sampling process to estimate N*
- *Strong assumptions about sampling process*
- *Leverage trends in sampled units over time to estimate population depletion*

Unknown population: Approaches

1. *Assume N known*
2. *Estimate N*
3. *Ratio Estimator (design-based)*
4. *Condition on part of sample*

Salganik, M.J. and D.D. Heckathorn, 2004, “Sampling and estimation in hidden populations using *respondent-driven sampling*.” *Sociological Methodology*, 34, 193-239.

Volz, E. and D. D. Heckathorn, 2008, “Probability Estimation Theory for *Respondent Driven Sampling*,” *Journal of Official Statistics*, 24, 79-97.

$$\hat{\mu} = \frac{\sum_i S_i \frac{x_i}{\pi_i}}{\sum_i S_i \frac{1}{\pi_i}}$$

- *Requires π_i only up to proportionality*
- *Random Walk stationary distribution probabilities proportional to degrees.*

Unknown population: Approaches

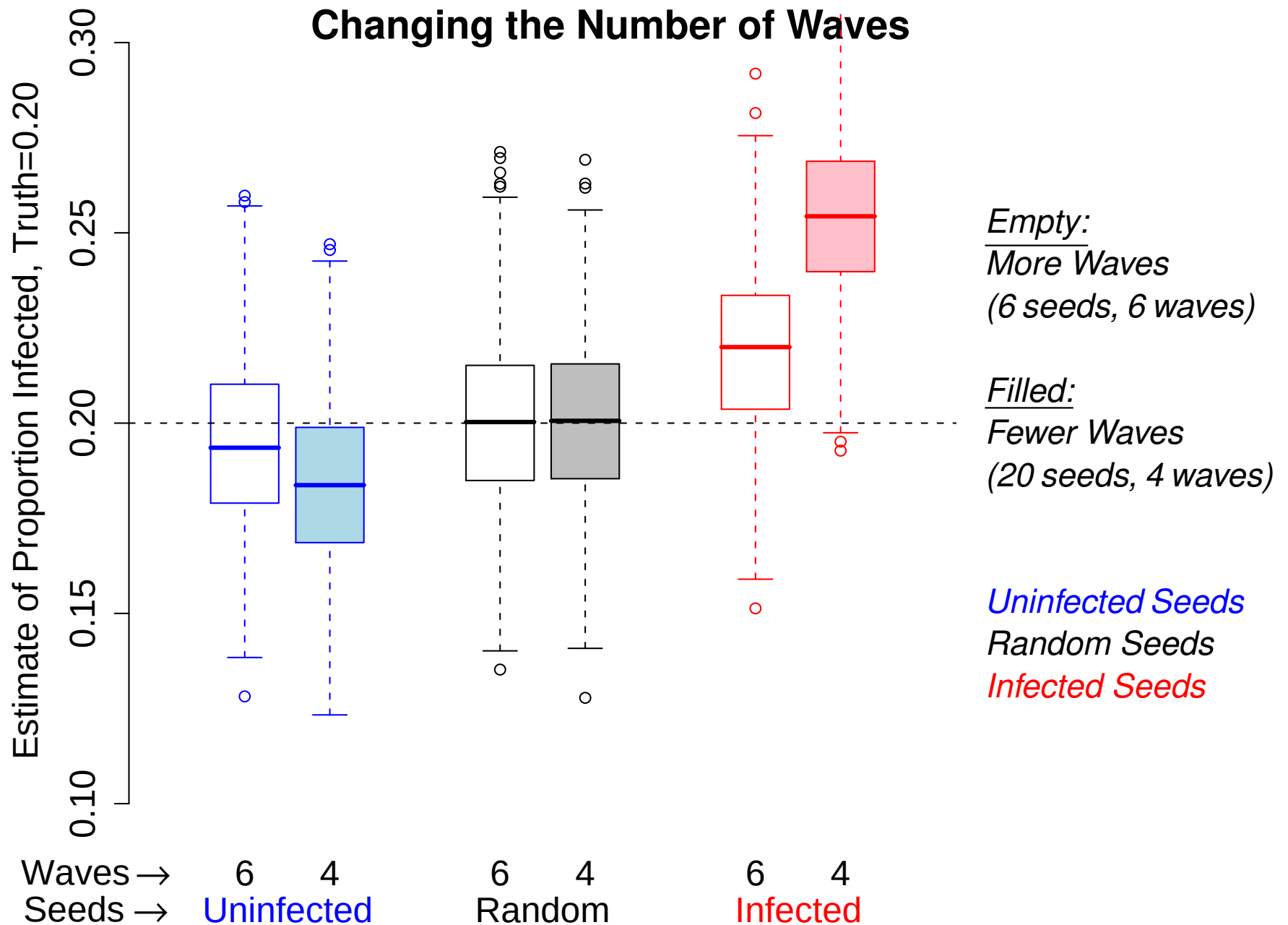
1. *Assume N known*
2. *Estimate N*
3. *Ratio Estimator (design-based)*
4. *Condition on part of sample*

Pattison, P., 2009, “Modelling large social networks: statistical issues.” , Presentation. Workshop Statistical Network Modeling, Nuffield College, Oxford, November, joint with G. Robins, G Daraganova, P. Wang, J. Koskinen, and T. Snijders.

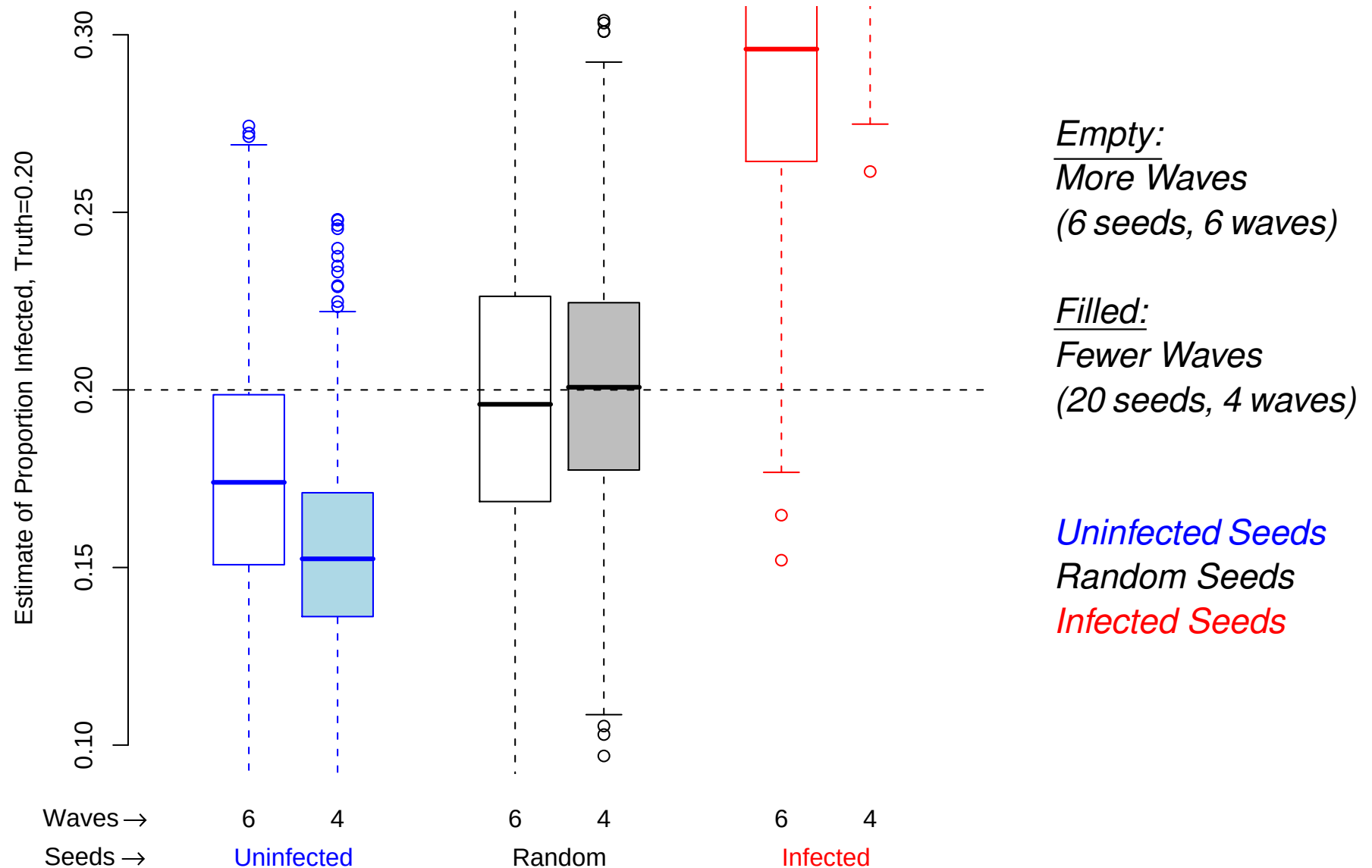
- *Exploit conditional independence feature of exponential random graph models (Snijders 2010)*
- *Fit network model to observed subset of data only, conditional on link-tracing boundary.*

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*



Changing the Number of Waves, With Replacement

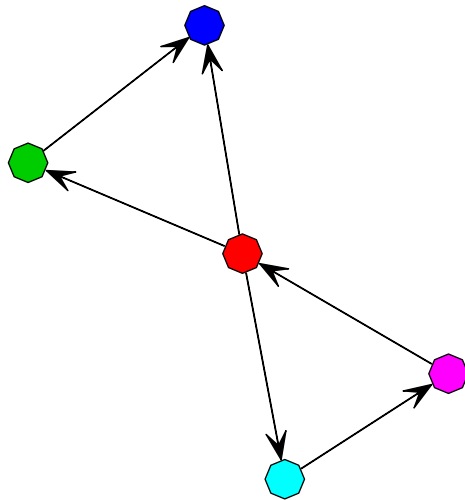


“Two Wrongs Make a Right:” Without replacement and list-wise sampling

- *Probabilities proportional to degree depends on Random Walk Model*
 - *With replacement*
 - *Draw-wise sampling probabilities (i.e. “at this step”)*
- *RDS violates both... and it helps!*
 - *Without replacement*
 - *List-wise sampling probabilities (i.e. “ever included”)*

Social Networks

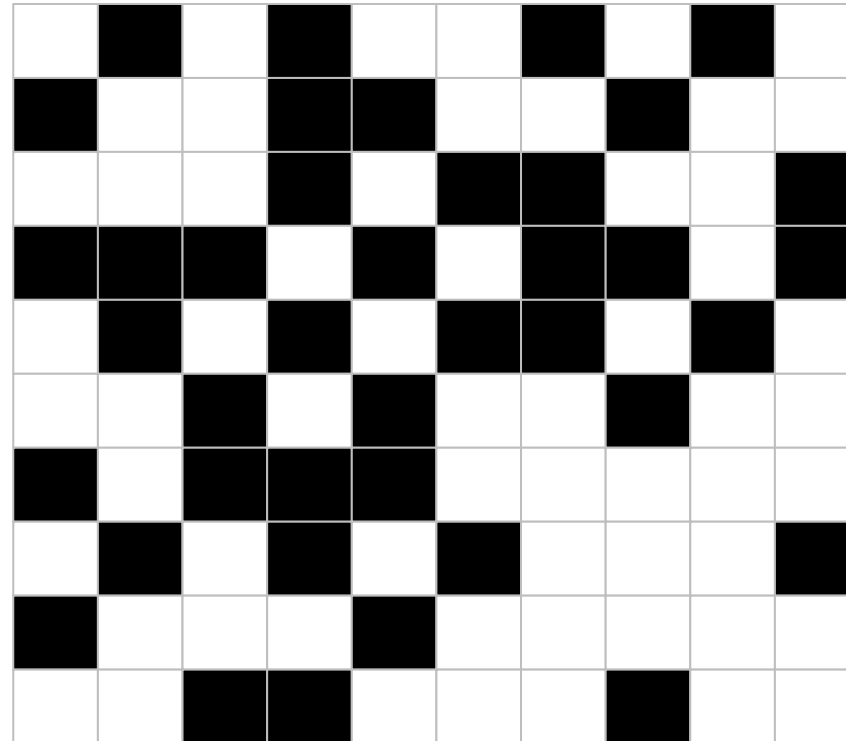
- *Social Network: Tool to formally represent and quantify relational social structure.*
- *Represent mathematically as a sociomatrix, Y , where y_{ij} = the value of the relationship from i to j*



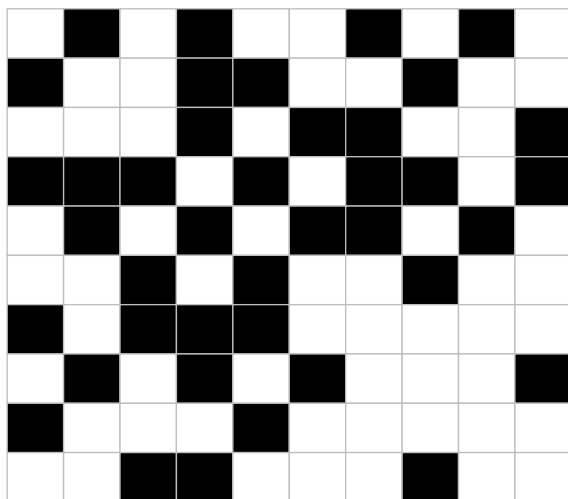
(a) Sociogram

	Red	Green	Blue	Cyan	Magenta
Red	0	1	1	1	0
Green	0	0	1	0	0
Blue	0	0	0	0	0
Cyan	0	0	0	0	1
Magenta	1	0	0	0	0

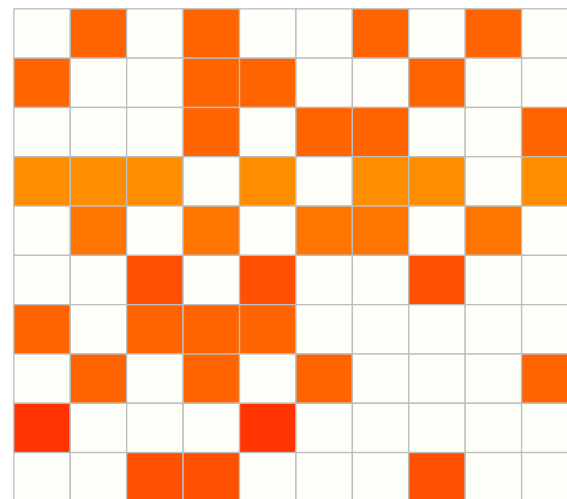
(b) Sociomatrix



0	1/16	1/8	3/16
1/4	5/16	3/8	7/16
1/2	9/16	5/8	11/16
3/4	13/16	7/8	15/16

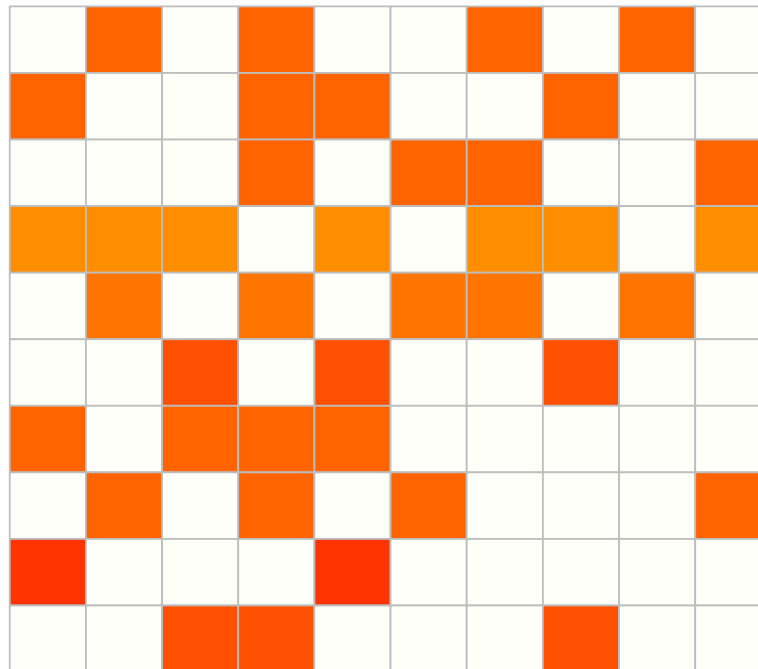


(c) Sociomatrix

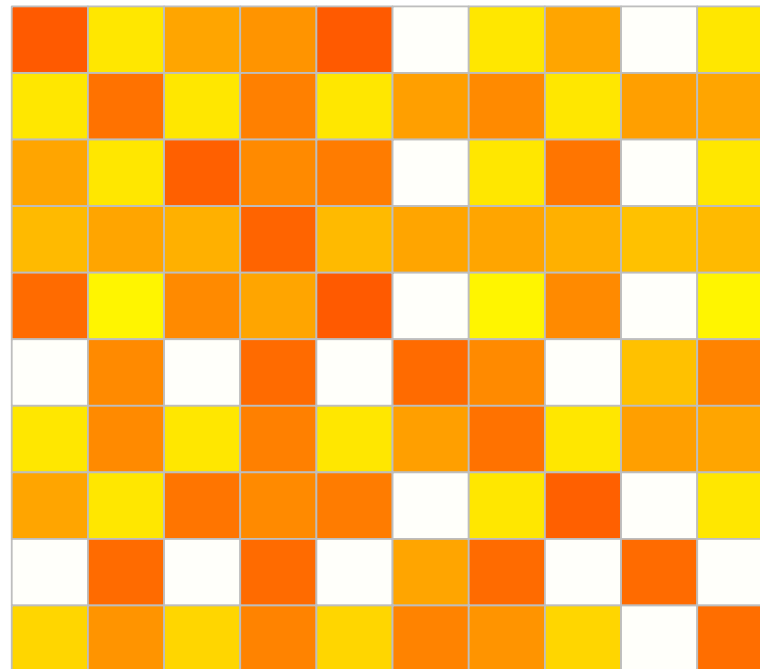


(d) First Wave Probabilities

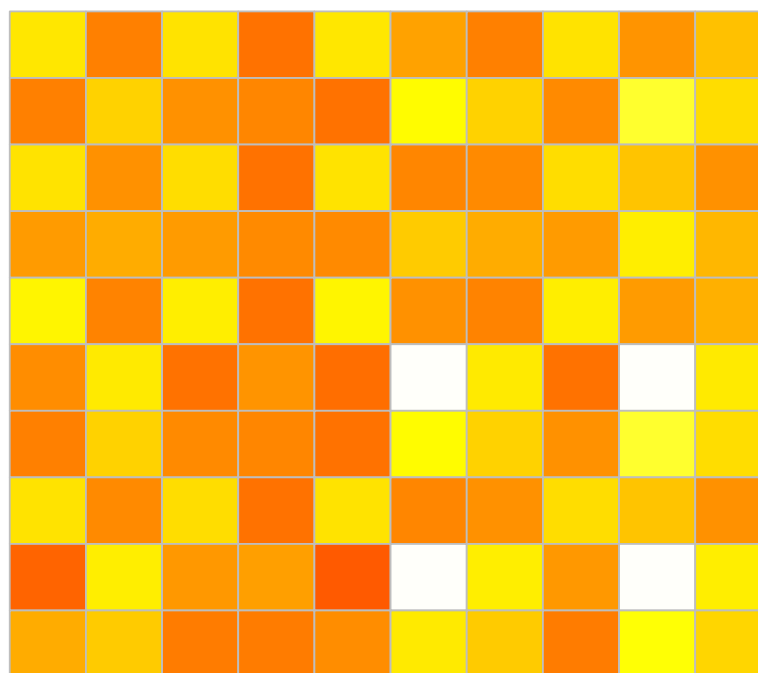
1



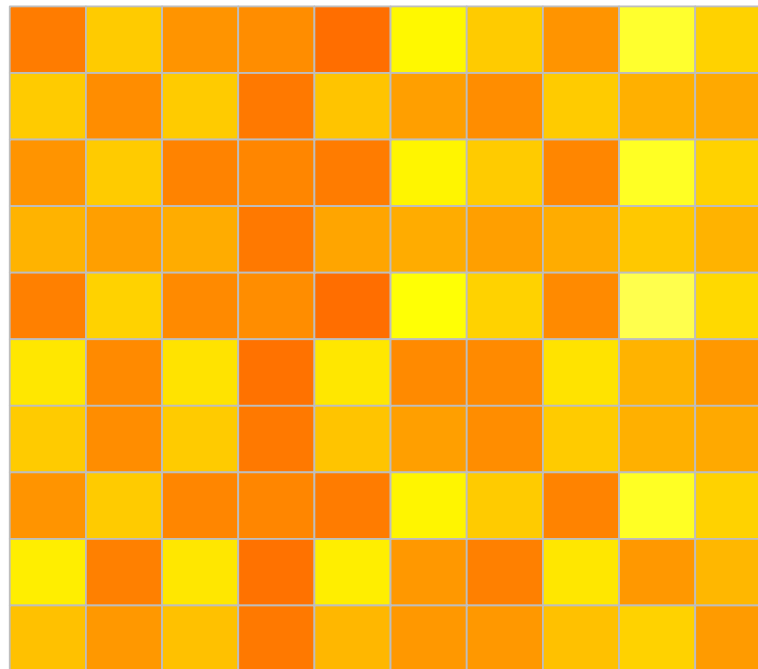
2



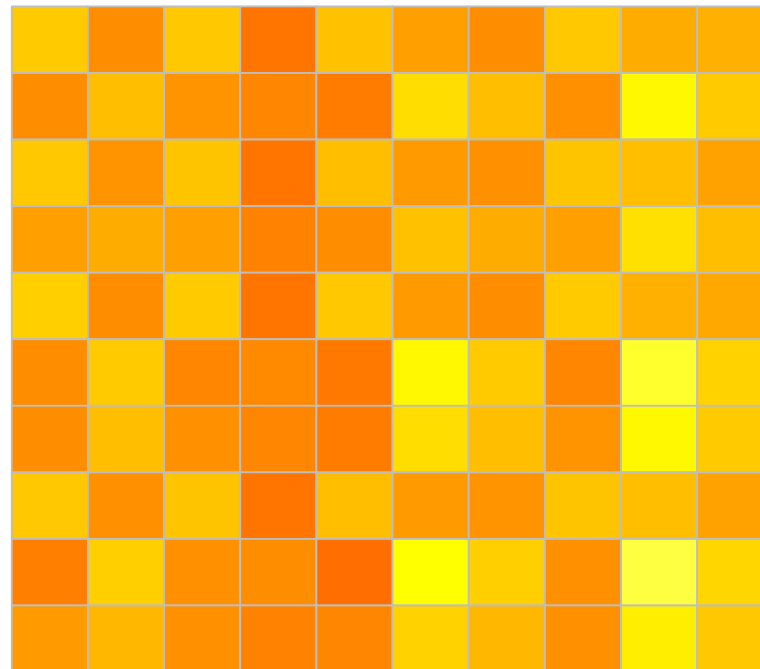
3



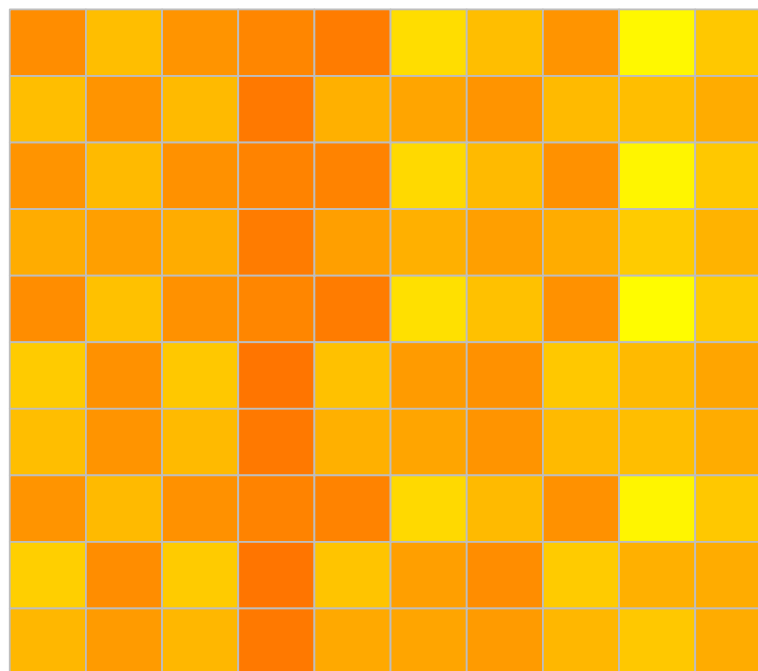
4



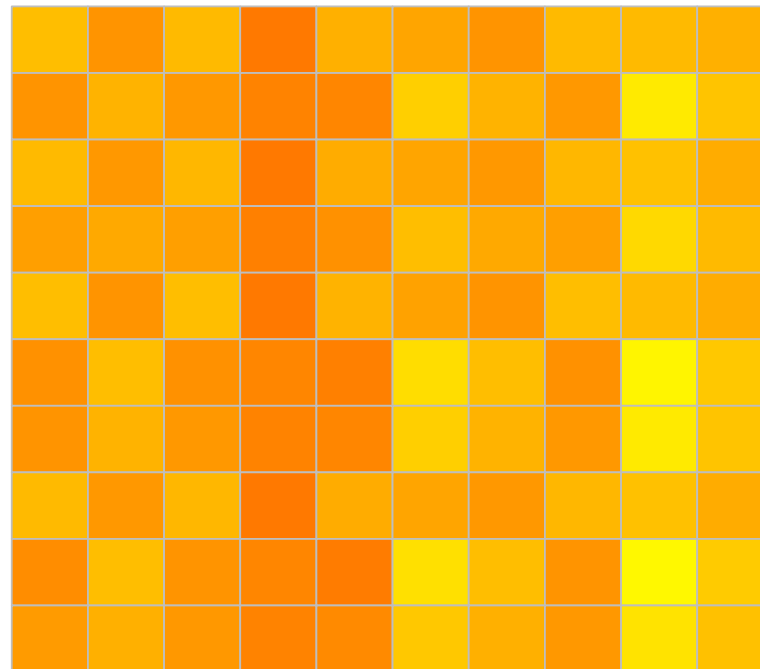
5



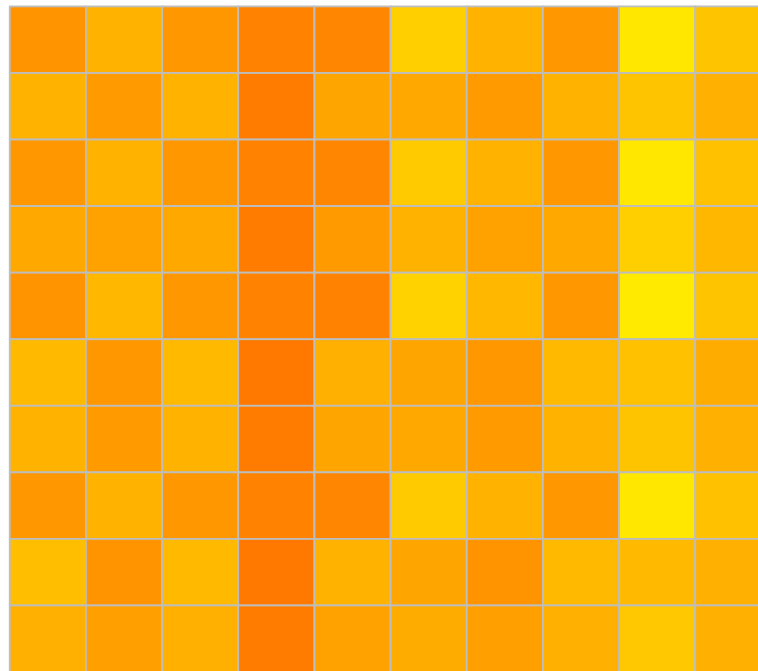
6



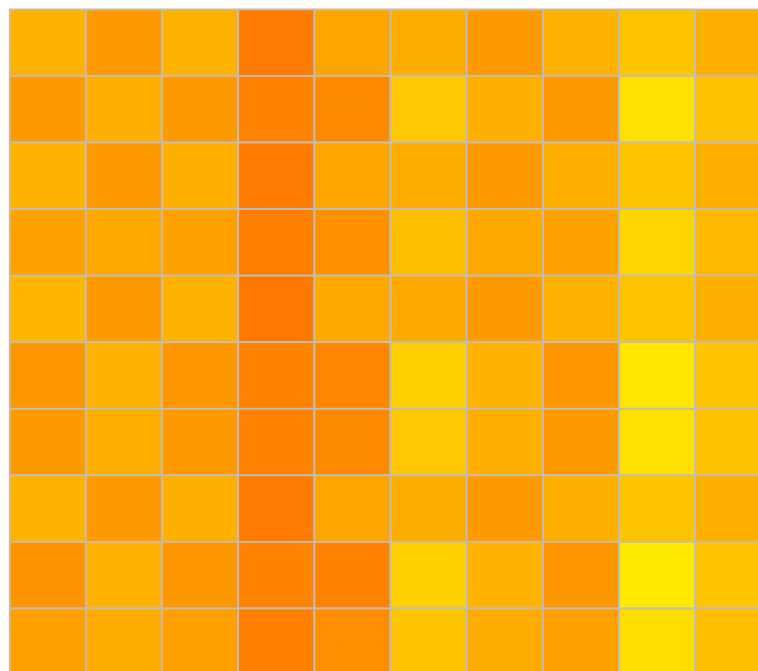
7



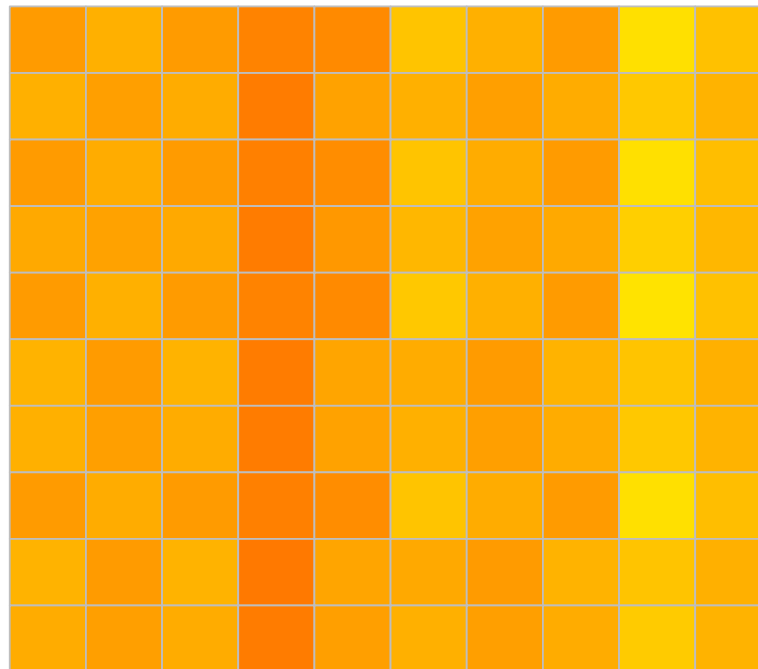
8



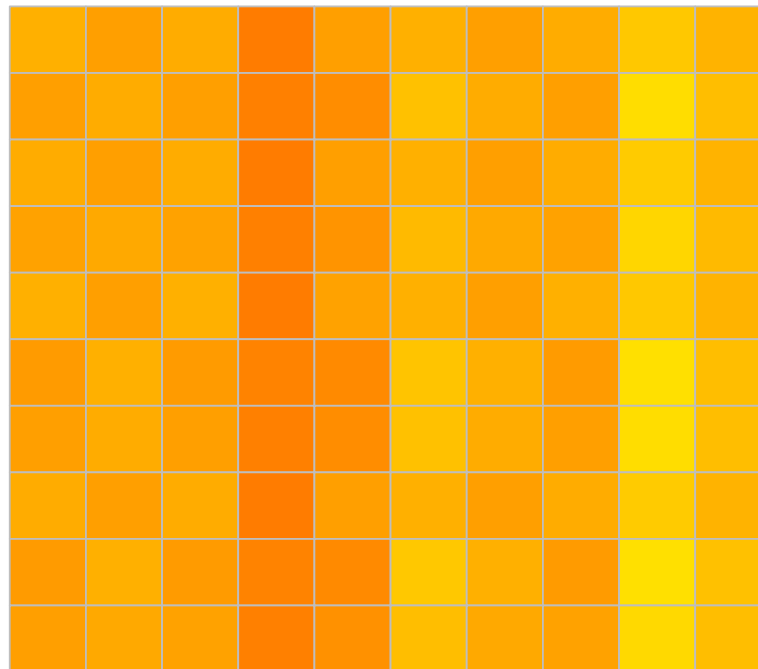
9



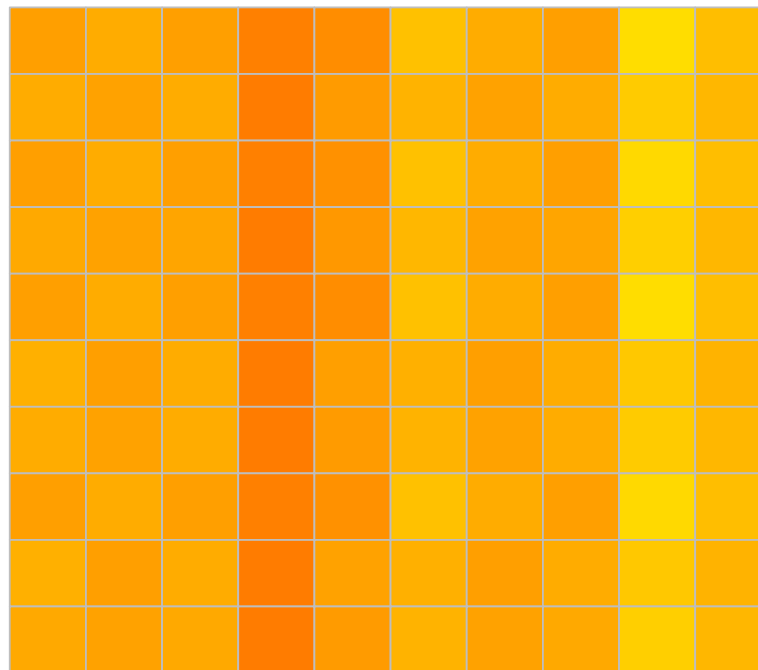
10



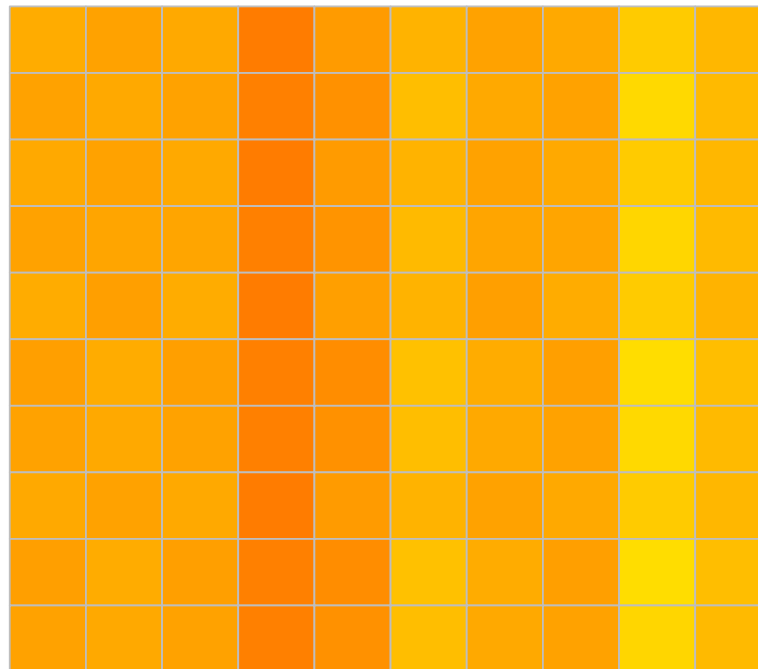
11



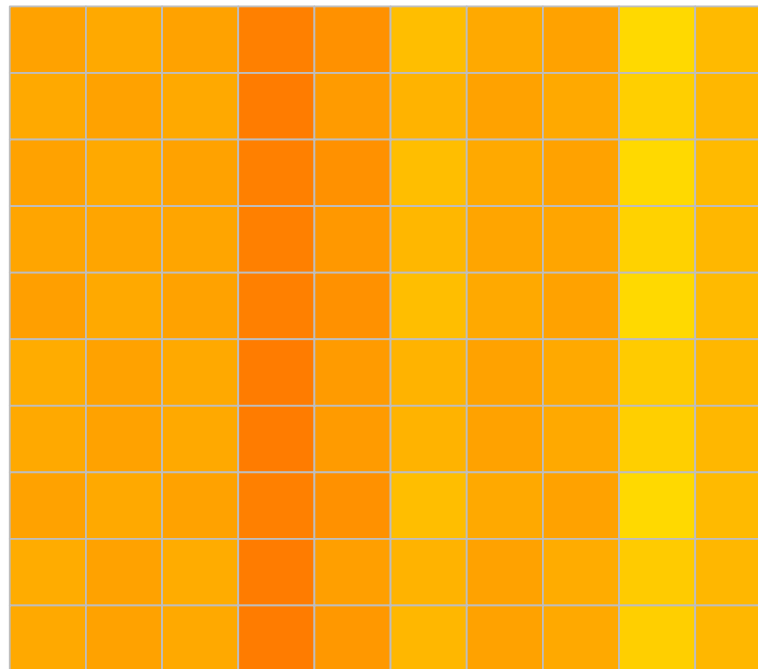
12



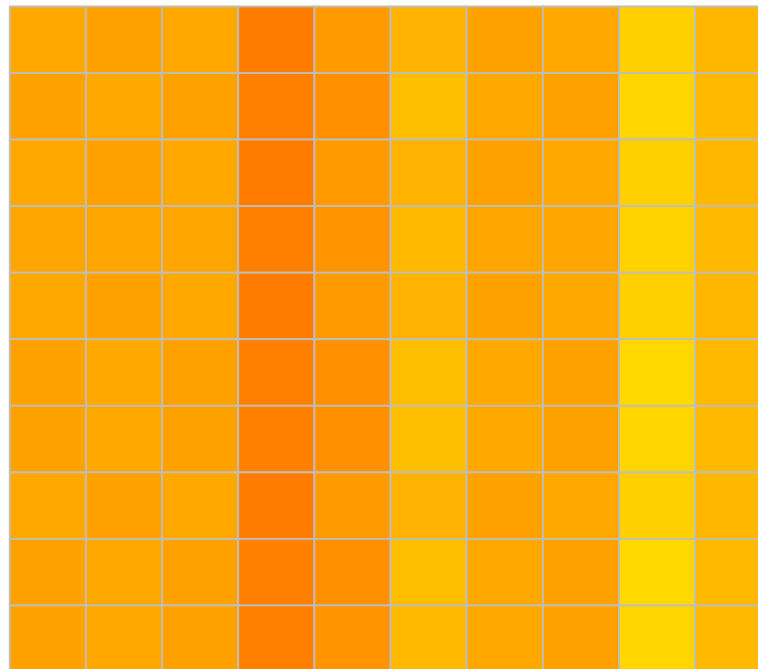
13

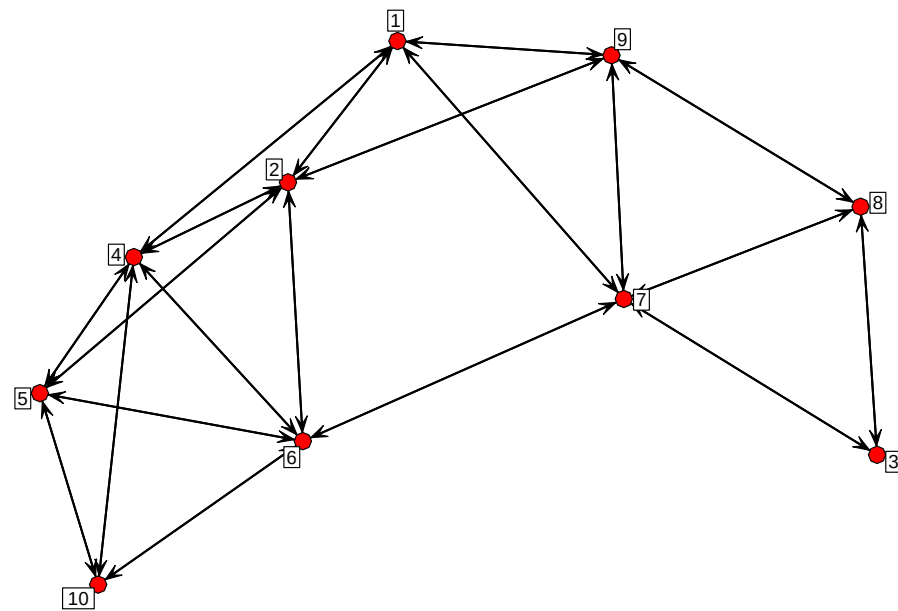


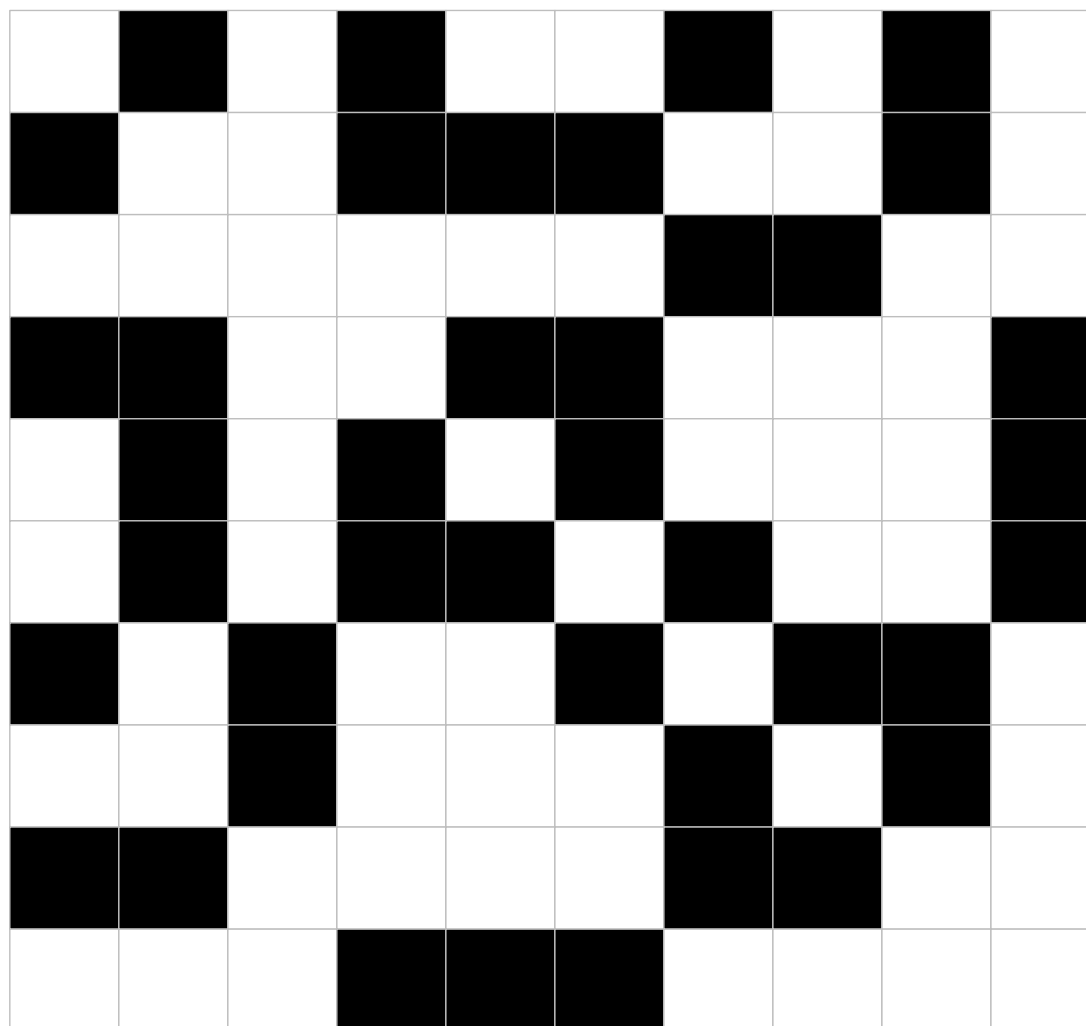
14

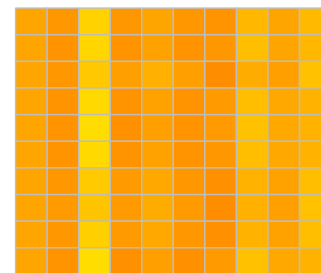
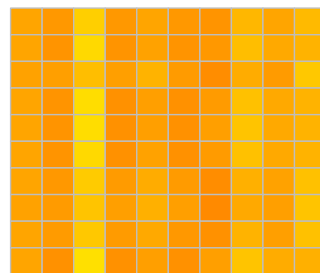
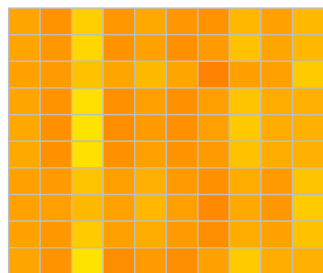
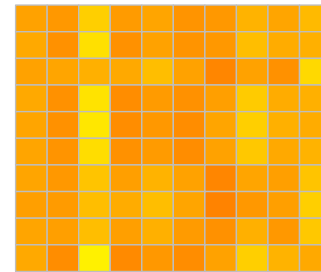
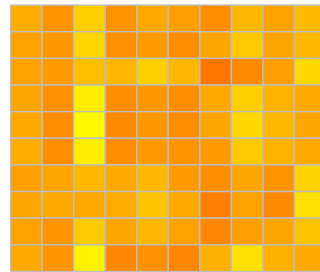
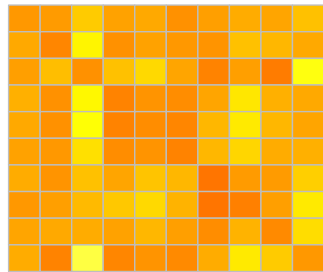
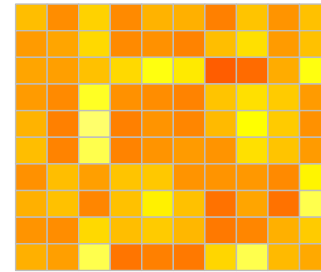
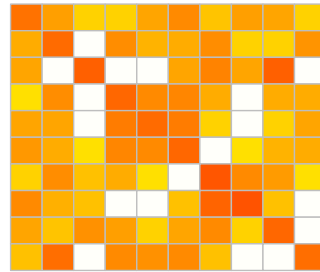
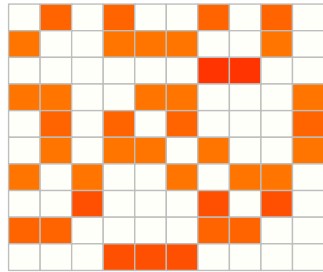


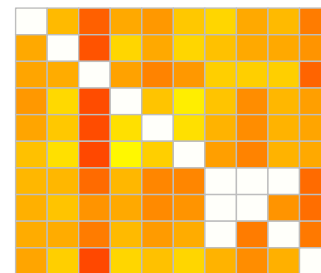
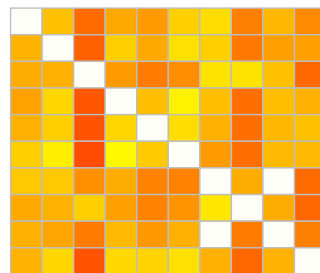
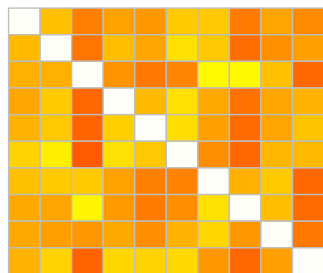
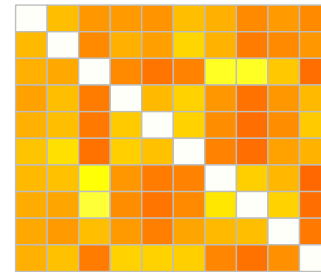
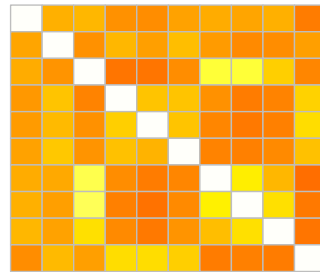
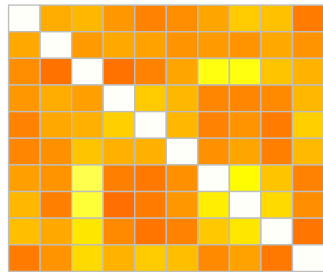
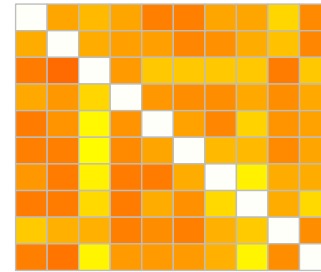
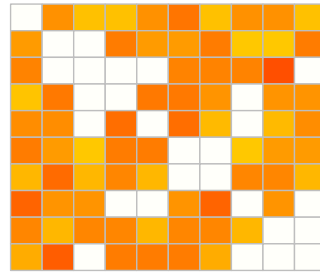
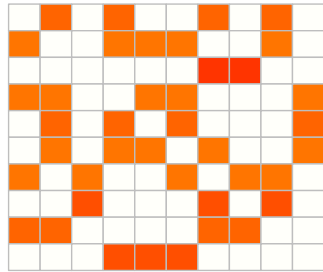
15

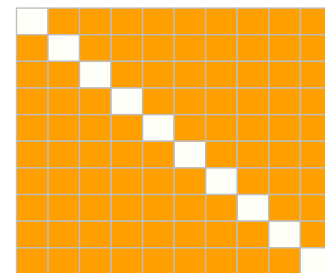
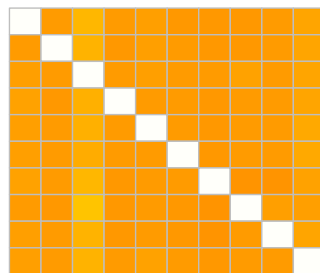
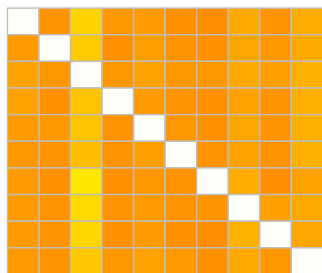
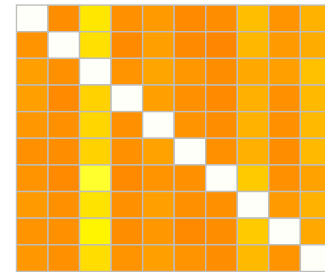
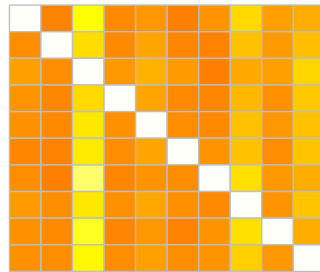
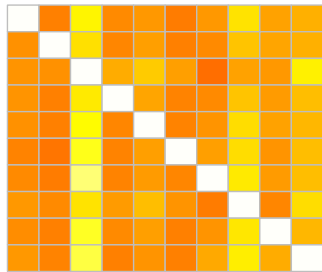
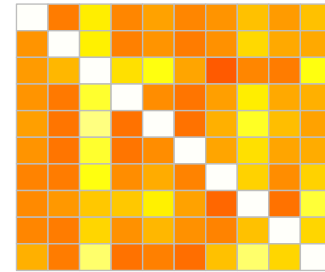
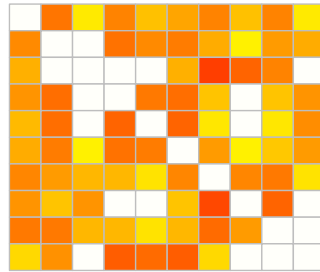
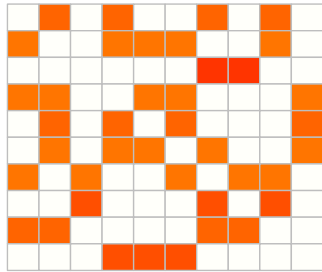


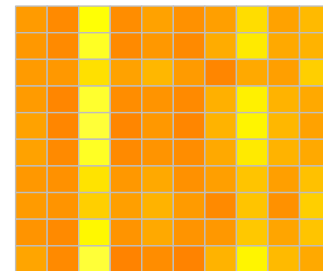
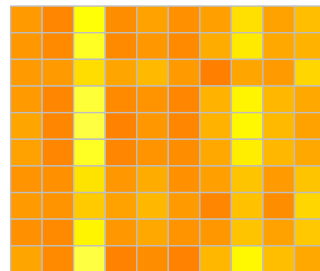
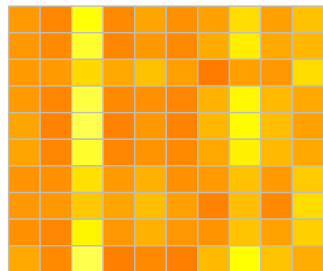
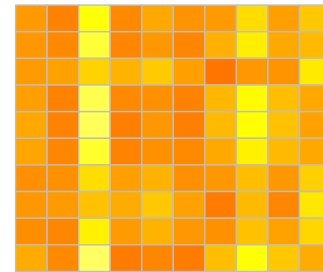
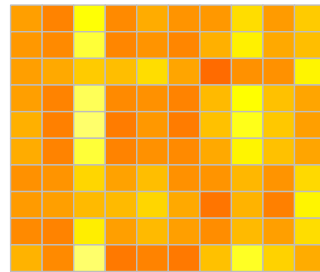
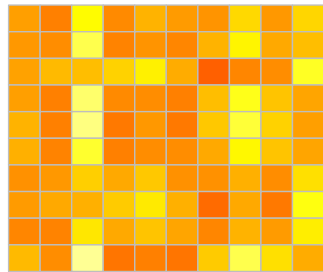
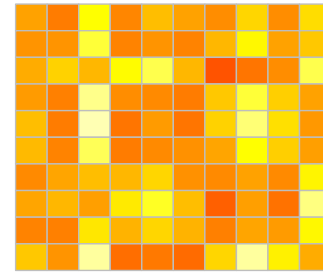
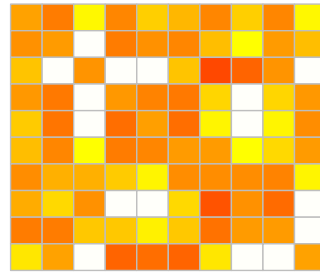
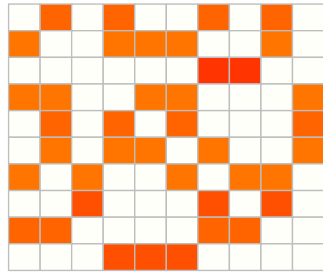


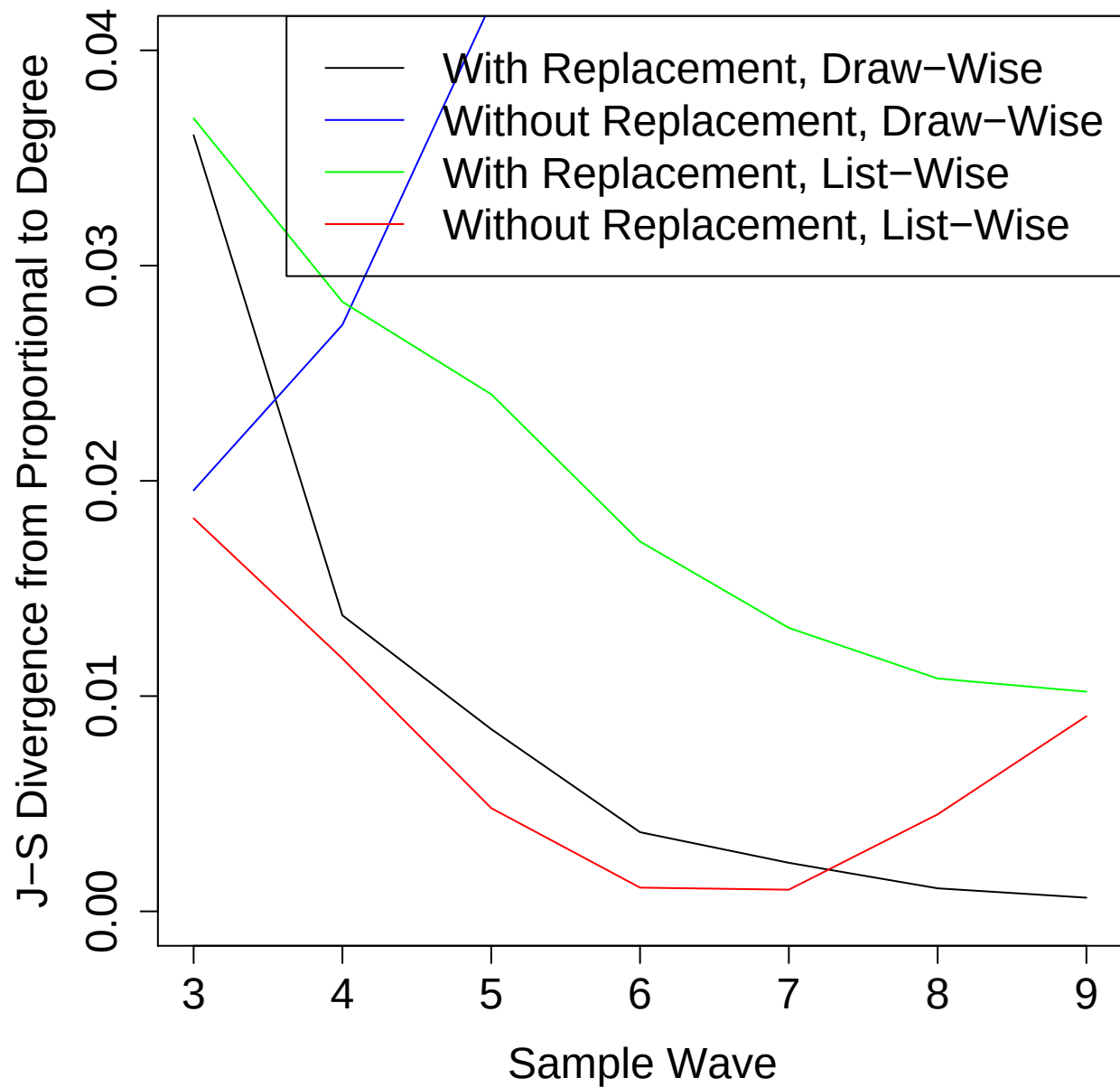




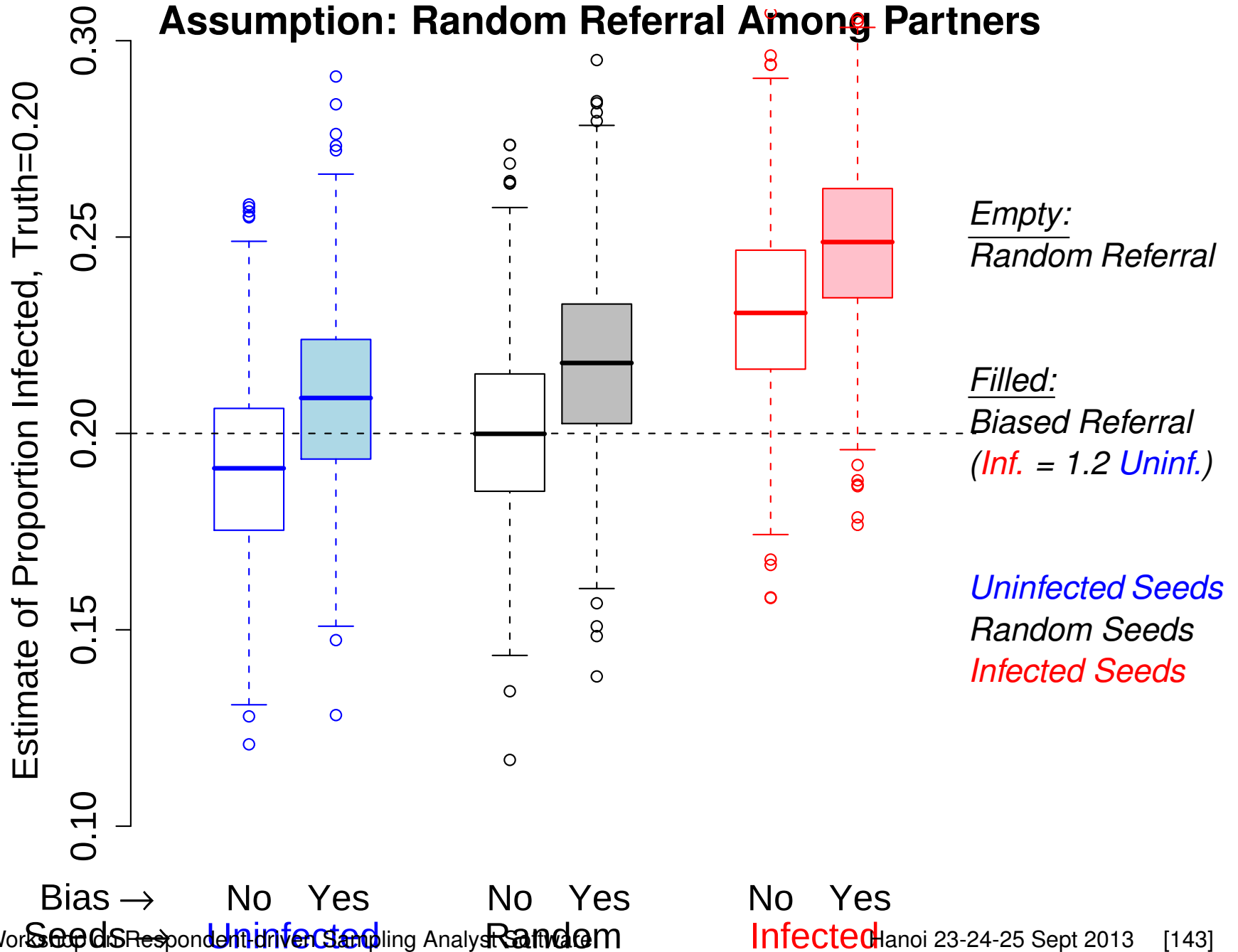




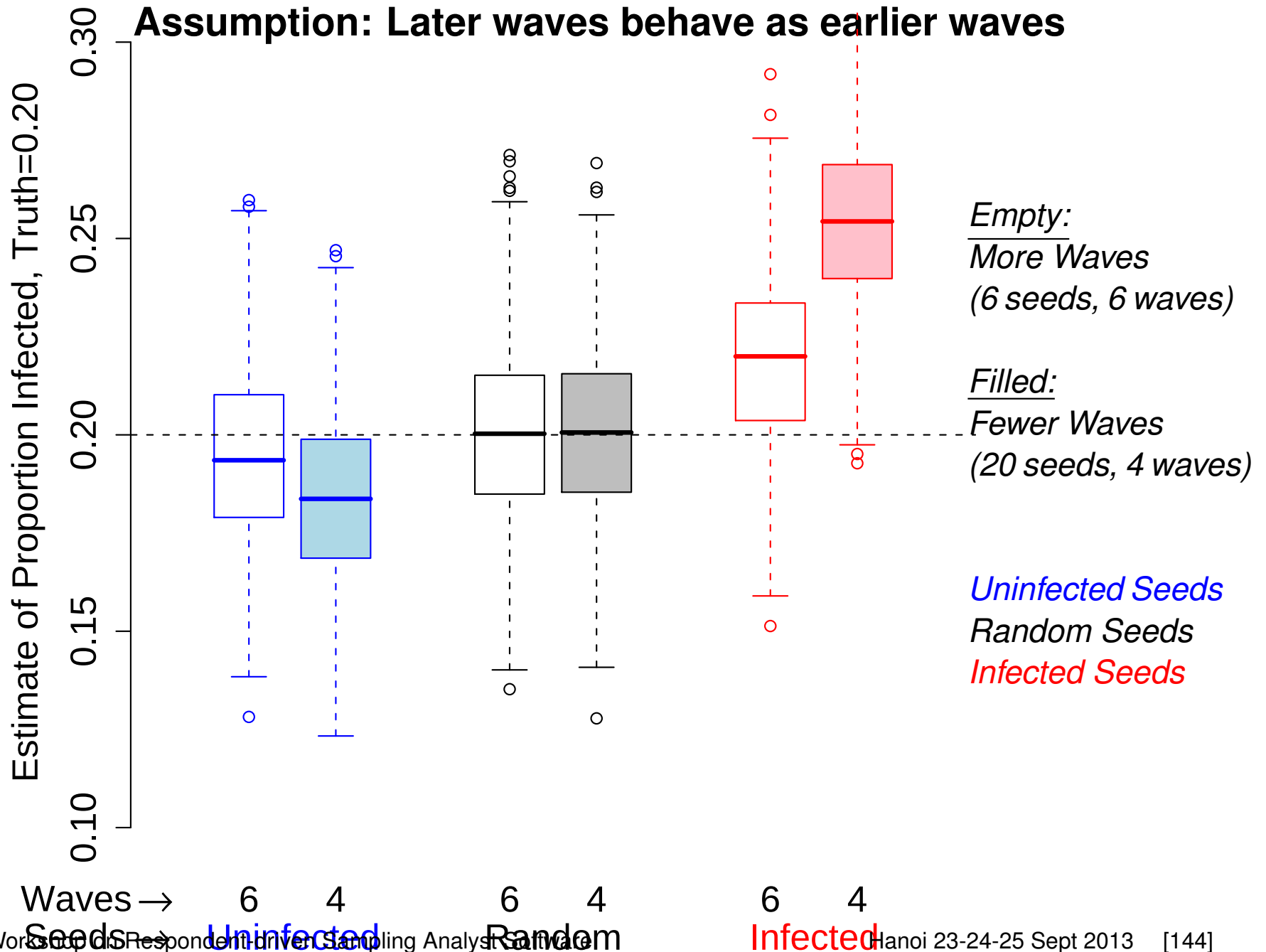




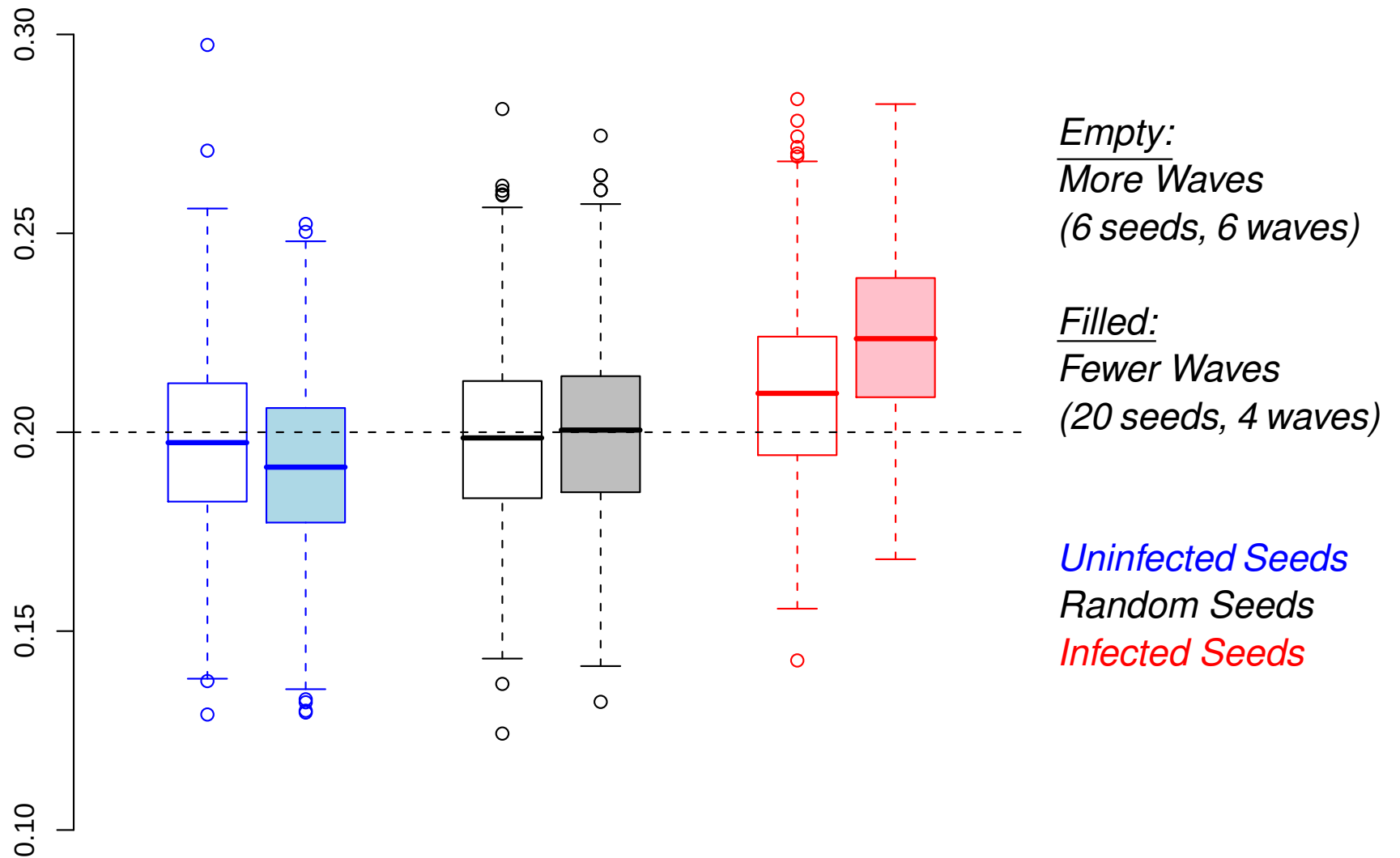
Assumption: Random Referral Among Partners



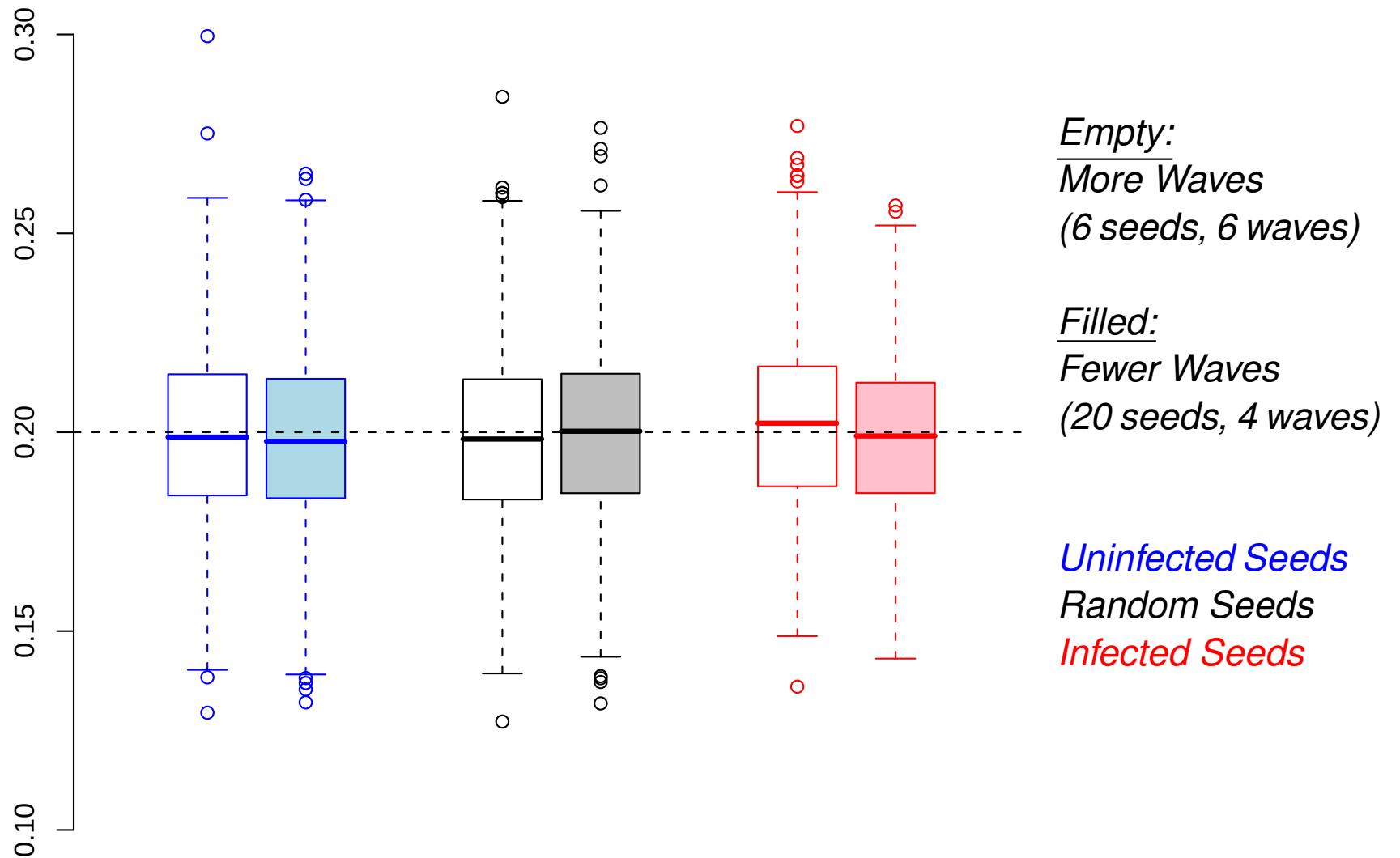
Assumption: Later waves behave as earlier waves



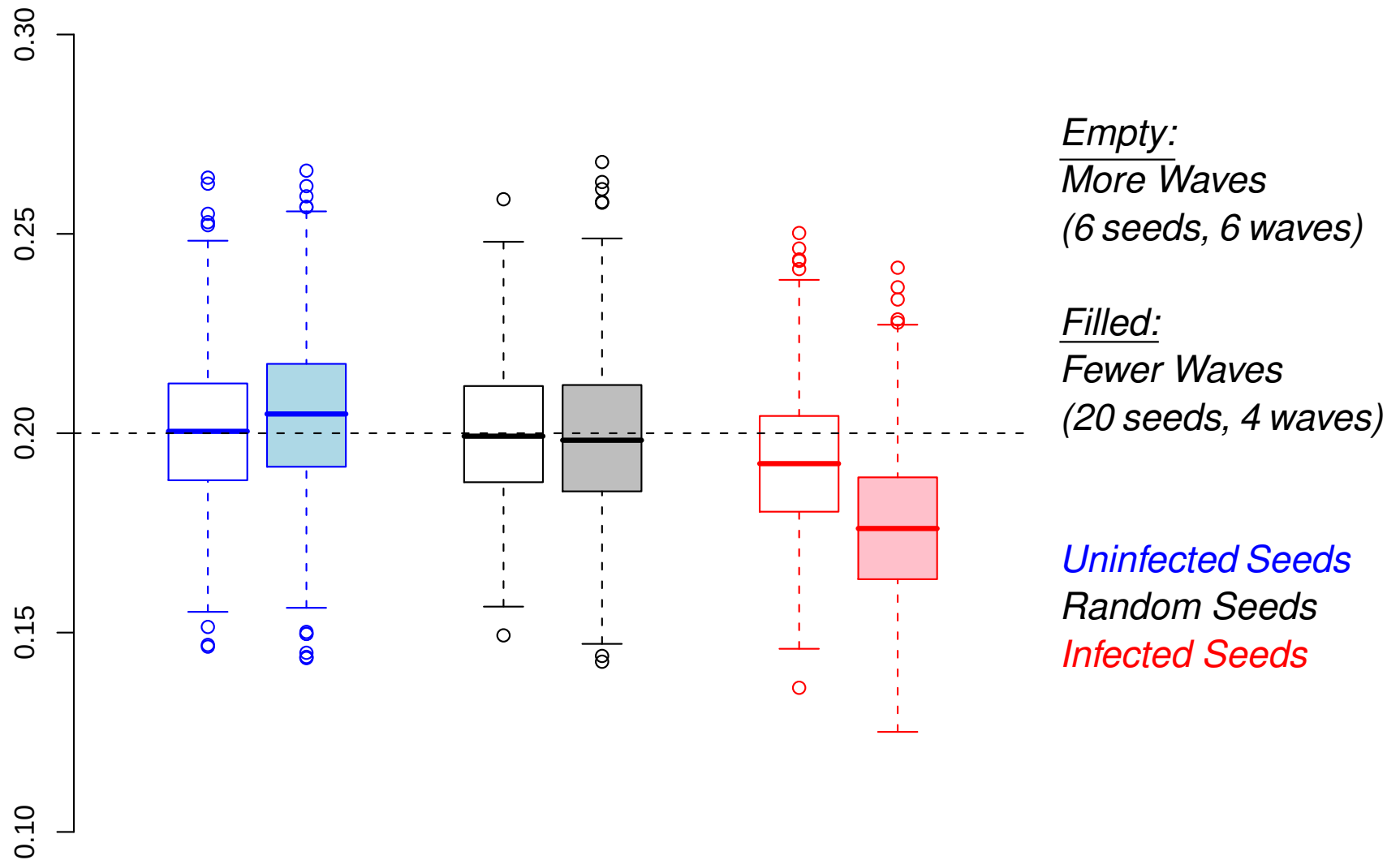
Drop Seeds



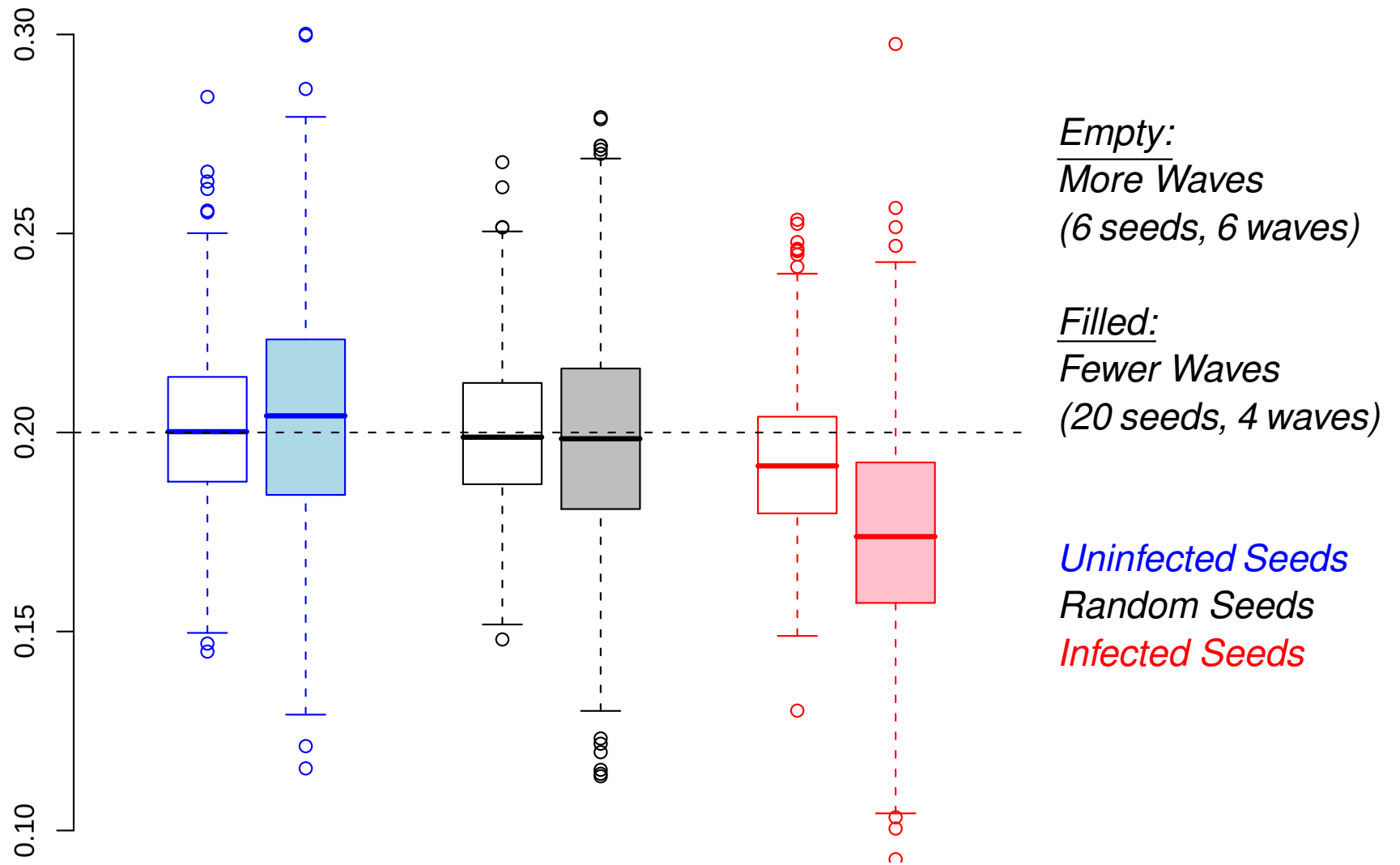
Drop Seeds, Wave 1



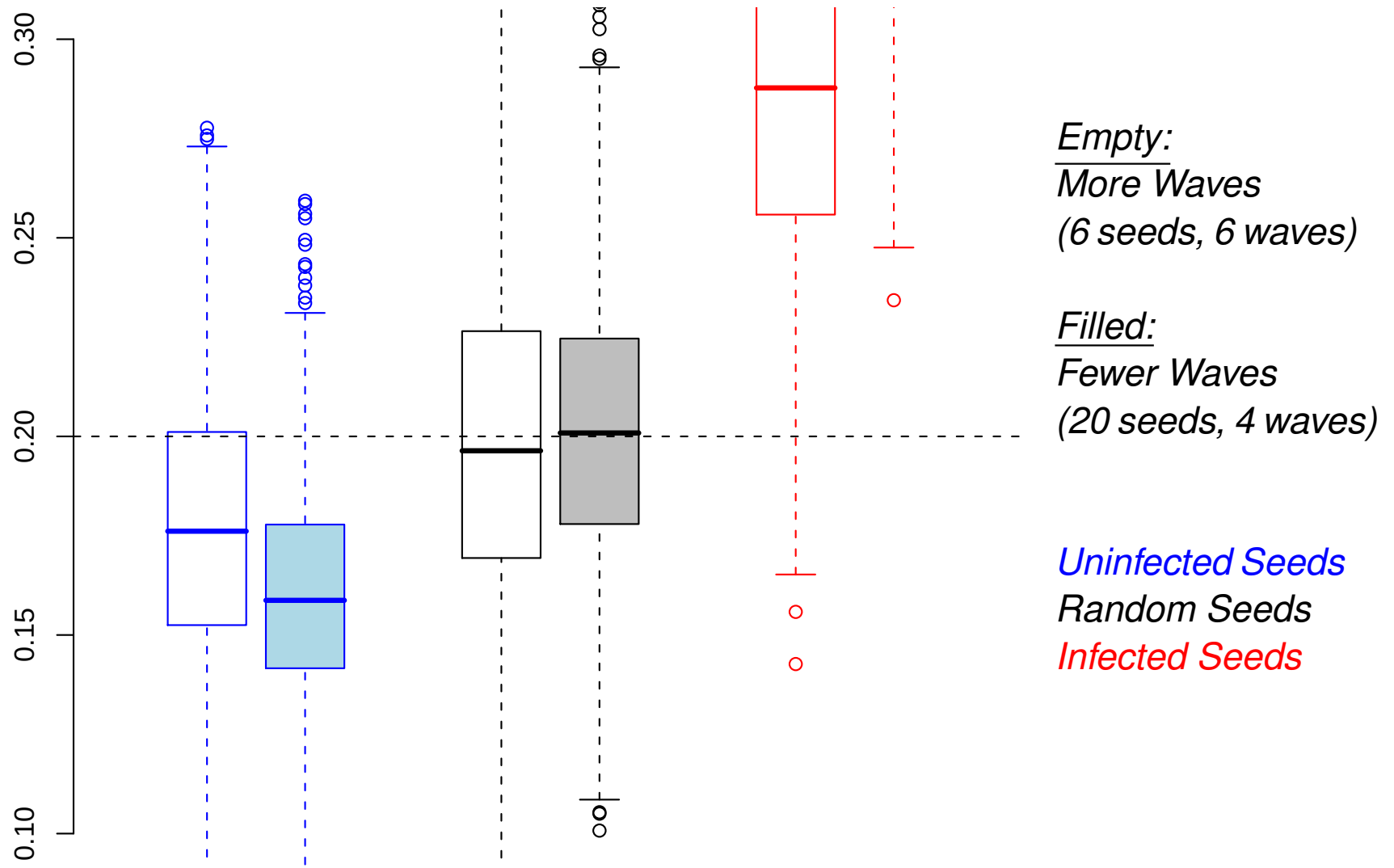
Drop Seeds, Wave 1, Wave 2



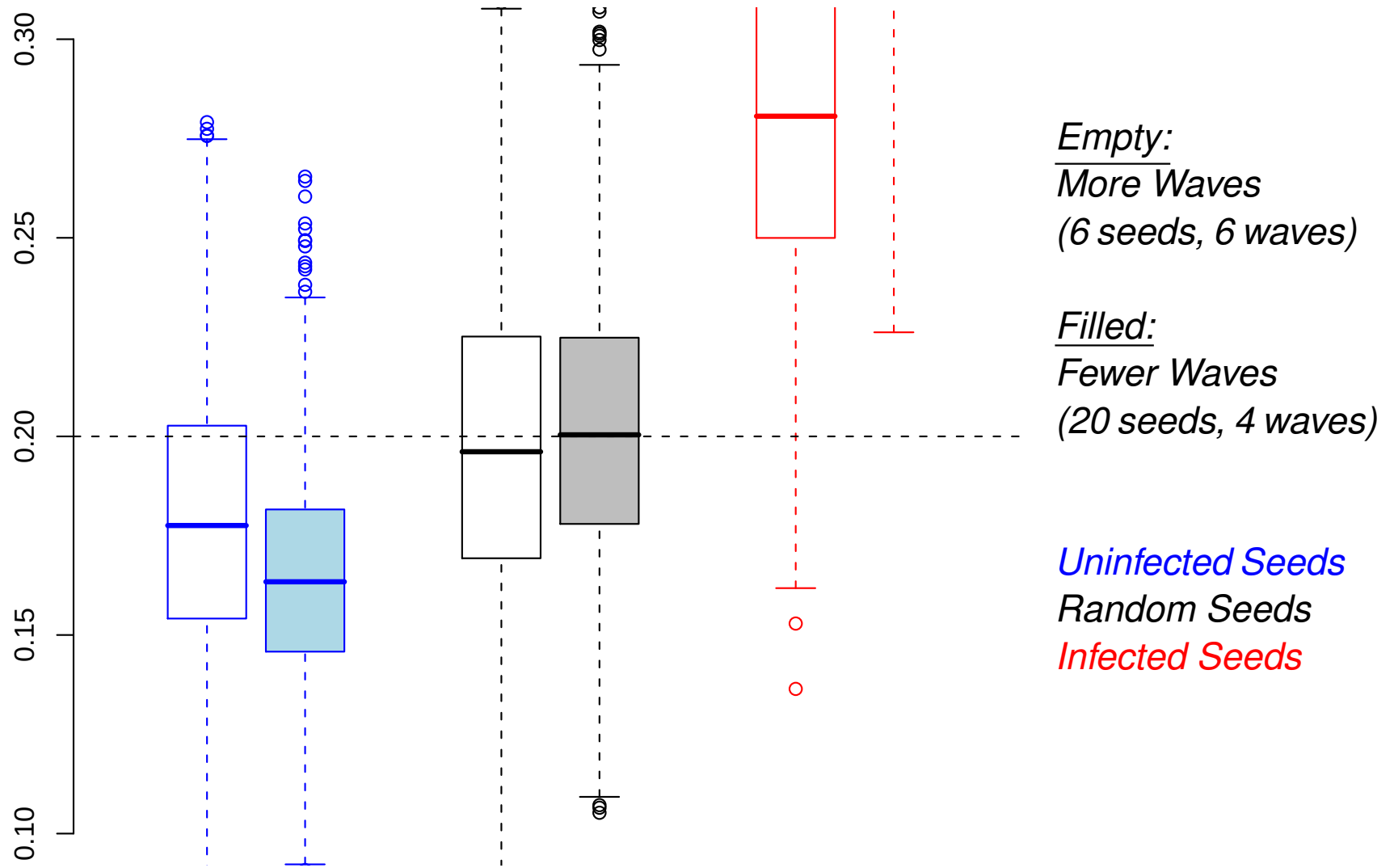
Drop Seeds, Wave 1, Wave 2, Wave 3



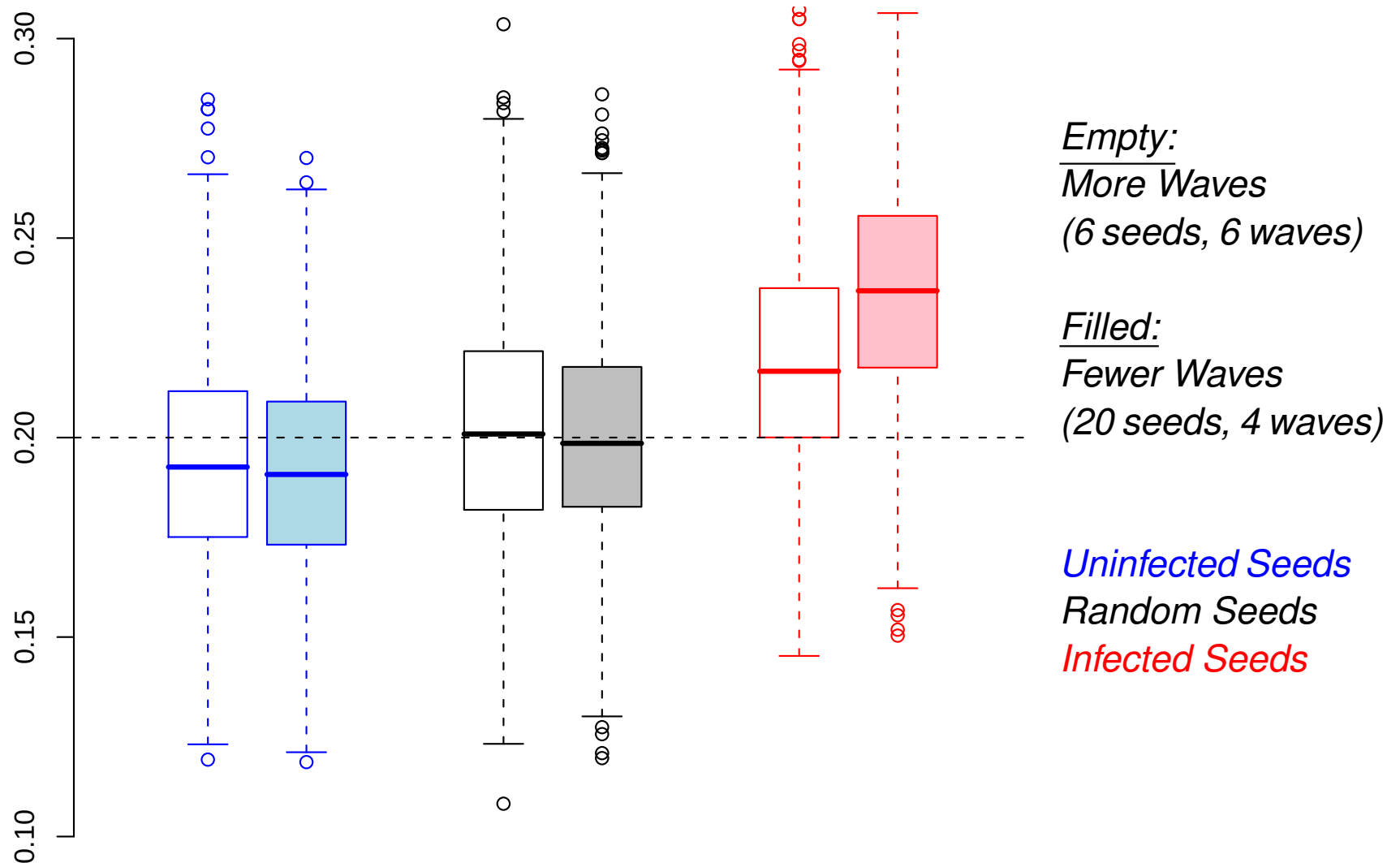
Assumption: With-Replacement Sampling, Drop Seeds



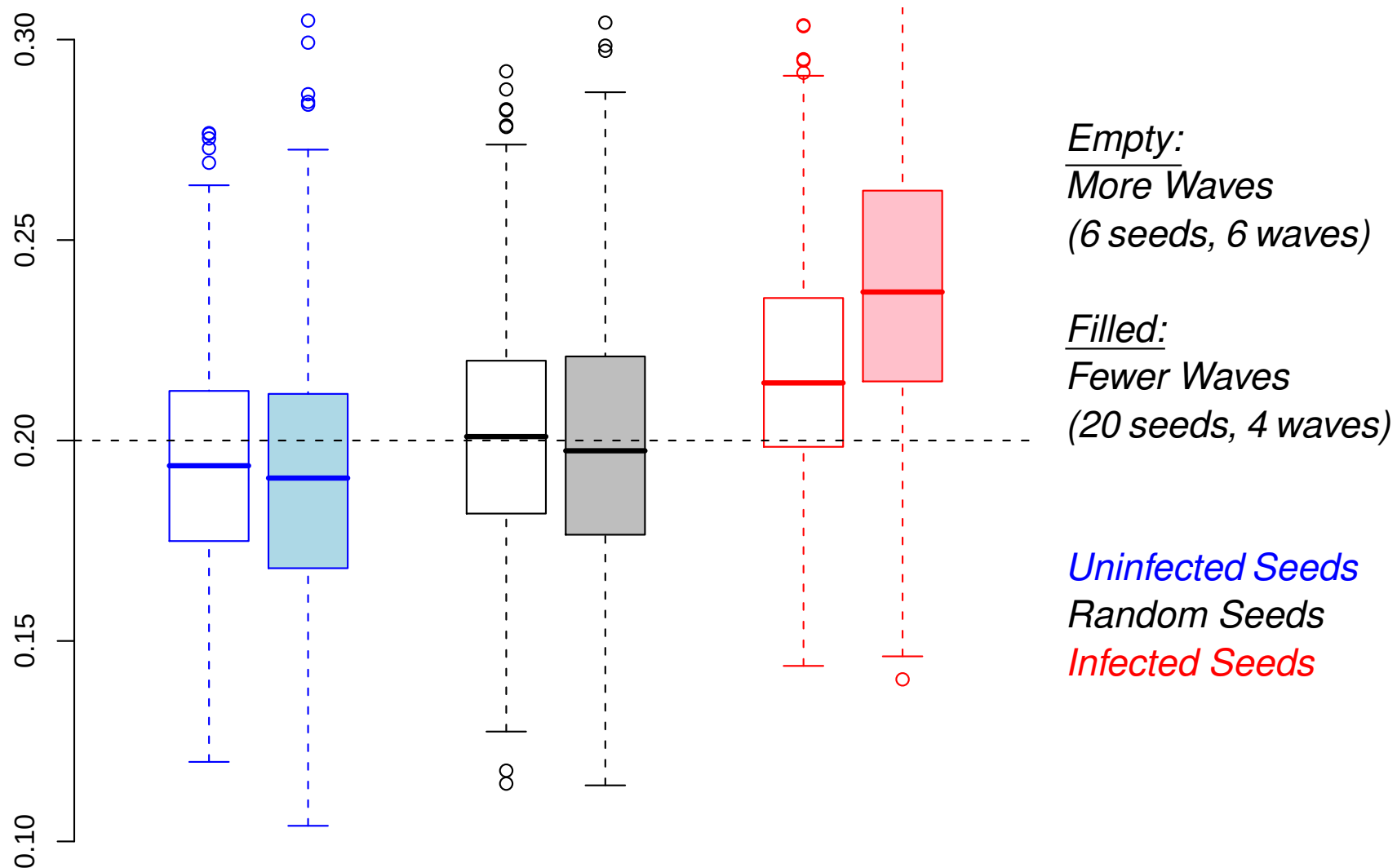
With-Replacement, Drop Seeds, Wave 1



With-Replacement, Drop Seeds, Wave 1, Wave 2



With-Replacement, Drop Seeds, Wave 1, Wave 2, Wave 3



Extra Slides follow

References

- Bernhardt, A., D. Heckathorn, R. Milkman, and N. Theodore, “Documenting Unregulated Work: A Survey of Workplace Violations in New York City,” New York University, 2008.
- Frank, O. and T.A.B. Snijders, “Estimating the size of hidden populations using snowball sampling,” *Journal of Official Statistics*, 10, 1994.
- Gile, K. J. “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., and M.S. Handcock, “Respondent-Driven Sampling: An Assessment of Current Methodology,” *Sociological Methodology*, 40, 2010.
- Gile, K.J., and M.S. Handcock, “Network Model-Assisted Inference from Respondent-Driven Sampling Data.” *R&R Journal of the American Statistical Association*, 2011
- Handcock, M.S., D.R. Hunter, C.T. Butts, S.M. Goodreau, and M. Morris, “`statnet`: An R package for the Statistical Modeling of Social Networks.” , 2003.
- Heckathorn, D.D., and J. Jeffri, “Finding the Beat: Using Respondent-Driven Sampling to Study Jazz Musicians,” *Poetics*, 2000.
- Little, R.J. and D.B. Rubin, *Statistical Analysis with Missing Data.*, 2002.
- Salganik, M.J. and D.D. Heckathorn, “Sampling and estimation in hidden populations using respondent-driven sampling.” *Sociological Methodology*, 34, 193-239, 2004.
- Sarndal, C.E., B. Swensson, and J. Wretman, *Model Assisted Survey Sampling.*, 1992.
- Sugden, R. and T. Smith, “Ignorable and informative designs in survey sampling inference,” *Biometrika*, 71, 1984.
- Volz, E. and D. D. Heckathorn, “Probability Estimation Theory for Respondent Driven Sampling,” *Journal of Official Statistics*, 24:1, 2008.

-
- *Walters, K. L., “Health Survey of Two-Spirited Native Americans,” [5 R01 MH 65871-02] National Institute of Mental Health, 2002.*

RDS is Widely Used

- *US Centers for Disease Control: RDS for monitoring risk behavior in IDUs in 25 cities every 3 years*
- *Over 100 international studies of hard-to-reach, high-risk populations*
- *Many more studies in varied populations including*
 - *Jazz musicians in New York City (Heckathorn and Jeffri, 2000)*
 - *LGBT Native Americans in US Cities (Walters, 2002)*
 - *Unregulated workers in US Cities (Bernhardt et al., 2008)*

Under SS Model:

- *Estimation approach: Horvitz-Thompson estimation based on SS inclusion probabilities.*
- *Inclusion probability of node i , π_i determined by:*
 - *Degree of i , d_i*
 - *Population size N*
 - *Number of nodes of degree k , N_k , for $k \in 1 \dots K$*
 - *Sample size n*

Note: Inclusion probabilities not available in closed form for sizable n , but can be computed by simulation.

Two Vulnerabilities of SS Estimator

- *Seed Bias*
- *Unknown Population Size N*

Sequential Probability Proportional to Size Sampling for RDS

Difficulty:

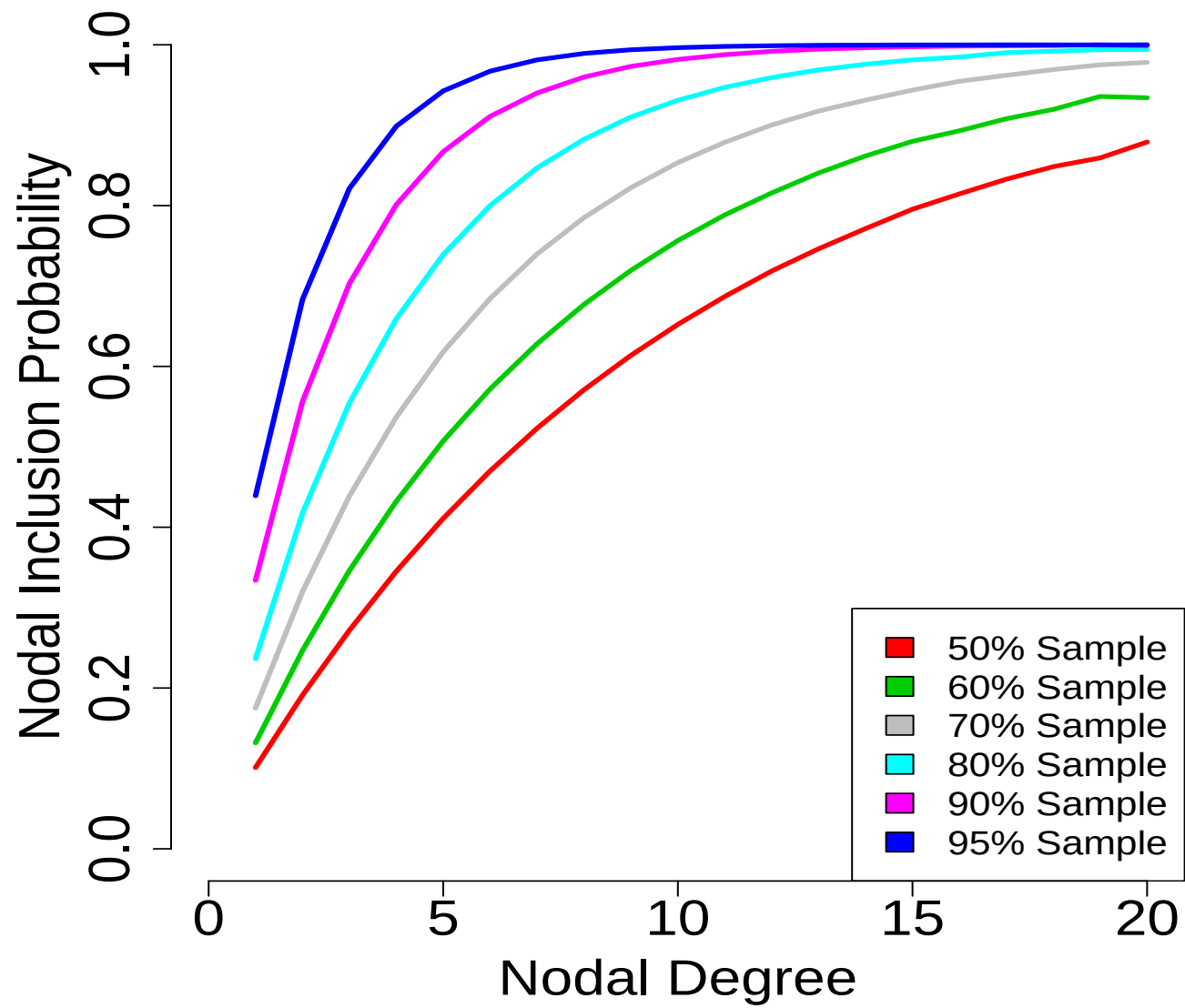
- *Population size N unknown*
- *Population degree distribution $\mathbb{N} = \{\mathbb{N}_1, \mathbb{N}_2, \dots, \mathbb{N}_K\}$ unknown*

Sequential Probability Proportional to Size Sampling for RDS

Difficulty:

- *Population size N unknown*
(Assume known. Sensitivity analysis.)
- *Population degree distribution $\mathbb{N} = \{\mathbb{N}_1, \mathbb{N}_2, \dots, \mathbb{N}_K\}$ unknown*
(Estimate iteratively)

Mapping $Q : d_i \rightarrow \pi_i$



Sequential Probability Proportional to Size Sampling for RDS

Difficulty:

- *Population size N unknown
(Assume known. Sensitivity analysis.)*
- *Population degree distribution $\mathbb{N} = \{\mathbb{N}_1, \mathbb{N}_2, \dots, \mathbb{N}_K\}$ unknown
(Estimate iteratively)*

Premise:

- *If $d_i = d_j$ then $\pi_i = \pi_j$.*
- *There is a mapping $Q_k(\mathbb{N}) : k \rightarrow \pi_i$ and can be computed for known population of degrees.*
- *For known Q , population proportion $\mathbb{N}_k$ can be estimated by inverse probability weighting.*

Approach:

- *Iterate to solve Method of Moments Equation to estimate inclusion probabilities*

Approach: Solve Design-Based Method of Moments Equation

- *Method of moments estimator for the unknown $\mathbb{N}$:*

$$\mathbb{E}[V_k; \mathbb{N}] = \mathbb{N}_k Q_k(\mathbb{N}) = v_k \quad k = 1, \dots, K$$

- V_k is the random variable for number of sampled nodes of degree k
- v_k is its observed value.

- *Note the expectation $\mathbb{E}[V_k : \mathbb{N}]$ is taken over realizations of the sampling process.*

SS Estimator: Algorithm

1. *Initial: $\pi_k(\mathbb{N}^0) \propto k$.*
2. *For $i = 1 \dots r$:*
 - (a) *Estimate $\mathbb{N}^i$:*

$$\mathbb{N}_k^i = N \cdot \frac{\frac{v_k}{\pi_k(\mathbb{N}^{i-1})}}{\sum_{l \in 1 \dots K} \frac{v_l}{\pi_l(\mathbb{N}^{i-1})}} \quad k = 1, \dots, K$$

where v_k is the observed number of sample units with degree k .

- (b) *Compute $Q_k(\mathbb{N}^i)$ by simulation:*

- i. *Simulate M SS samples from $\mathbb{N}^i$*
- ii.

$$\pi_k(\mathbb{N}^i) = \frac{\mathbb{E}[V_k; \mathbb{N}^i]}{\mathbb{N}_k^i} \approx \frac{U_k + 1}{M \cdot \mathbb{N}_k^i + 1},$$

where U_k is the number of observed units of size k in the M simulations.

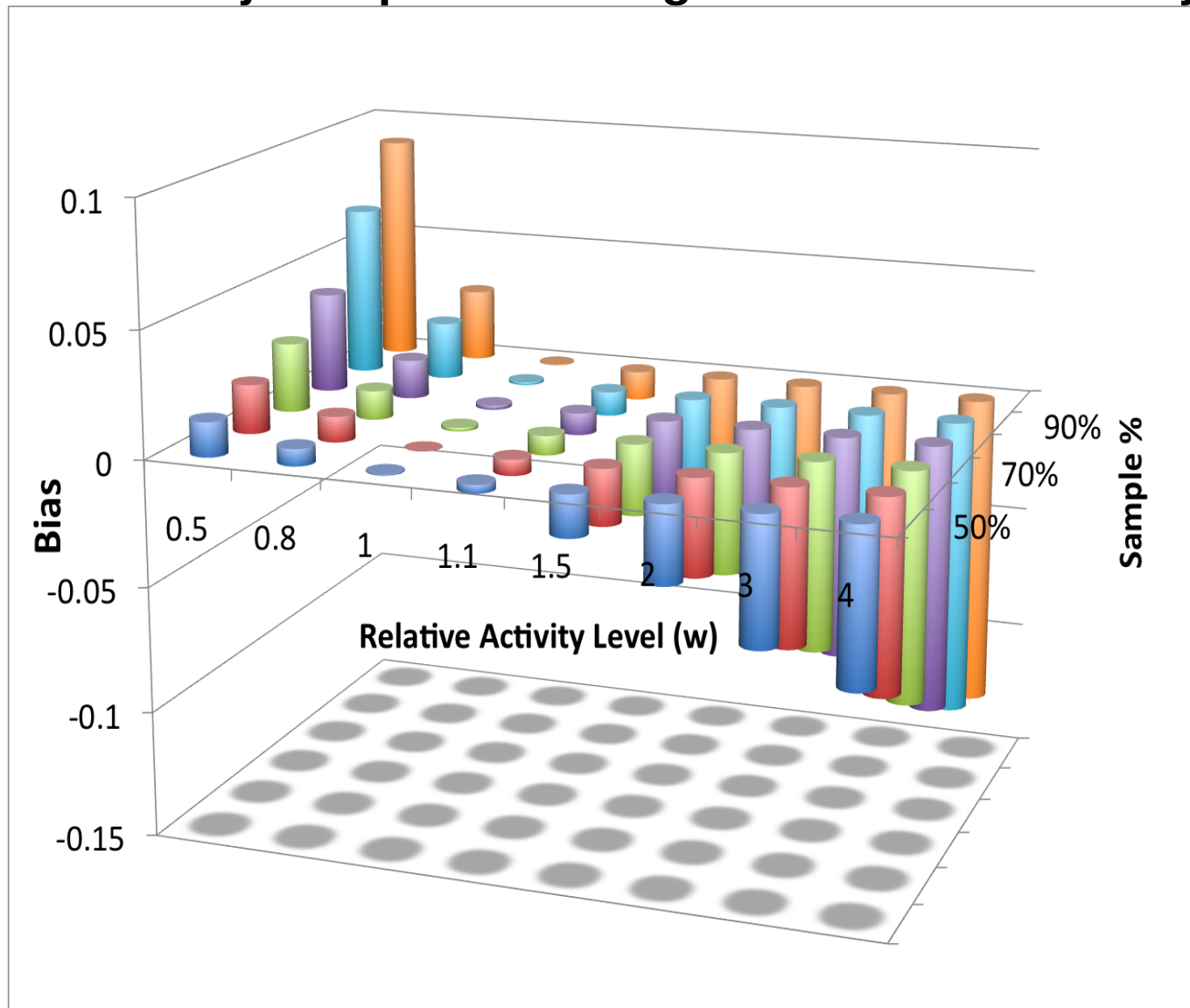
3. *Use $\hat{Q} = Q(\mathbb{N}^r)$ to estimate μ :*

$$\hat{\mu}_M = \frac{\sum_{j: S_j=1} \frac{z_j}{\hat{Q}_{d_j}}}{\sum_{j: S_j=1} \frac{1}{\hat{Q}_{d_j}}}.$$

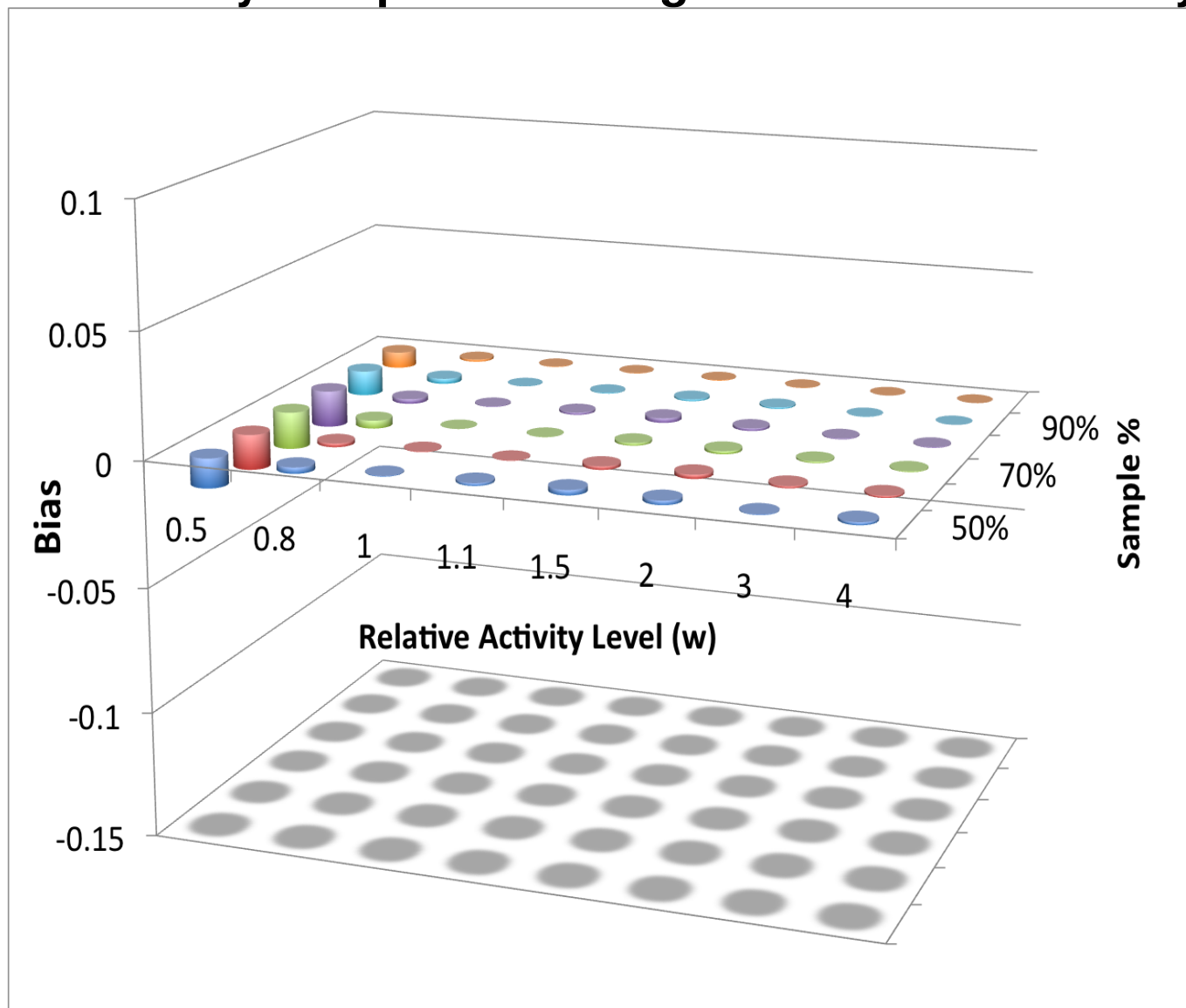
Properties of Algorithm

- *Subject of ongoing study*
- *Basic Points:*
 - *Converges very quickly (2-3 iterations)*
 - *Solves method of moments equation*
 - *Believe to be a fixed point process*

VH Bias by Sample Percentage and Relative Activity



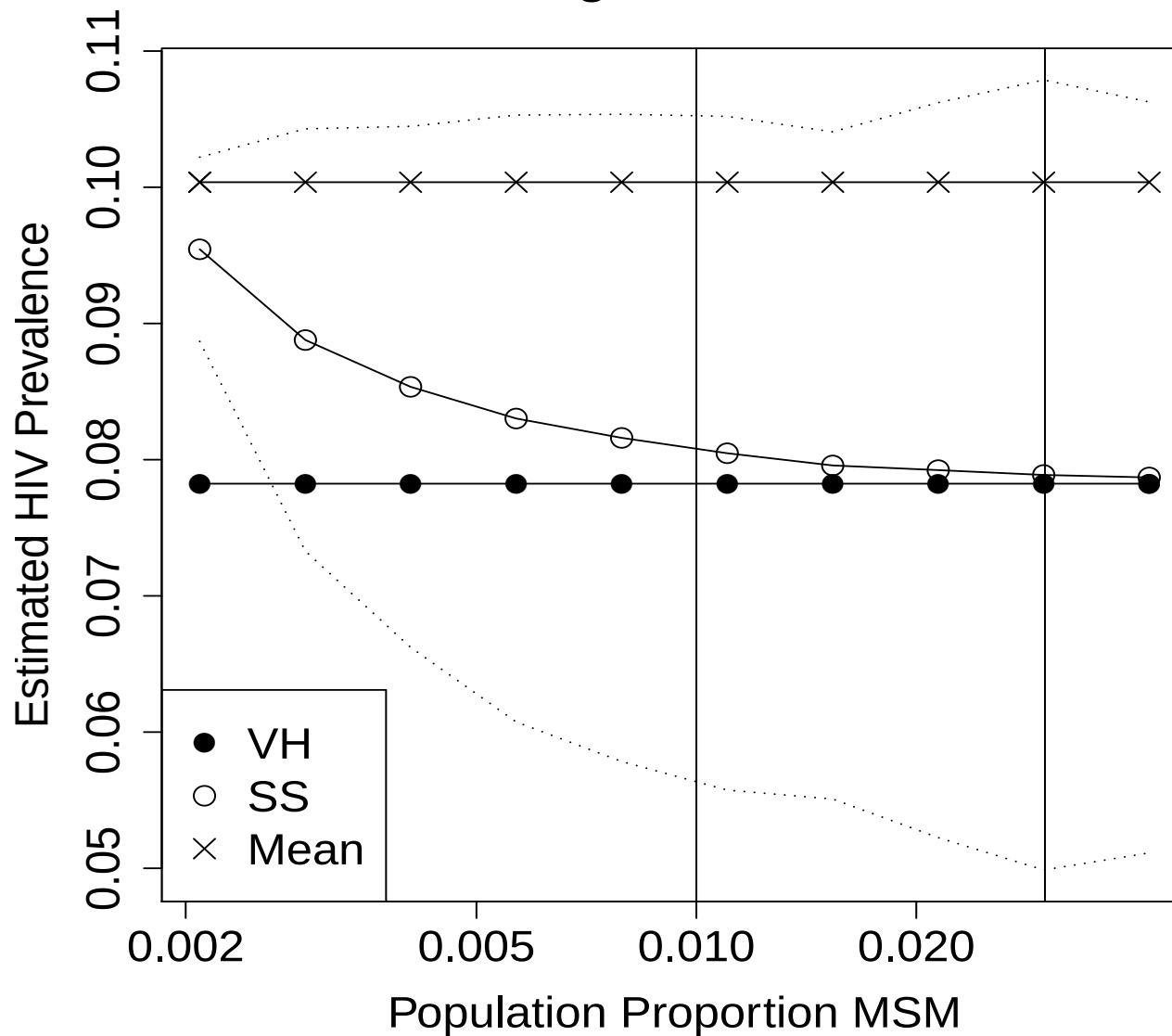
SS Bias by Sample Percentage and Relative Activity



Two Vulnerabilities of SS Estimator: Sensitivity Analysis

- *Seed Bias*
 - *No worse than Volz-Heckathorn (VH)*
- *Unknown Population Size N*
 - *For moderate error, still out-performs VH*
 - *Typically between VH and sample mean*

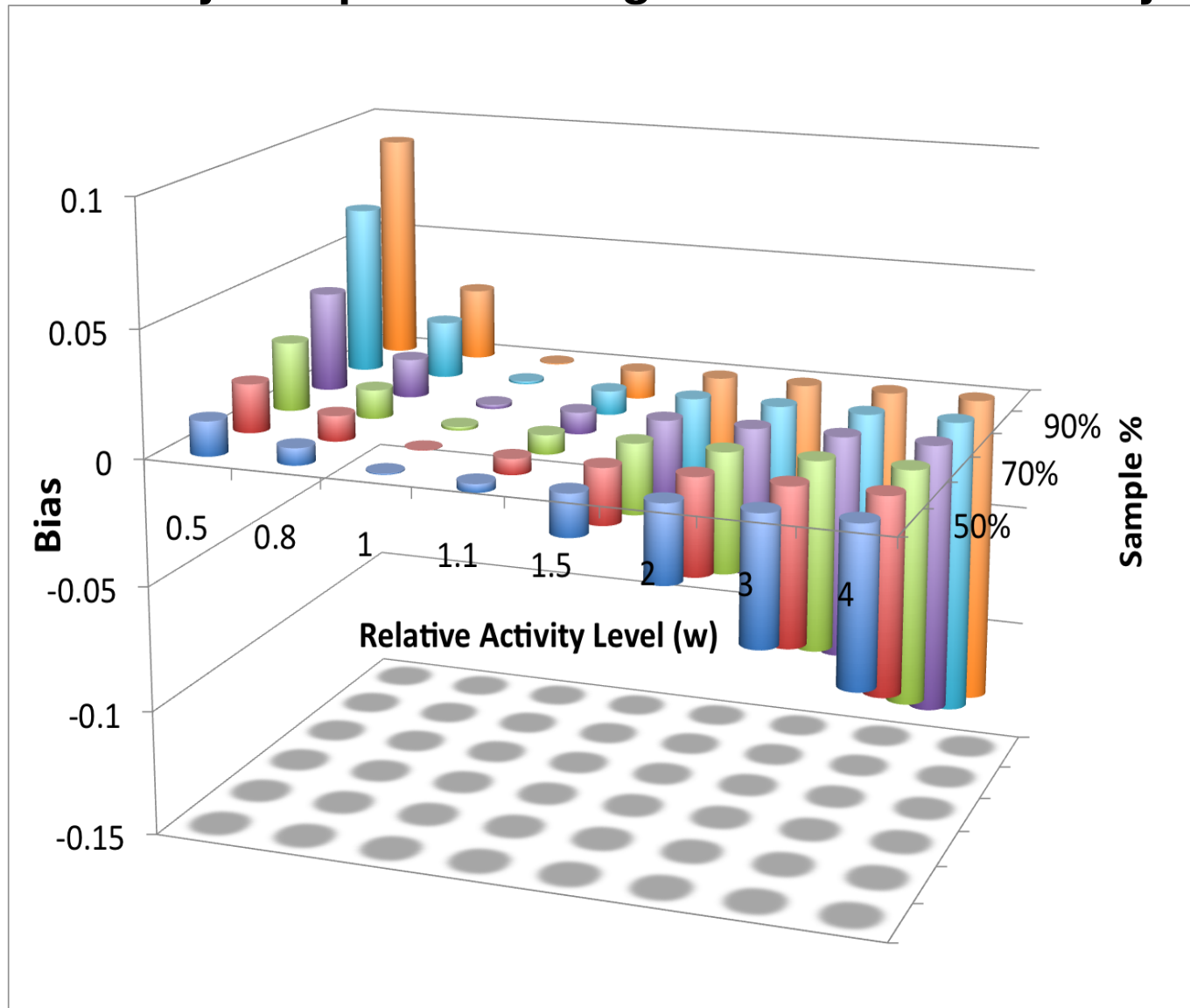
HIV Prevalence among MSM in a Caribbean City



Summary of Findings:

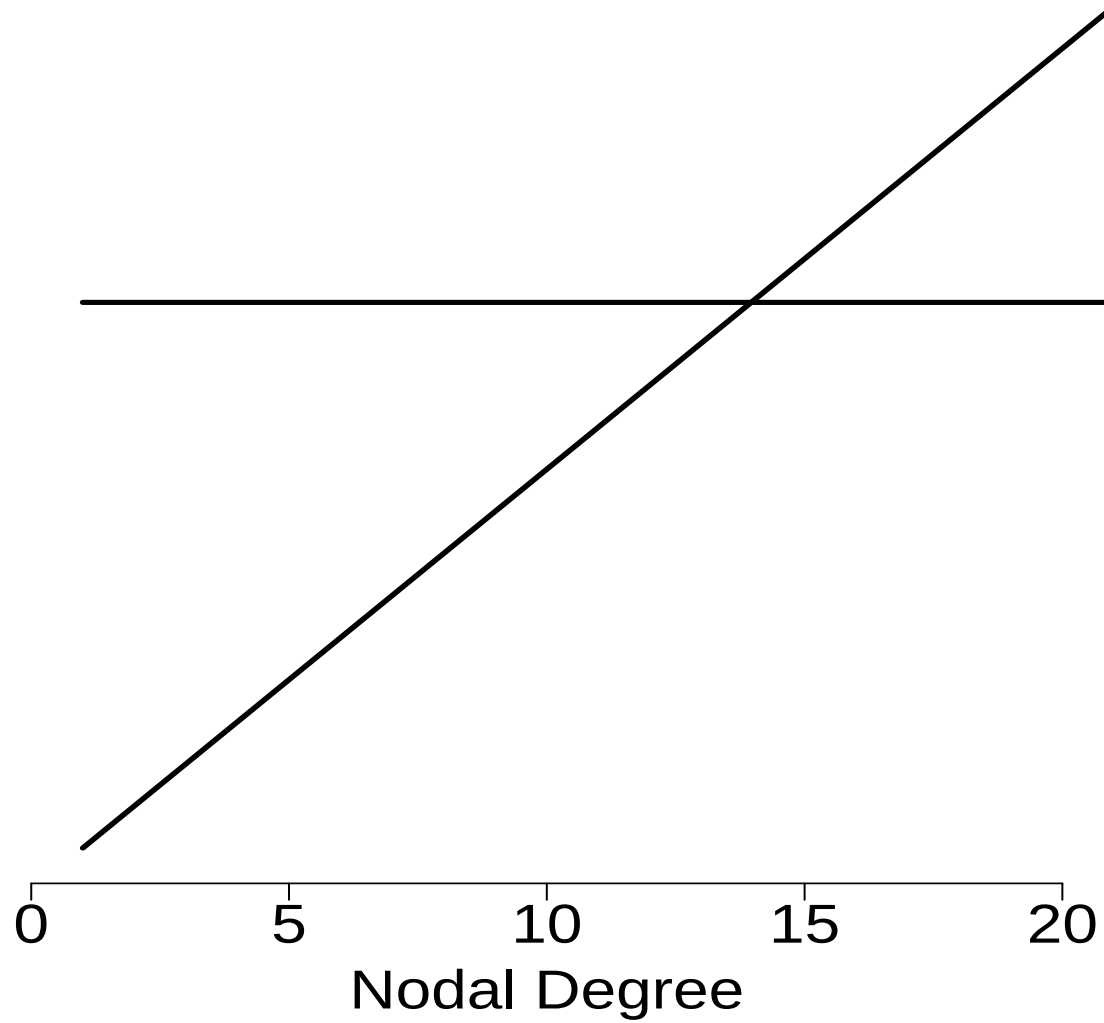
- *SS out-performs VH in all circumstances studied*
- *Still subject to seed bias*
- *We introduce a further improvement*

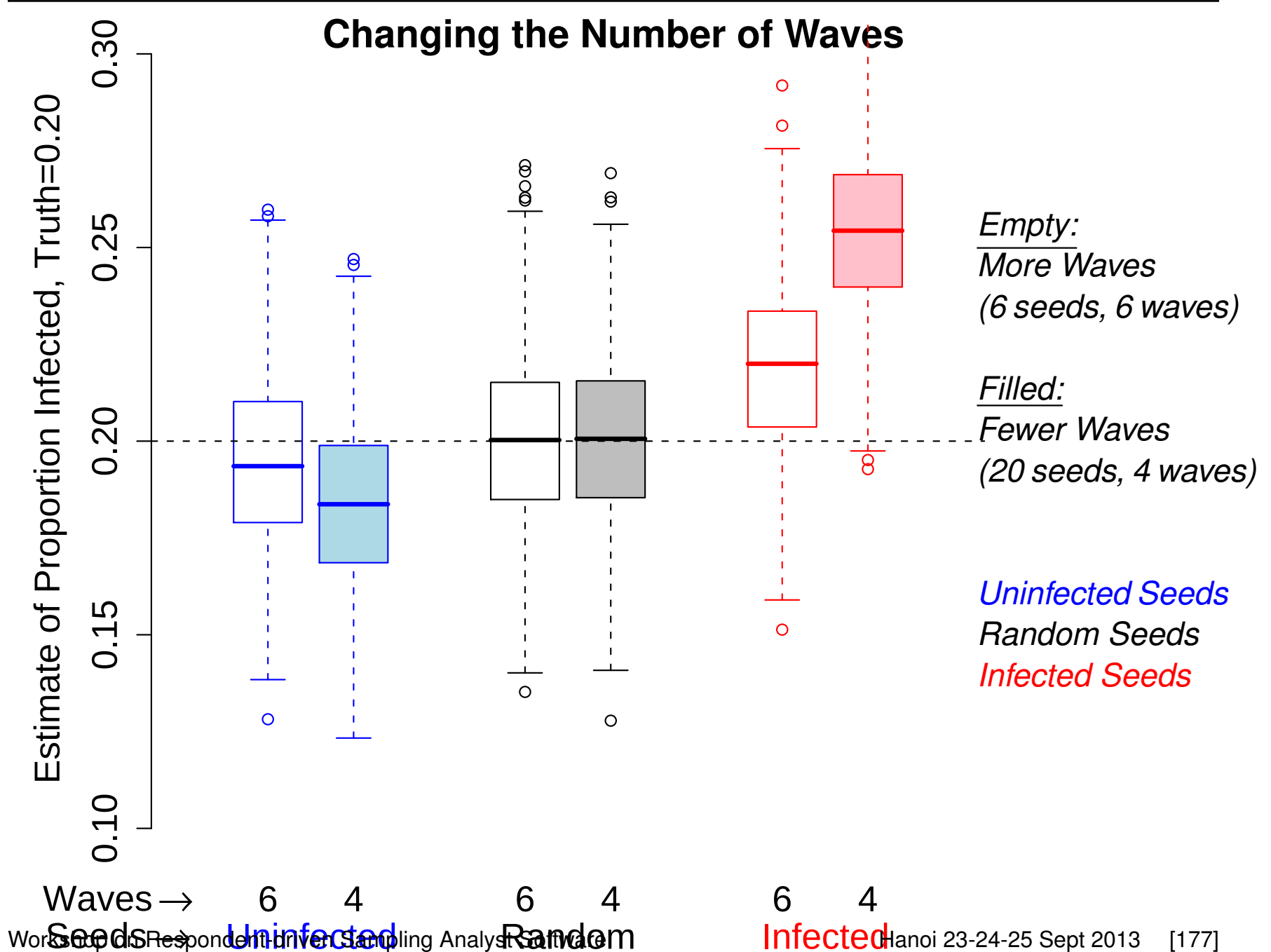
Bias by Sample Percentage and Relative Activity



Mapping from Degree to Probability

Nodal Inclusion Probability





Further Extension: Network-Based Estimator

Premise:

- *Given:*
 - *True network form β (ERGM homophily parameter)*
 - *True sampling structure $\mathcal{S}$ (e.g. seed characteristics)*
- *Then:*
 - *Nodal equivalence classes based on degrees k and infection status z*
 - *Exists a mapping $\pi_{k,z}^*(\beta, \mathcal{S}) : \{k, z\} \rightarrow \pi$, from the nodal degree k and value z , to nodal inclusion probabilities, π .*

Furthermore,

- *Given inclusion probabilities, $\{\pi_i\}$, use Horvitz-Thompson estimators to estimate the network and sampling structures, $\{\beta, \mathcal{S}\}$.*
- *Can simulate networks according to β .*
- *Can simulate samples according to $\mathcal{S}$.*
- *Use simulations to estimate mapping, $Q_{k,z}^*(\beta, \mathcal{S})$.*

Network-Based Algorithm

- Estimate Q^* proportional to degree k .
- Iterate the following steps:
 - Estimate population distribution of degrees k by infection status z
 - Estimate network features β based on observed data and weights
 - Estimate sampling features $\mathcal{S}$ based on observed data and weights
 - Simulate M networks, and samples from networks. Estimate weights.
- Use the resulting estimated probabilities, $\hat{Q}^*$, to form weighted estimator.

$$\hat{\mu}_M^* = \frac{\sum_{j:S_j=1} \frac{z_j}{\hat{Q}_{d_j, z_j}^*}}{\sum_{j:S_j=1} \frac{1}{\hat{Q}_{d_j, z_j}^*}}.$$

Configuration Model: VH

- *V-H estimator assumes step-wise selections proportional to degree, based on convergence argument.*
- *Without convergence, it also might work if:*

Configuration Model

- *Given a population of N nodes, with degree distribution.*
- *Randomly select pairs of edge-ends and connect them. (Molloy and Reed, 1995)*
(Include self-ties and multiple ties, but fewer for larger networks with lower maximum degrees.)

Then the probability sampled node g_{j-1} refers node g_j is:

$$\begin{cases} \frac{d_{g_j}}{2E-1} & g_j \neq g_{j-1} \\ \frac{d_{g_j}-1}{2E-1} & g_j = g_{j-1}, \end{cases}$$

where $E = \sum_i d_i$.

Nearly proportional to d_{gj} - so V-H accurate without convergence.

Configuration Model: SS

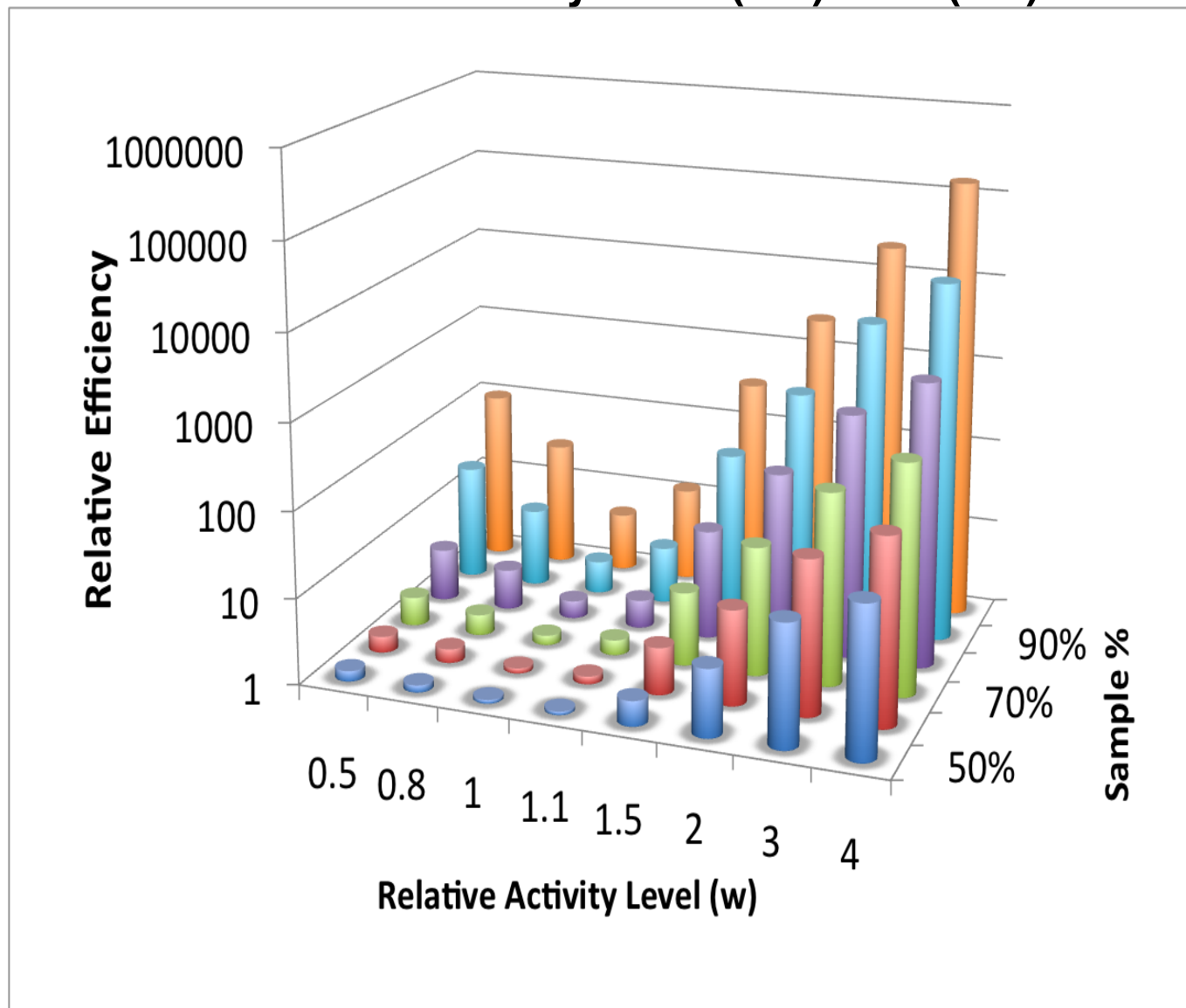
Now consider a self-avoiding random walk under the configuration model.

Then the probability sampled node g_{j-1} refers node g_j is:

$$\begin{cases} \frac{d_{g_j}}{2E - \sum_{i=1}^{j-1} d_{g_i}} & g_j \notin g_1 \dots g_{j-1} \\ 0 & g_j \in g_1 \dots g_{j-1}. \end{cases}$$

Exactly the SS procedure.

Relative efficiency: $MSE(VH)/MSE(SS)$



Two Vulnerabilities of SS Estimator

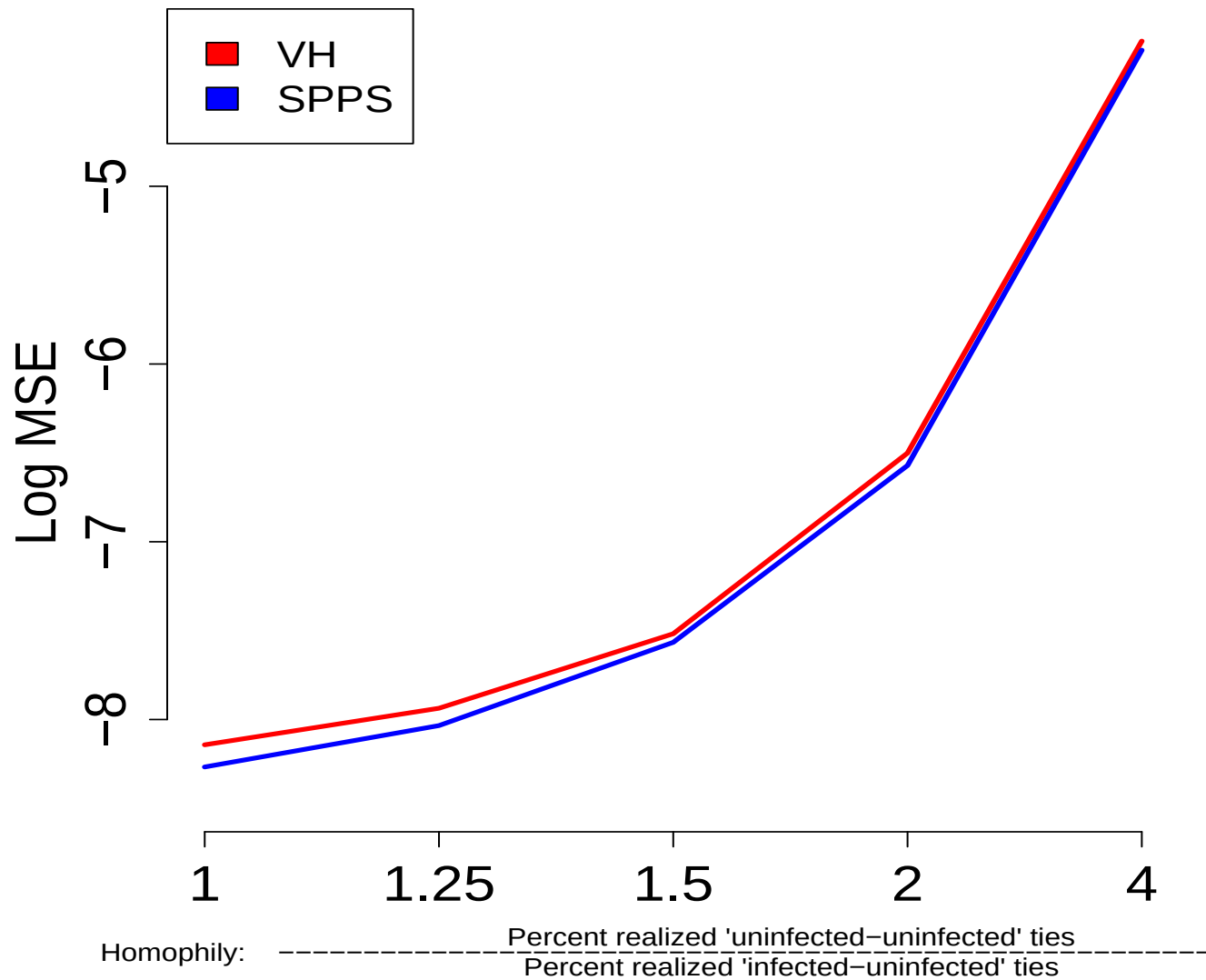
- *Seed Bias*
 - *Compare MSE of VH and SS estimator when all seeds infected*
- *Unknown N*
 - *Set $\hat{N}$ too small - decrease by $0.5 * (N - n)$*

$$\hat{N} = N - 0.5 * (N - n)$$

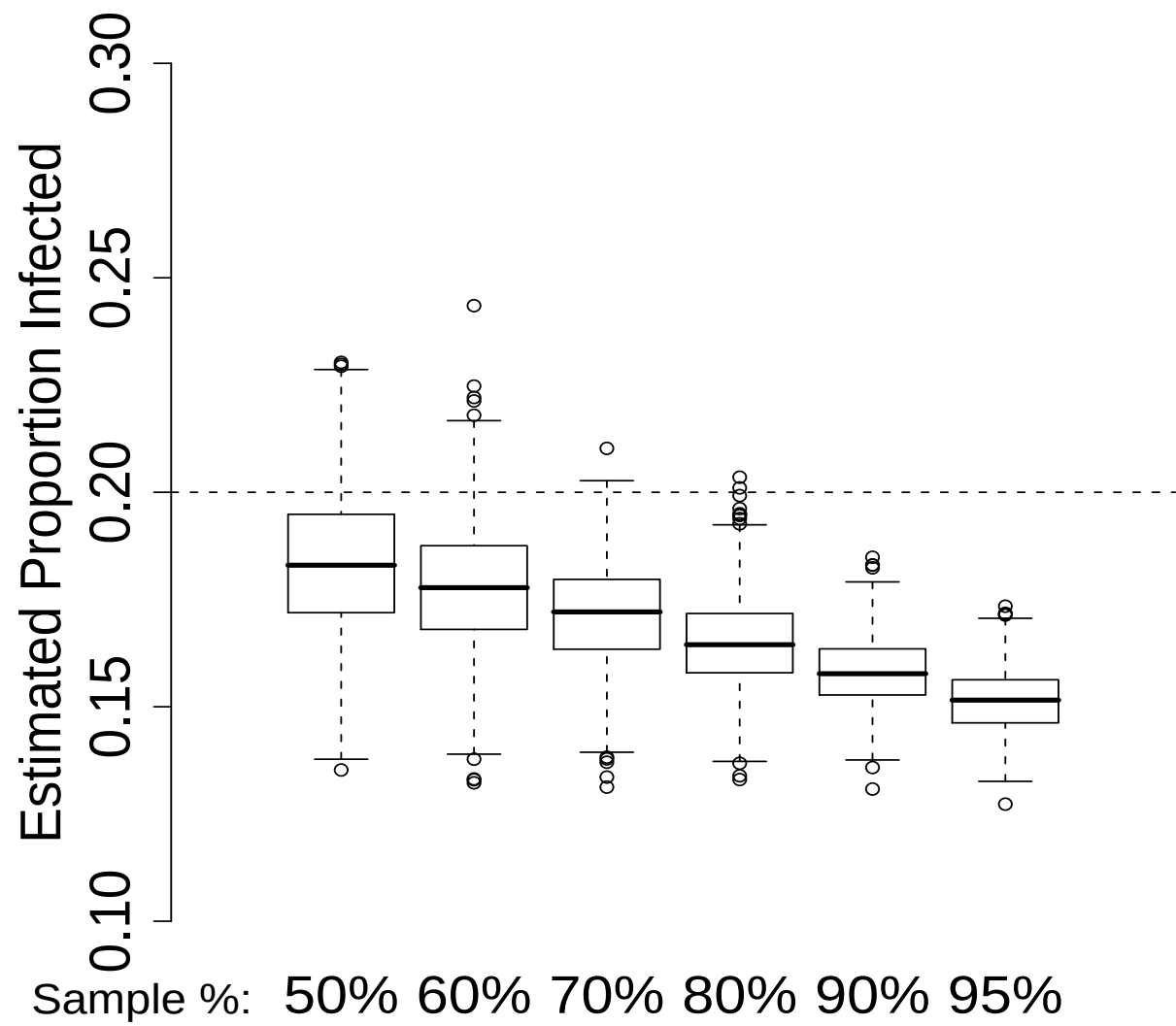
- *Set $\hat{N}$ too large - increase by $0.5 * (N - n)$*

$$\hat{N} = N + 0.5 * (N - n)$$

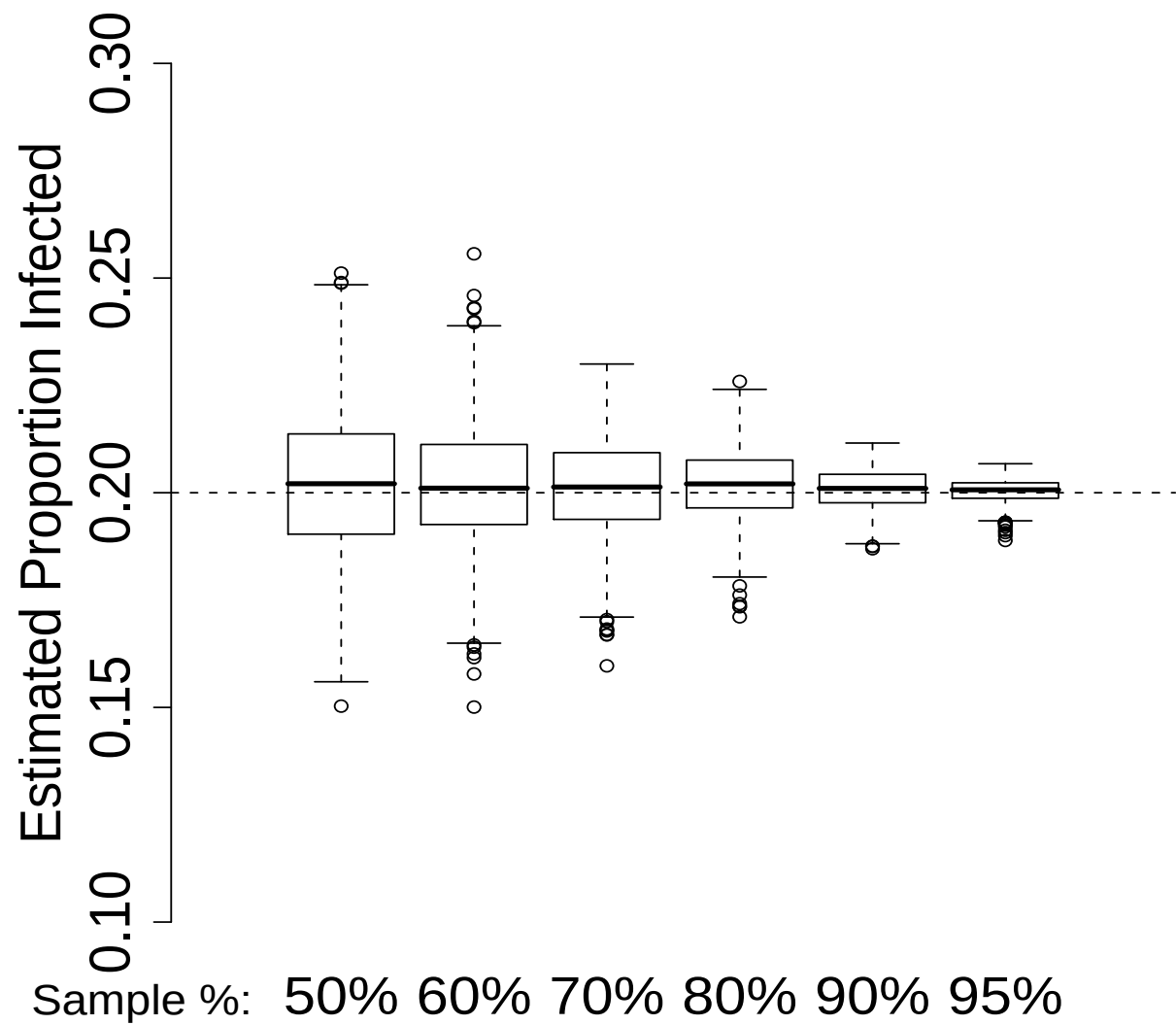
Log MSE by Homophily, All Infected Seeds, $N=1000$, $w=1$



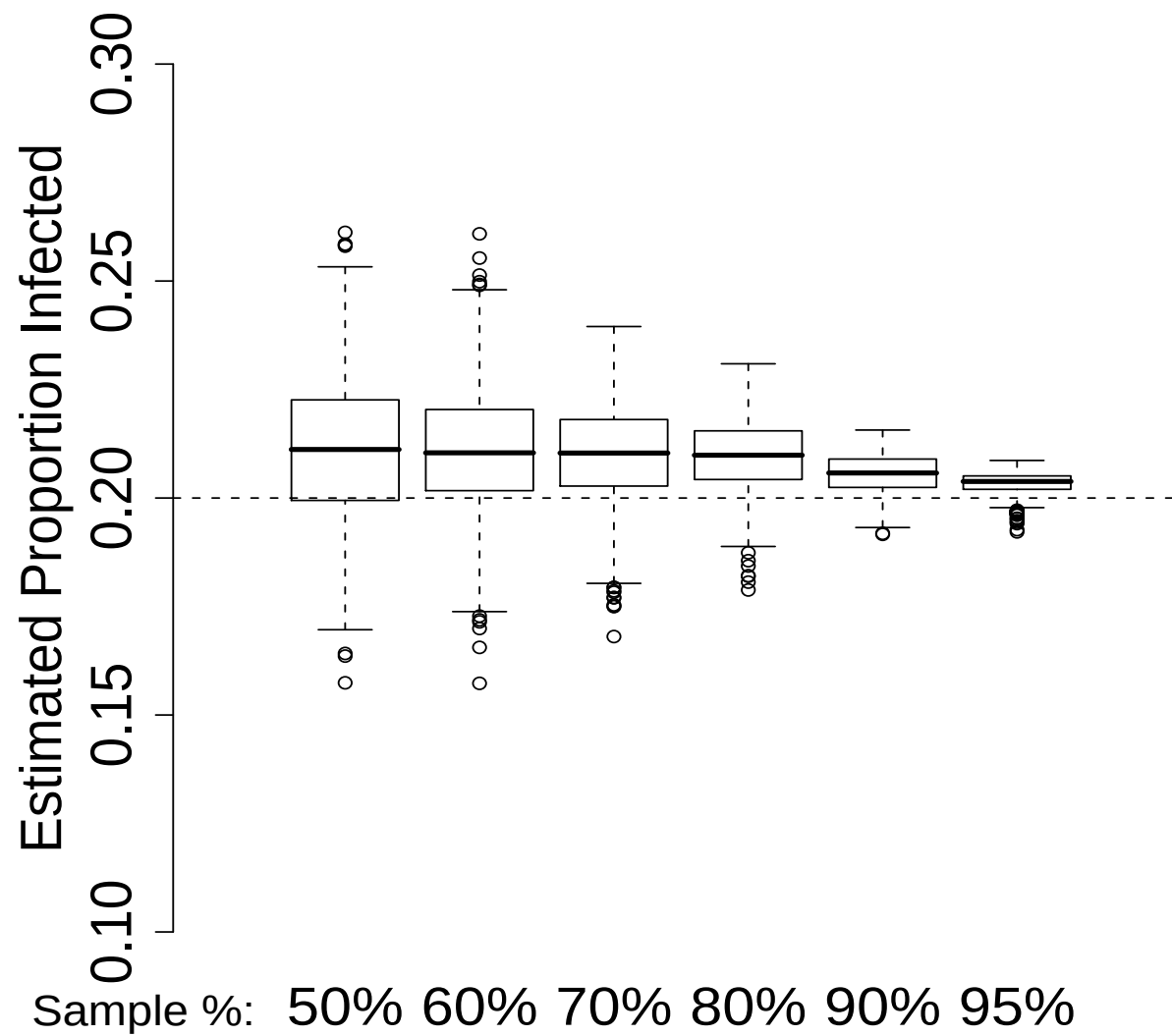
Volz-Heckathorn, $w=1.5$



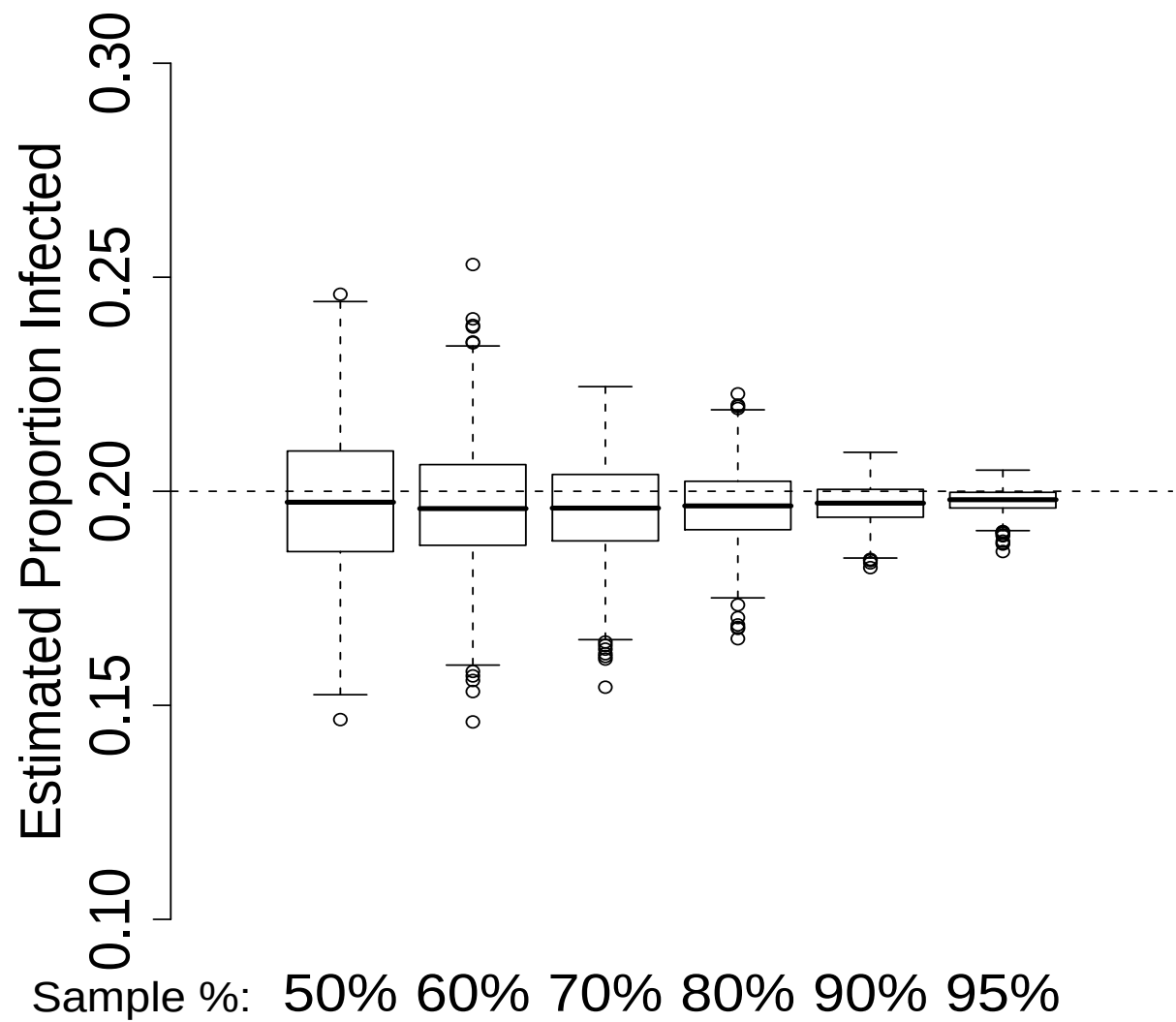
SS, $w=1.5$



$$\mathbf{SS}, w=1.5, \hat{N} < N$$



$$\mathbf{SS}, w=1.5, \hat{N} > N$$



Further Extension: Network-Based Estimator

Premise:

- *Given:*
 - *True network form β (ERGM homophily parameter)*
 - *True sampling structure $\mathcal{S}$ (e.g. seed characteristics)*

Further Extension: Network-Based Estimator

Premise:

- *Given:*
 - *True network form β (ERGM homophily parameter)*
 - *True sampling structure $\mathcal{S}$ (e.g. seed characteristics)*
- *Then:*
 - *Nodal equivalence classes based on degrees k and infection status z*
 - *Exists a mapping $\pi_{k,z}^*(\beta, \mathcal{S}) : \{k, z\} \rightarrow \pi$, from the nodal degree k and value z , to nodal inclusion probabilities, π .*

Further Extension: Network-Based Estimator

Premise:

- *Given:*
 - *True network form β (ERGM homophily parameter)*
 - *True sampling structure $\mathcal{S}$ (e.g. seed characteristics)*
- *Then:*
 - *Nodal equivalence classes based on degrees k and infection status z*
 - *Exists a mapping $\pi_{k,z}^*(\beta, \mathcal{S}) : \{k, z\} \rightarrow \pi$, from the nodal degree k and value z , to nodal inclusion probabilities, π .*

Furthermore,

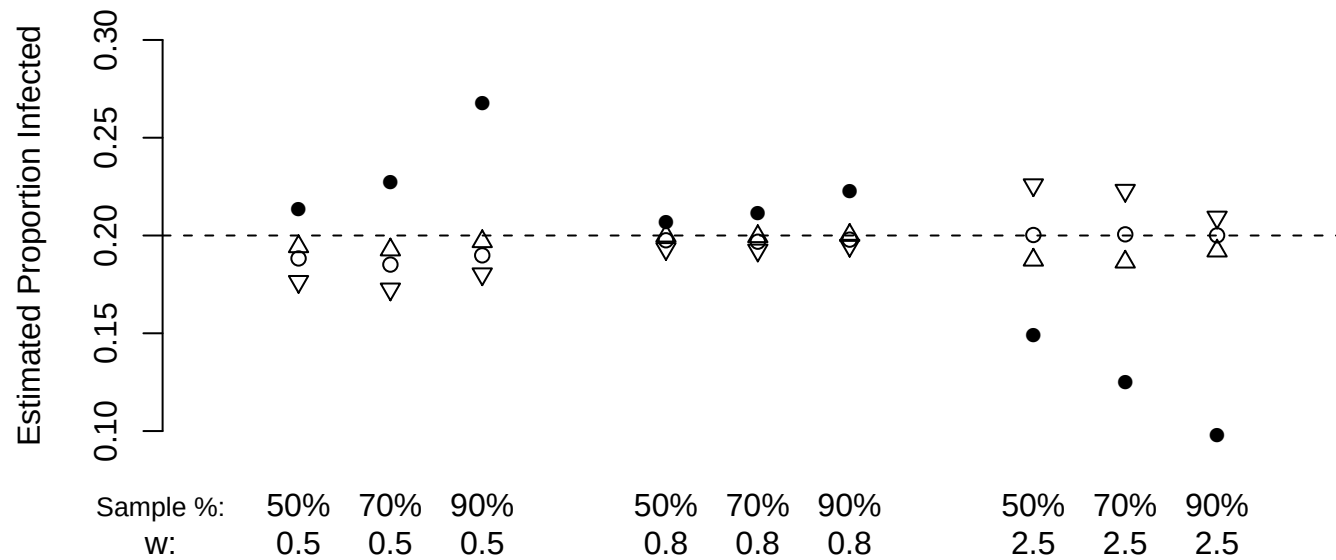
- *Given inclusion probabilities, $\{\pi_i\}$, use Horvitz-Thompson estimators to estimate the network and sampling structures, $\{\beta, \mathcal{S}\}$.*
- *Can simulate networks according to β .*
- *Can simulate samples according to $\mathcal{S}$.*
- *Use simulations to estimate mapping, $Q_{k,z}^*(\beta, \mathcal{S})$.*

Network-Based Algorithm

- Estimate Q^* proportional to degree k .
- Iterate the following steps:
 - Estimate population distribution of degrees k by infection status z
 - Estimate network features β based on observed data and weights
 - Estimate sampling features $\mathcal{S}$ based on observed data and weights
 - Simulate M networks, and samples from networks. Estimate weights.
- Use the resulting estimated probabilities, $\hat{Q}^*$, to form weighted estimator.

$$\hat{\mu}_M^* = \frac{\sum_{j:S_j=1} \frac{z_j}{\hat{Q}_{d_j, z_j}^*}}{\sum_{j:S_j=1} \frac{1}{\hat{Q}_{d_j, z_j}^*}}.$$

Population Size and w



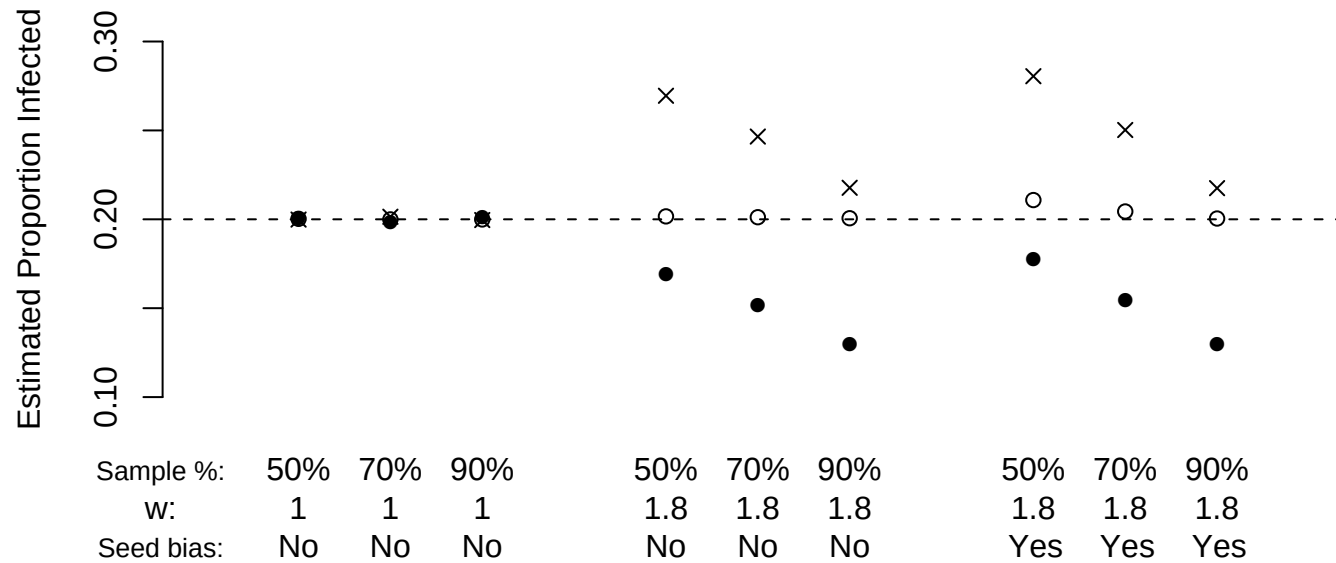
Solid Dot: VH

Open Dot: SS

Up Triangle: $\hat{N}$ too big

Down Triangle: $\hat{N}$ too small

V-H, SS, and Mean

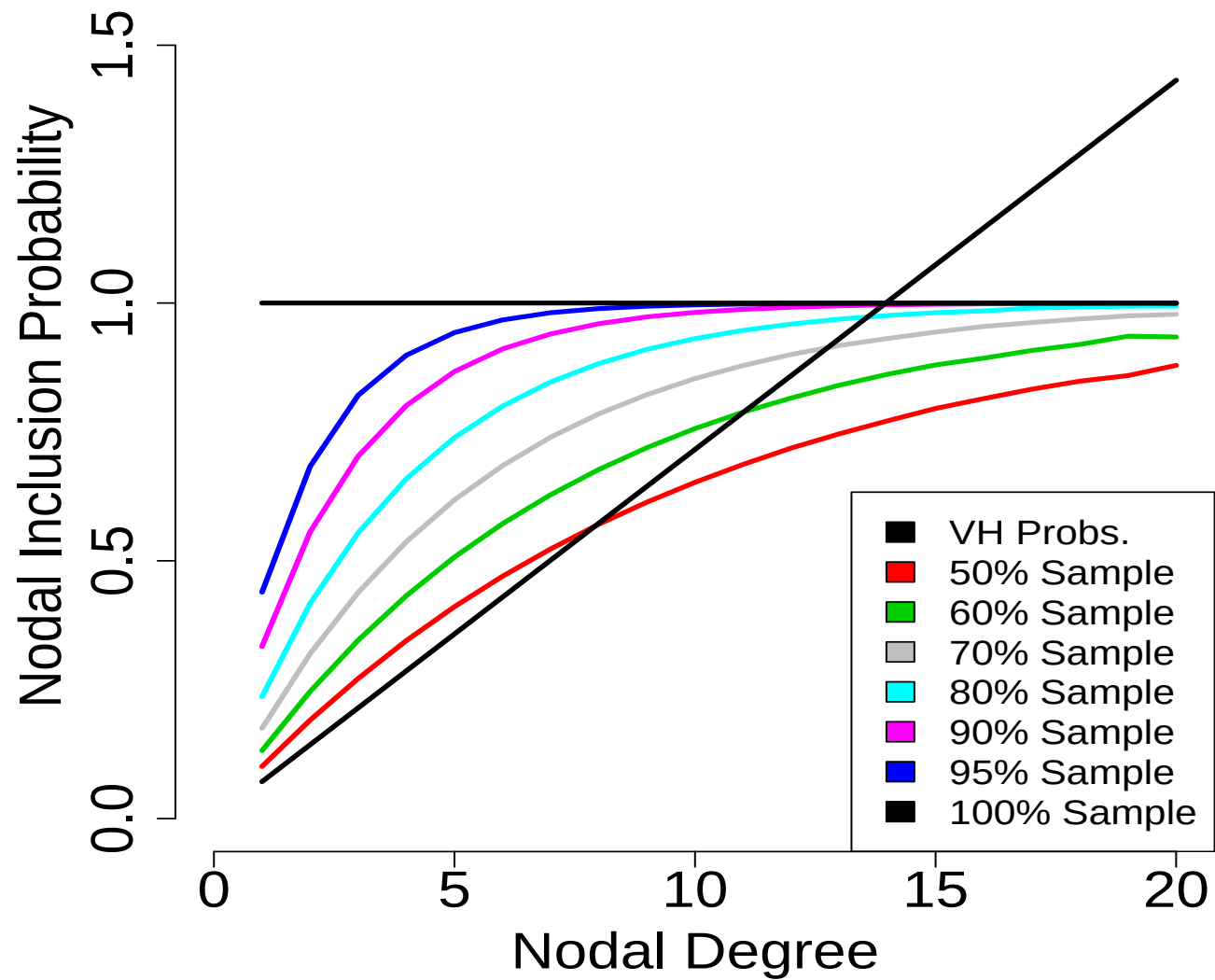


Solid Dot: VH

Open Dot: SS

x: sample mean

Mapping $Q_k(\mathbb{N}) : k \rightarrow \pi$

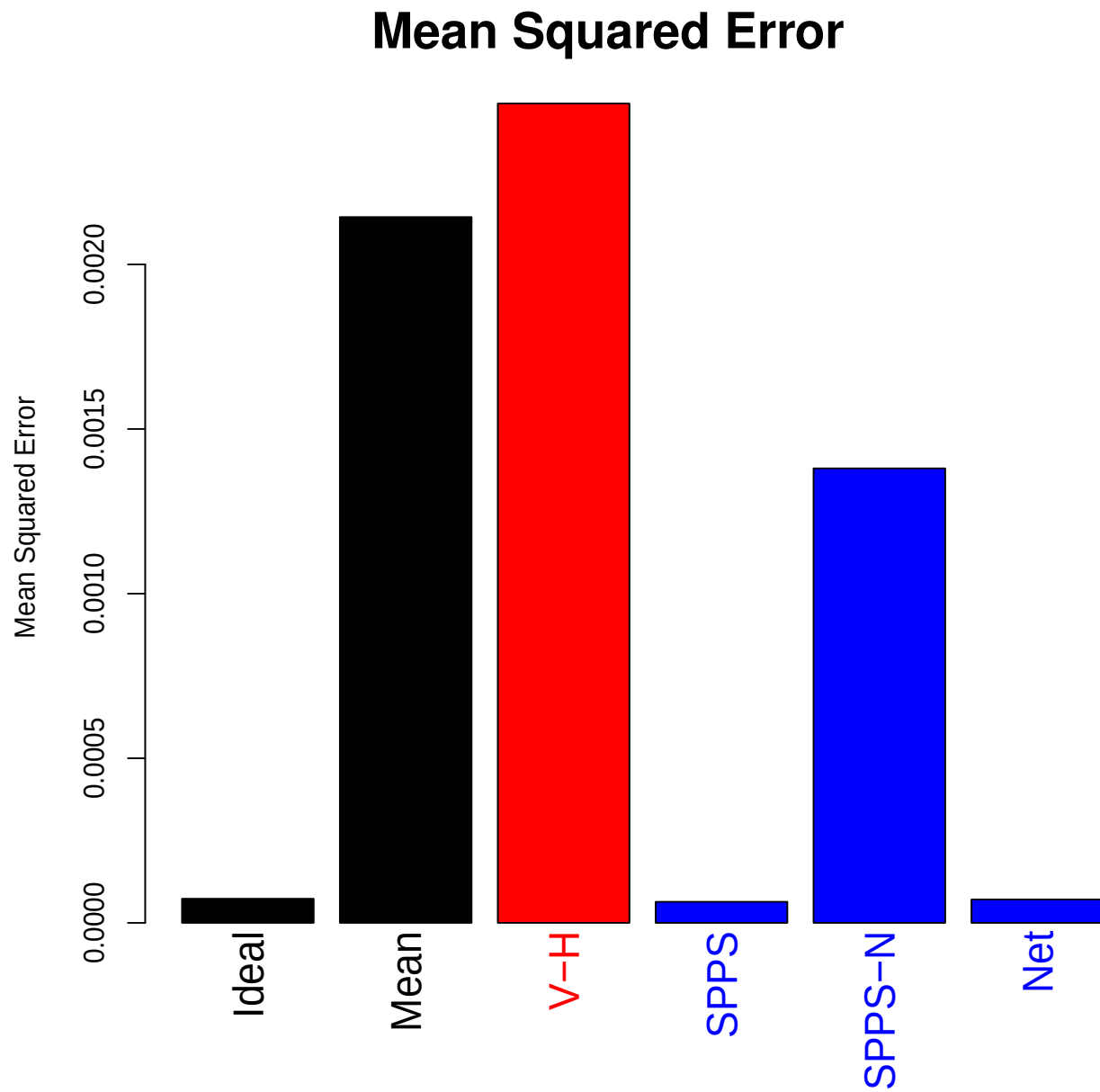


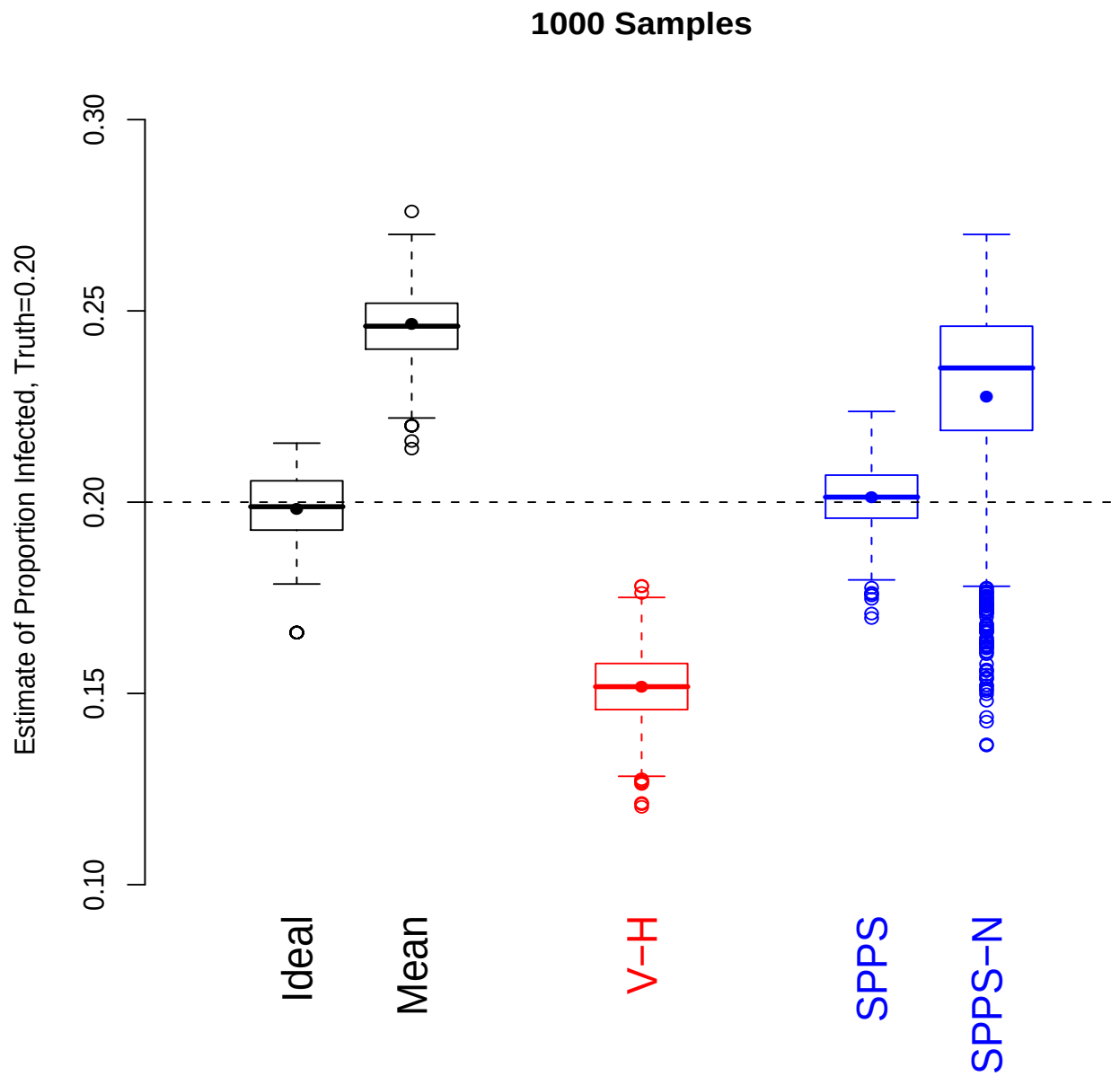
Current RDS Estimation Assumptions

	<i>Network Structure Assumptions</i>	<i>Sampling Assumptions</i>
<i>Random Walk Model</i>	<i>Network size large ($N \gg n$)</i>	<i>Sampling with replacement Single non-branching chain</i>
<i>Remove Seed Dependence</i>	<i>Homophily weak enough Connected graph</i>	<i>Sufficiently many sample waves</i>
<i>To Estimate Probabilities</i> <i>($\pi_i \propto \tilde{d}_i$)</i>	<i>All ties reciprocated</i>	<i>Degree accurately measured Random referral</i>

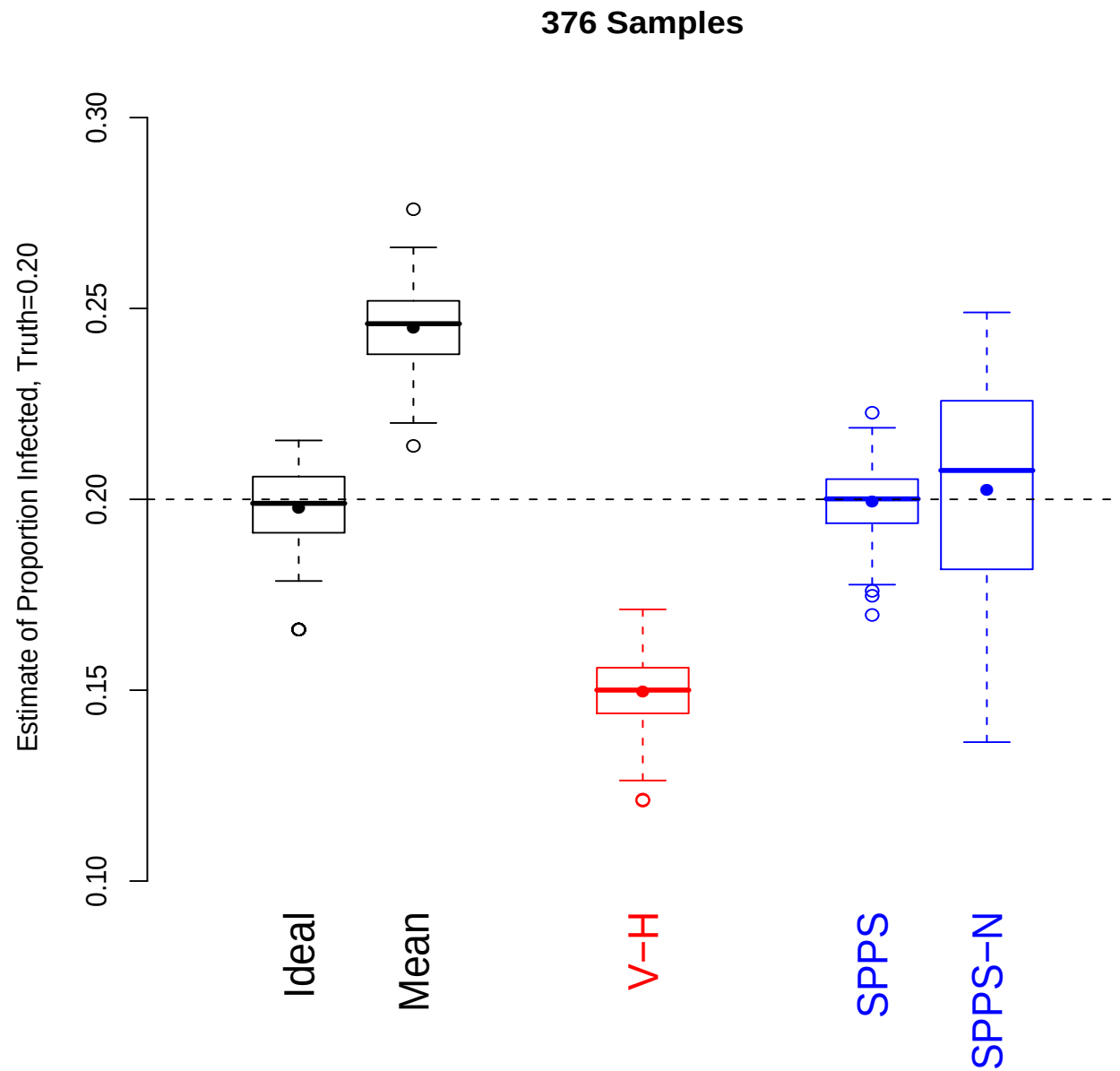
Estimating Population Size

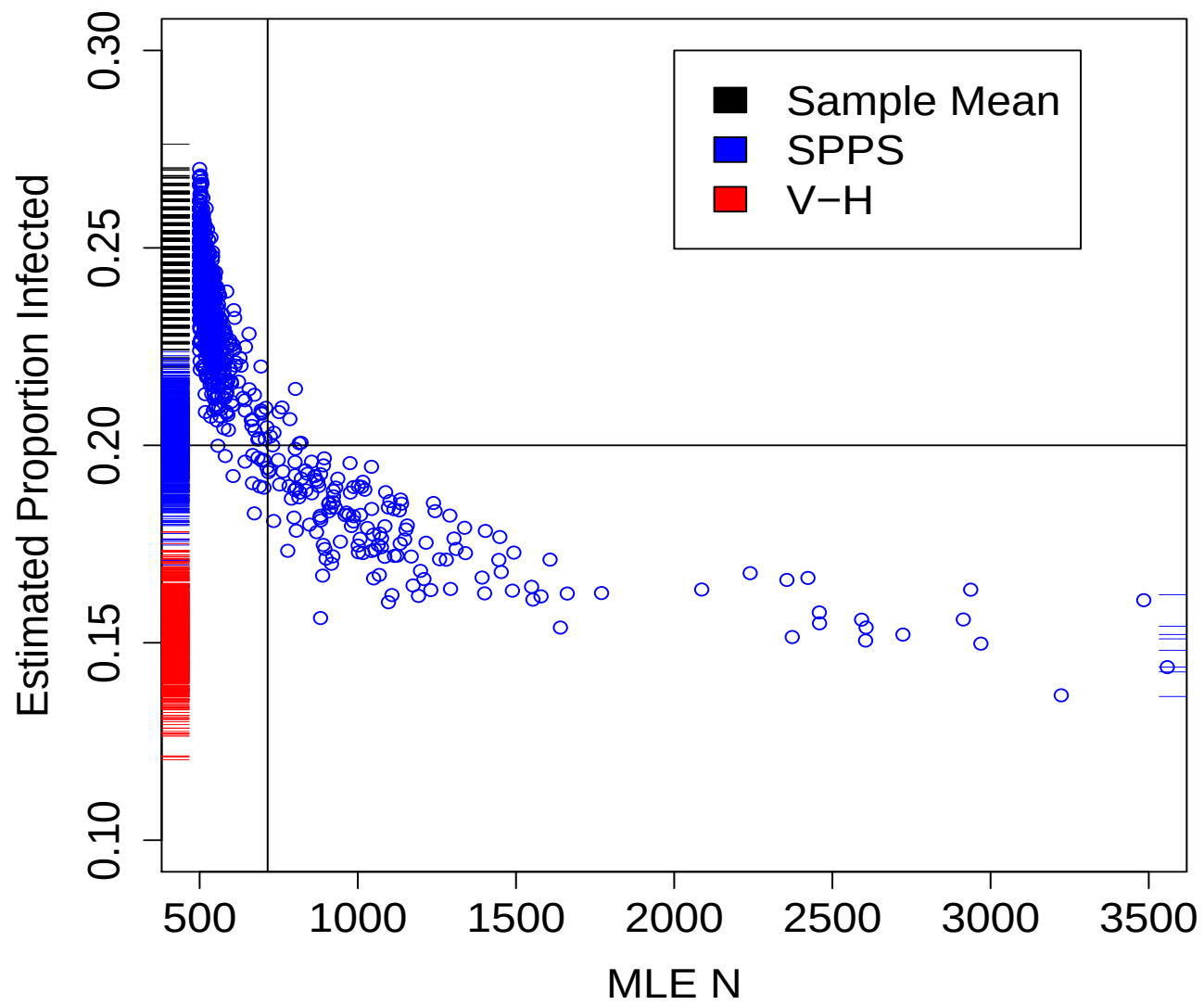
- *Use SS model*
- *Premise:*
 - *High-degree nodes tend to be sampled first.*
 - *Time-decreasing degree of sampled nodes provides information on population size*
- *First approach: MLE. Next: Bayesian.*





Consider $\hat{N} > 540$ only

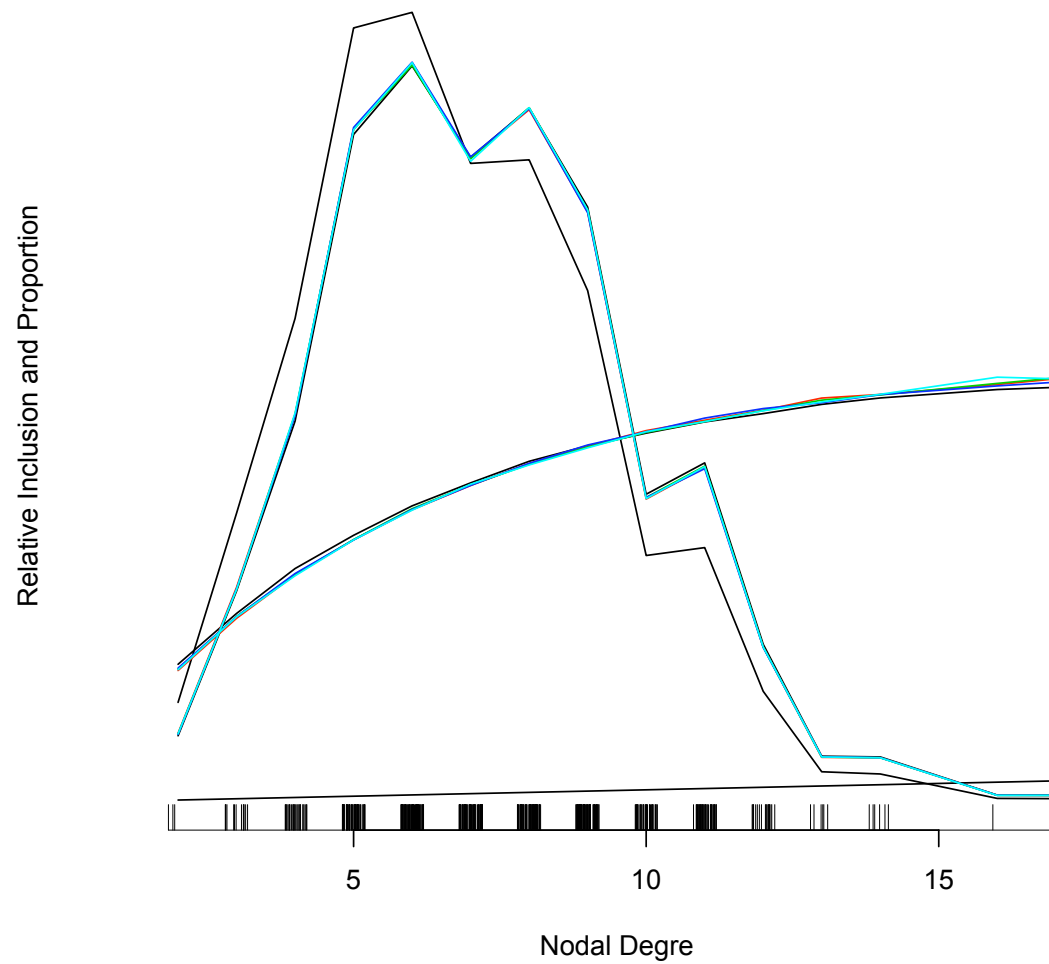




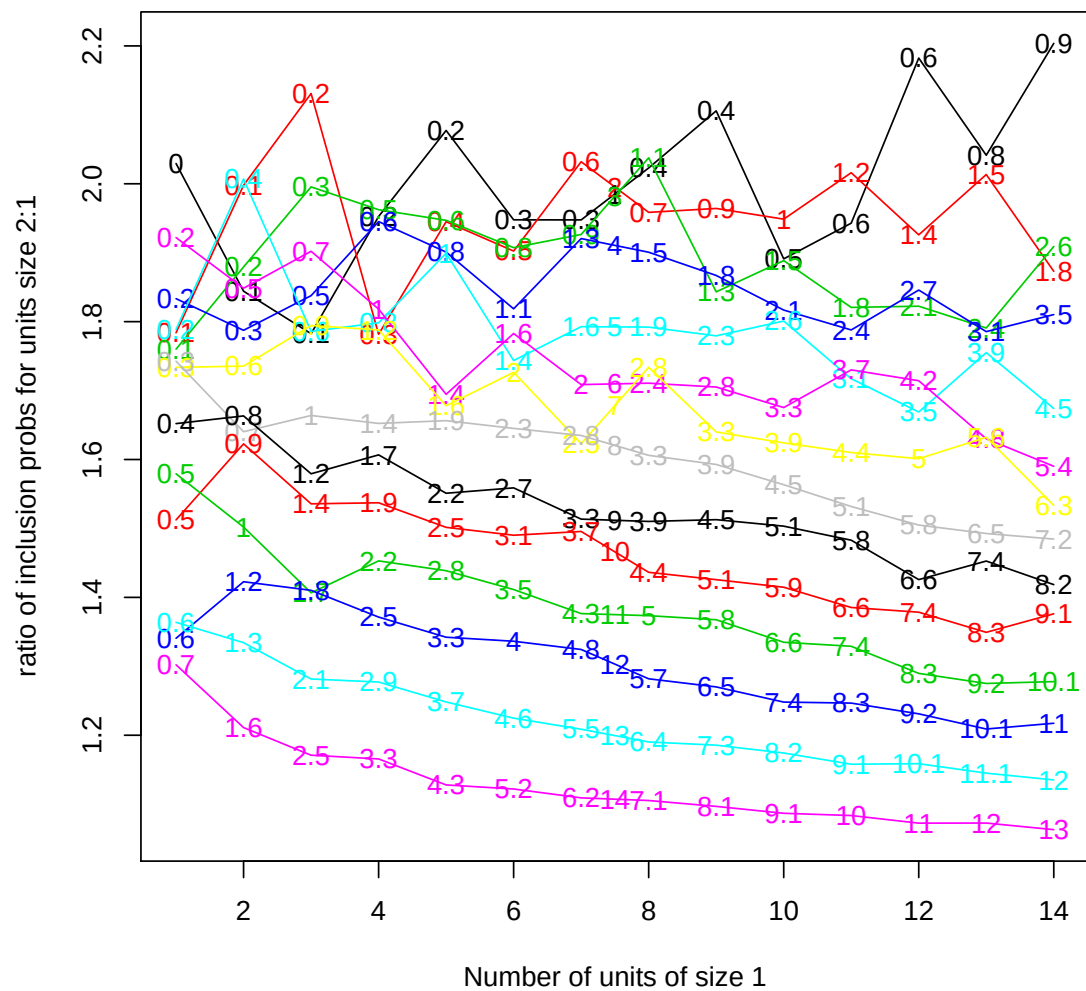
Summary of Findings:

- *The data are somewhat informative*
- *The estimation procedure requires improvement*
- *Are the data informative enough to be useful?*

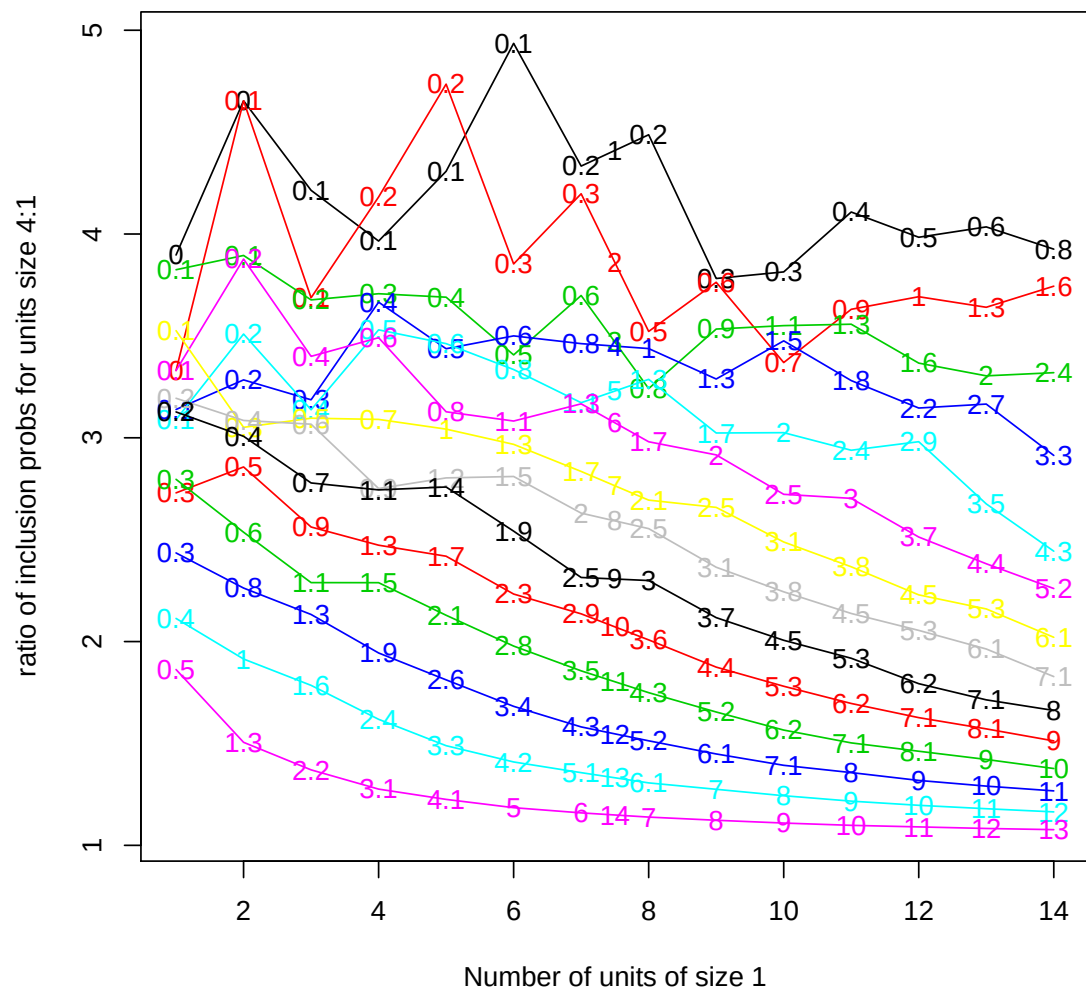
Algorithm-Related Plots



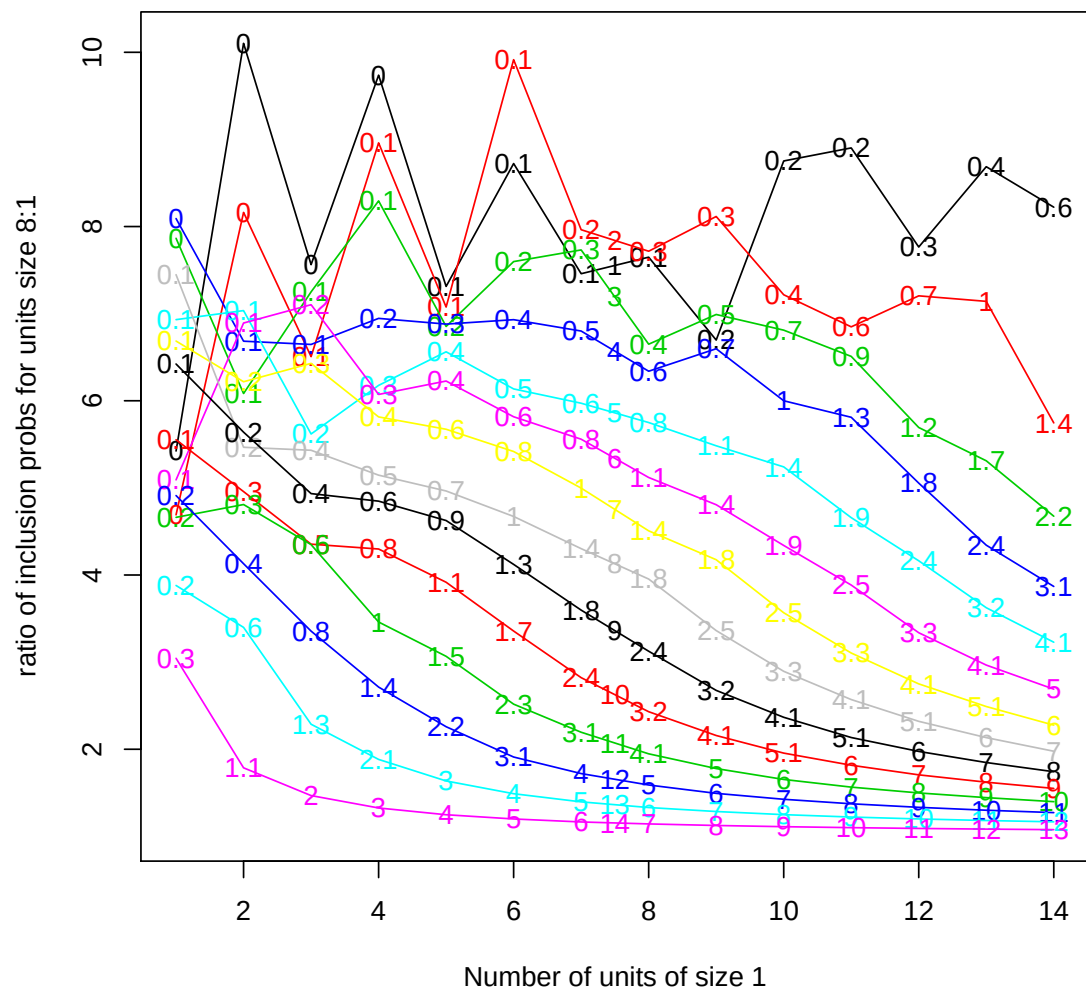
Estimated SPPS Inclusion Probabilities based on 1000 samples
 For sample size given by color and colored number, from population size 15
 Two nodal sizes:(1,2), number of 1's given by x axis



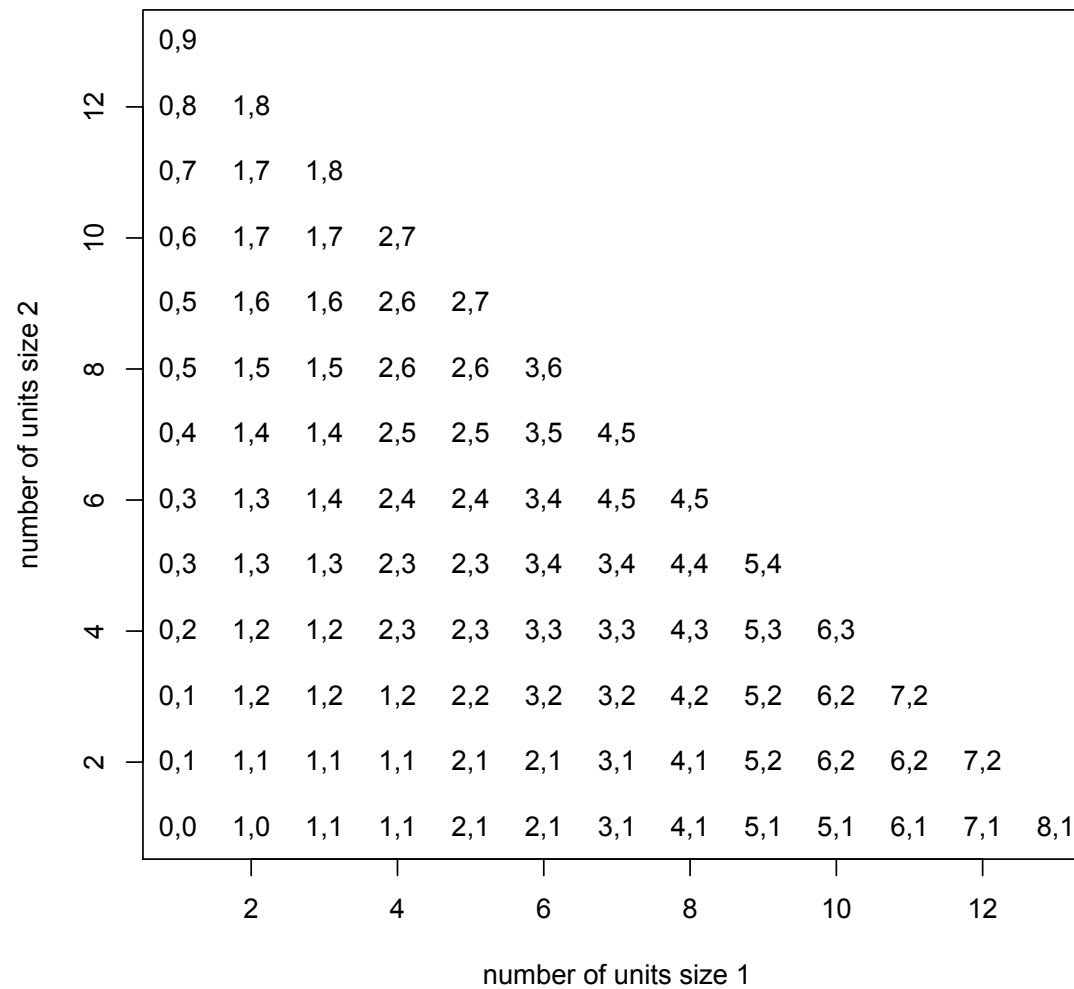
Estimated SPPS Inclusion Probabilities based on 1000 samples
 For sample size given by color and colored number, from population size 15
 Two nodal sizes:(1,4), number of 1's given by x axis



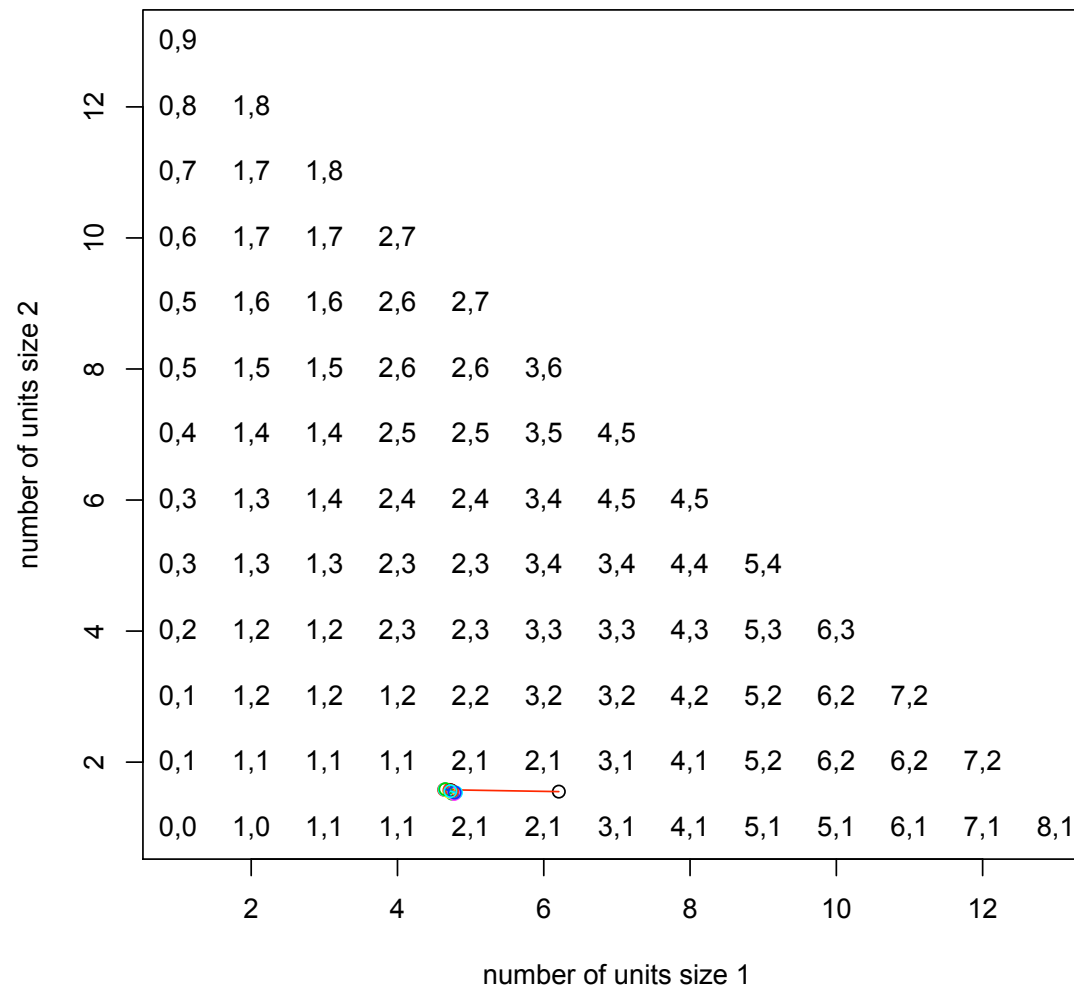
Estimated SPPS Inclusion Probabilities based on 1000 samples
 For sample size given by color and colored number, from population size 15
 Two nodal sizes:(1,8), number of 1's given by x axis

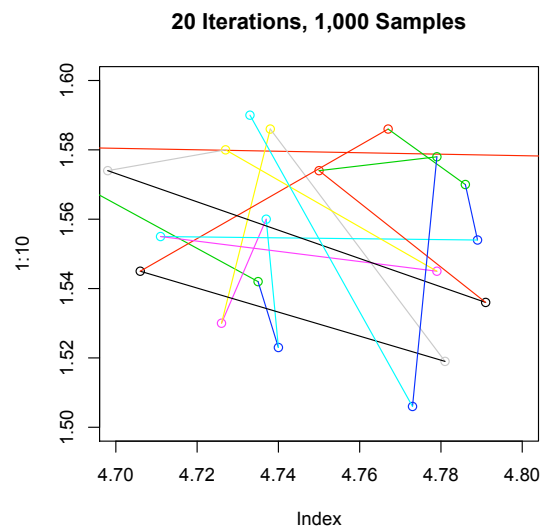


Expected Counts, N=15, n=10, d={1,2,3}

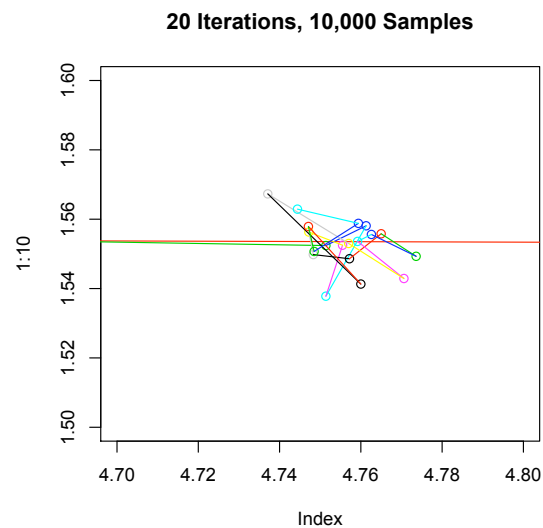


Estimated Population, 20 Iterations, Observed: 2 1's, 1 2's, 7 3's

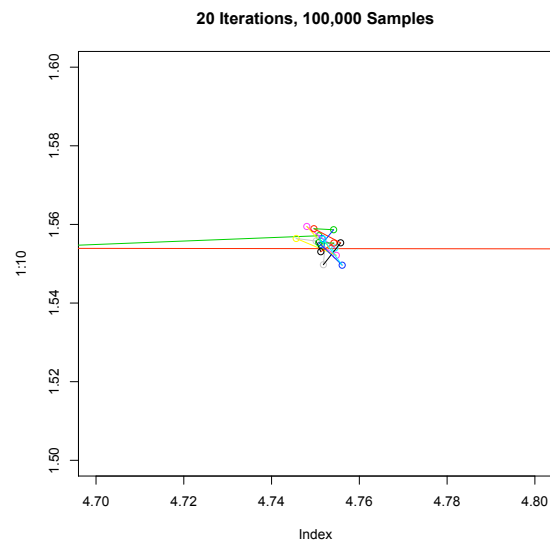




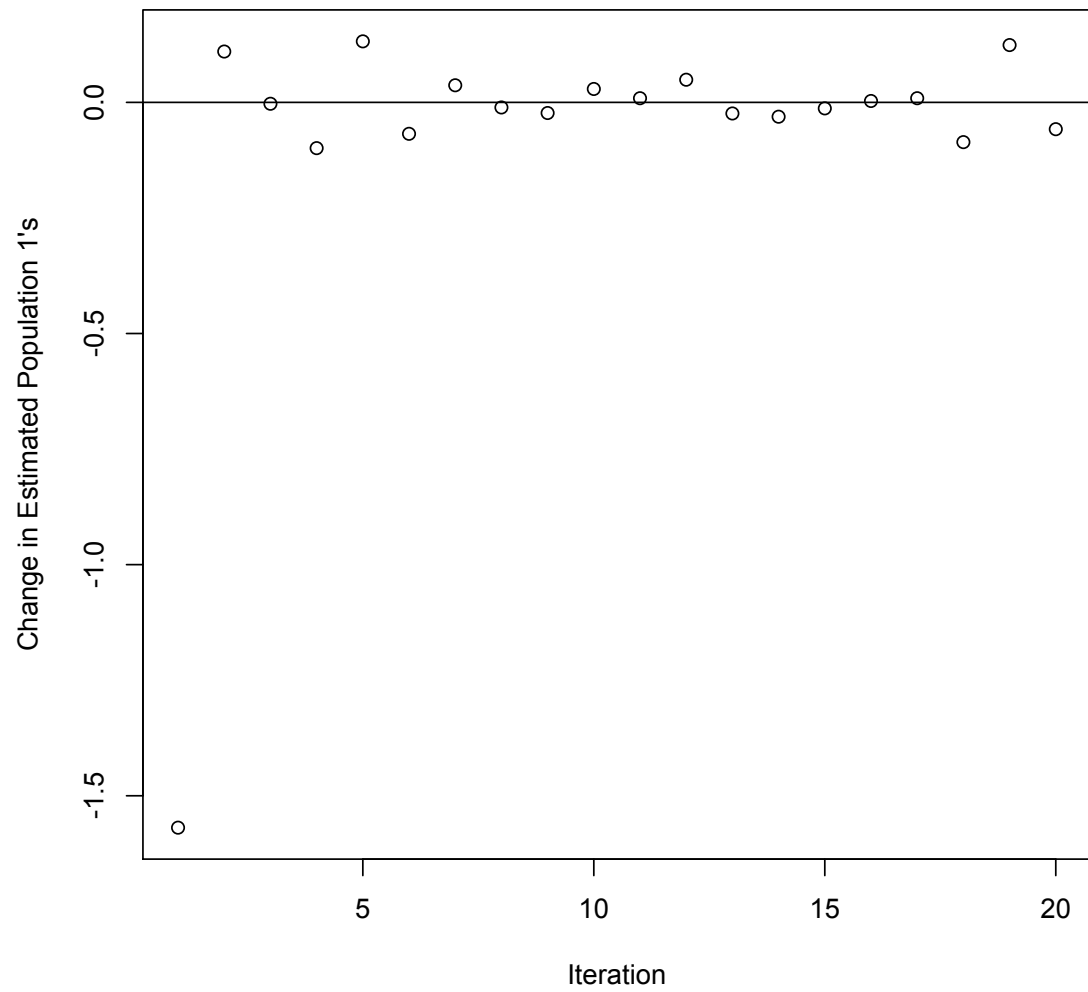
(e) 1000 Samples



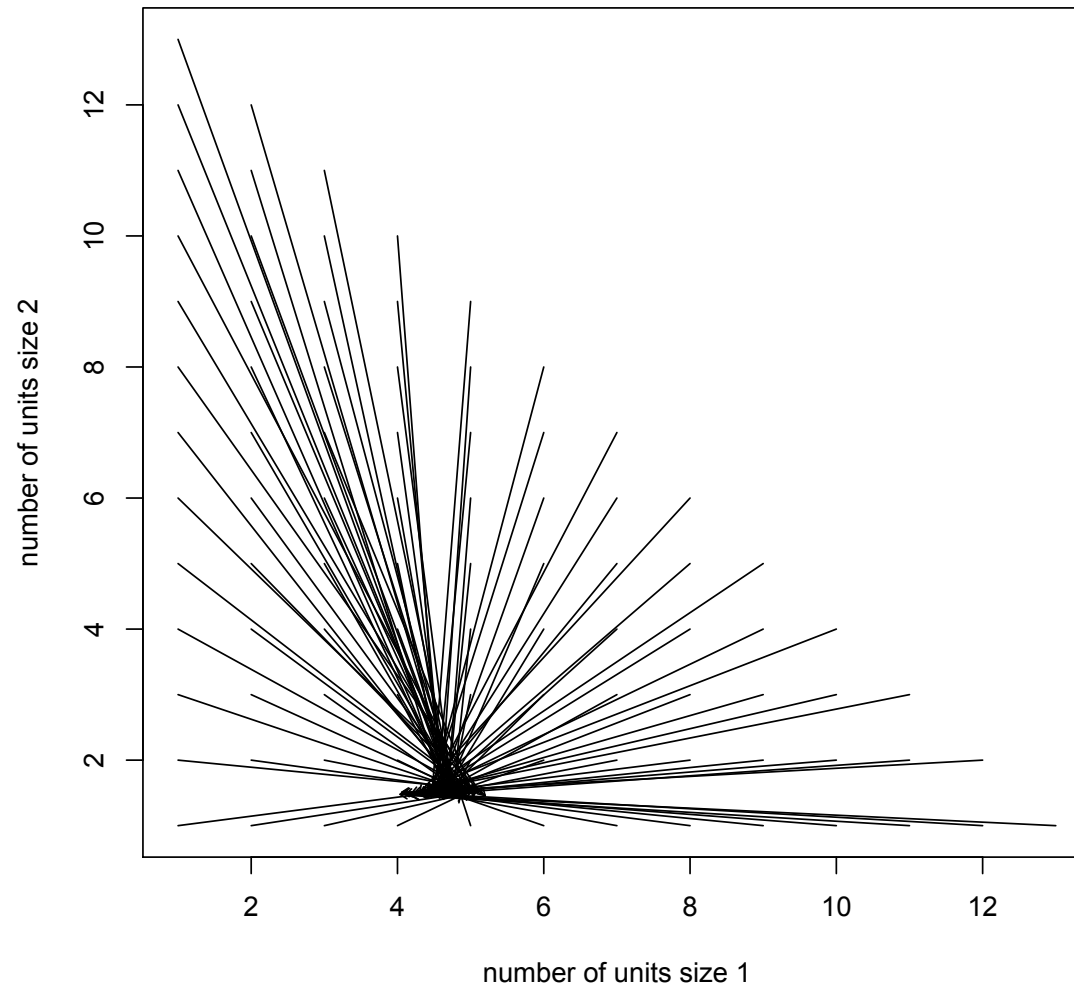
(f) 10,000 Samples



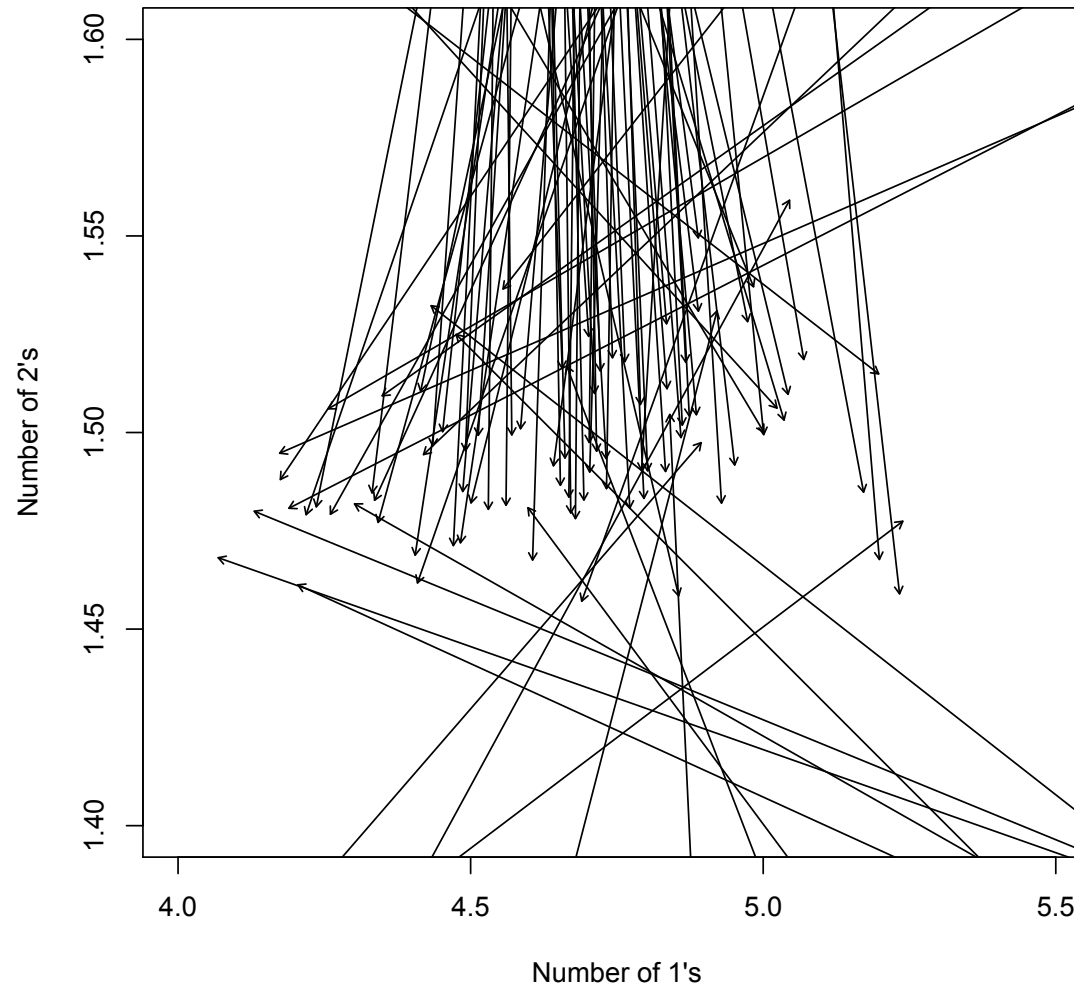
(g) 100,000 Samples



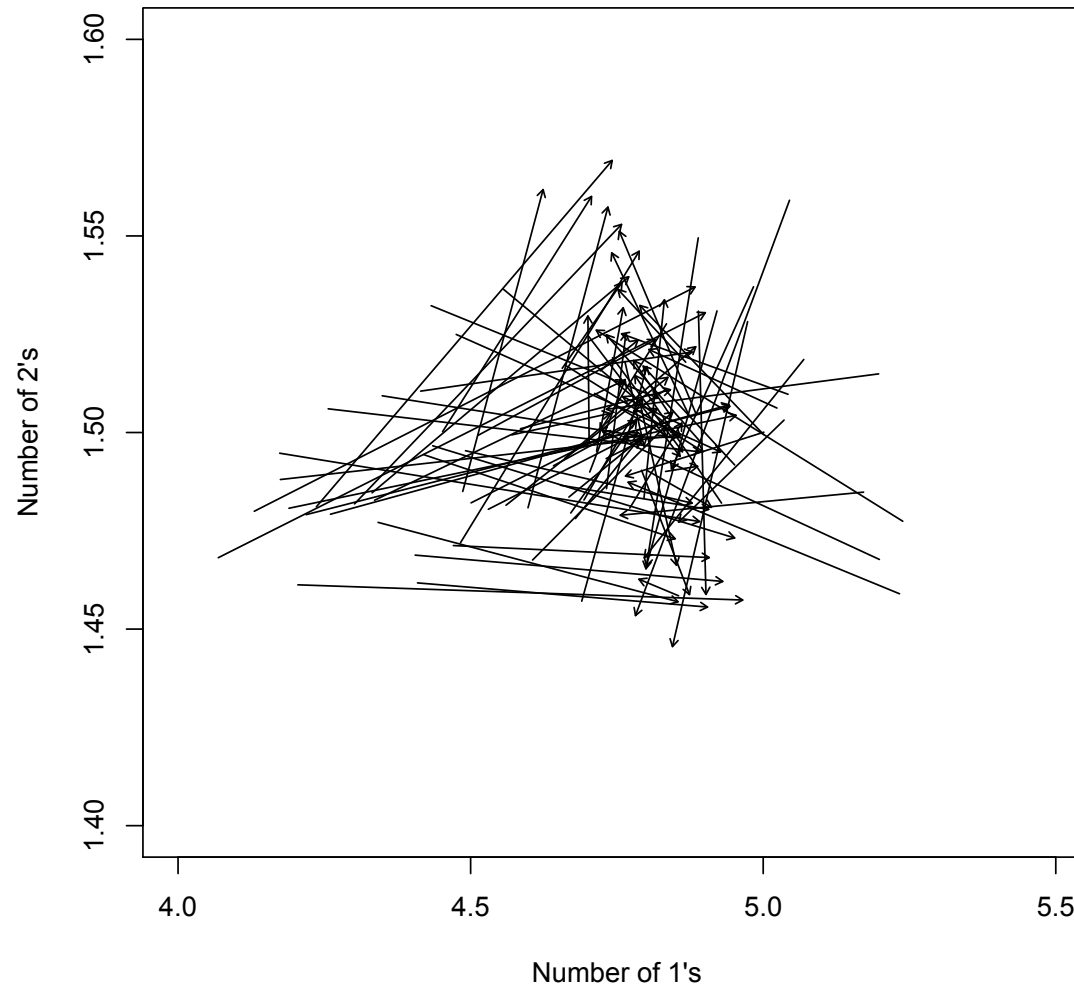
Mapping from Population to Population, Observe 2 1's, 1 2's, 7 3's



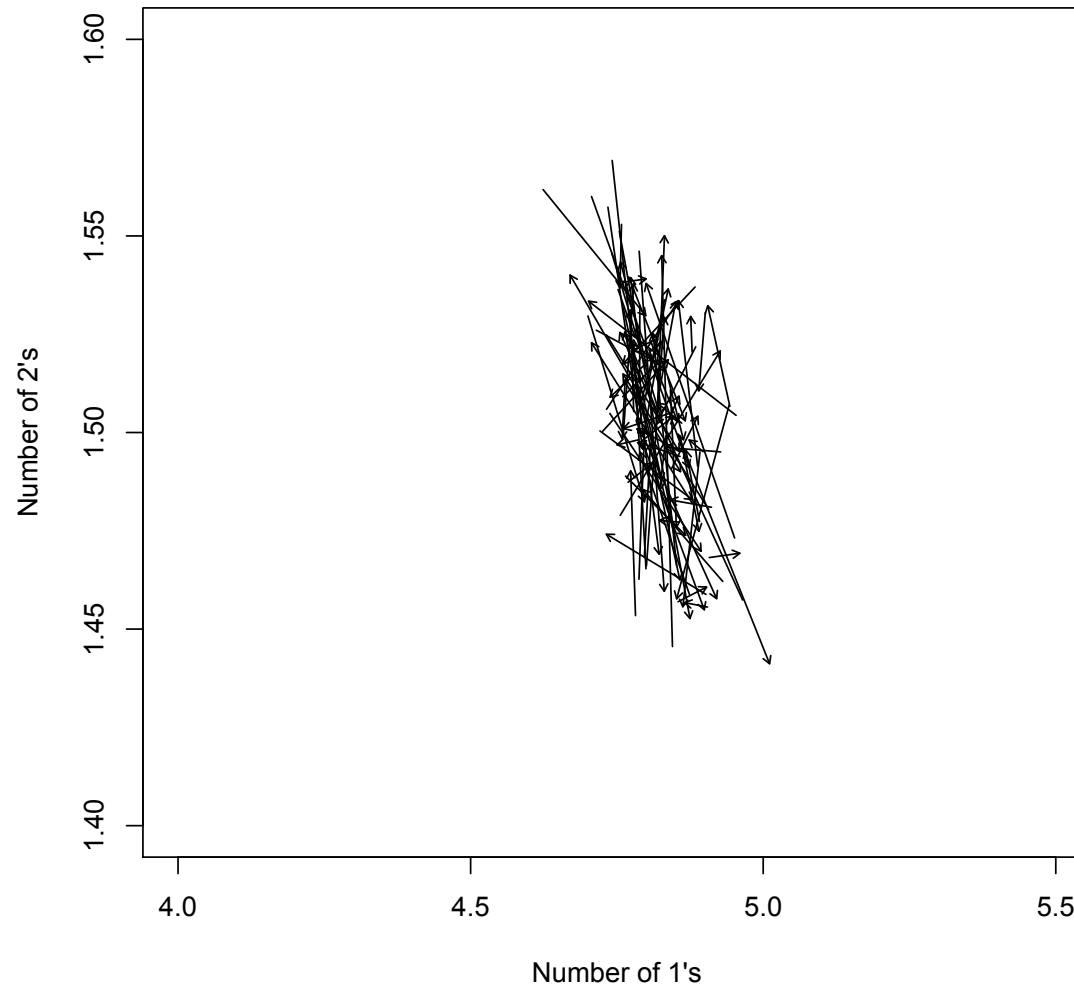
Mapping, Smaller Range, 1,000 Samples



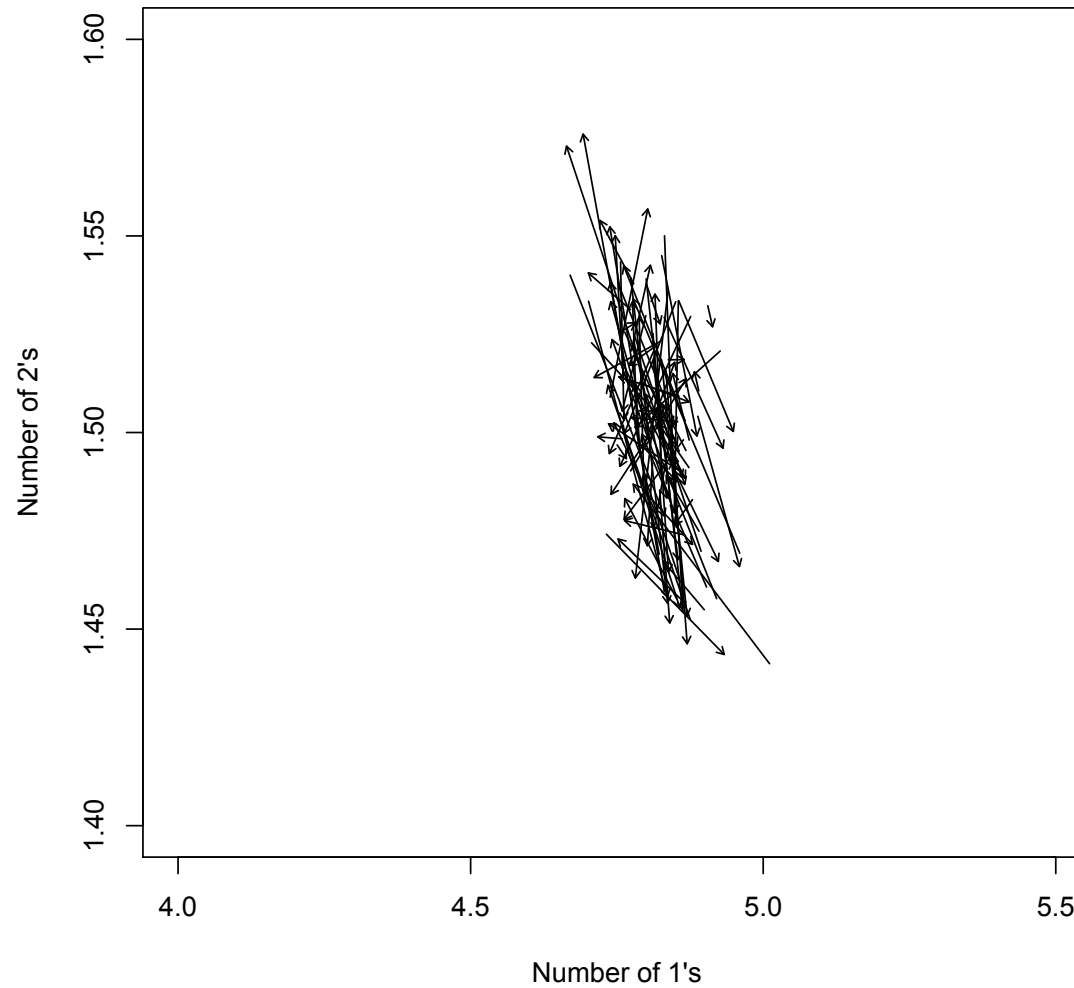
Mapping, Smaller Range, 1,000 Samples, Iteration 2



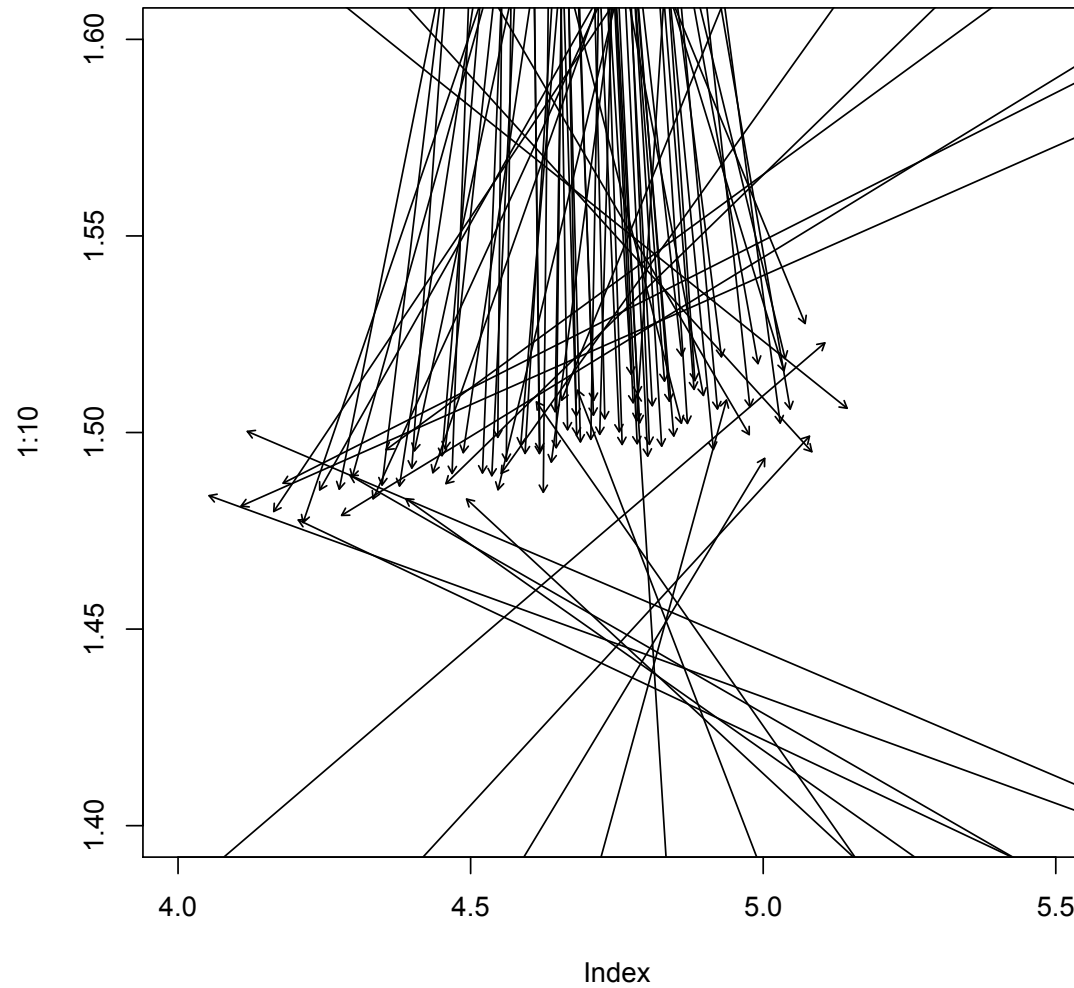
Mapping, Smaller Range, 1,000 Samples, Iteration 3

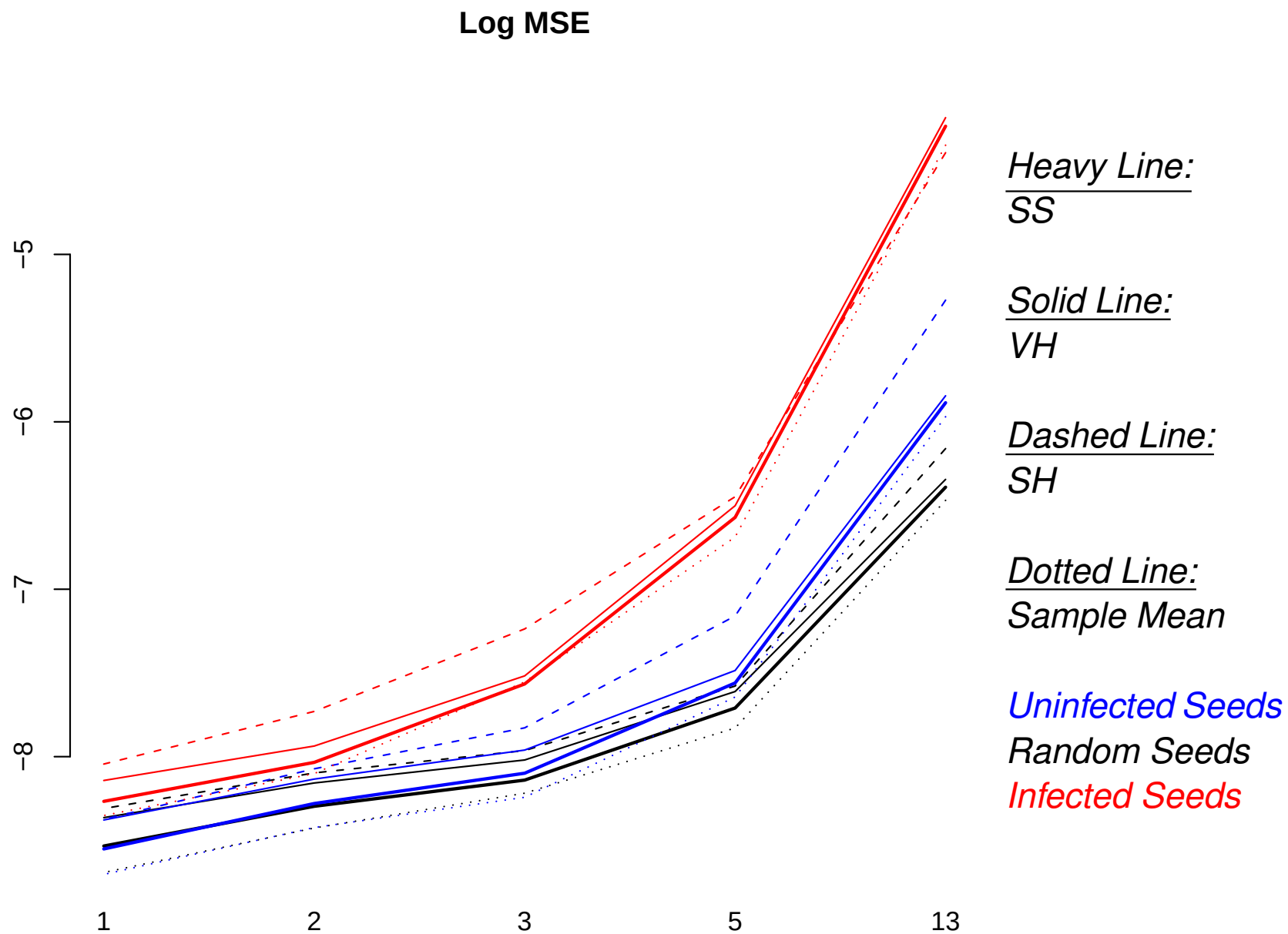


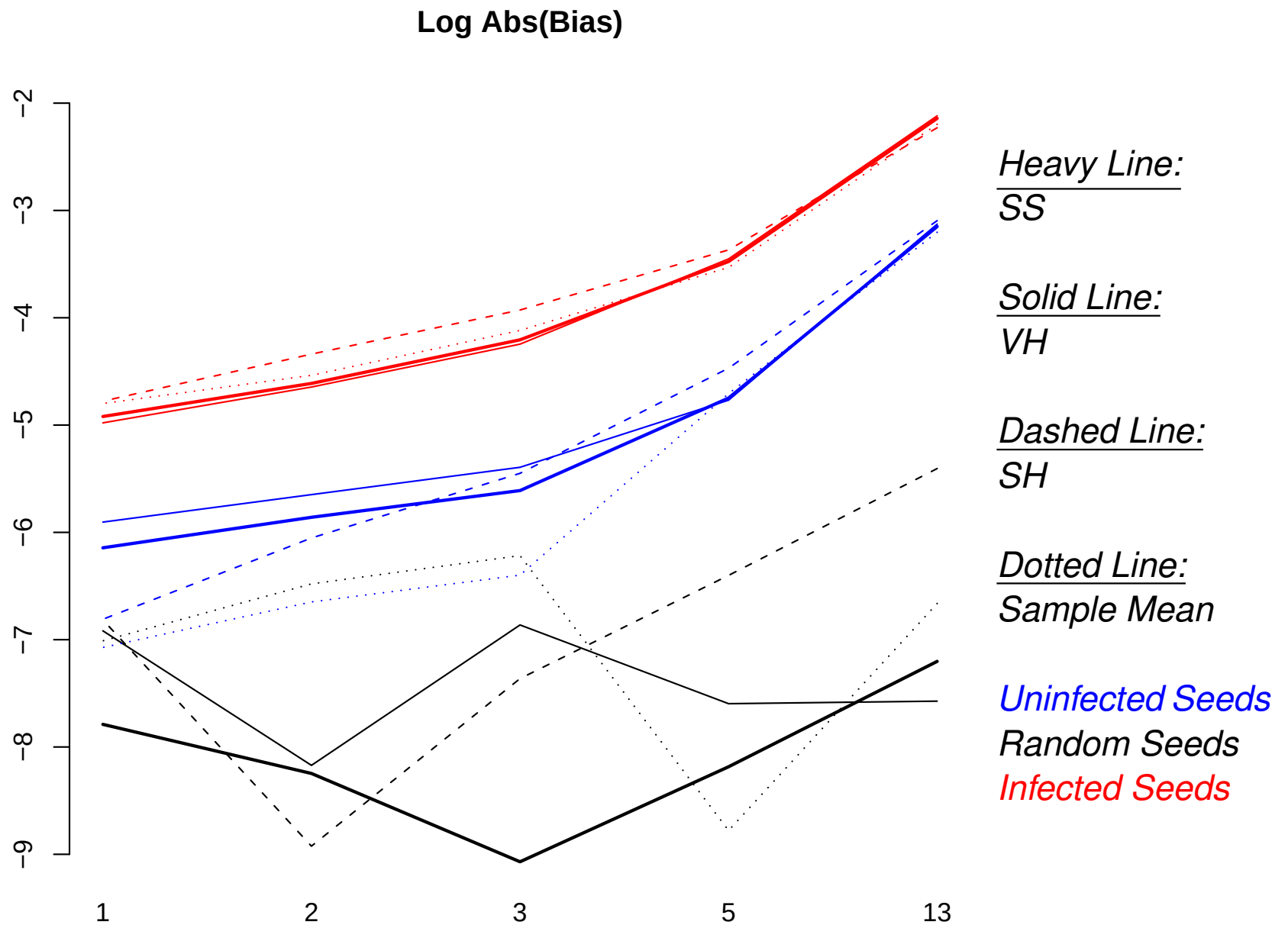
Mapping, Smaller Range, 1,000 Samples, Iteration 4



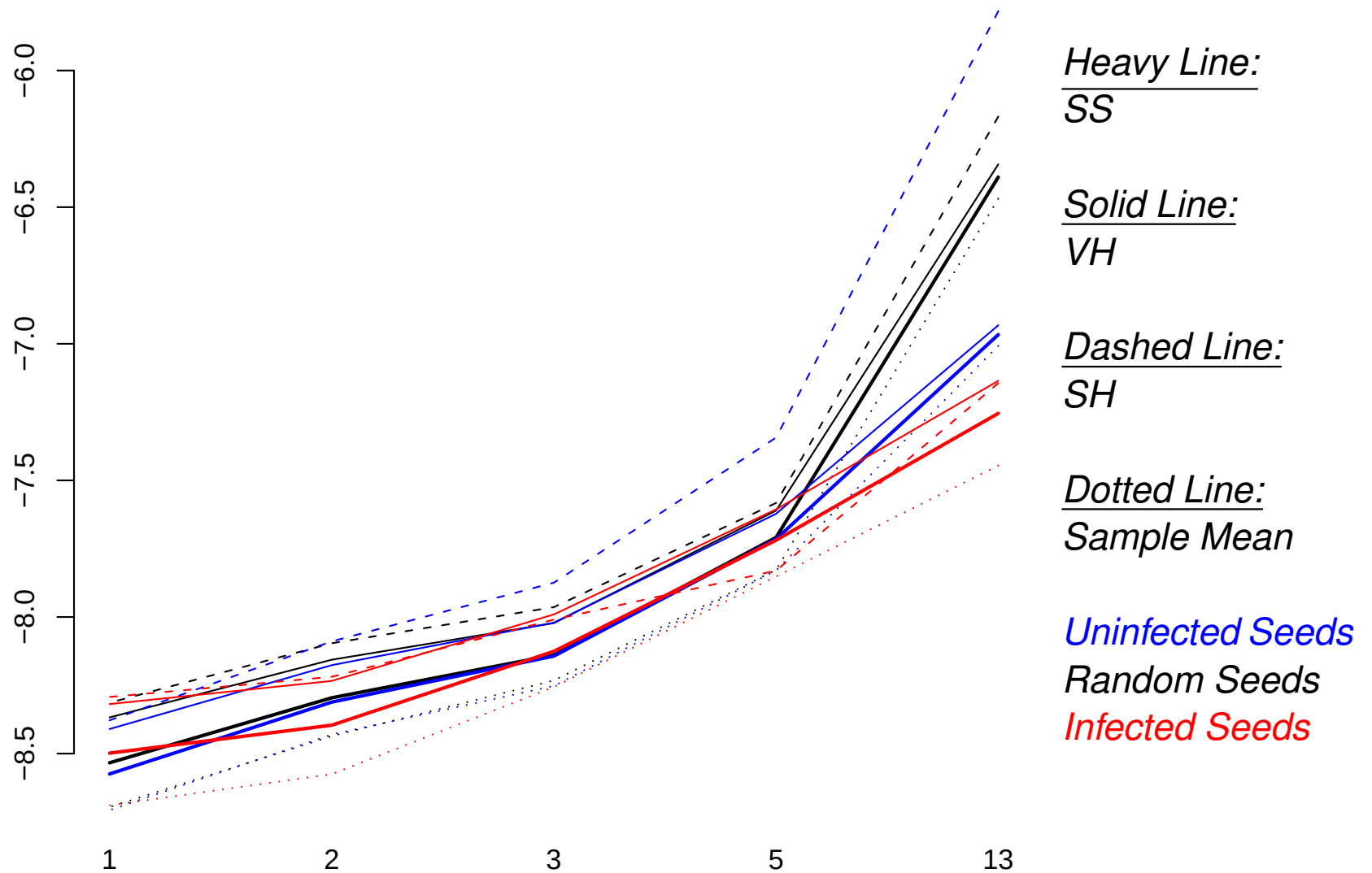
Mapping, Smaller Range, 10,000 Samples

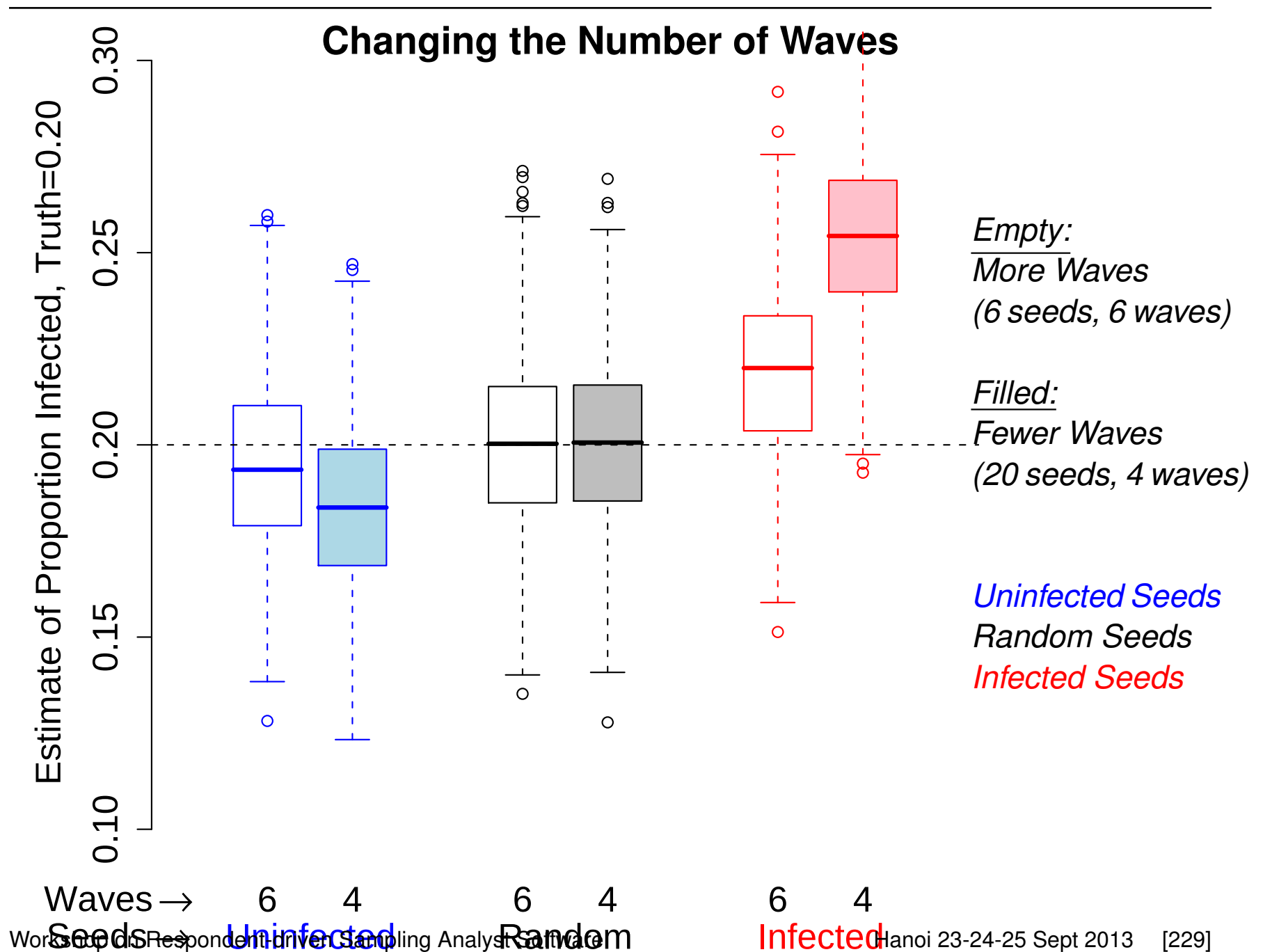




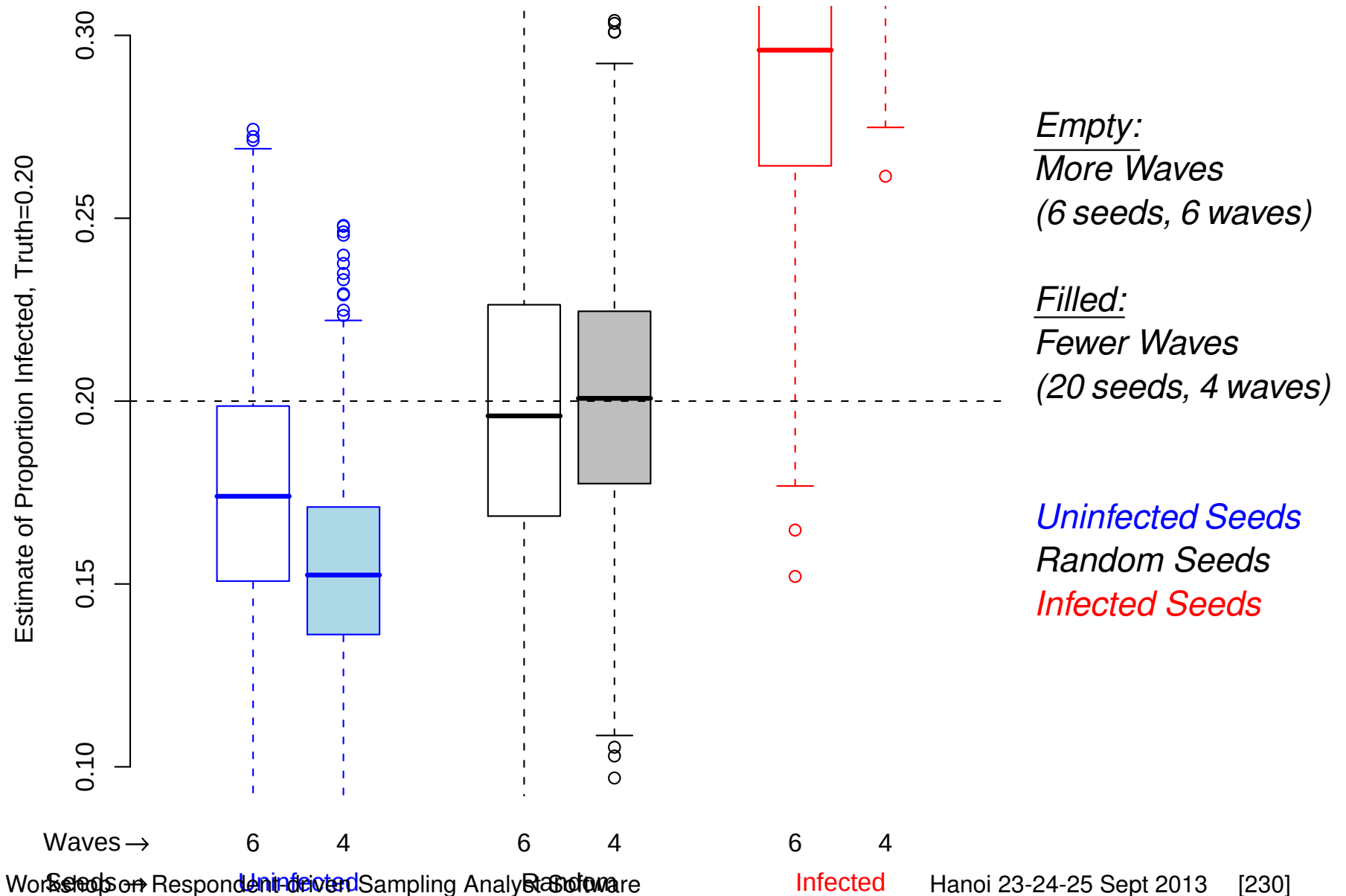


Log Variance





Changing the Number of Waves, With Replacement

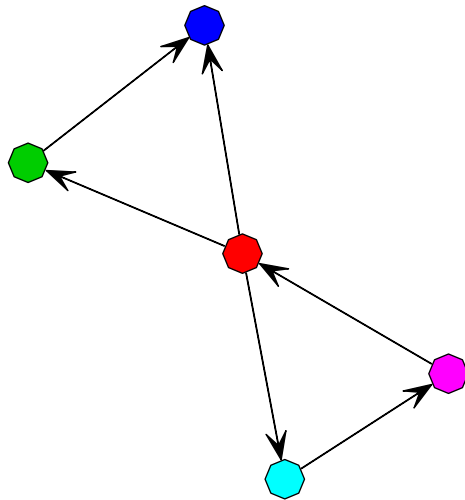


“Two Wrongs Make a Right:” Without replacement and list-wise sampling

- *Probabilities proportional to degree depends on Random Walk Model*
 - *With replacement*
 - *Draw-wise sampling probabilities (i.e. “at this step”)*
- *RDS violates both... and it helps!*
 - *Without replacement*
 - *List-wise sampling probabilities (i.e. “ever included”)*

Social Networks

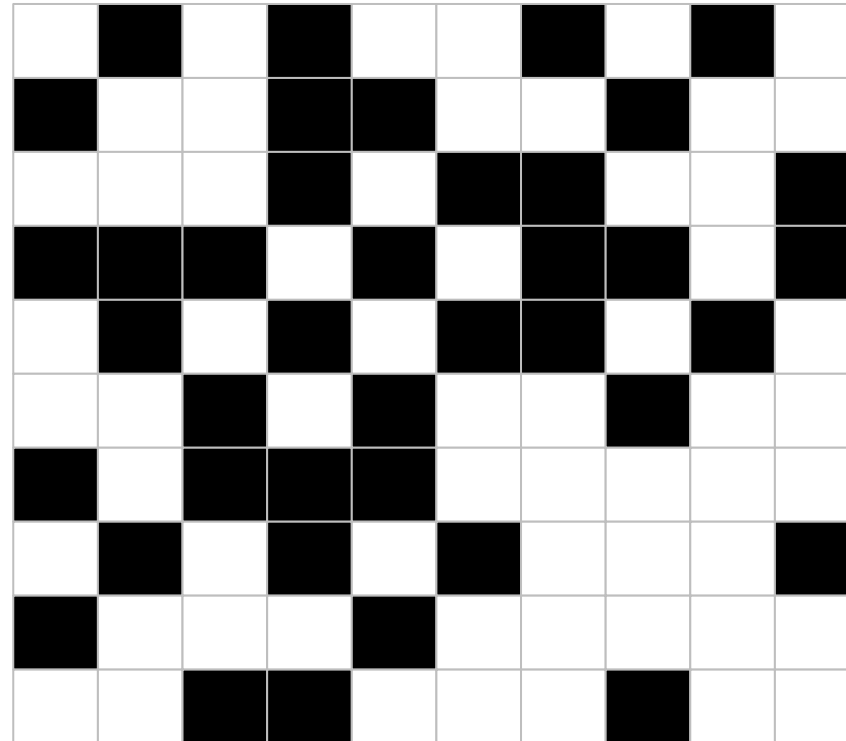
- *Social Network: Tool to formally represent and quantify relational social structure.*
- *Represent mathematically as a sociomatrix, Y , where y_{ij} = the value of the relationship from i to j*



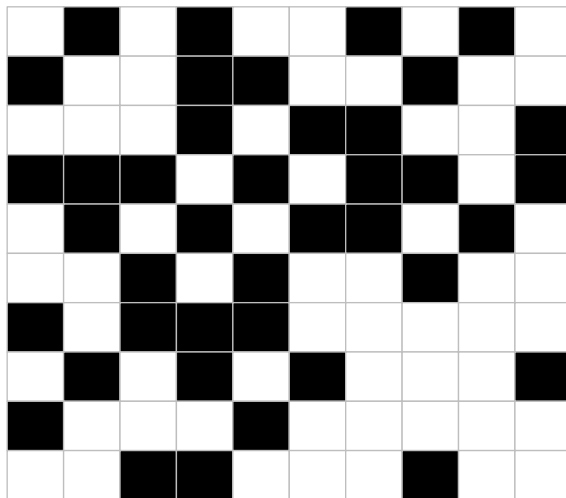
(h) Sociogram

	Red	Green	Blue	Cyan	Magenta
Red	0	1	1	1	0
Green	0	0	1	0	0
Blue	0	0	0	0	0
Cyan	0	0	0	0	1
Magenta	1	0	0	0	0

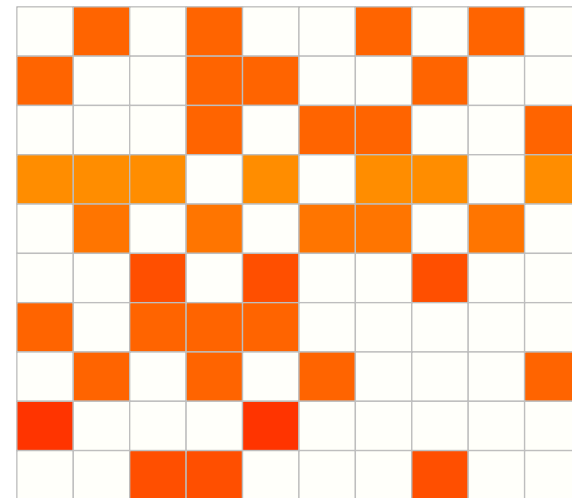
(i) Sociomatrix



0	1/16	1/8	3/16
1/4	5/16	3/8	7/16
1/2	9/16	5/8	11/16
3/4	13/16	7/8	15/16

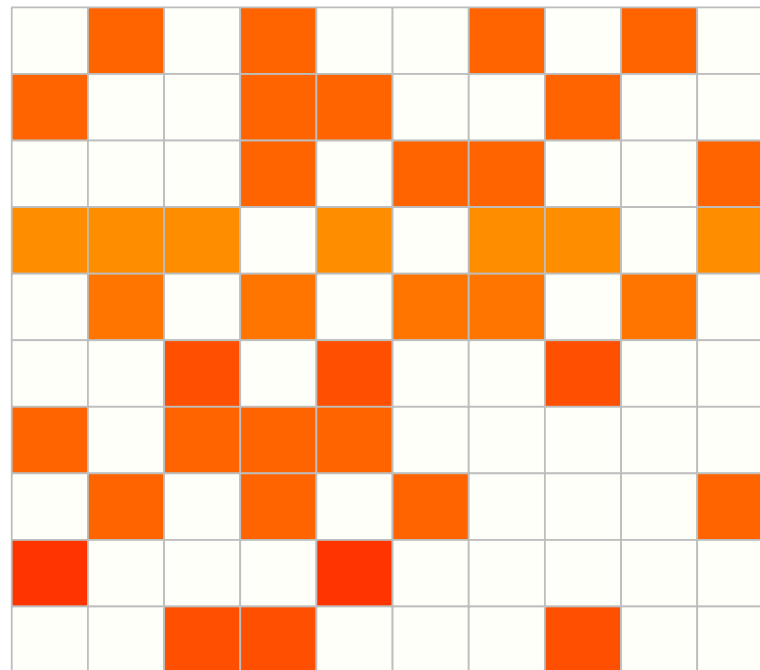


(j) Sociomatrix

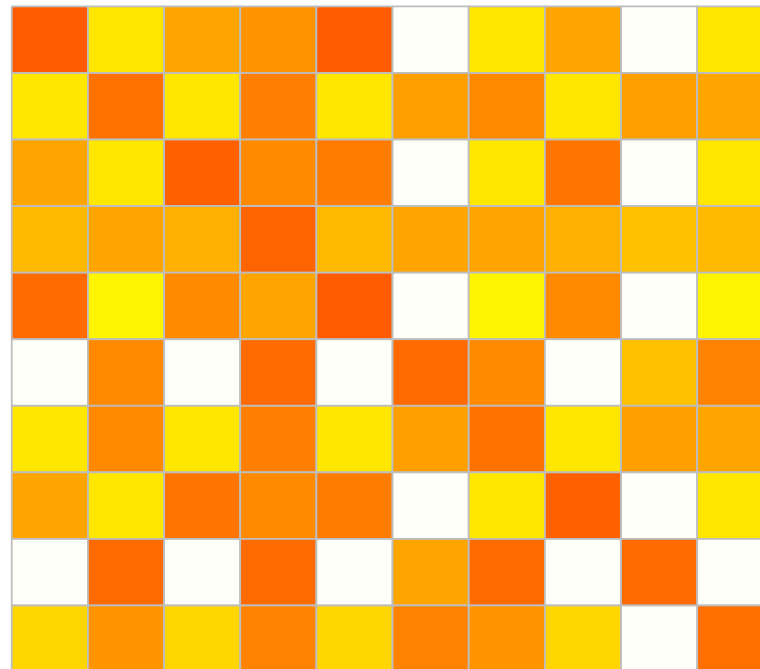


(k) First Wave Probabilities

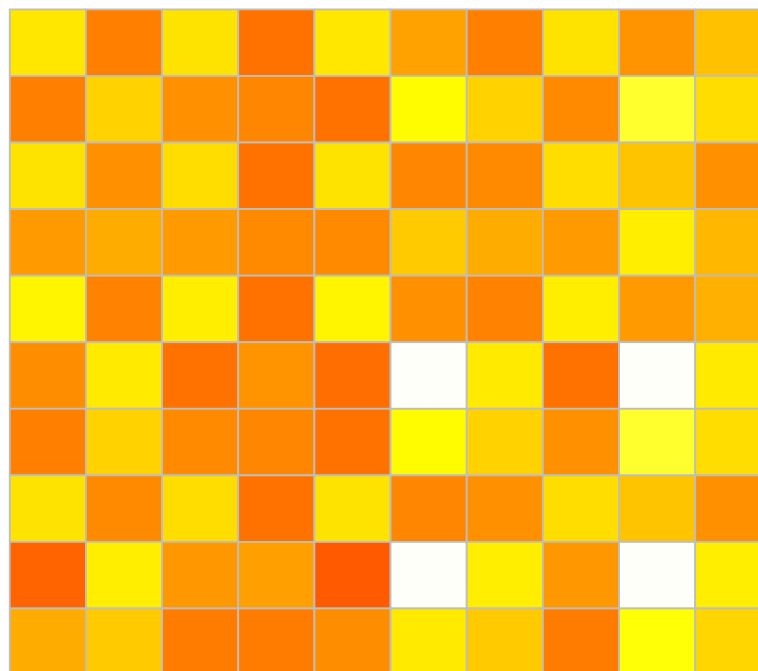
1



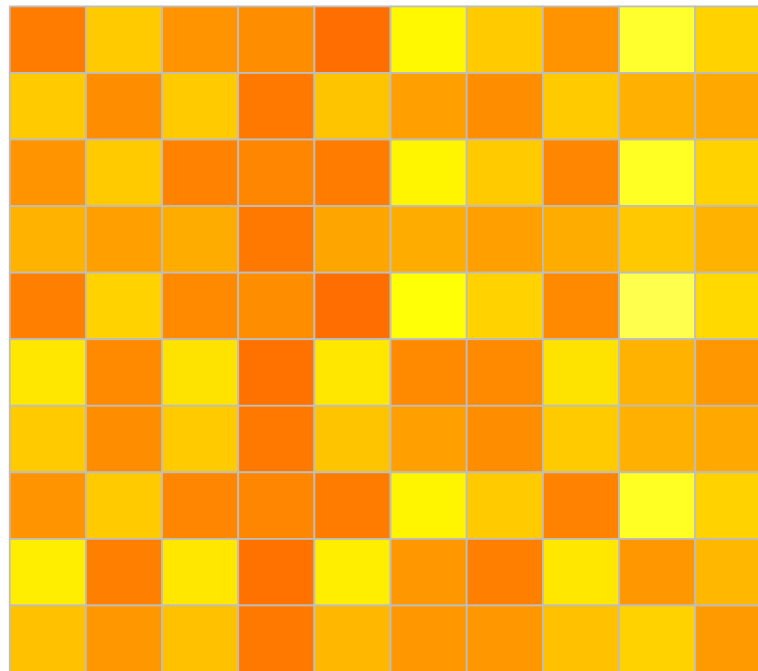
2



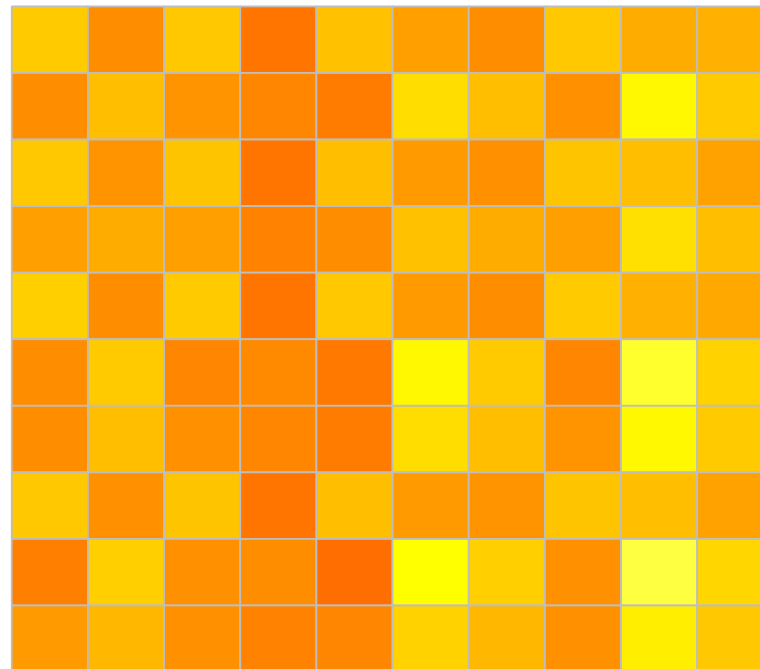
3



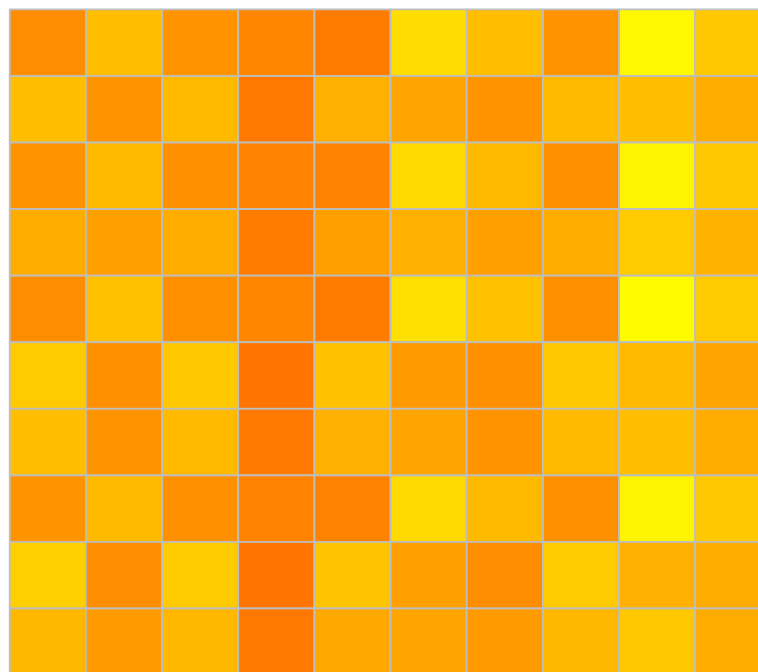
4



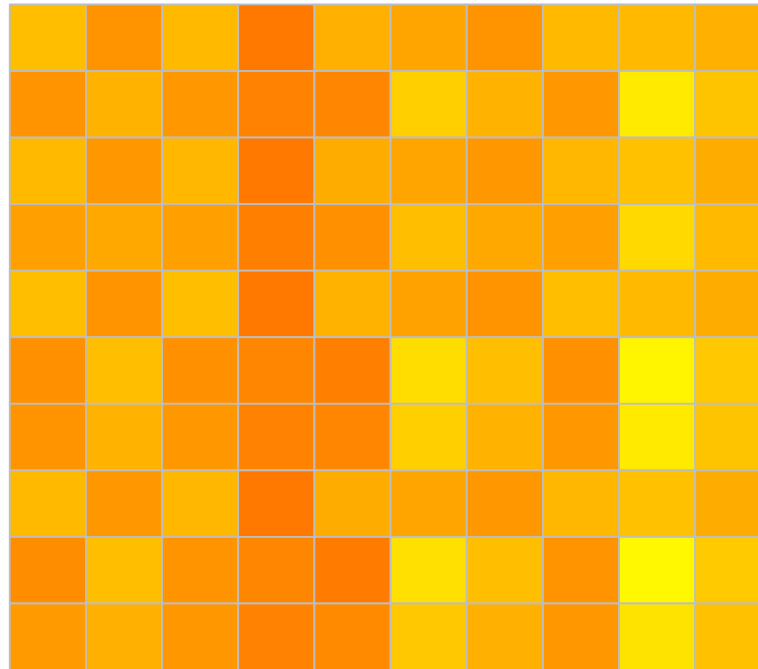
5



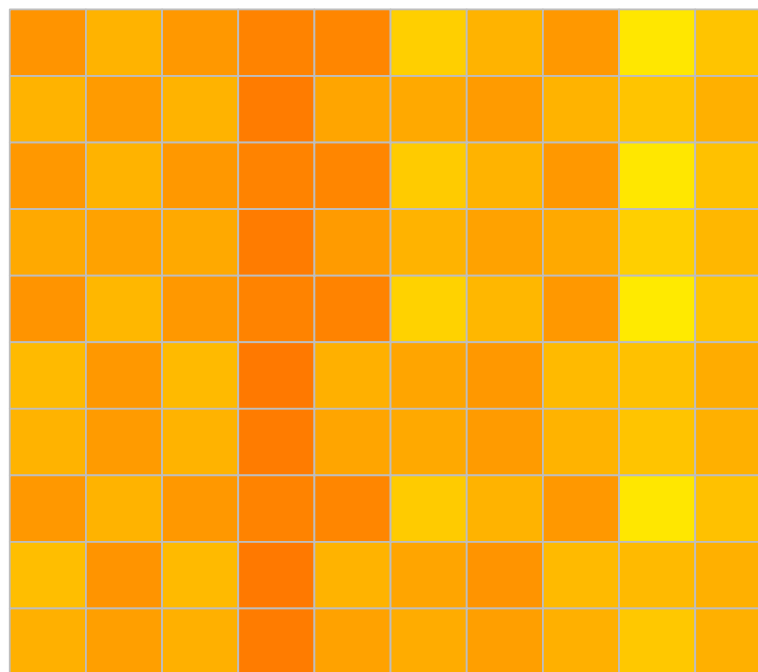
6



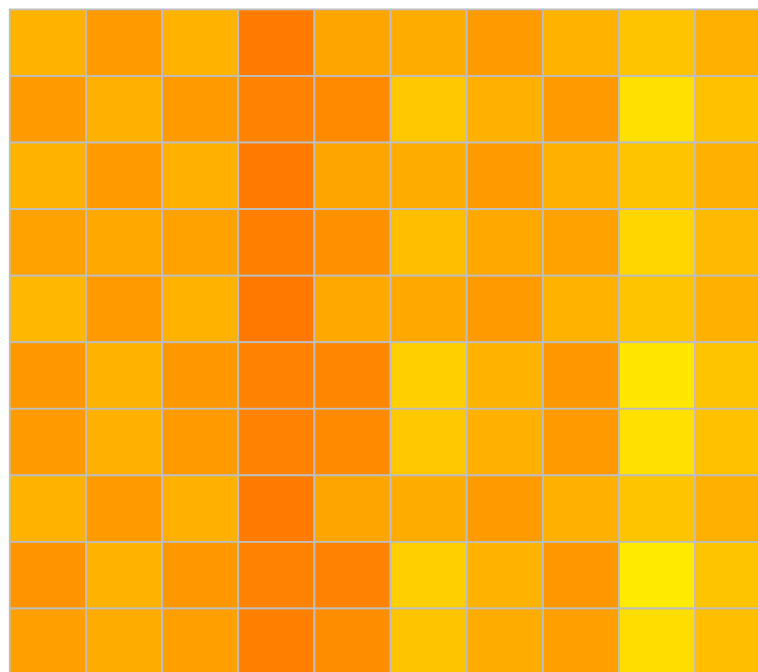
7



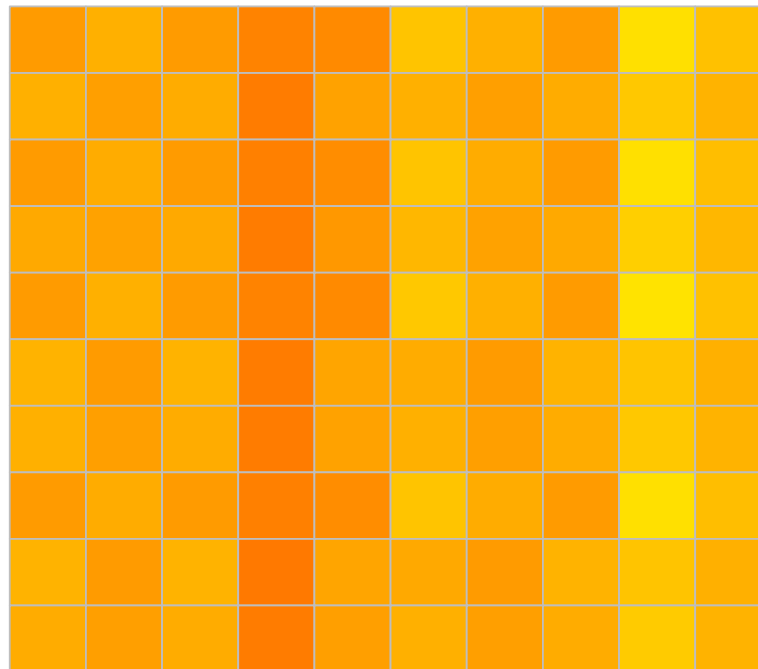
8



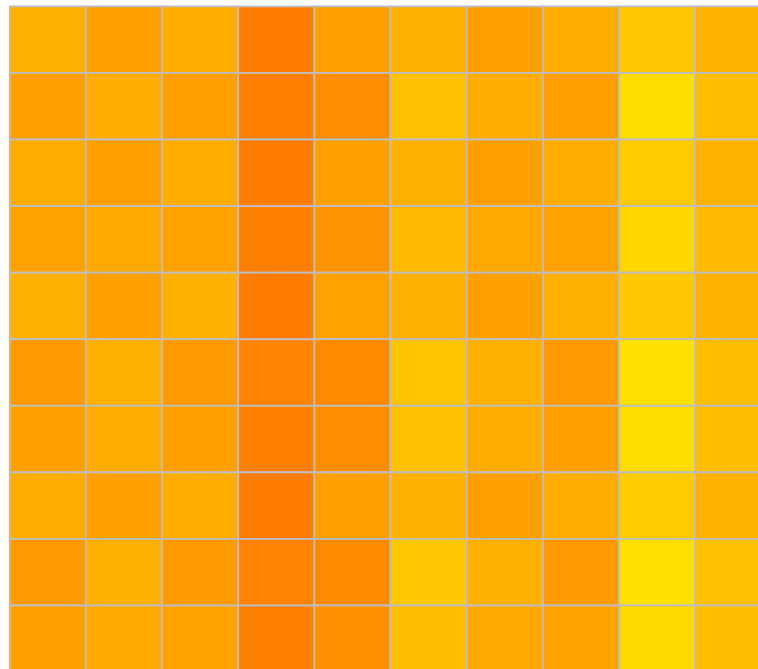
9



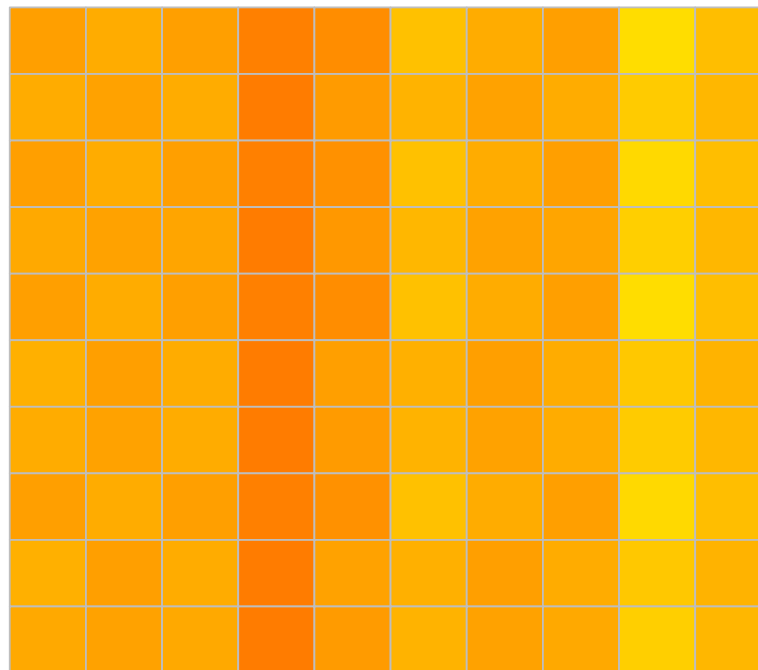
10



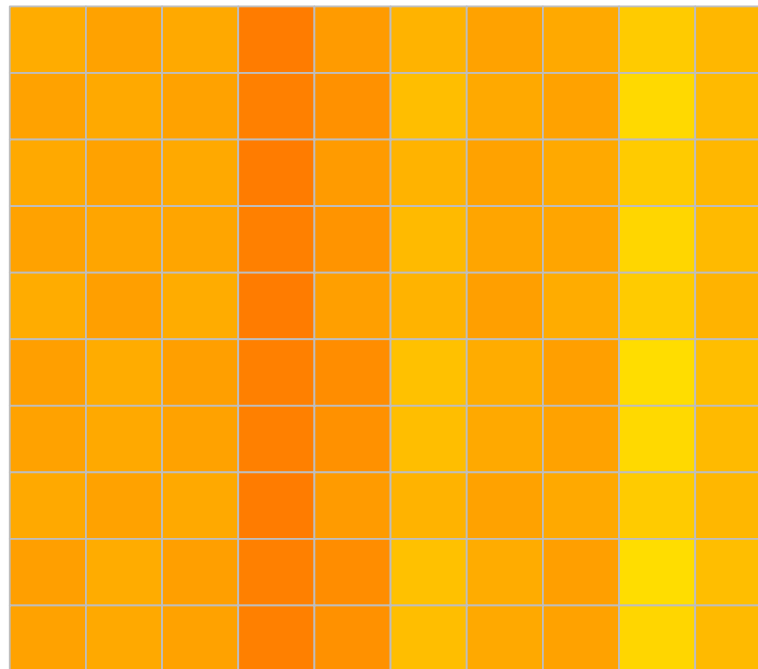
11



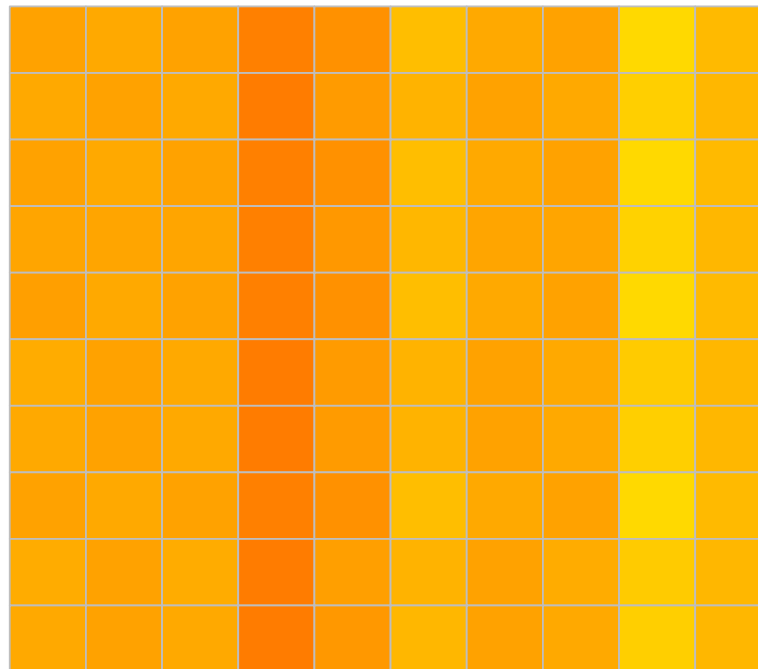
12



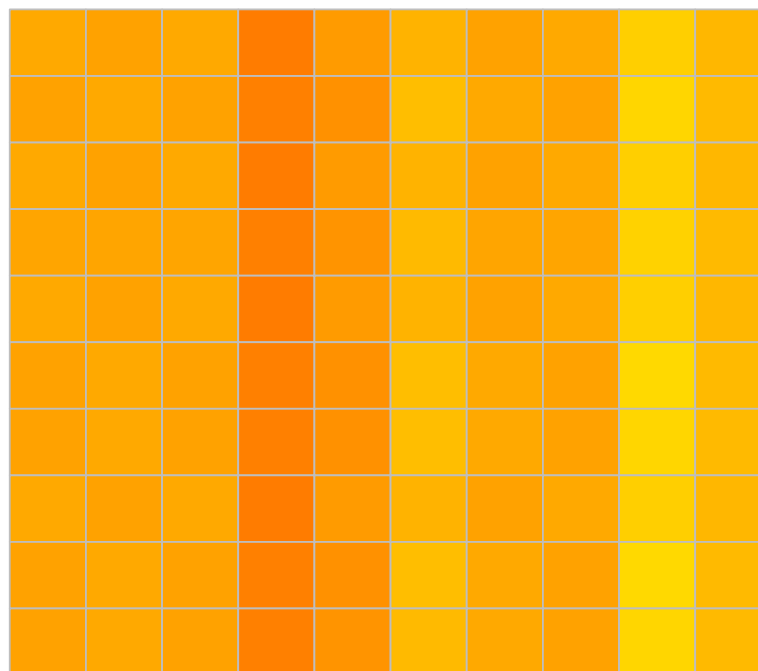
13

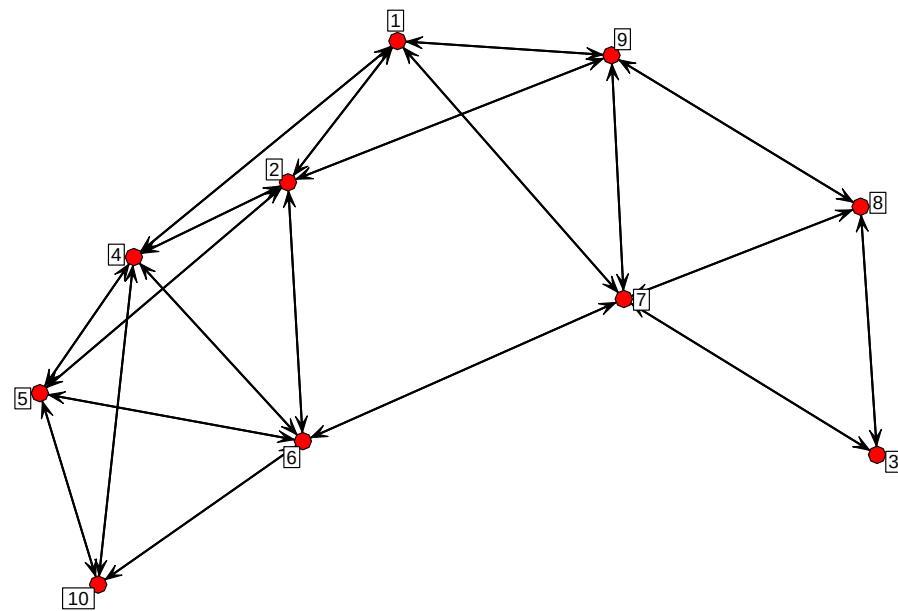


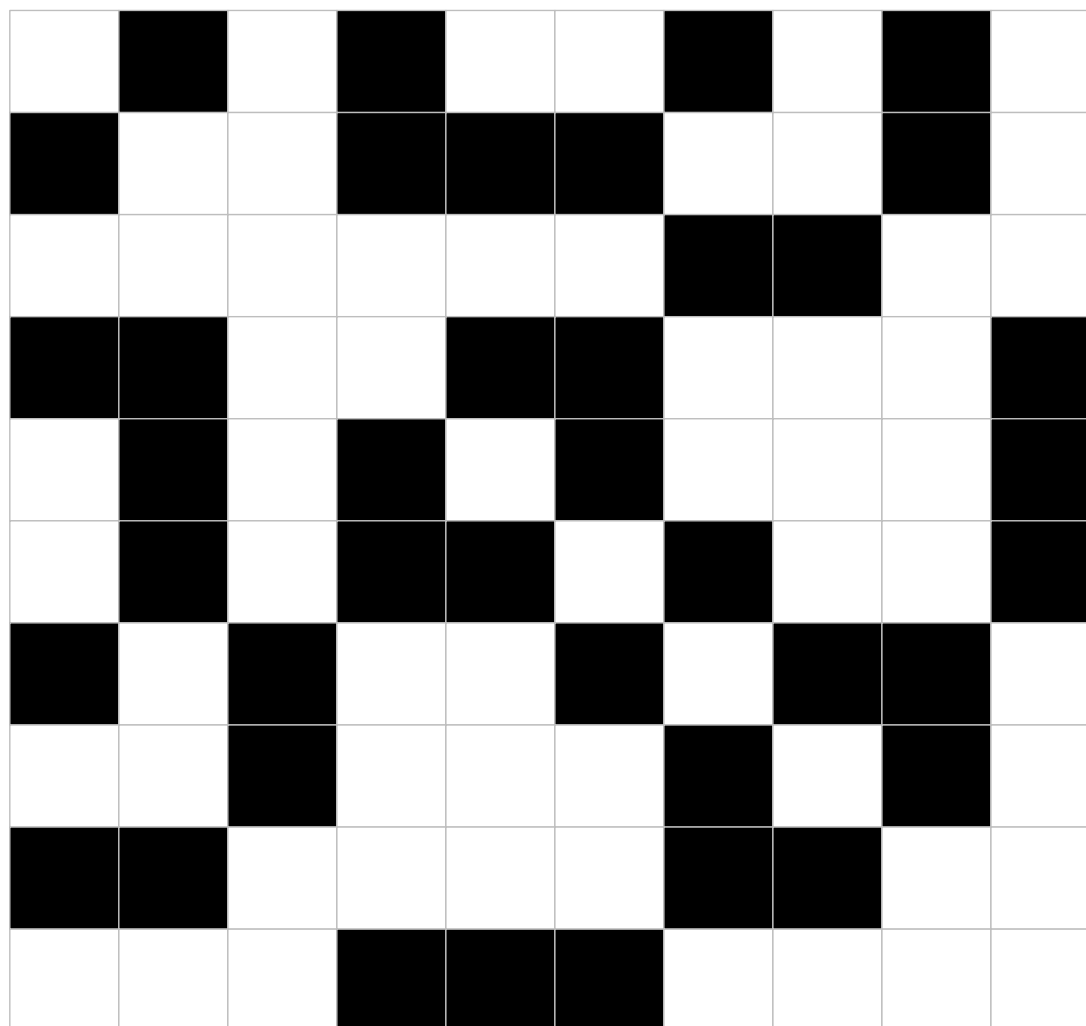
14

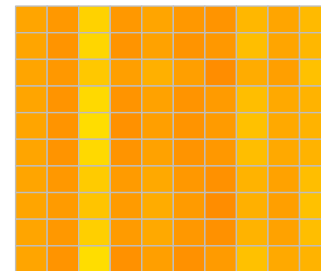
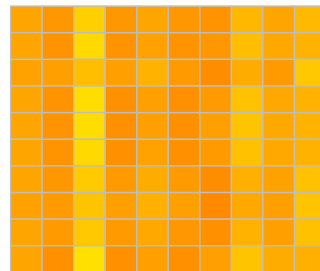
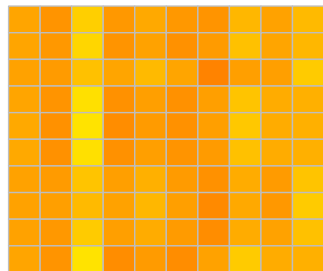
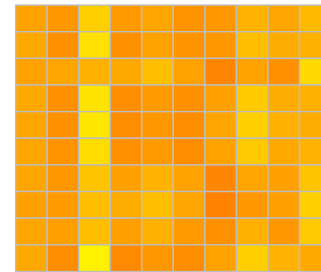
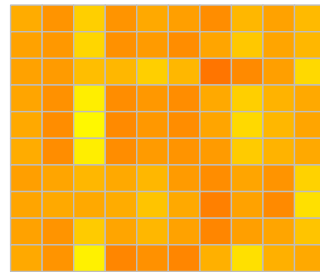
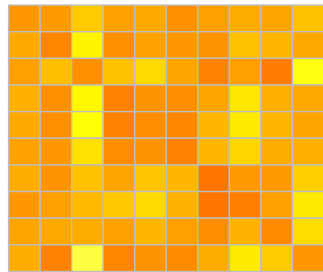
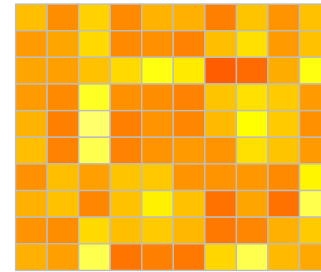
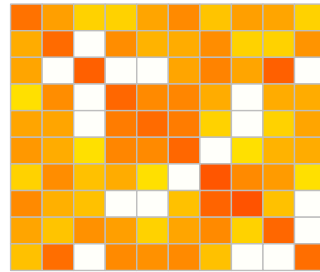
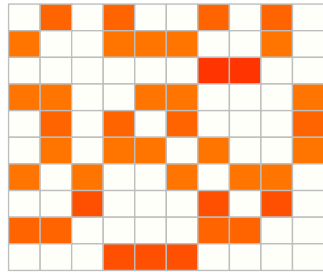


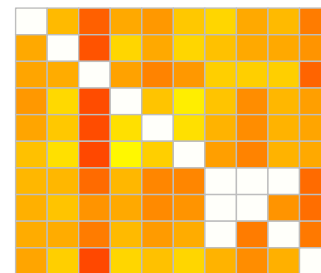
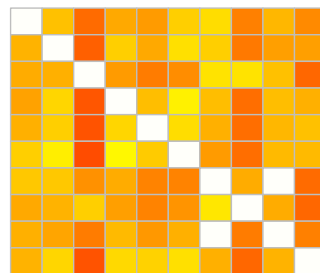
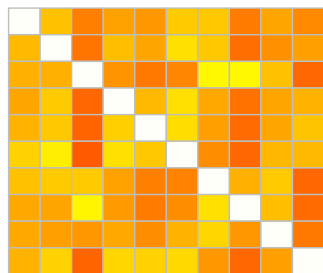
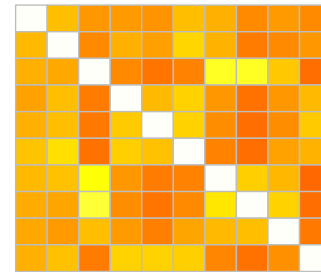
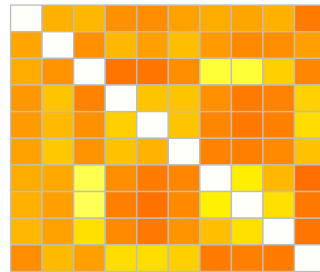
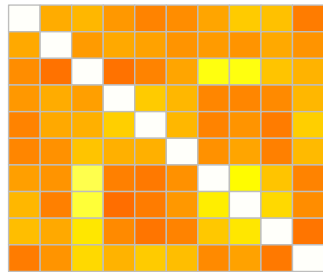
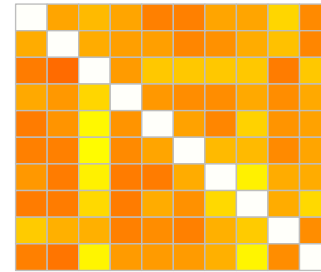
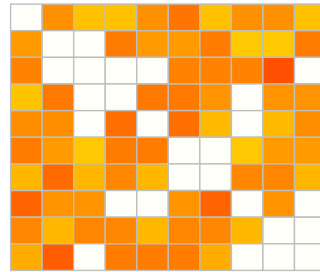
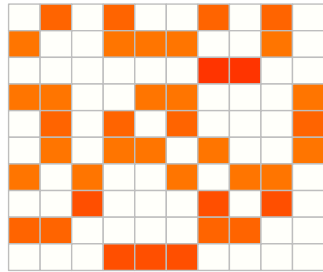
15

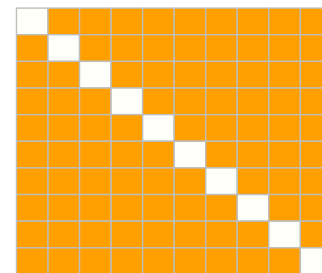
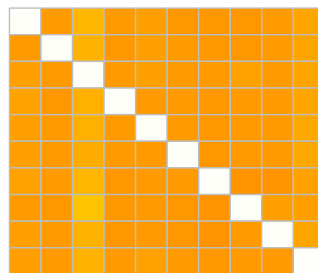
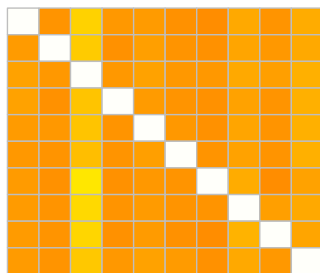
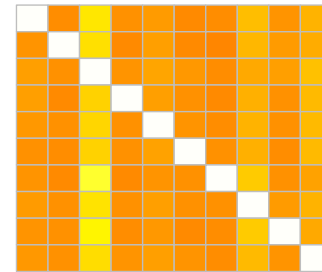
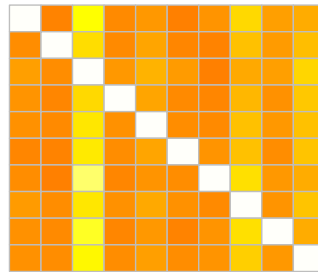
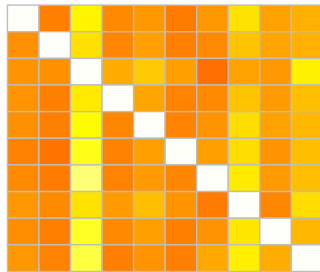
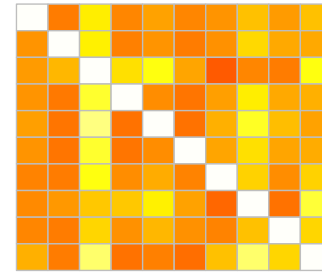
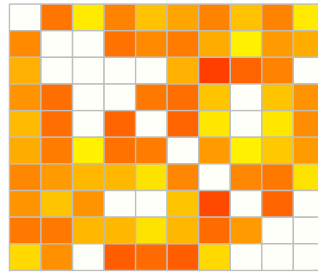
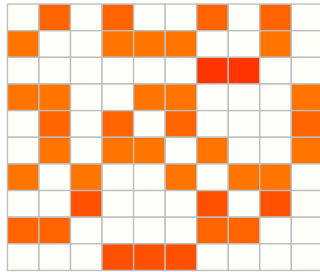


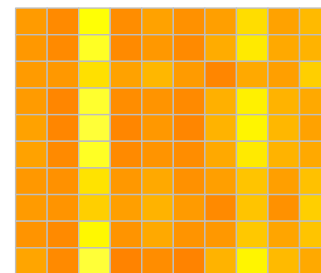
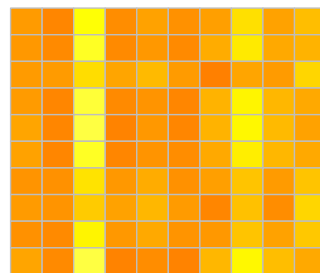
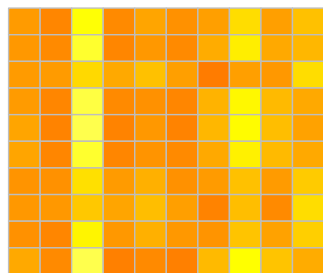
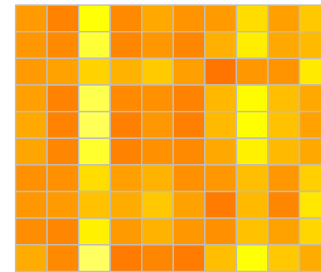
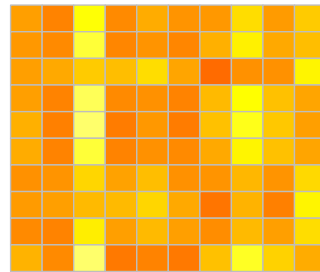
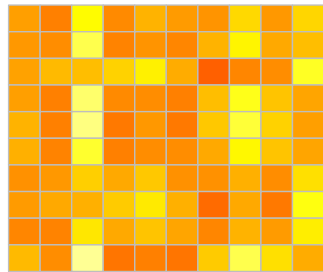
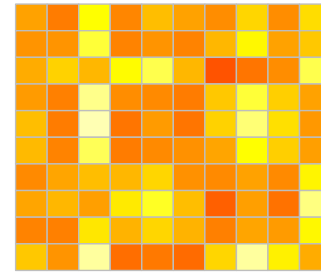
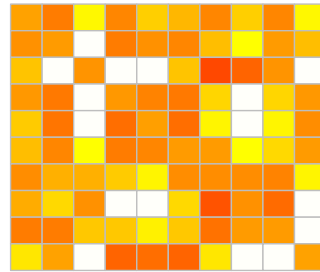
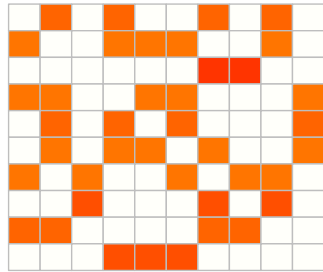


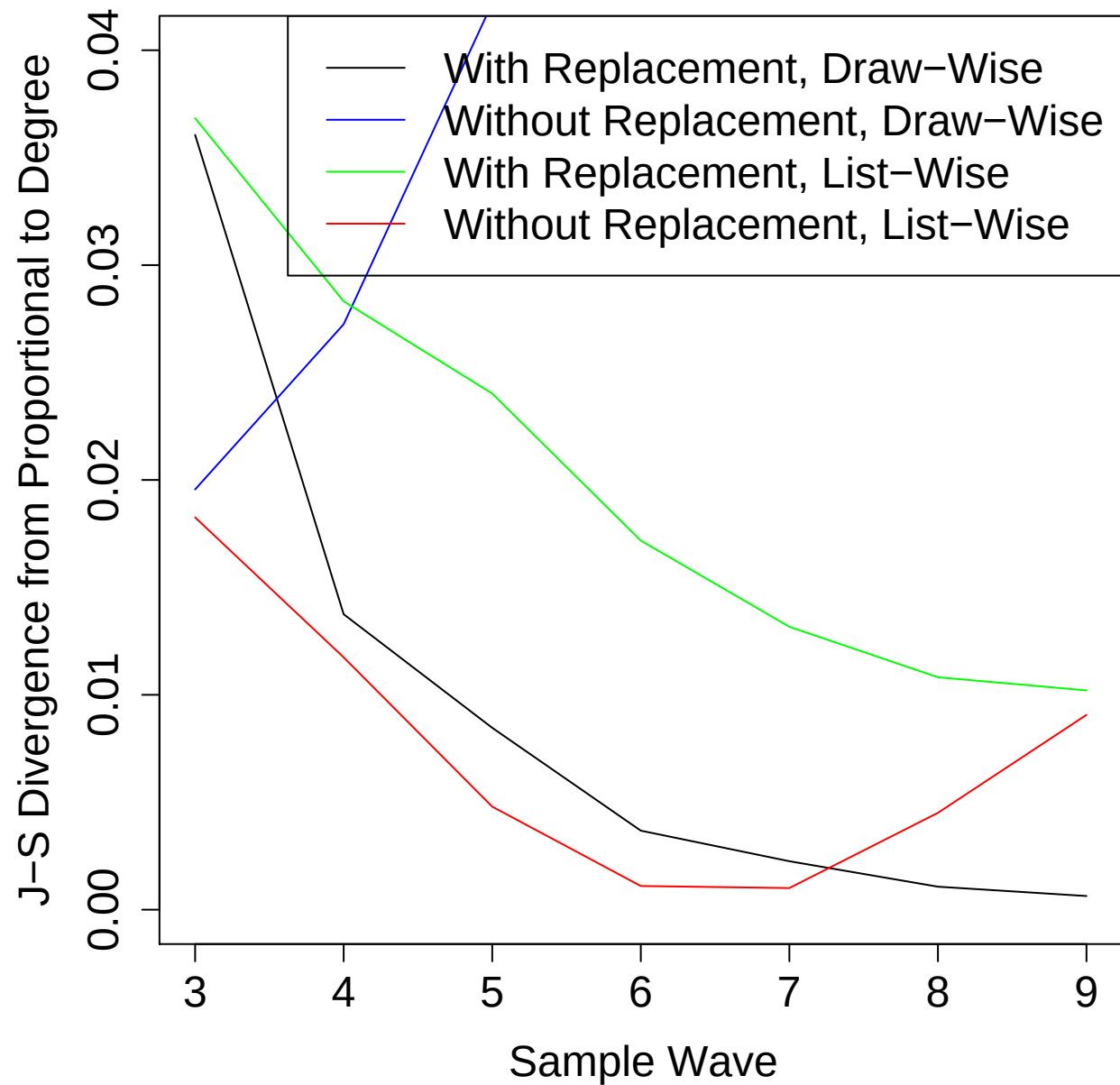








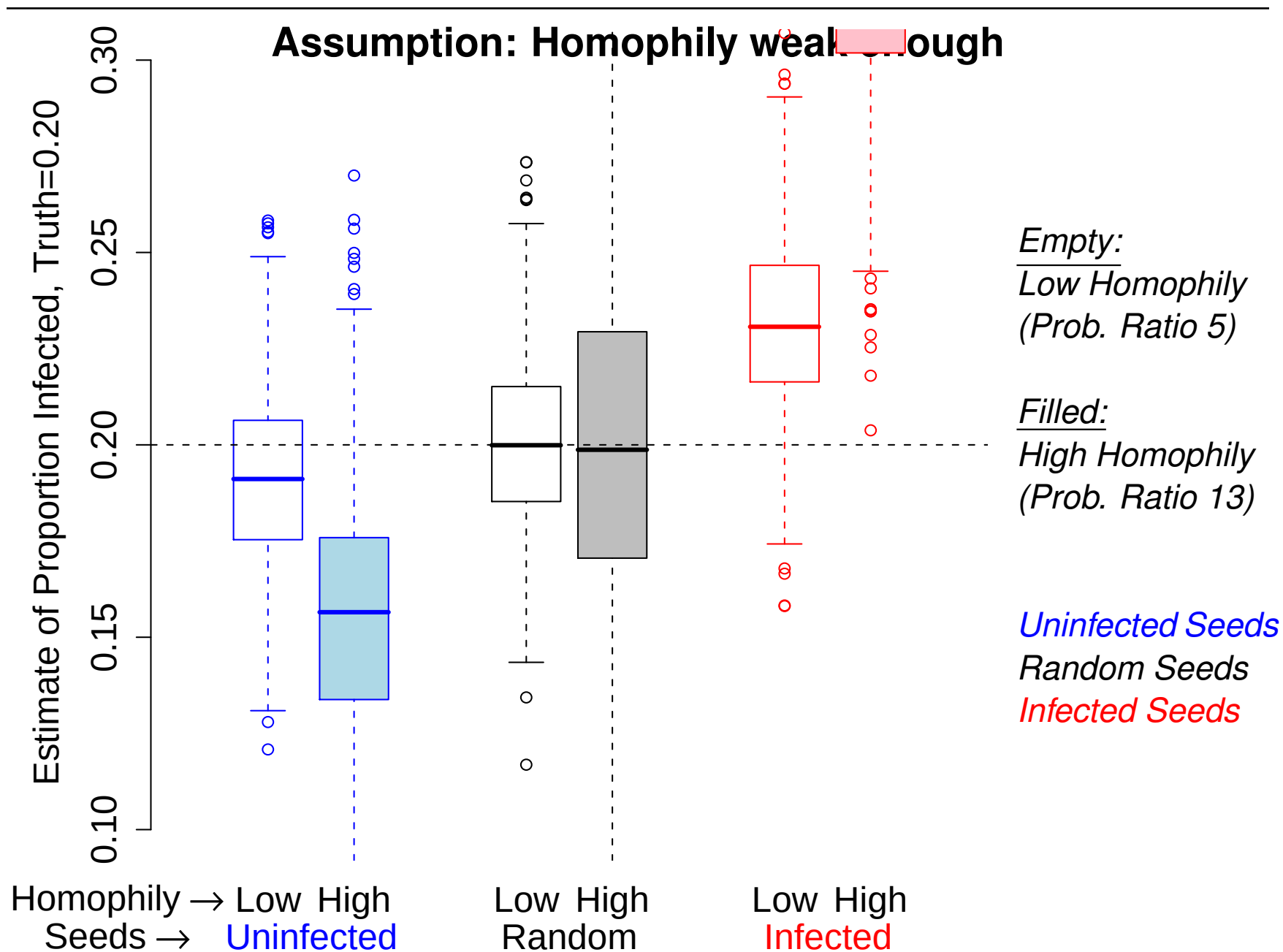


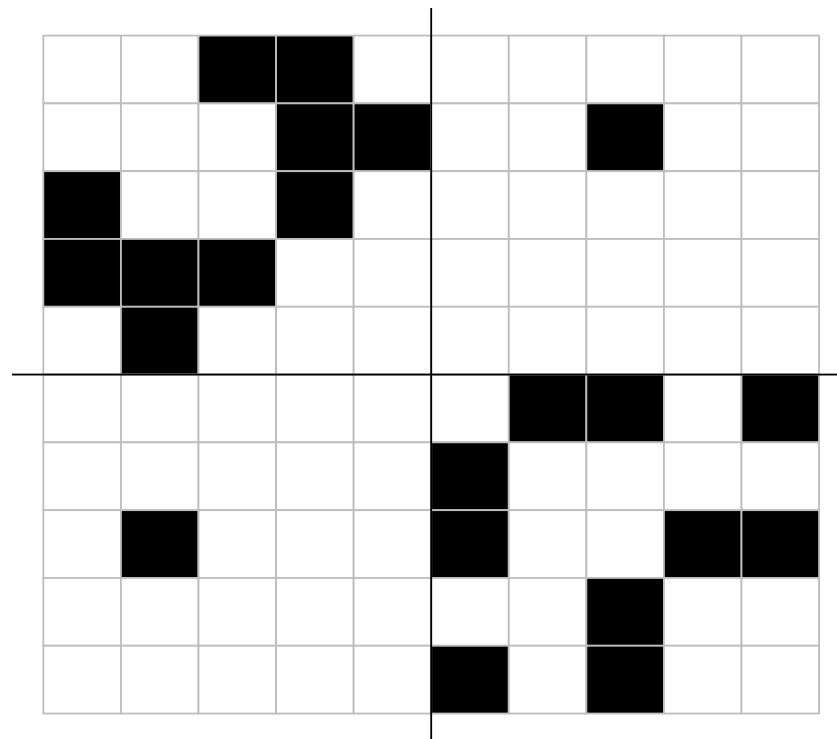


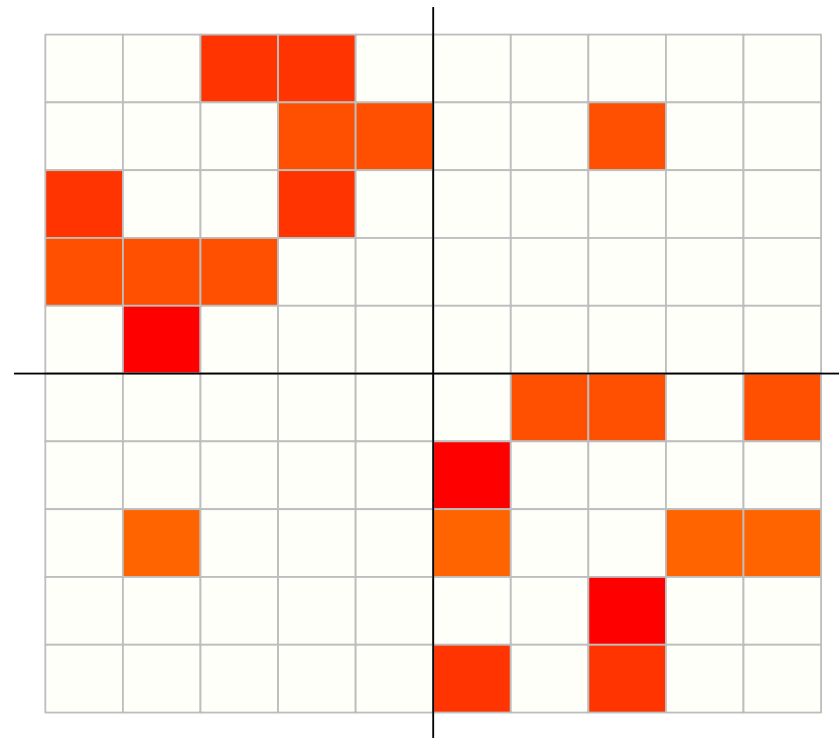
End of Two Wrongs Development

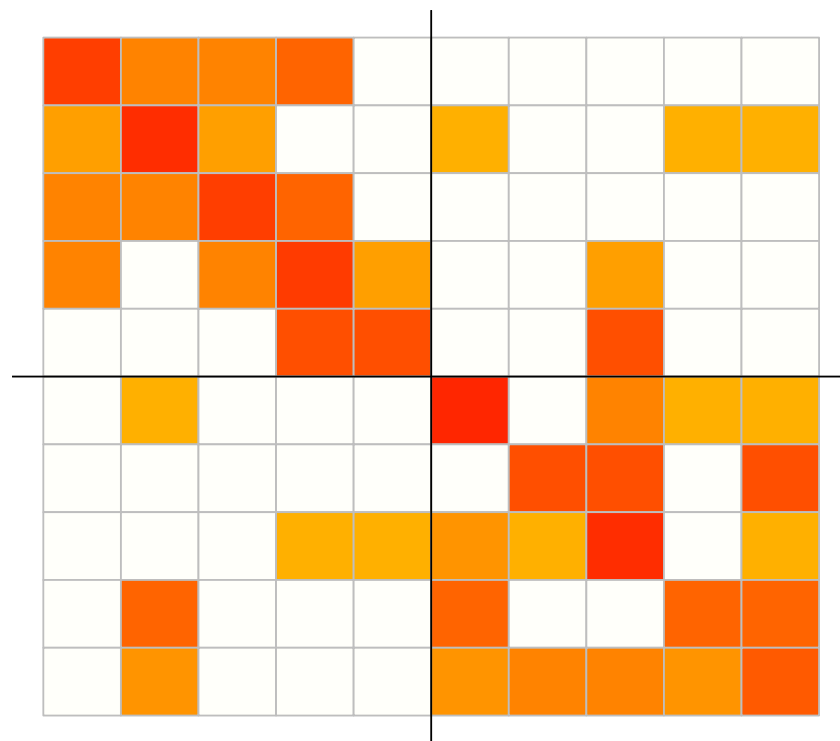
Placement of This Work

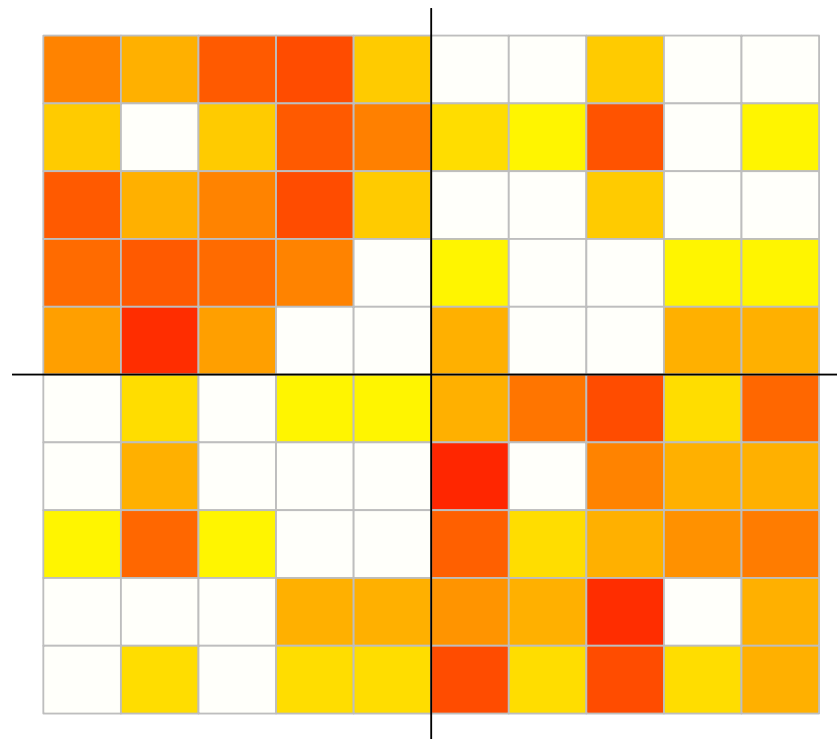
	<i>Describe Structure (finite population)</i>	<i>Describe Mechanism (super-population)</i>
<i>Fully Observed Data</i>	(1) Nodes: Individual summary statistics (not network analysis) (2) Relations: Centrality, Transitivity, etc. (3) Both: Homophily etc.	(1) Nodes: Spatial statistics, dependent data (not network analysis) (2) Relations: ERGMs, Latent Variable Models, inner product models, etc. (3) Both: Social Selection, Social Influence, Actor-oriented, Dynamic Disease models
<i>Partially Observed Data</i>	(1) Nodes: Network Sampling Respondent-Driven Sampling (RDS) <i>Chapter 2, Chapter 5, Chapter 6</i> (2) Relations: Network Sampling <i>Chapter 2</i> (3) Both: Model-Assisted Estimator for RDS <i>Chapter 6</i>	(1) Nodes: (currently, not in network frame) (2) Relations: Adaptive Network Sampling <i>Chapter 3</i> (3) Both: Disease Modeling, Contact Tracing <i>Chapter 4</i>

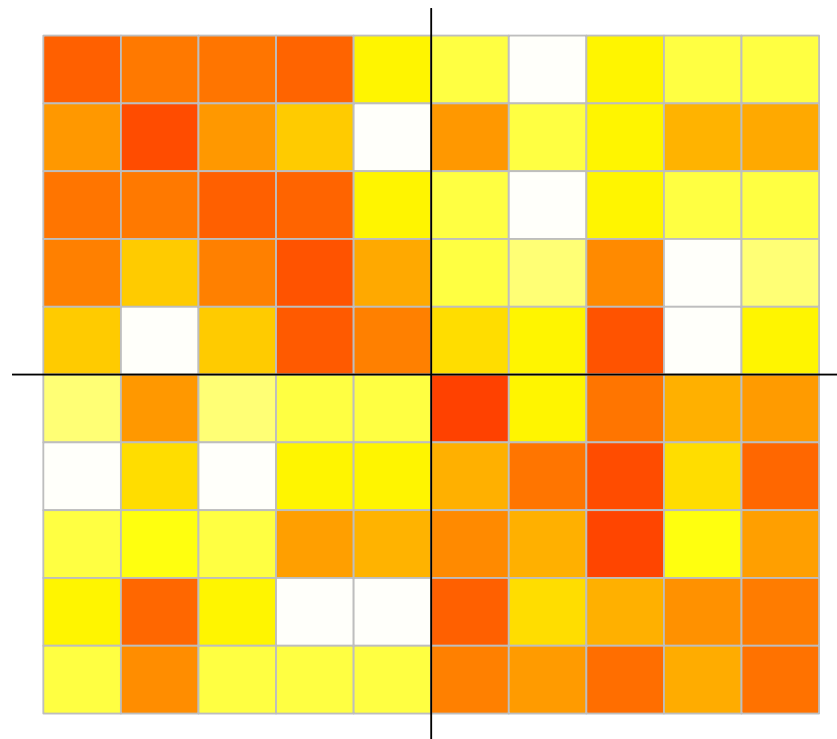


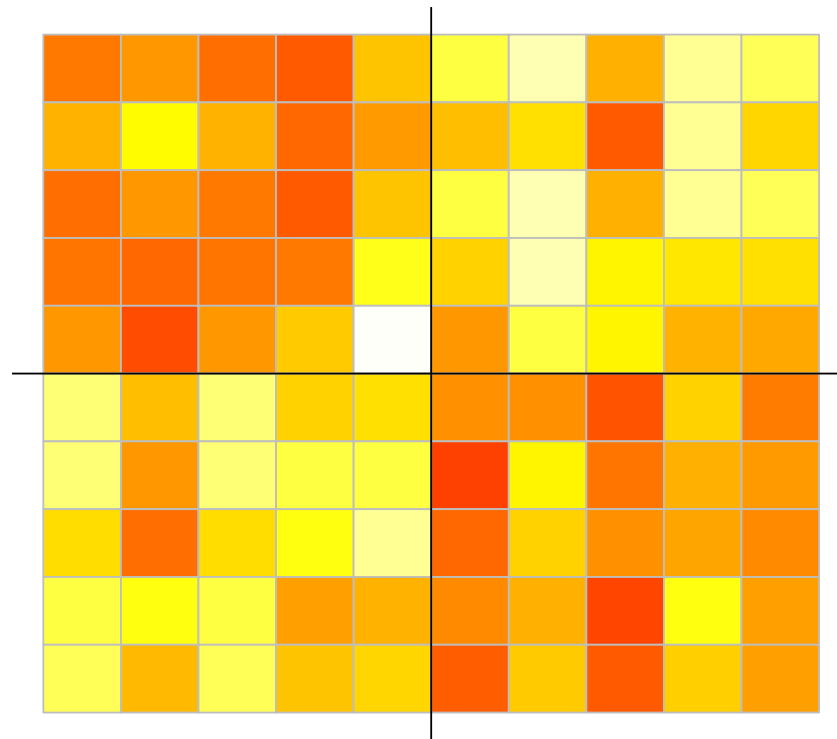


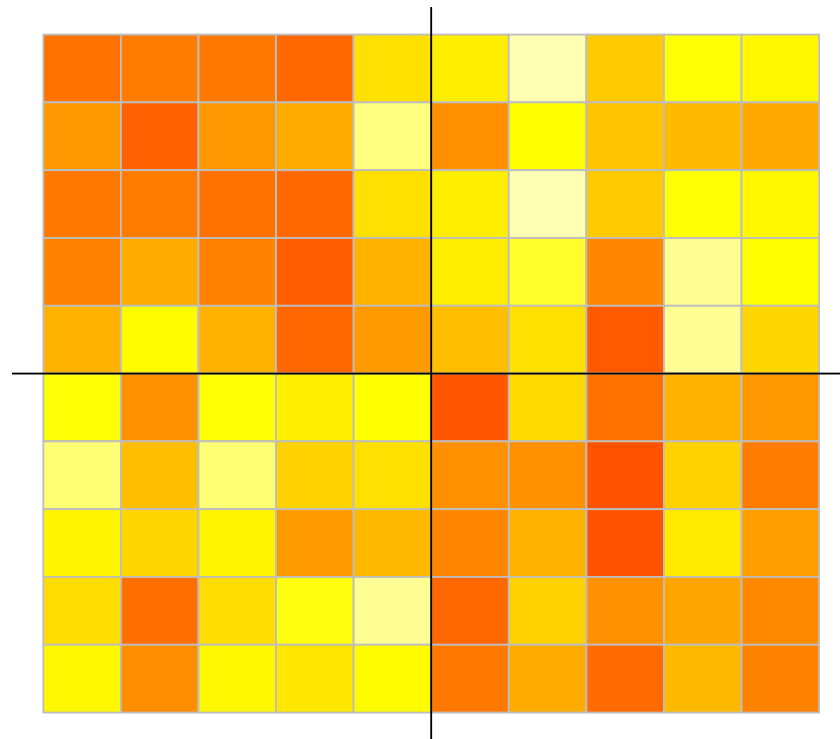


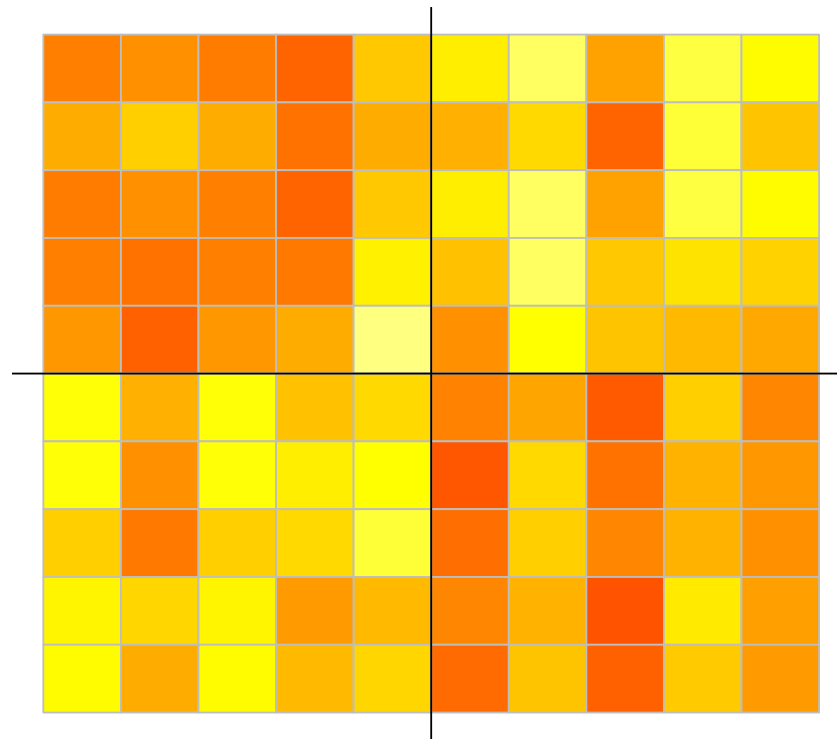


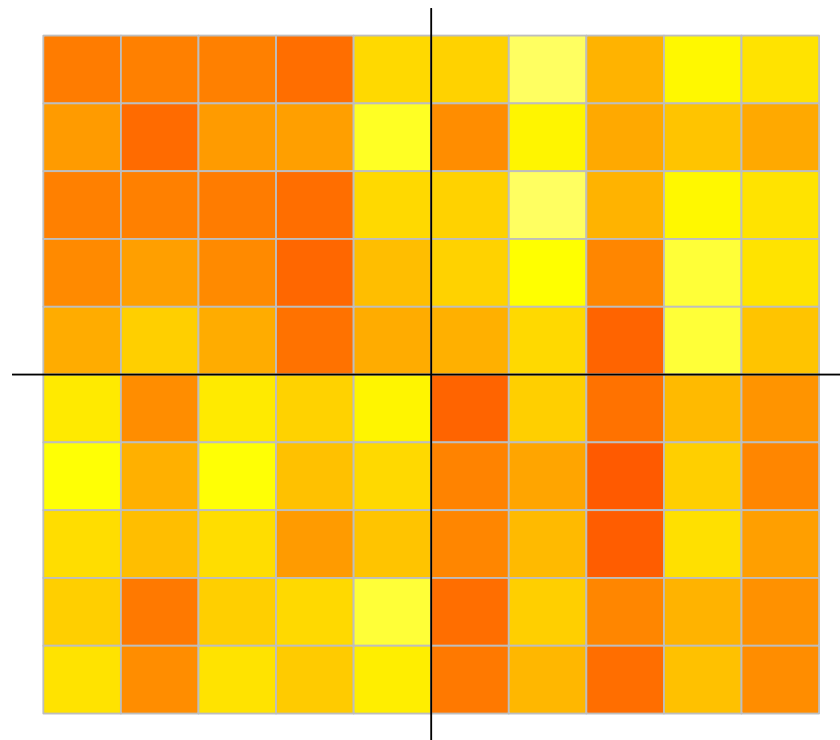


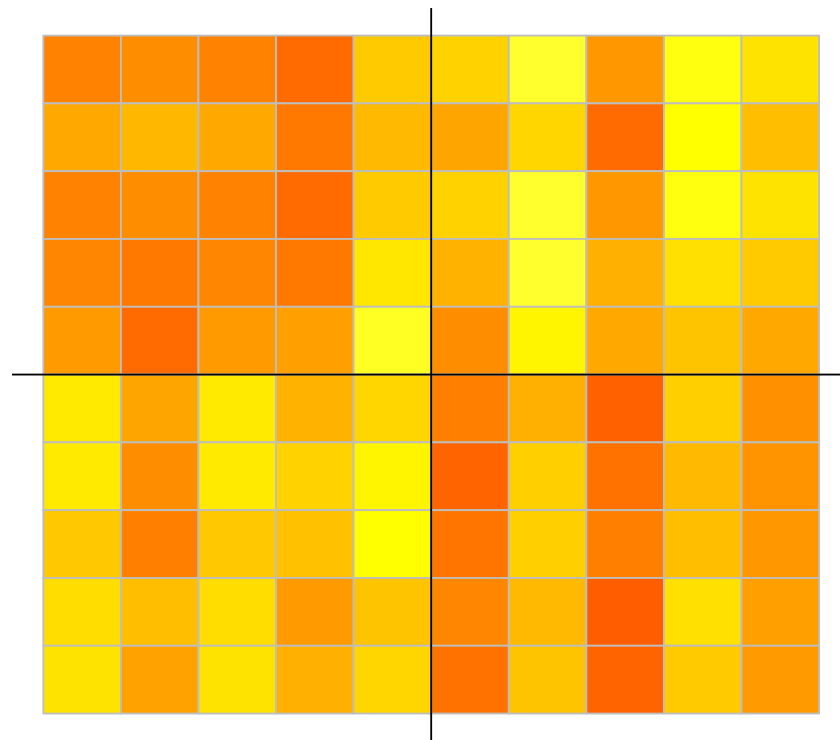


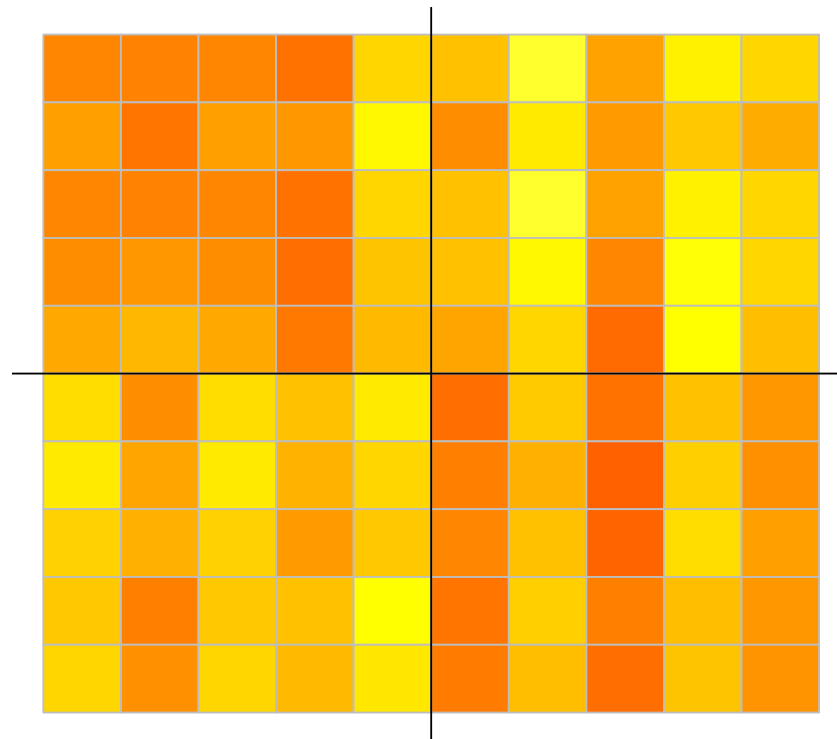


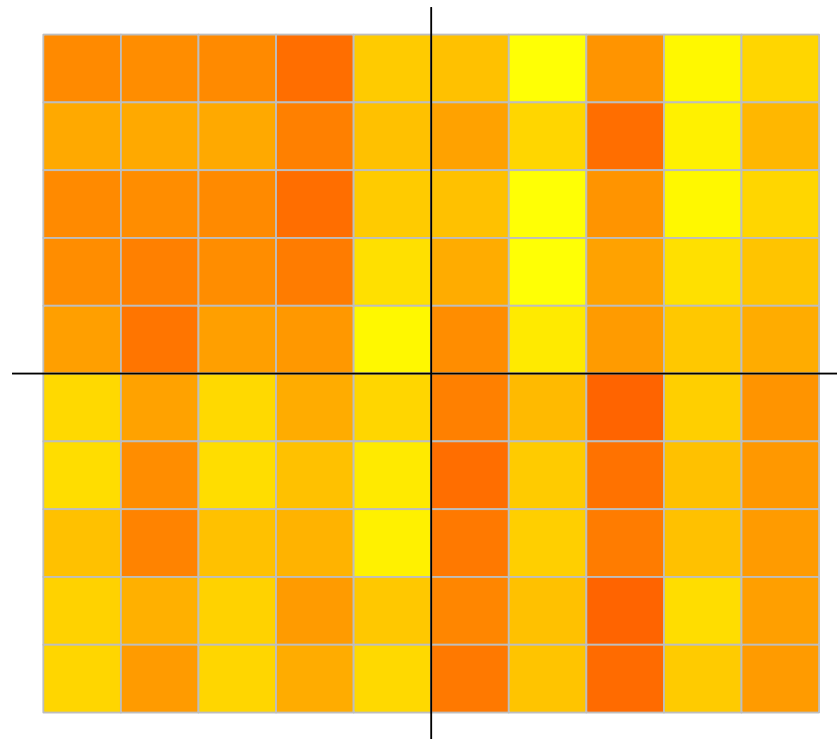


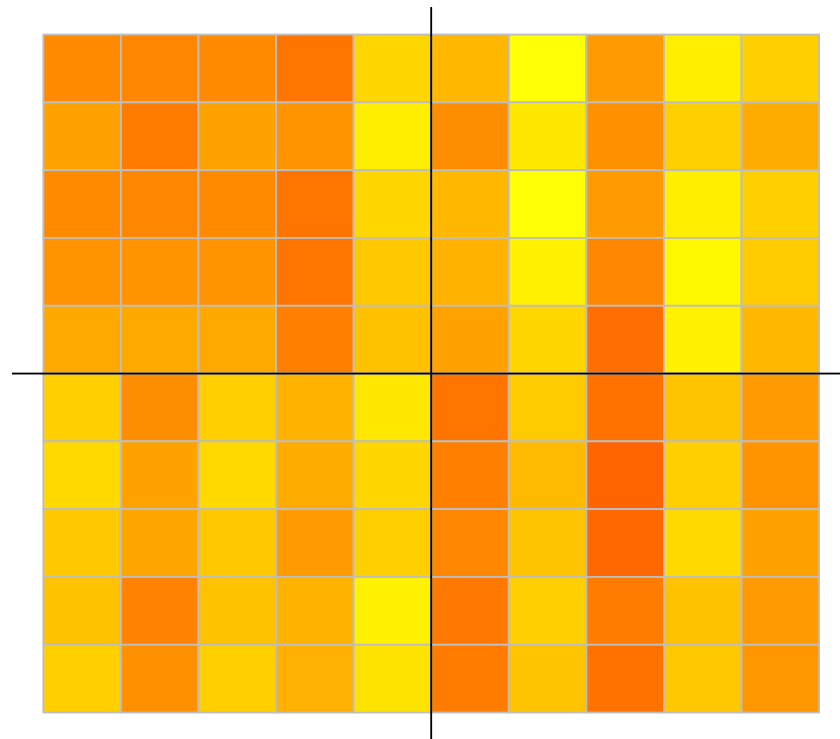


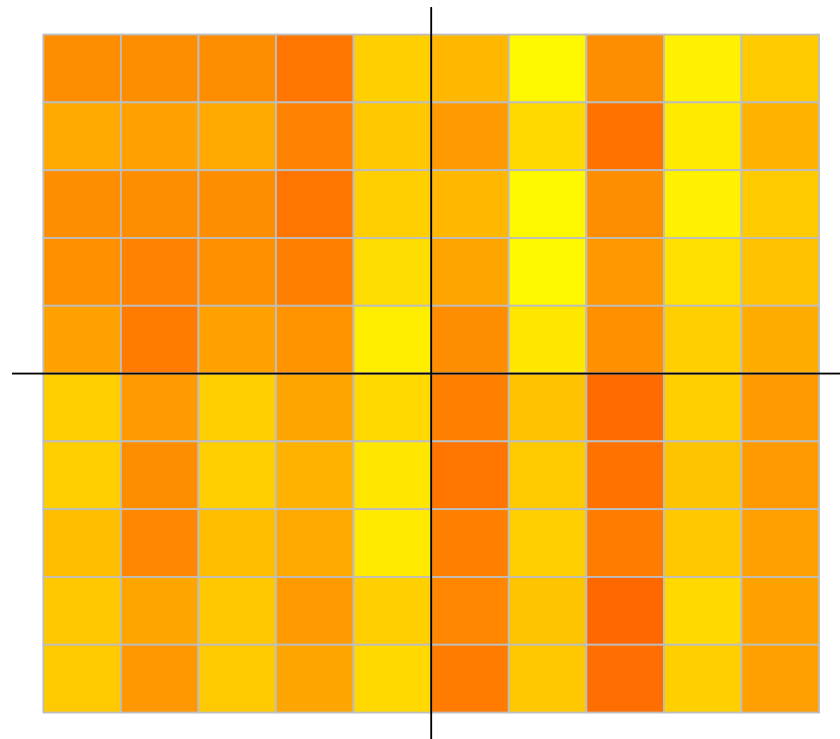


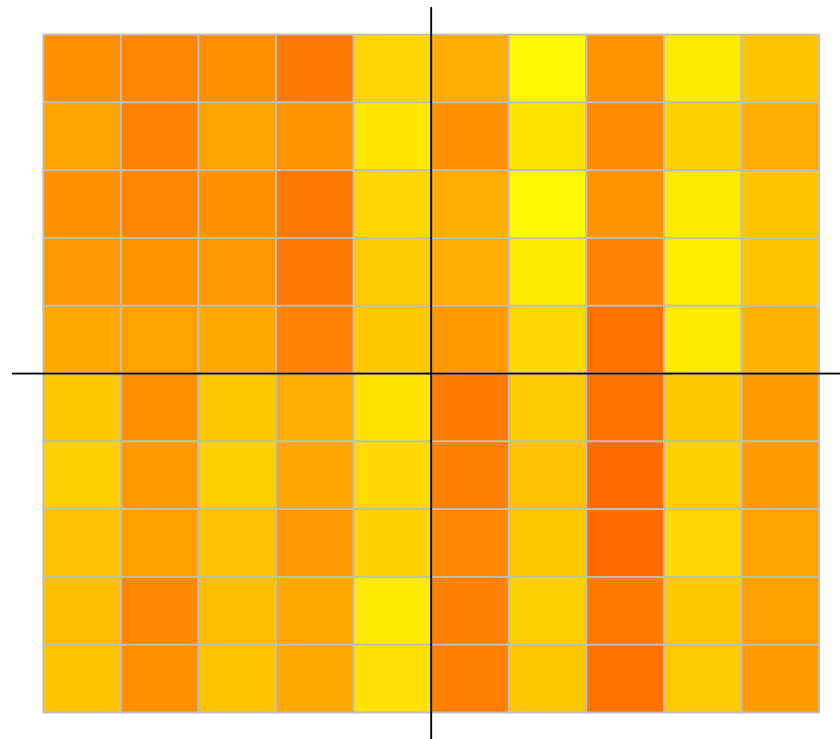


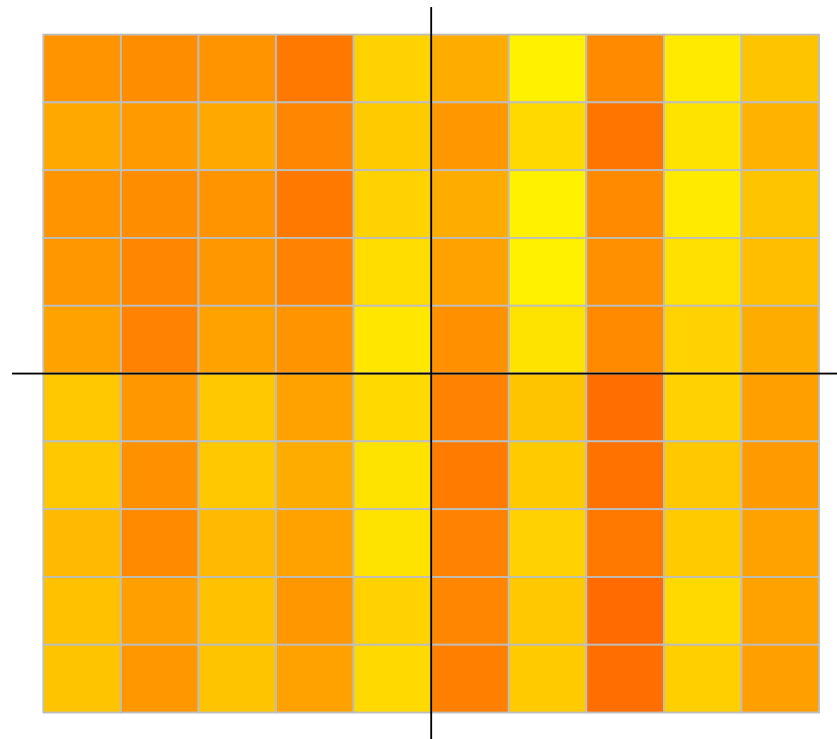


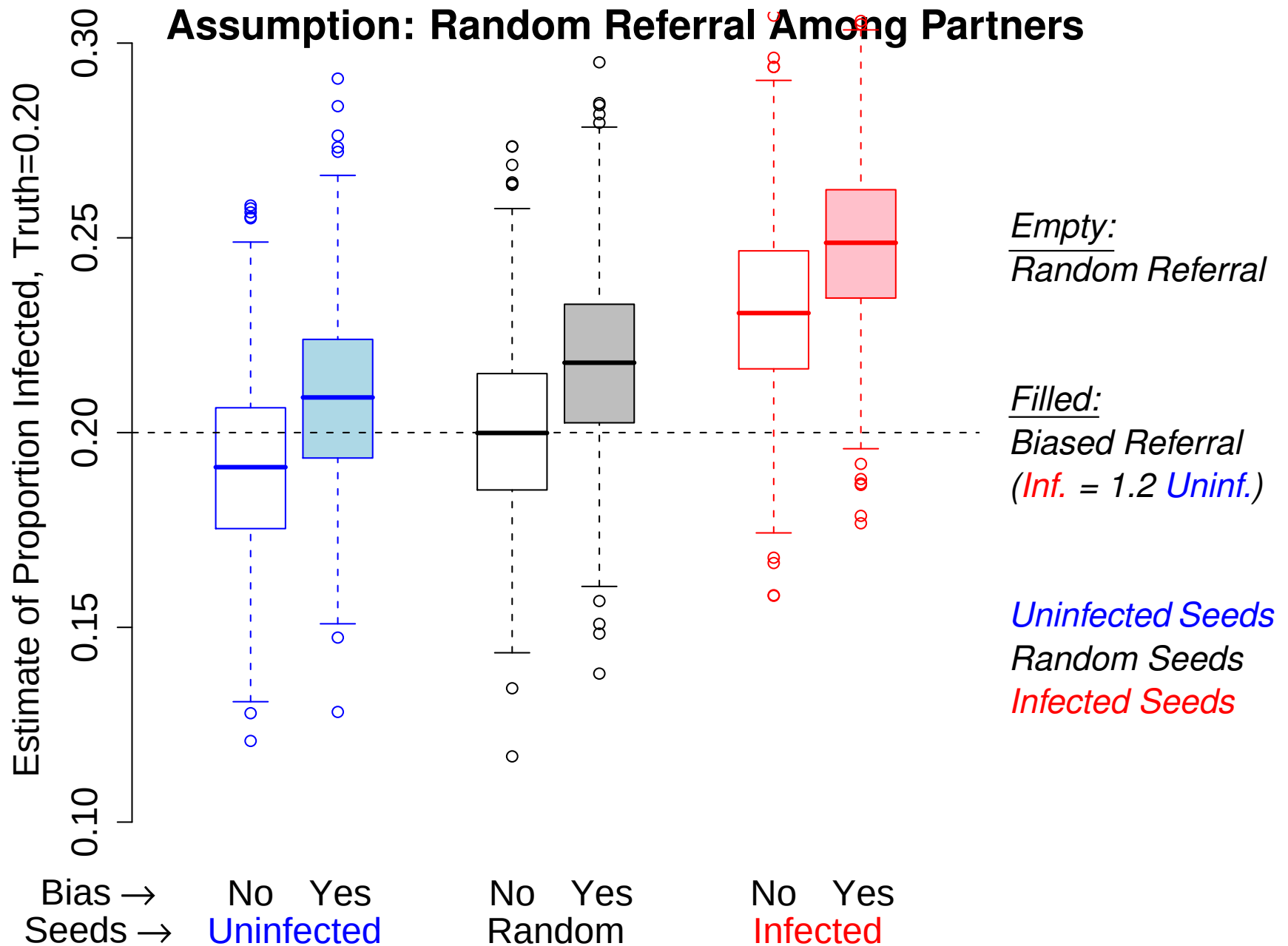


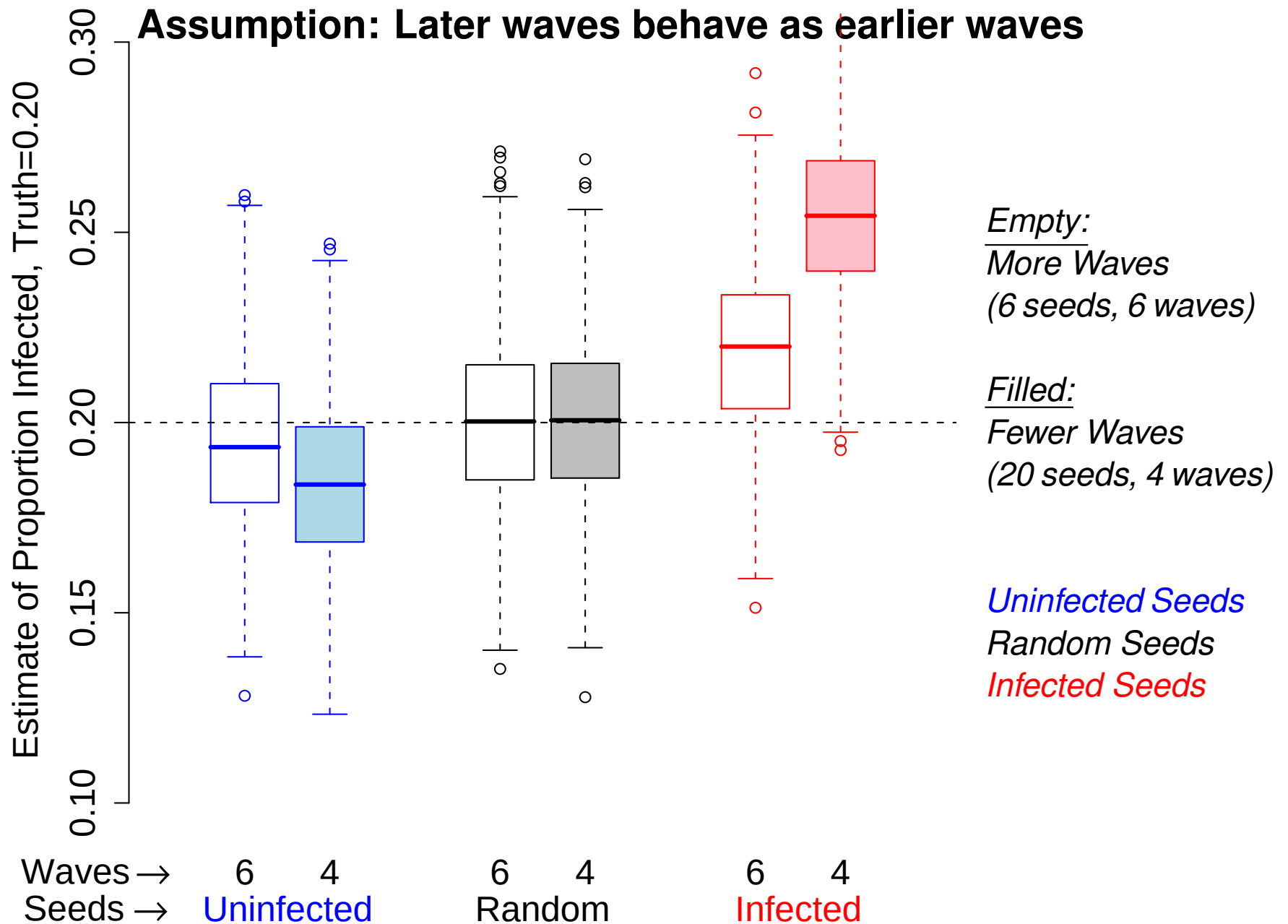




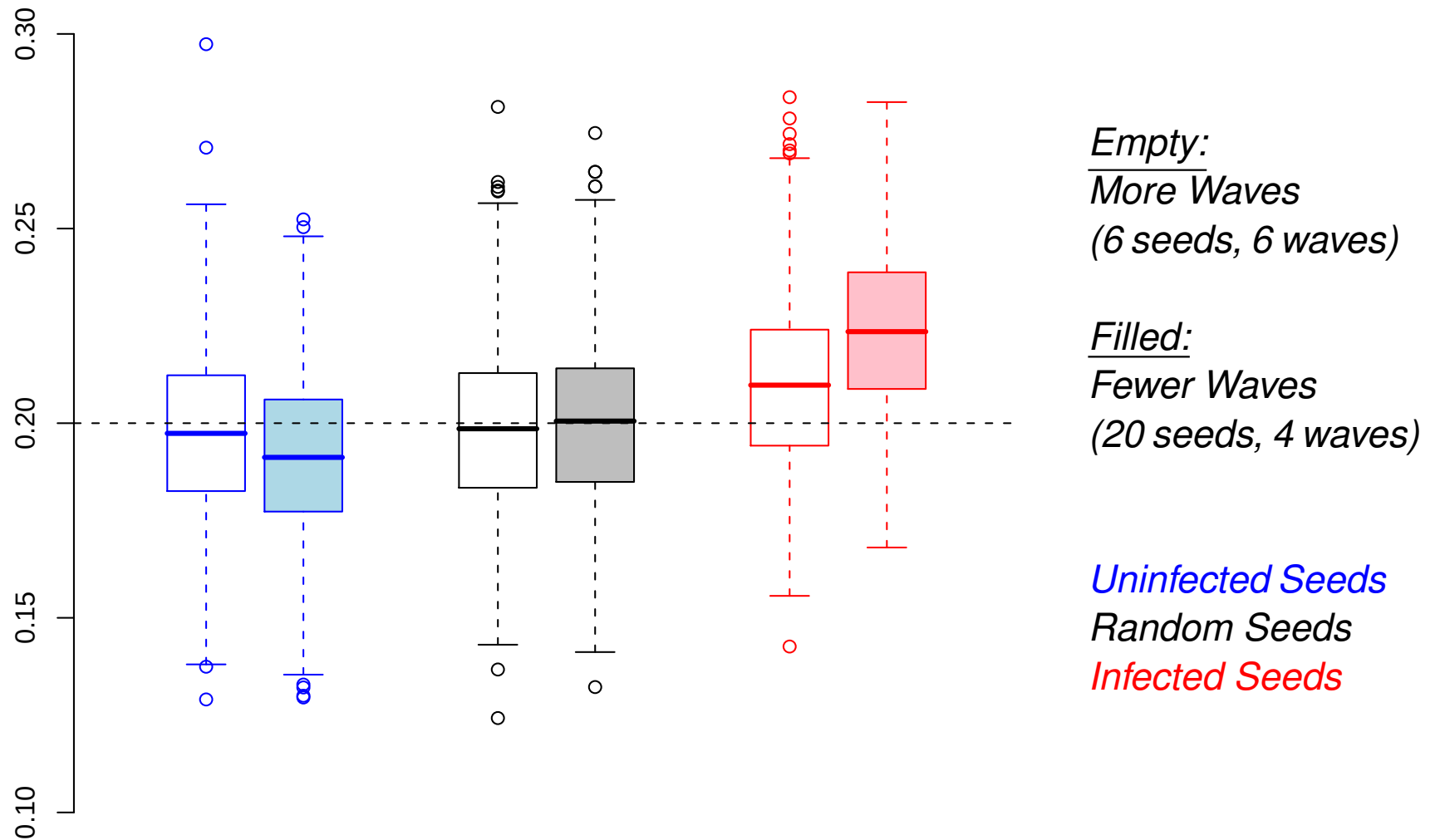




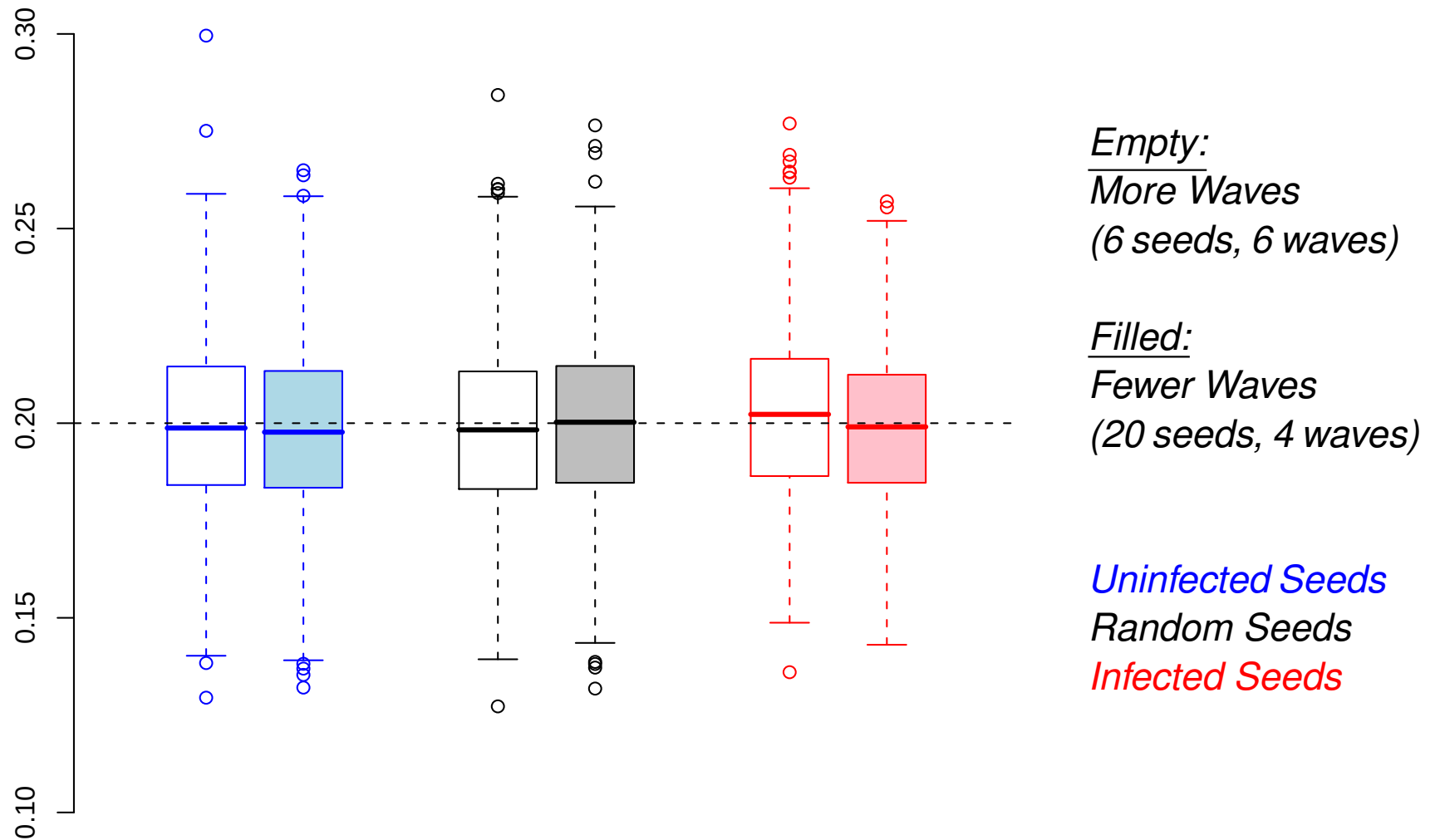




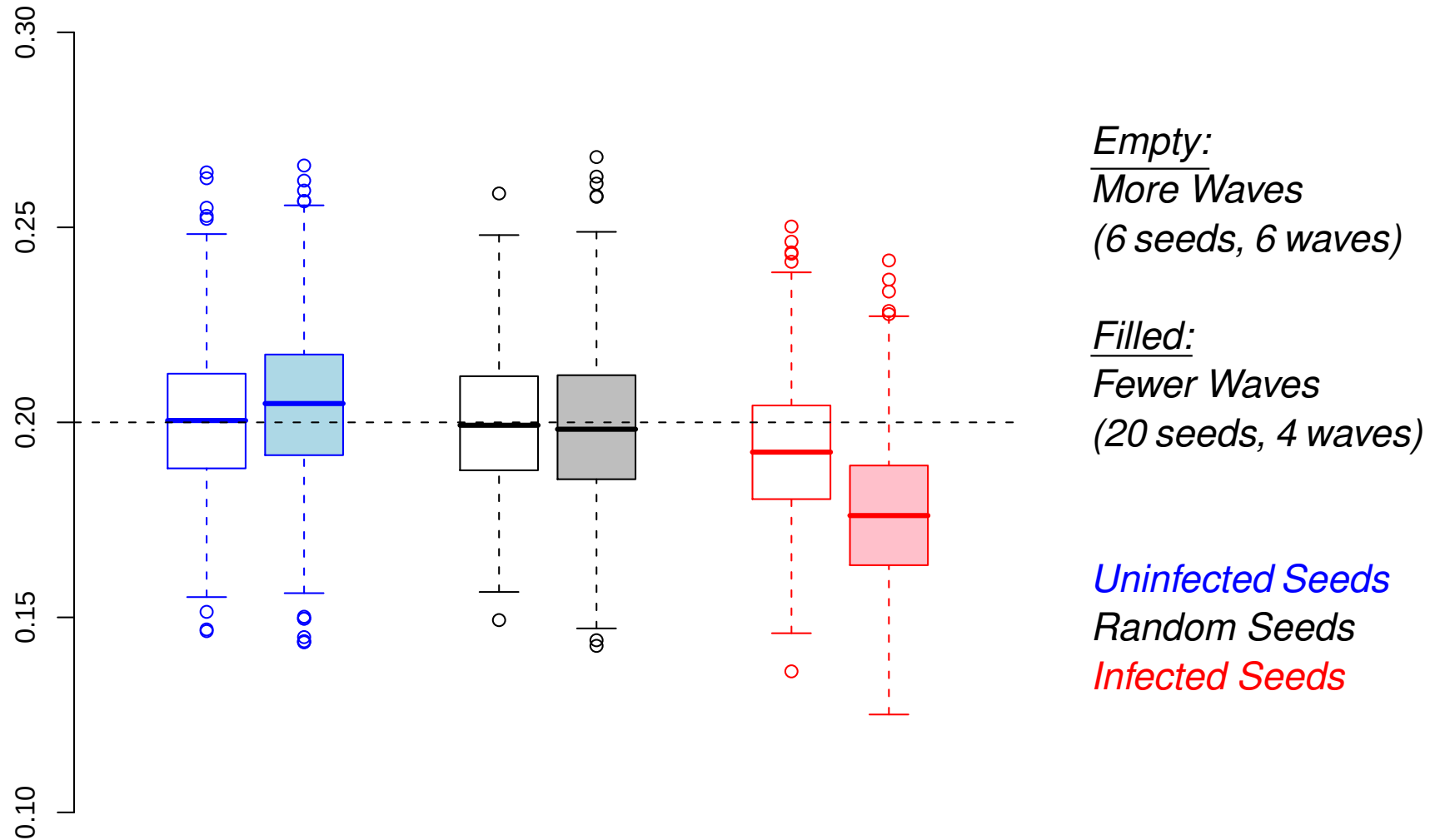
Drop Seeds



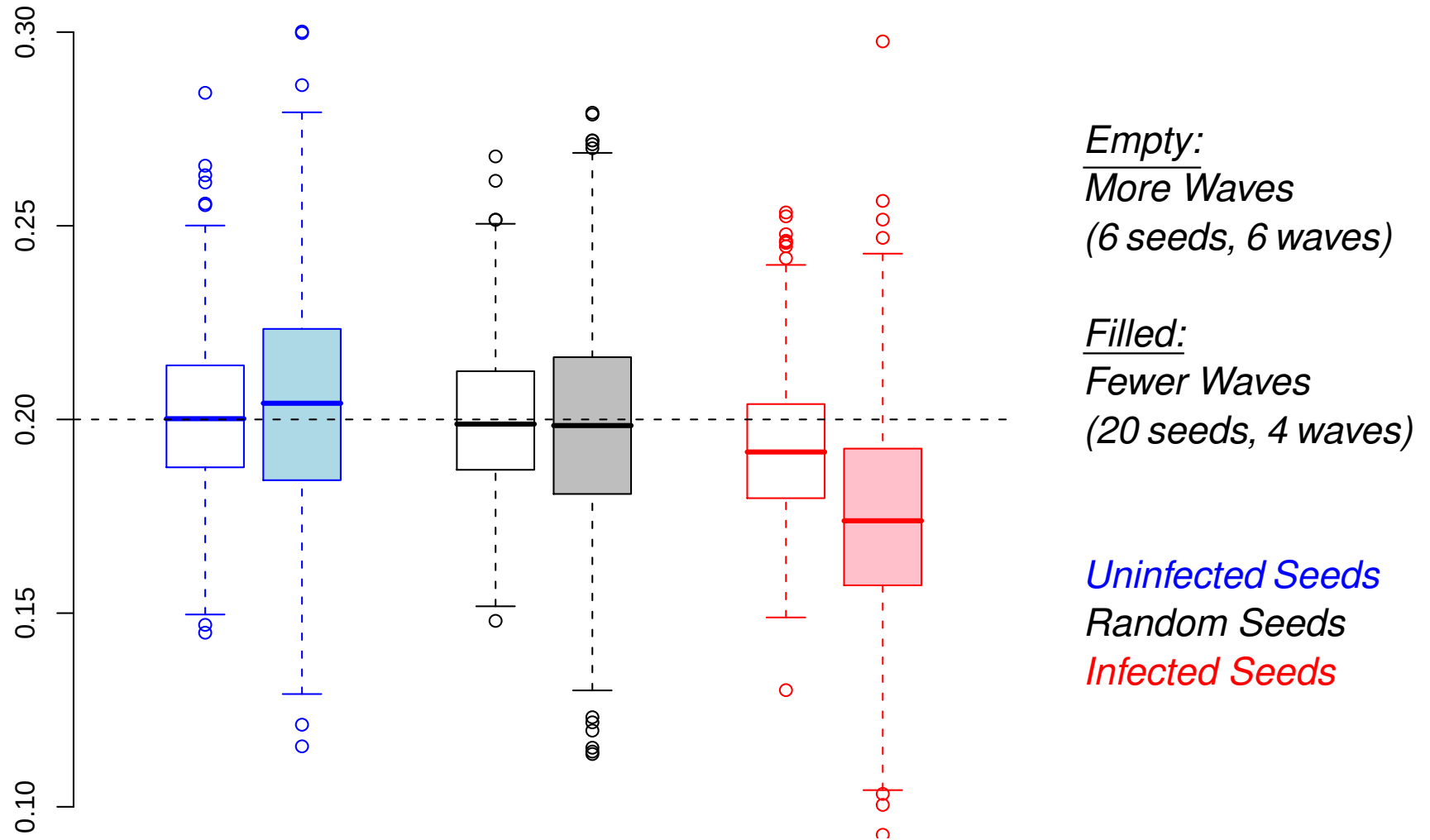
Drop Seeds, Wave 1



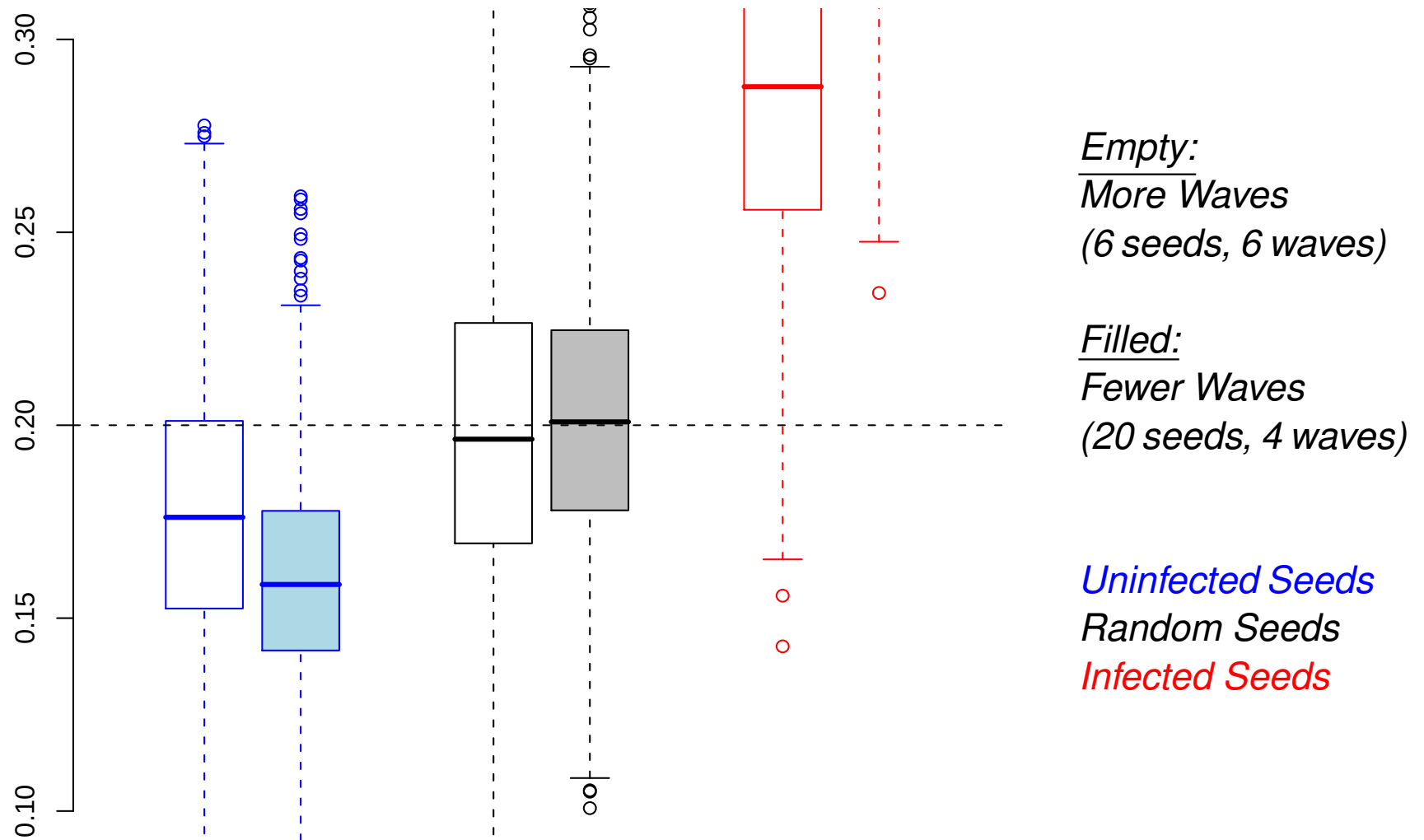
Drop Seeds, Wave 1, Wave 2



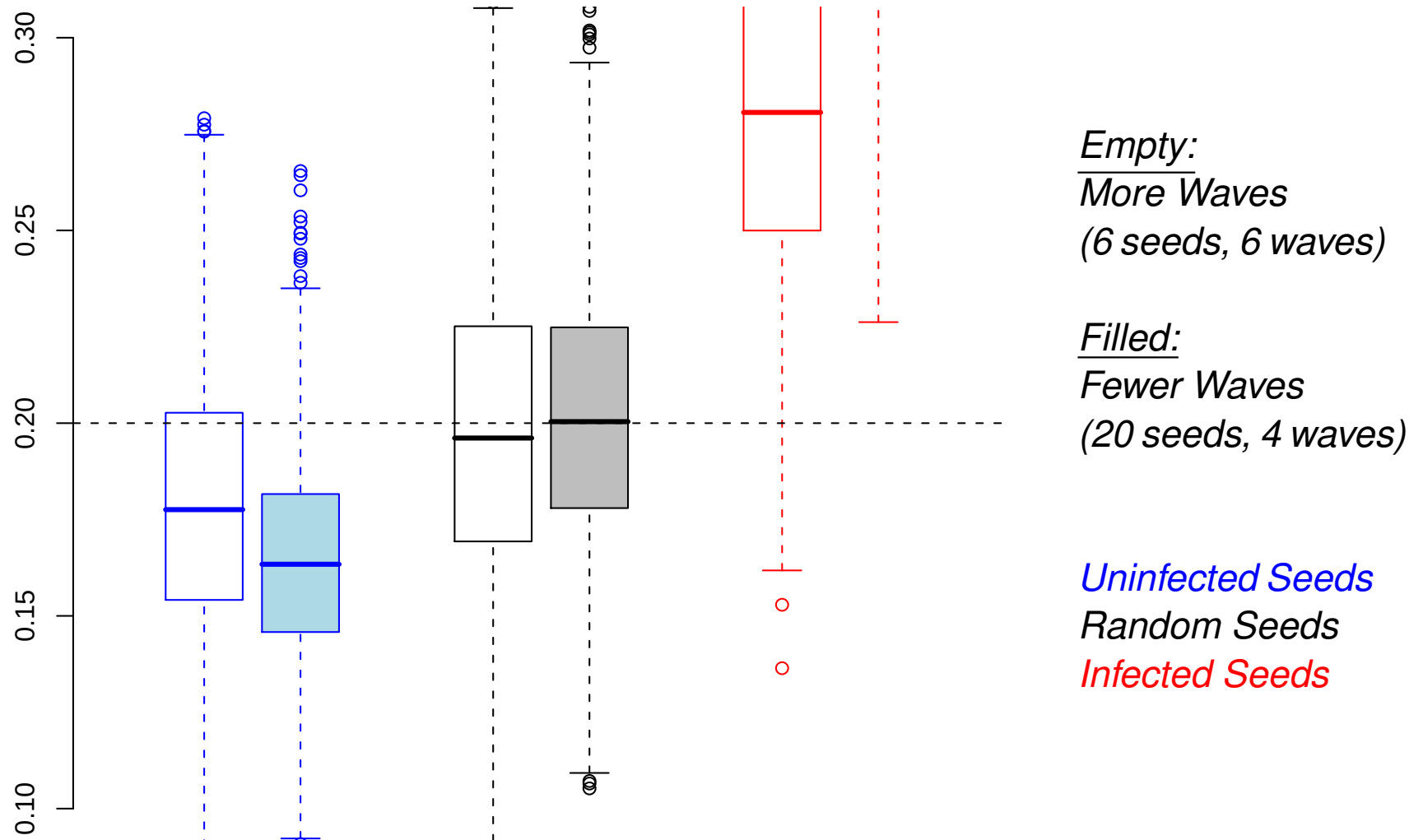
Drop Seeds, Wave 1, Wave 2, Wave 3



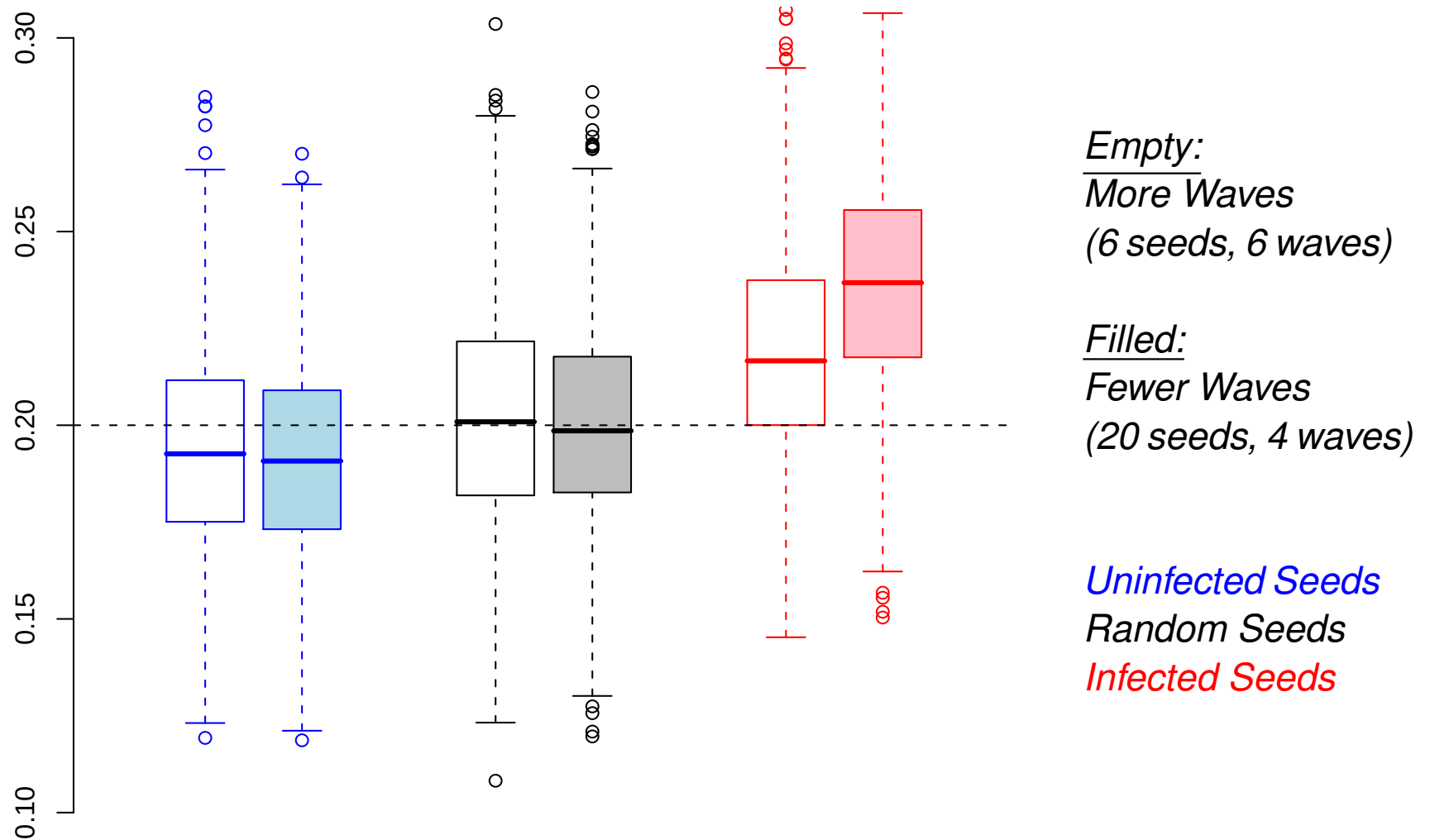
Assumption: With-Replacement Sampling, Drop Seeds



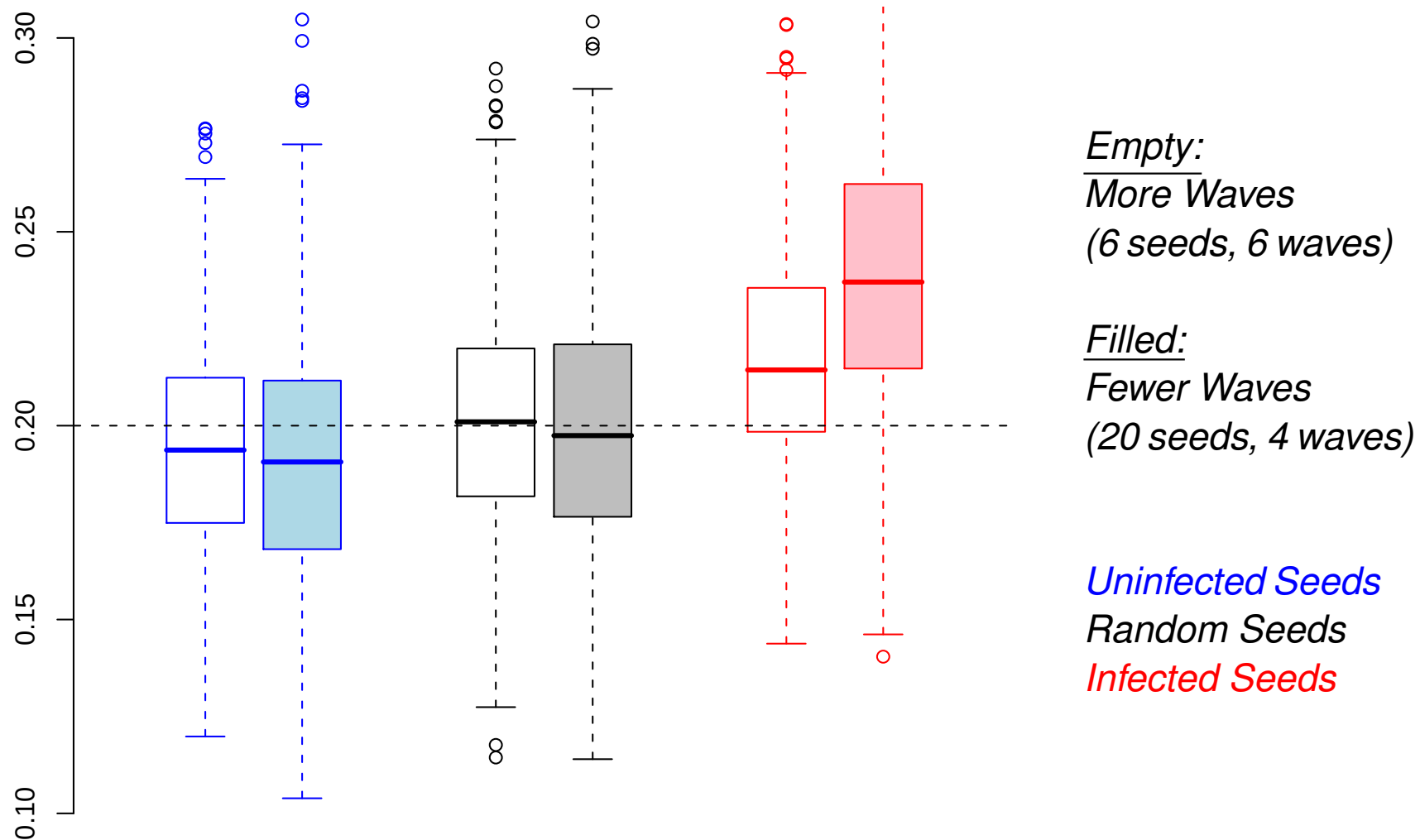
With-Replacement, Drop Seeds, Wave 1



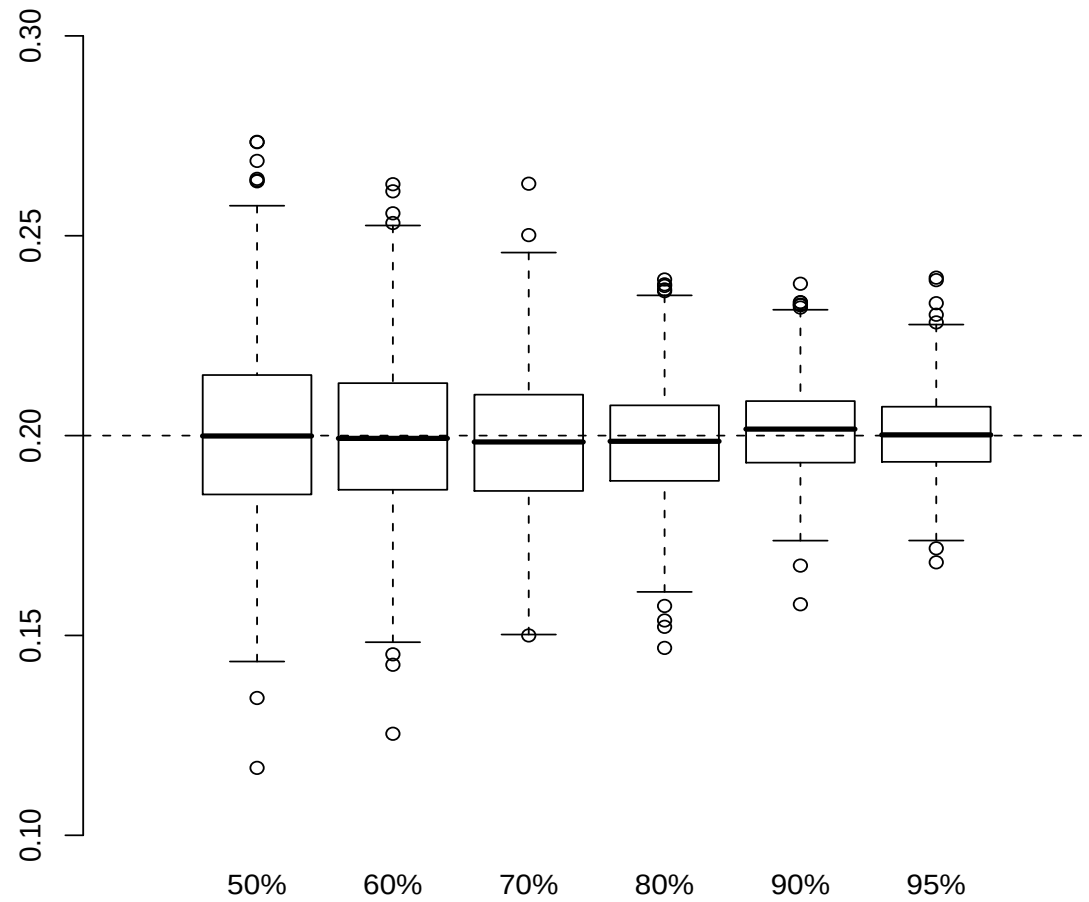
With-Replacement, Drop Seeds, Wave 1, Wave 2



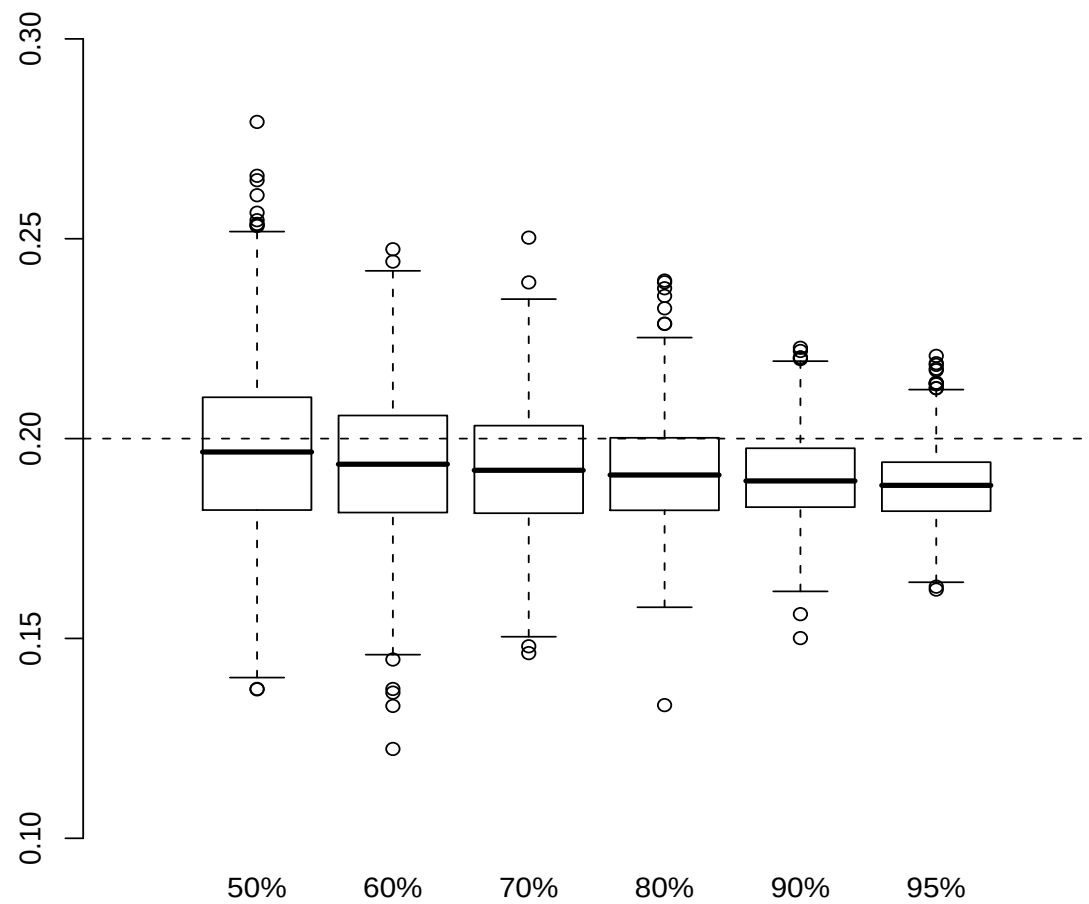
With-Replacement, Drop Seeds, Wave 1, Wave 2, Wave 3



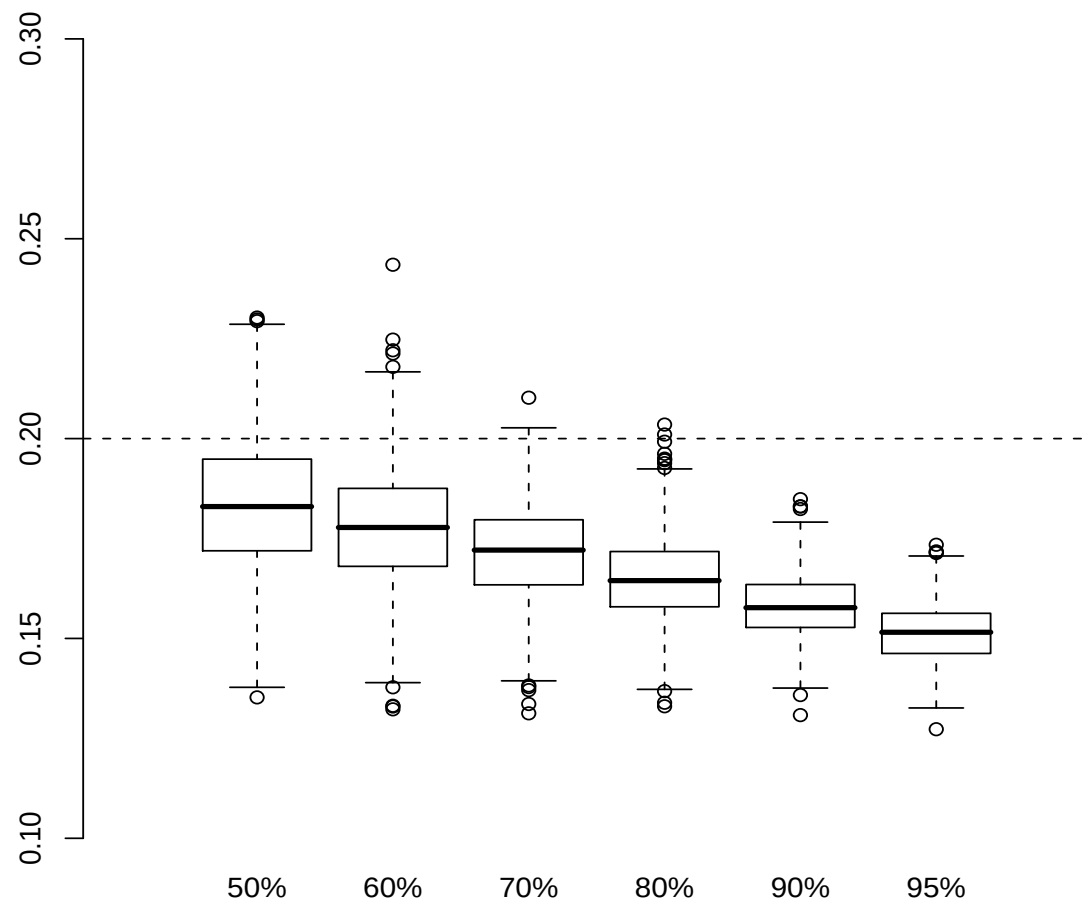
$w=1$



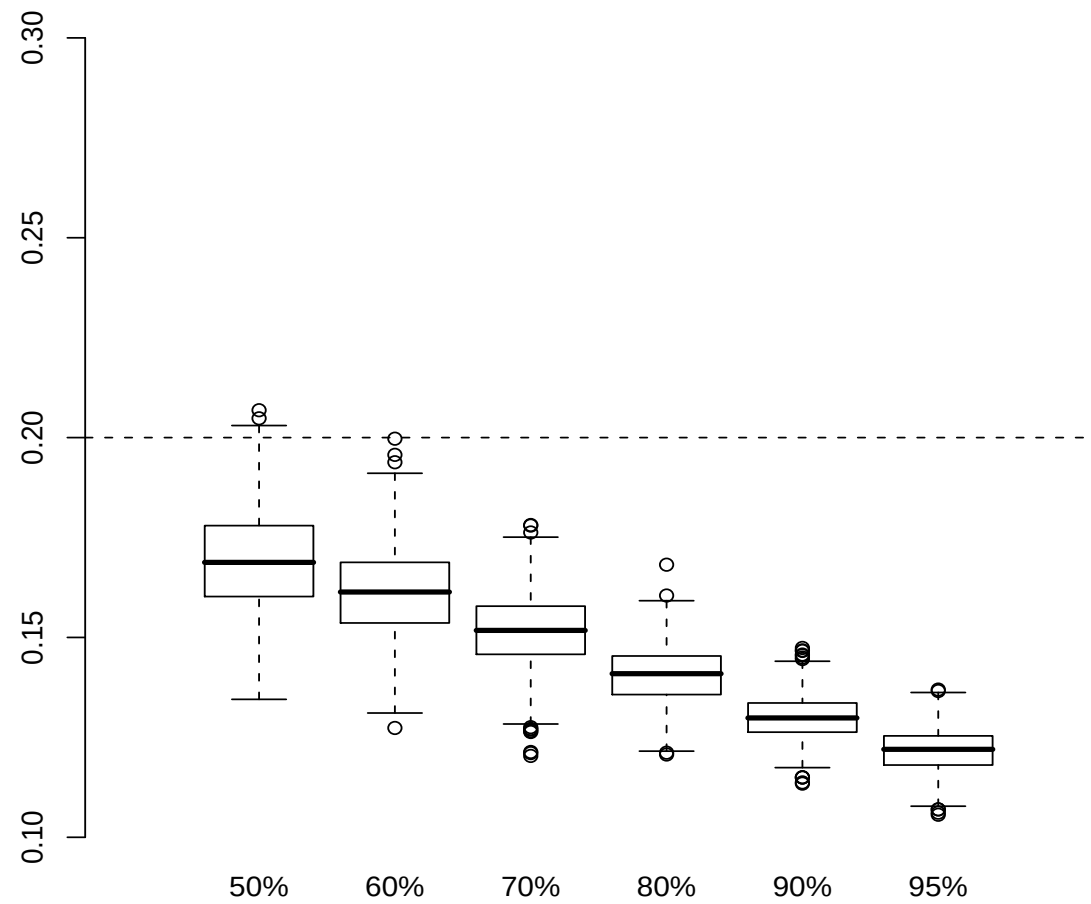
$w=1.1$



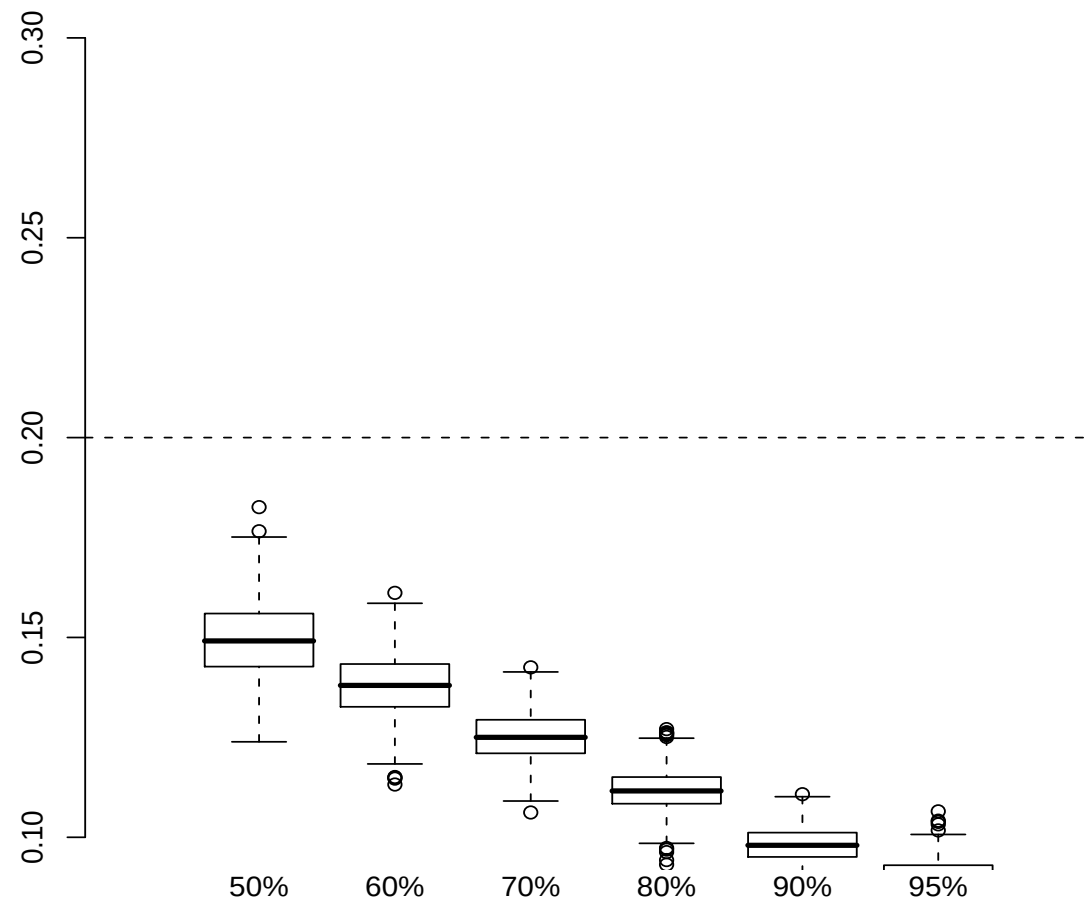
$w=1.5$



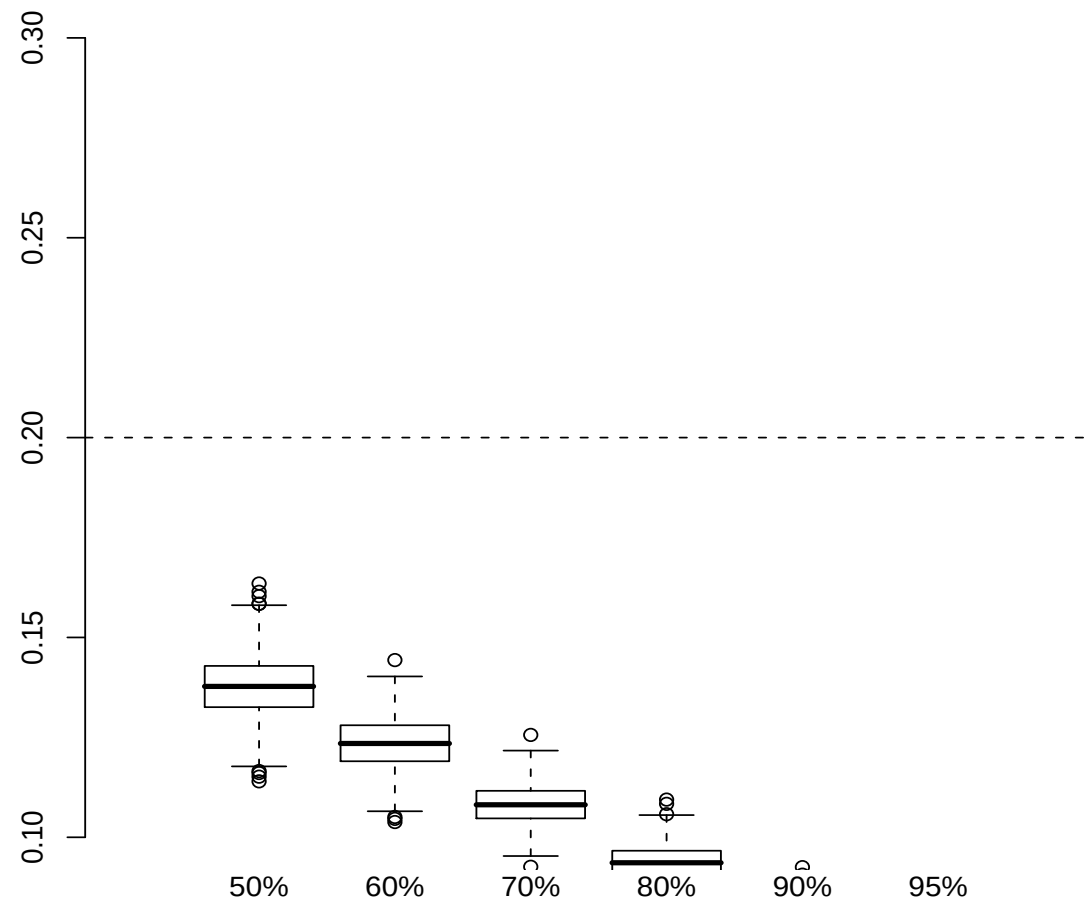
$w=2$



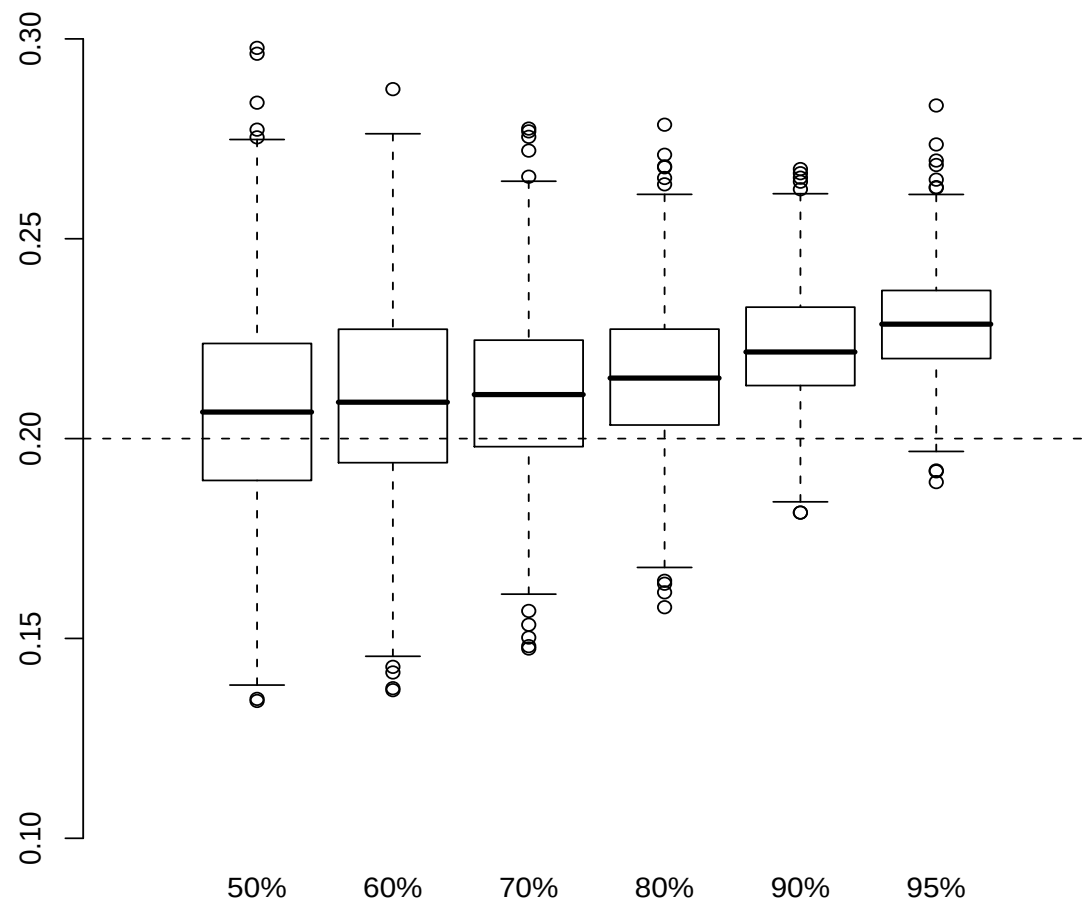
$w=3$



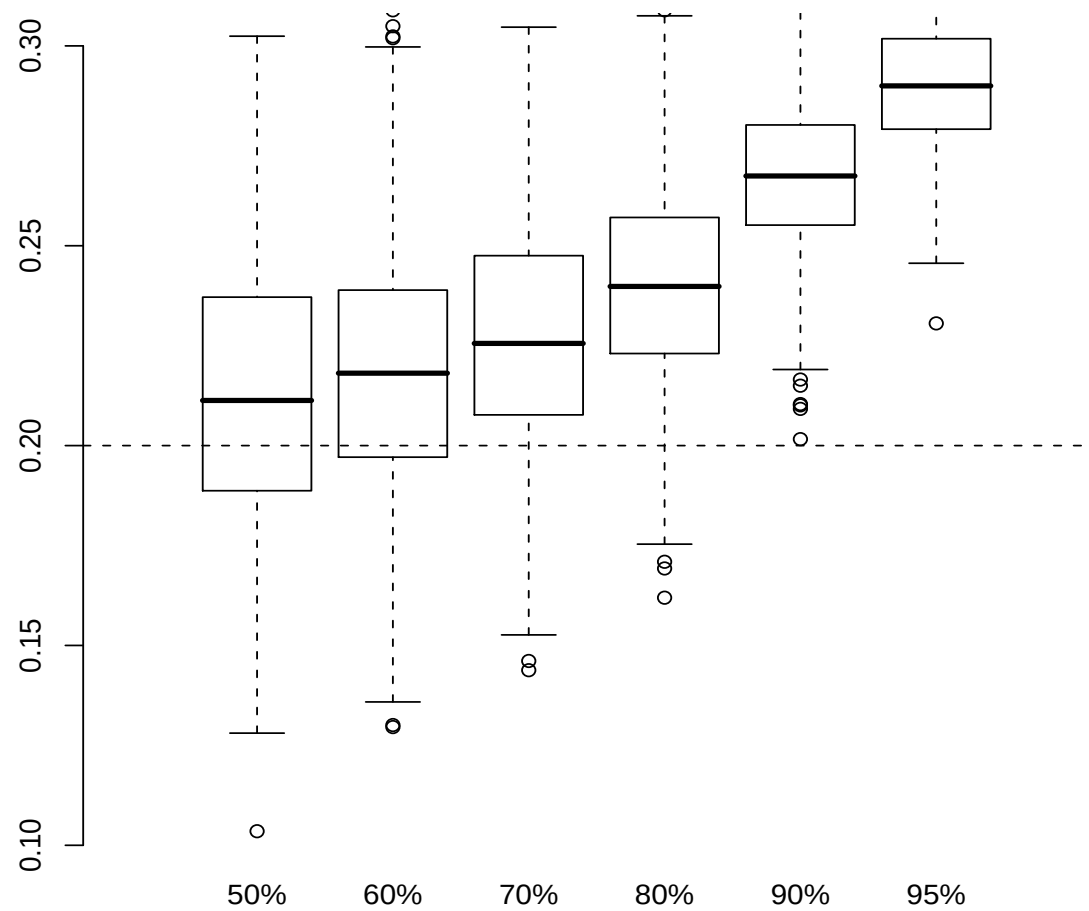
$w=4$



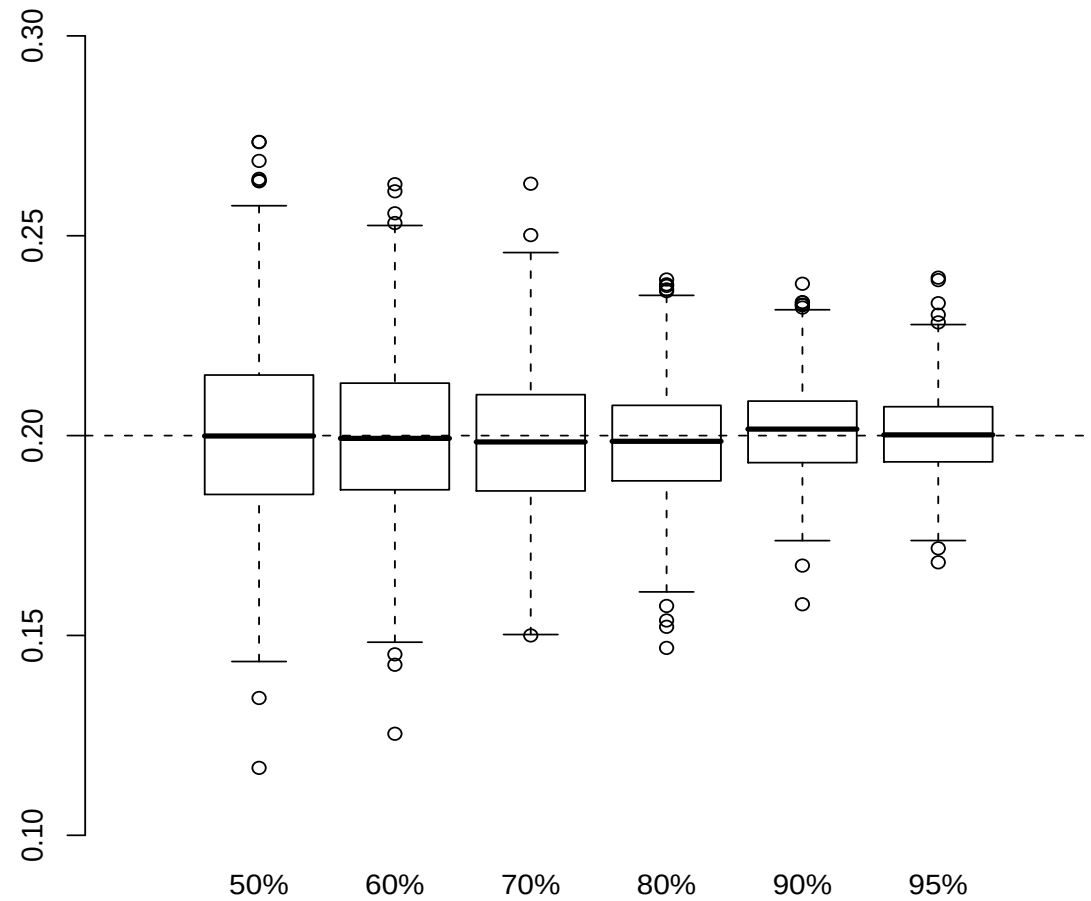
$w=0.8$



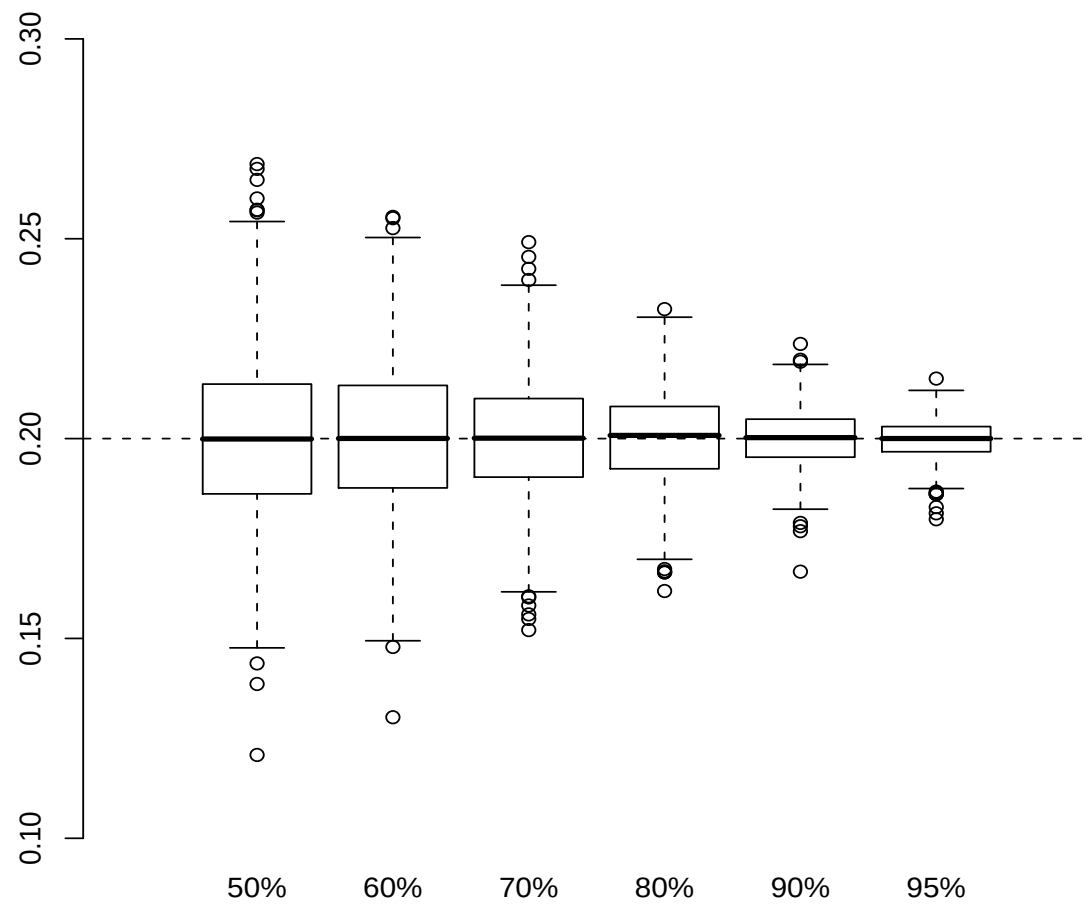
$w=0.5$



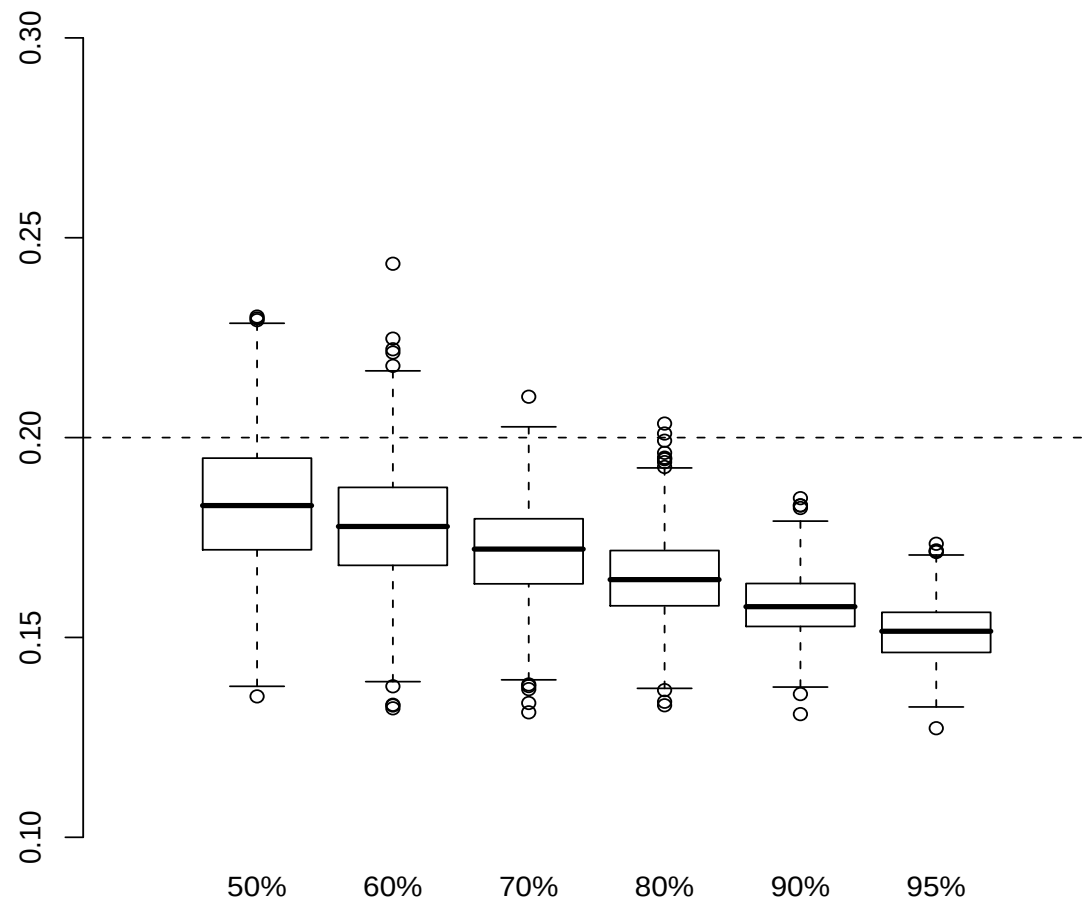
Volz-Heckathorn, $w=1$



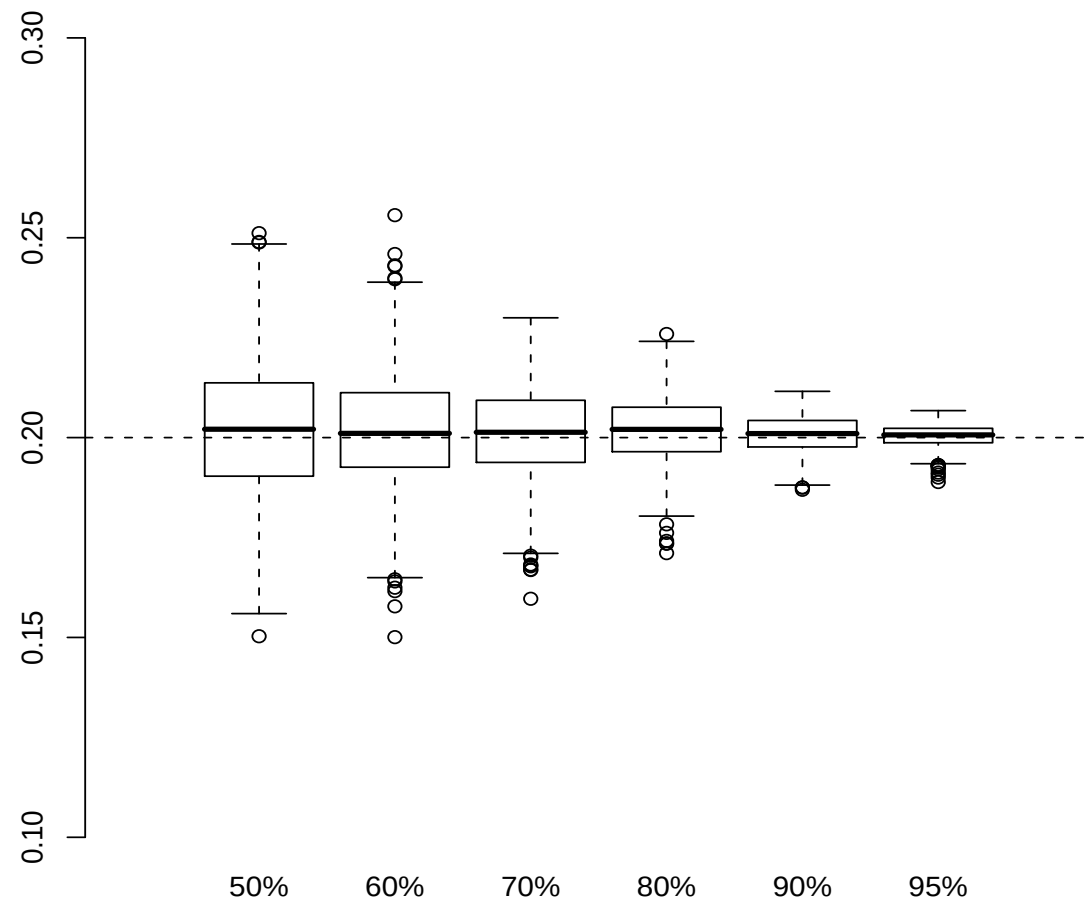
SS, $w=1$



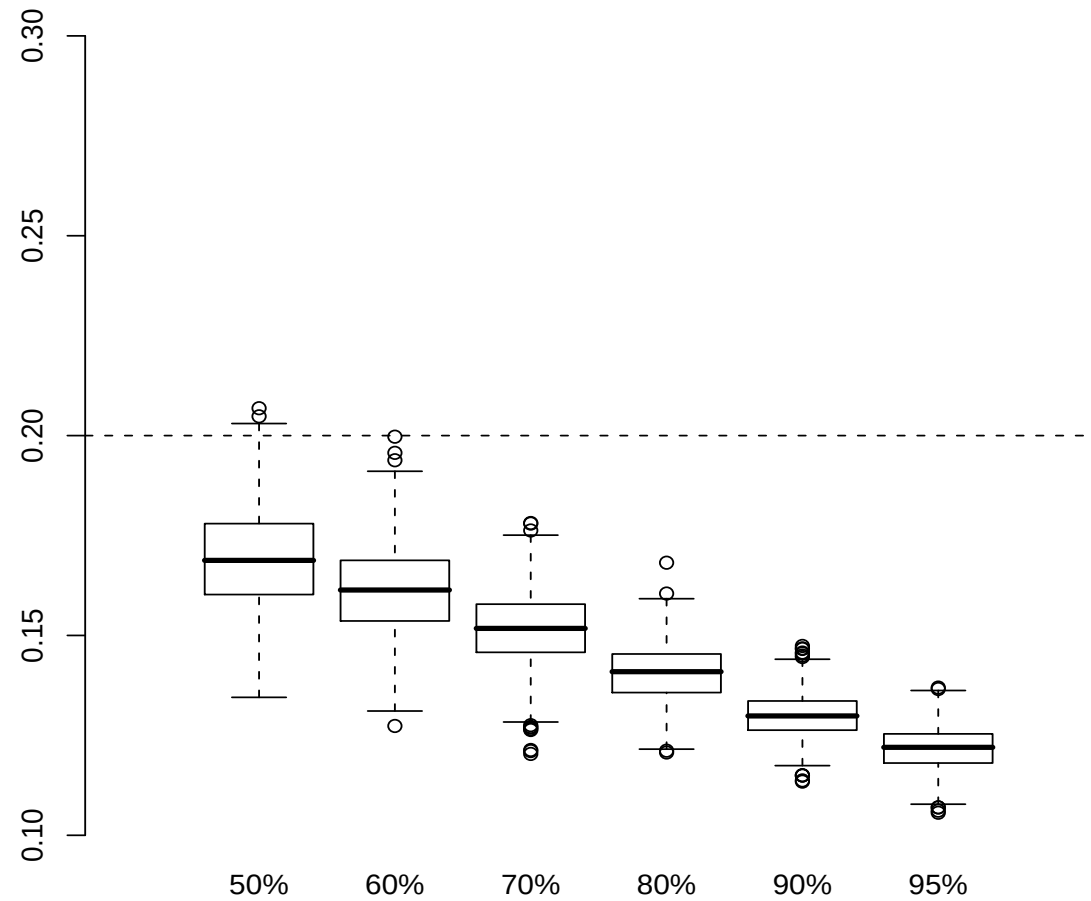
Volz-Heckathorn, $w=1.5$



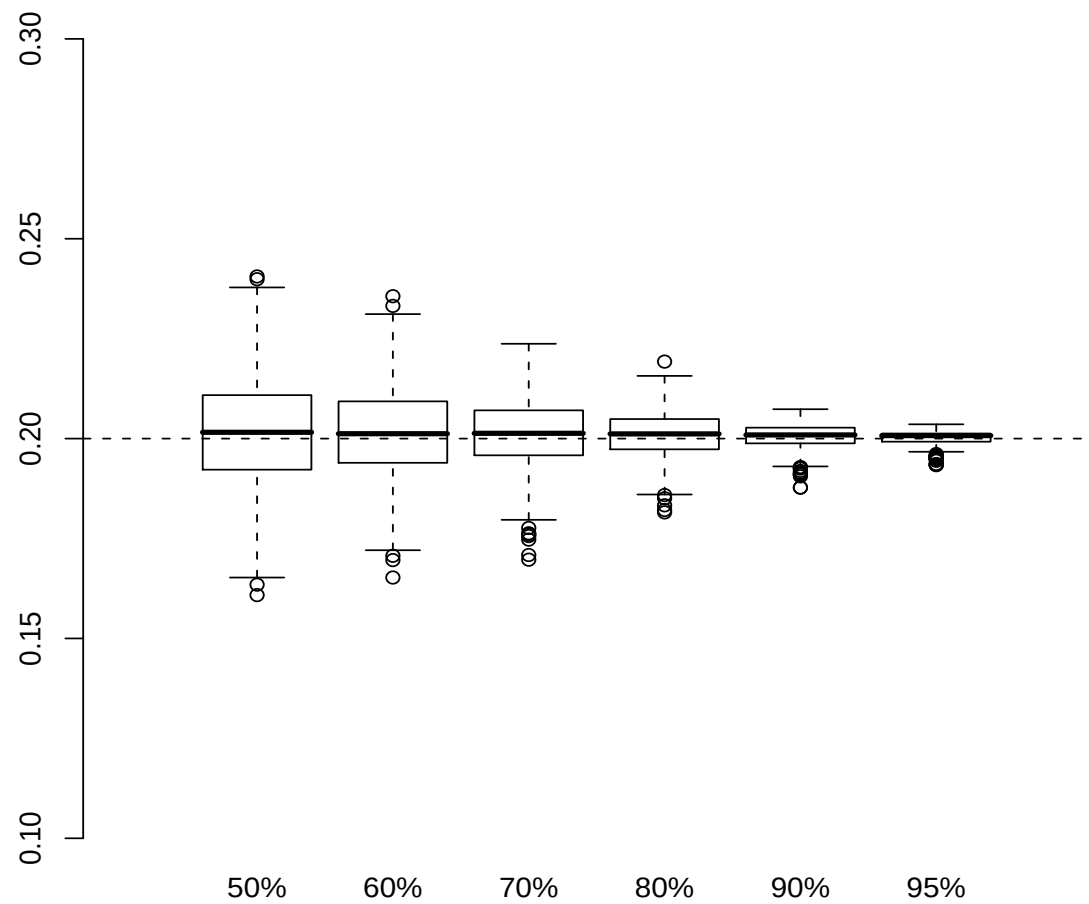
SS, $w=1.5$



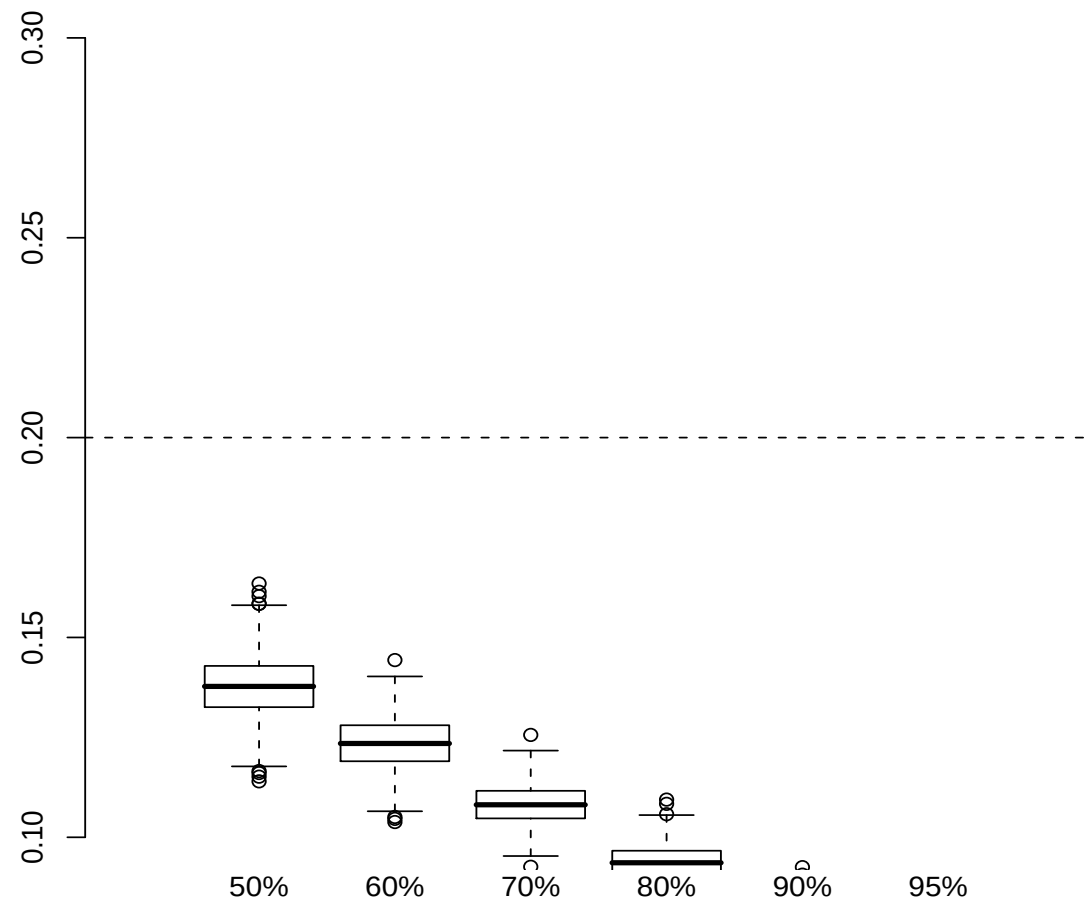
Volz-Heckathorn, $w=2$



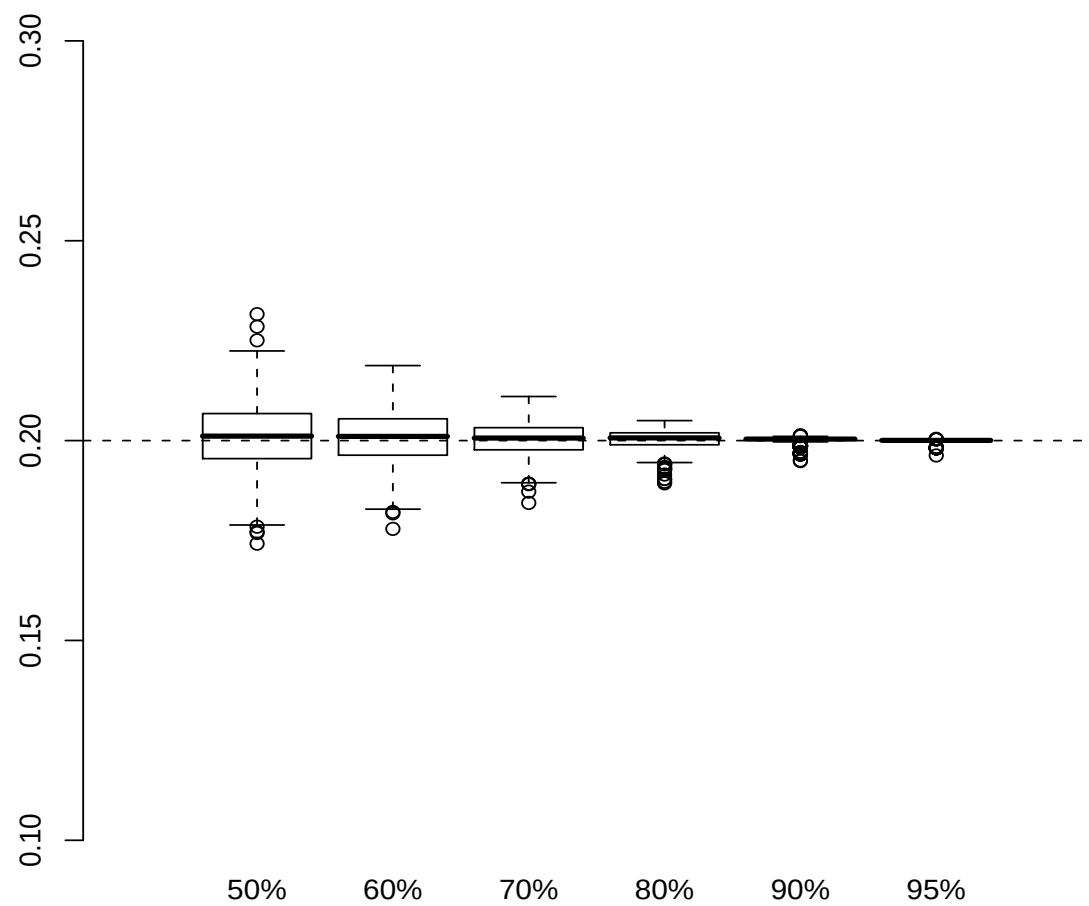
SS, $w=2$



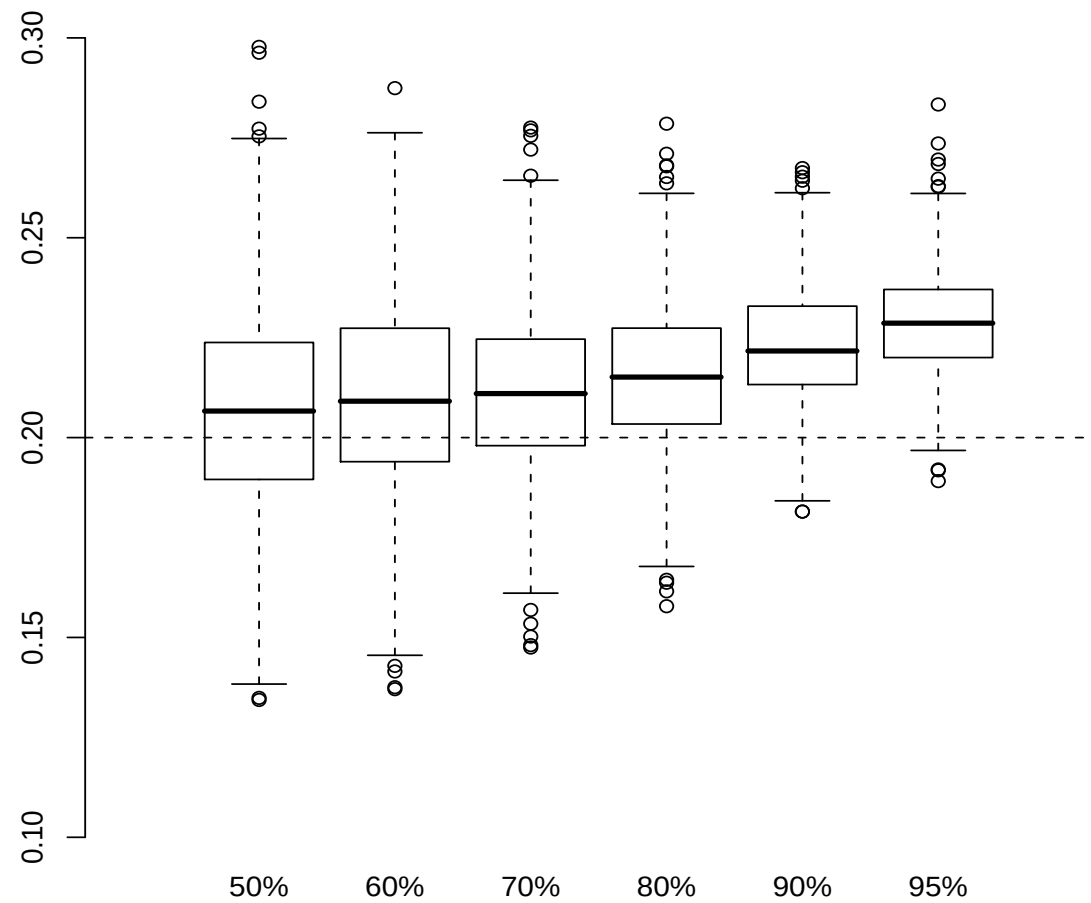
Volz-Heckathorn, $w=4$



SS, $w=4$

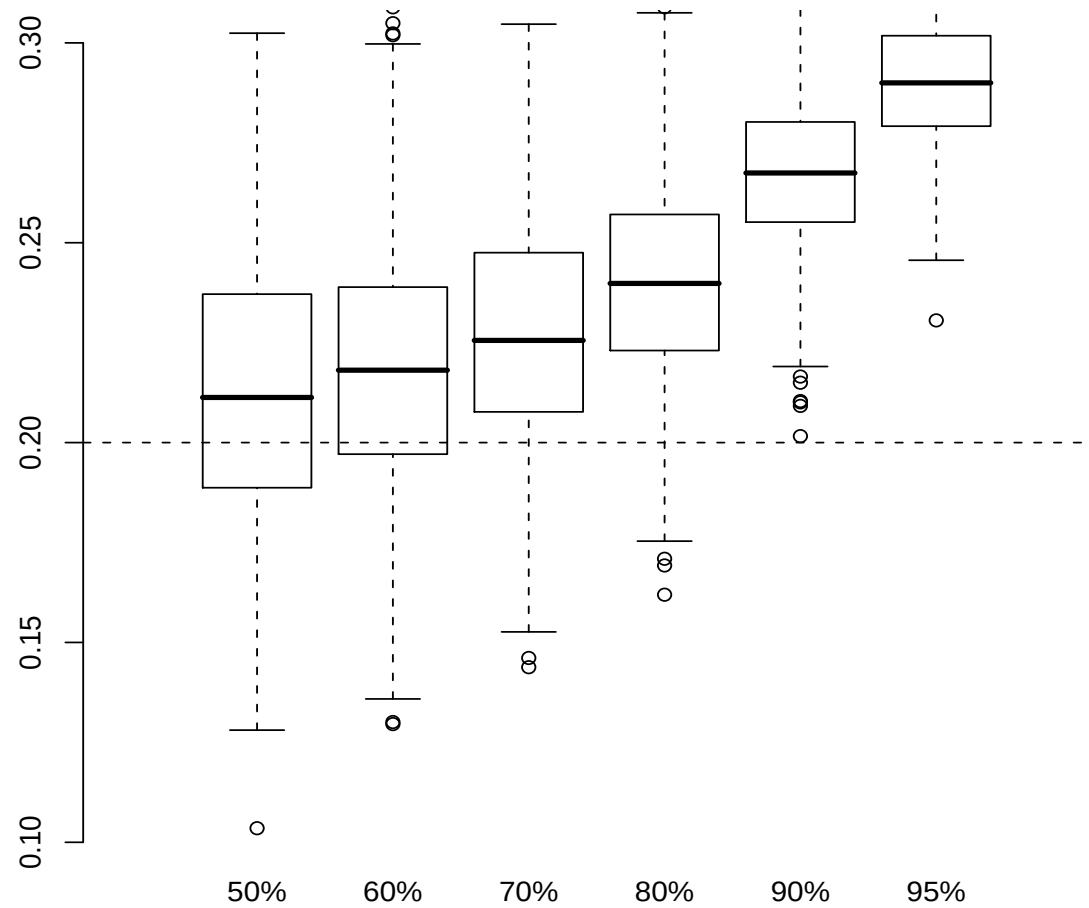


Volz-Heckathorn, $w=0.8$

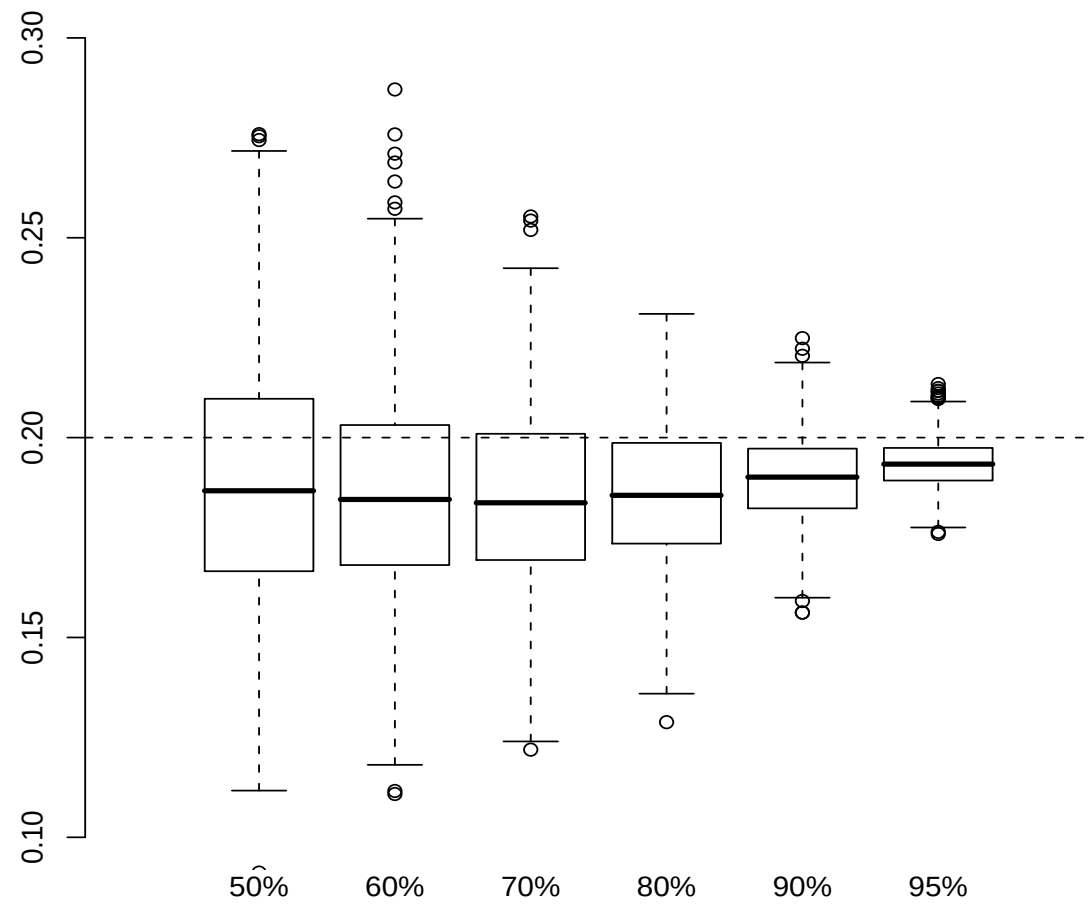




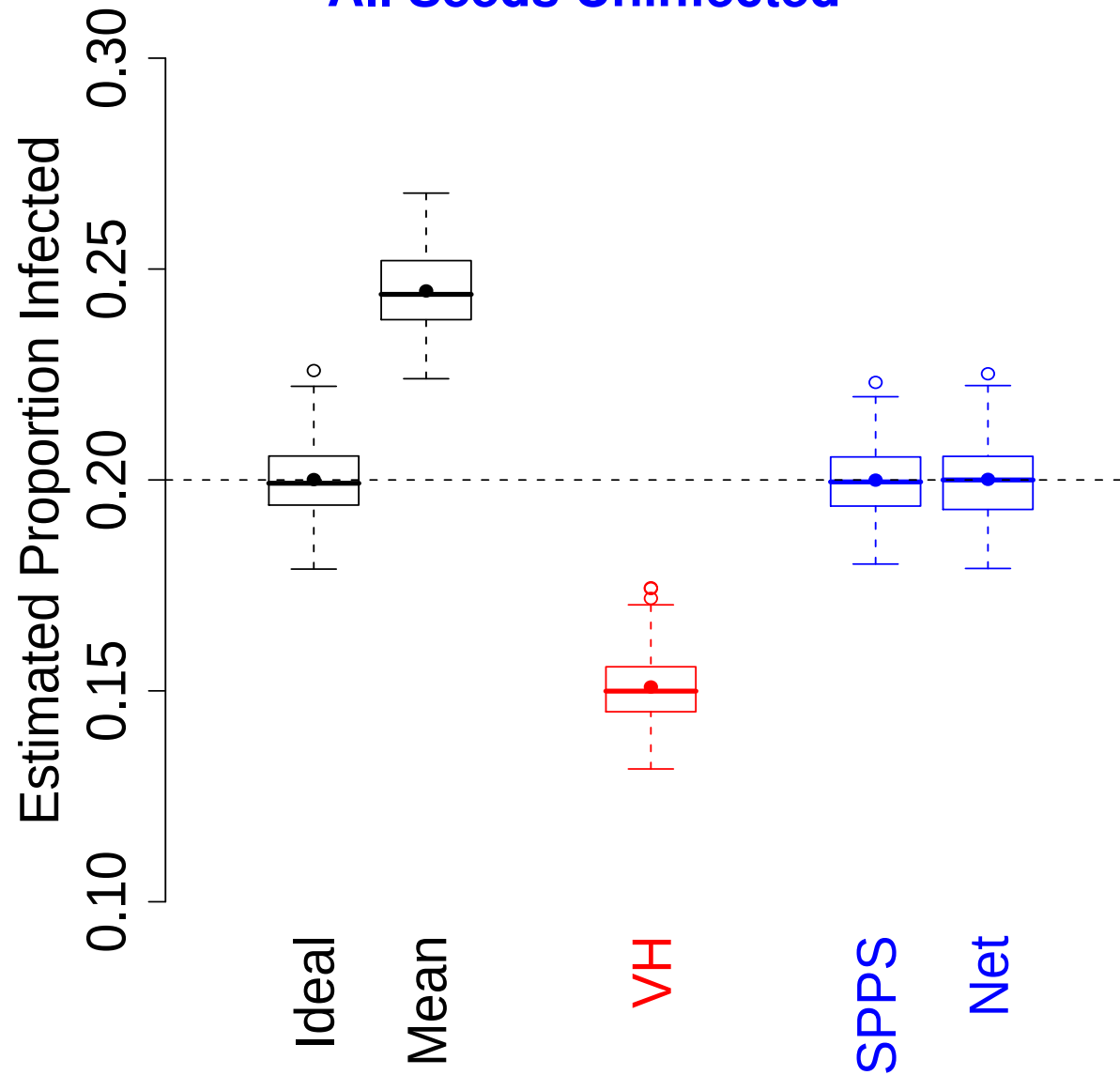
Volz-Heckathorn, $w=0.5$



SS, $w=0.5$



All Seeds Uninfected



Simulation Parameters

<i>Parameter</i>	<i>Homophily</i>	<i>Waves</i>	<i>Sampling Proportion</i>	<i>Bias</i>
<i>Number of Nodes</i>	<i>1000</i>	<i>1000</i>	<i>1000 to 526</i>	<i>1000</i>
<i>Mean Degree</i>	<i>7</i>	<i>7</i>	<i>7</i>	<i>7</i>
<i>Proportion Infected</i>	<i>0.2</i>	<i>0.2</i>	<i>0.2</i>	<i>0.2</i>
<i>Clustering Factor</i>	<i>2 and 4</i>	<i>2</i>	<i>2</i>	<i>2</i>
<i>Increased Infected Activity</i>	<i>1</i>	<i>1</i>	<i>.4</i>	<i>1</i>
<i>Initial Samples</i>	<i>10</i>	<i>6 and 20</i>	<i>10</i>	<i>10</i>
<i>Coupons (Branches)</i>	<i>2</i>	<i>2</i>	<i>2</i>	<i>2</i>
<i>Coupon Bias</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1 and 1.2</i>
<i>Number Samples</i>	<i>500</i>	<i>500</i>	<i>500</i>	<i>500</i>
<i>Number of Trials</i>	<i>1000</i>	<i>1000</i>	<i>1000</i>	<i>1000</i>

The original document ended here

Current RDS Estimation

- *Estimate sampling probability $\pi_i \propto d_i$, where*

$d_i = \text{degree of node } i$

- *Volz-Heckathorn (VH) Estimator:*

$$\hat{\mu}_{VH} = \frac{\sum_{i:S_i=1} \frac{1}{d_i} z_i}{\sum_{i:S_i=1} \frac{1}{d_i}}$$

- *$\pi_i \propto d_i$ based on stationary distribution of random walk on nodes.*

*Erik Volz and Douglas D. Heckathorn, “Probability Estimation Theory for Respondent Driven Sampling,” *Journal of Official Statistics*, 24:1, 2008.*

Challenges for Estimation

- *Hajek Estimator*
- *A leading existing RDS estimator*
- *Two Concerns*
- *Simulation Study*

Hajek Estimator

- *Goal: Estimate proportion “infected” :*

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

where

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

- *Hajek Estimator:*

$$\hat{\mu} = \frac{\sum_{i:S_i=1} \frac{1}{\pi_i} z_i}{\sum_{i:S_i=1} \frac{1}{\pi_i}}$$

where

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

Key Point: Requires $\pi_i \forall i : S_i = 1$

Current RDS Estimation

- *Estimate sampling probability $\pi_i \propto d_i$, where*

$d_i = \text{degree of node } i$

- *Volz-Heckathorn (VH) Estimator:*

$$\hat{\mu}_{VH} = \frac{\sum_{i:S_i=1} \frac{1}{d_i} z_i}{\sum_{i:S_i=1} \frac{1}{d_i}}$$

- *$\pi_i \propto d_i$ based on stationary distribution of random walk on nodes.*

Erik Volz and Douglas D. Heckathorn, “Probability Estimation Theory for Respondent Driven Sampling,” Journal of Official Statistics, 24:1, 2008.

Two Concerns:

- **Concern:** *Actual walk is without-replacement - no stationarity.*
- **Simulations Show:** *Substantial bias for mean degree unequal across groups and substantial sample fraction.*
- **Concern:** *Sampling chains may not be long enough to sufficiently reduce seed dependence - seed bias.*
- **Simulations Show:** *Substantial bias for all infected seeds, exacerbated by shorter sampling chains and stronger homophily.*

Gile, K.J., and M.S. Handcock, “Respondent-Driven Sampling: An Assessment of Current Methodology,” Sociological Methodology, 40, 2010, available on arXiv.

Successive Sampling (SS)

(aka PPSWOR)

- *Select first unit with probability proportional to degree (“size”)*
- *Select each subsequent unit with probability proportional to degree from among the previously unsampled units*
- *Use resulting weights to form new estimator.*

Krista J. Gile, “Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation,” Journal of the American Statistical Association, 106 (493), 135-146.

Network Model-Assisted Estimator

Premise:

- *Given:*
 - *True network form β (ERGM homophily parameter)*
 - *True sampling structure $\mathcal{S}$ (e.g. seed characteristics)*

Network-Model Based Estimator

Premise:

- *Given:*
 - *True network form β (ERGM homophily parameter)*
 - *True sampling structure $\mathcal{S}$ (e.g. seed characteristics)*
- *Then:*
 -

Network-Model Based Estimator

Premise:

- *Given:*
 - *True network form β (ERGM homophily parameter)*
 - *True sampling structure $\mathcal{S}$ (e.g. seed characteristics)*
- *Then:*
 - *Nodal equivalence classes based on degrees k and infection status z*
 - *Exists a mapping $\pi_{k,z}^*(\beta, \mathcal{S}) : \{k, z\} \rightarrow \pi$, from the nodal degree k and value z , to nodal inclusion probabilities, π .*

Furthermore,

- *Given inclusion probabilities, $\{\pi_i\}$, use Horvitz-Thompson estimators to estimate the network and sampling structures, $\{\beta, \mathcal{S}\}$.*
- *Can simulate networks according to β .*
- *Can simulate samples according to $\mathcal{S}$.*
- *Use simulations to estimate mapping, $Q_{k,z}^*(\beta, \mathcal{S})$.*

Specifics:

- *Network Model: ERGM with differential homophily conditioning on degrees by infection class*
- *Sampling Model: Replicate seed characteristics, offspring distribution, sample size*