Sampling Hard-to-Reach Populations

Mark S. Handcock

Department of Statistics
University of California - Los Angeles

UCLA

Some joint work with

Krista J. Gile Katherine R. McLaughlin Brian Kim
Ian E. Fellows Henry F. Raymond

Supported by US NIH NICHD Grant HD041877 and US NSF award MMS-0851555.

Statistics Canada
2022 International Methodology Symposium

Stylized Conventional Survey Sampling

Stylized description

- Choose a *population* of interest and a population characteristic of interest μ
- Determine the *sampling frame*: $i = 1, \ldots, N$ sample units.
- Choose variables to measure on them: *outcome variables* $y_i, i = 1, \ldots, N$.

Population

Stylized Conventional Survey Sampling

Stylized description

- Choose a *population* of interest and a population characteristic of interest μ
- Determine the *sampling frame*: $i = 1, \ldots, N$ sample units.
- Choose variables to measure on them: *outcome variables* $y_i, i = 1, \ldots, N$.

Population

Stylized Conventional Survey Sampling

Stylized description

- Choose a *population* of interest and a population characteristic of interest μ
- Determine the *sampling frame*: $i = 1, \ldots, N$ sample units.
- Choose variables to measure on them: *outcome variables* $y_i, i = 1, \ldots, N$.
- Choose a *sampling design*:
  e.g., simple random sampling, stratified sampling on covariates, stratified sampling on y
- Choose a sample of units $i = 1, \ldots, n$ and collect data on the sampled units

Sampled people (green)

Stylized Conventional Survey Sampling

Stylized description

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

Using the sample to describe the population

- Goal: Estimate the population mean of y:

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

where

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

Using the sample to describe the population

- Goal: Estimate the population mean of y:

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

- Sample indicators

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

- Inclusion probabilities

$$\pi_i = P(S_i = 1) \qquad i = 1, \ldots, N$$

  e.g. simple random sampling

$$\pi_i = n/N \qquad i = 1, \ldots, N$$

Classic Design-Based Inference

- Goal: Estimate proportion positive (e.g., addicted, COVID+):

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

- Hajek Estimator:

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

where

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

Challenges to Traditional Survey Sampling

- Eroding survey response rates and non-ignorable non-response
  - computation of inclusion probabilities is difficult
  - estimation of inclusion probabilities is required

- Allure of model-assisted and model-based modes of inference
  - models require assumptions
  - assumptions may be difficult to validate

Some types of non-probability samples

General types:

- Convenience sampling: no formal design, the goal is acquisition
- Online panels: using internet, social media, etc, to select and recruit
- Sample matching: stratify on important population characteristics
- Network sampling: applicable when the population is networked

Modes of inference for non-probability samples:

- Quasi-randomization: Design-based via a model for $\pi_i = P(S_i = 1)$
- Super-population: Model-based via a model for $P(Y_i)$

Non-probability survey sampling

We should not use it because:

- Accuracy relies on (often untestable) modeling assumptions
- The design is often opaque
- Validity depends on implementation and is case/application specific

We may use it because:

- It may be a “fit for purpose" design
- Sensitivity analyses for each case are possible
- We can insist on transparency / reproducibility

Hard-to-Reach Population Sampling: Motivating Questions

- What proportion of sub-Saharan migrants to Morocco have children?
- What proportion of semi-rural people are at high-risk for opioid addiction?
- What proportion of unregulated workers in New York City experience workplace violations of code?
- What proportion of Injecting Drug Users in Kampala are HIV Positive?
- What proportion of sex workers in rural China belong to ethnic minorities?

Limitation: No practical conventional sampling frame.

Adaptive Network Sampling

Suppose:

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

Networked Population

Adaptive Network Sampling

Suppose:

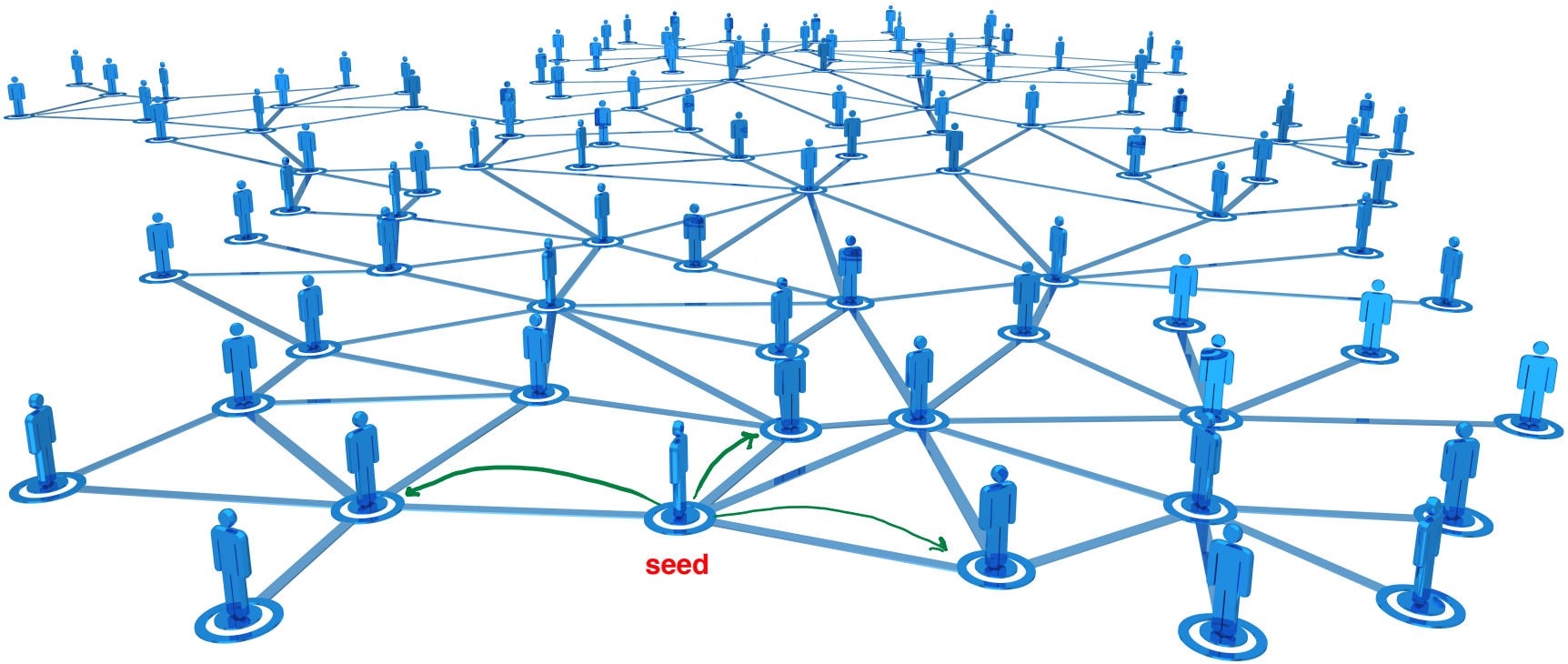
- The population is joined by informal social network of relationships.
- Researchers can access some members of the population.

Sampling design:

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

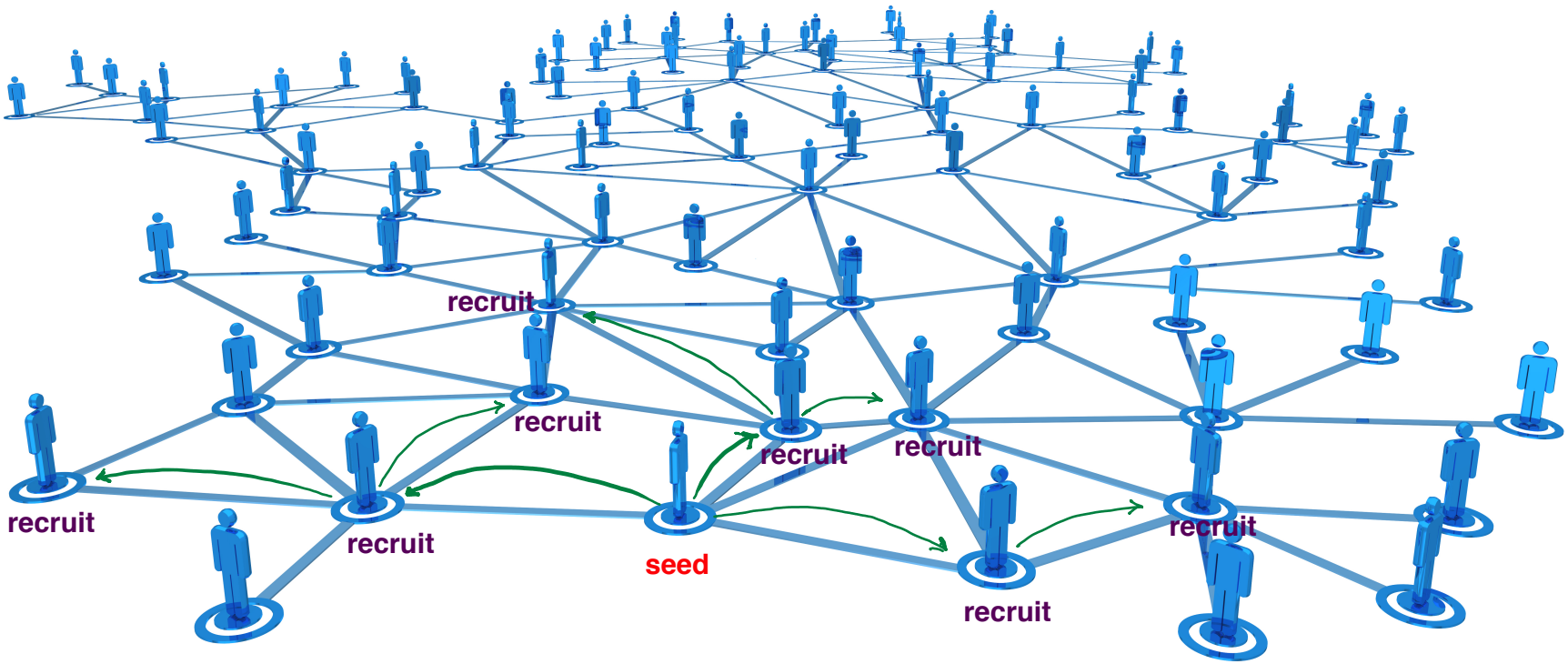
Start with a seed person

Contact other people via the seed's social network



Contact other people via the seed's social network

Contact other people via the seed's social network



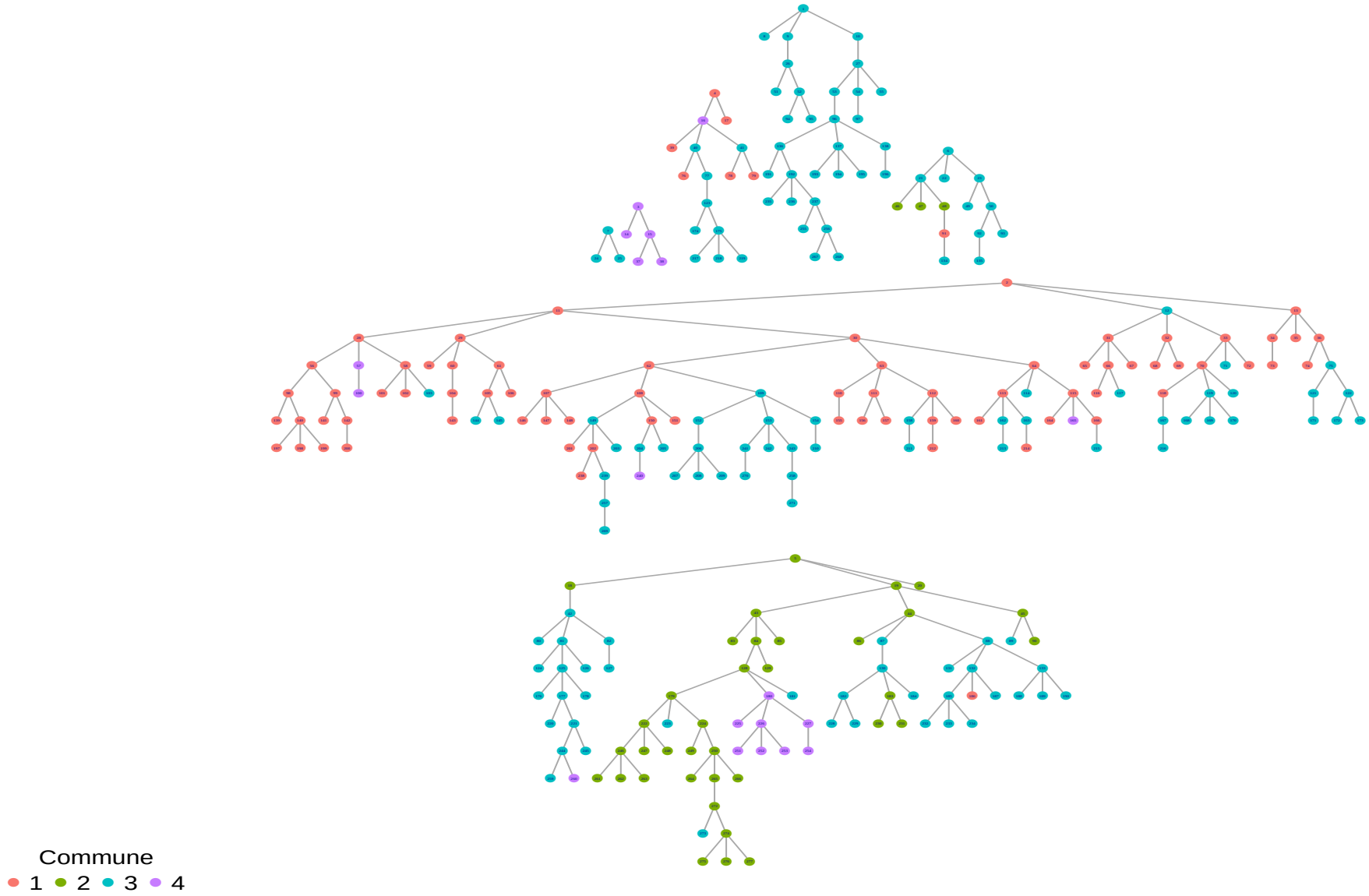
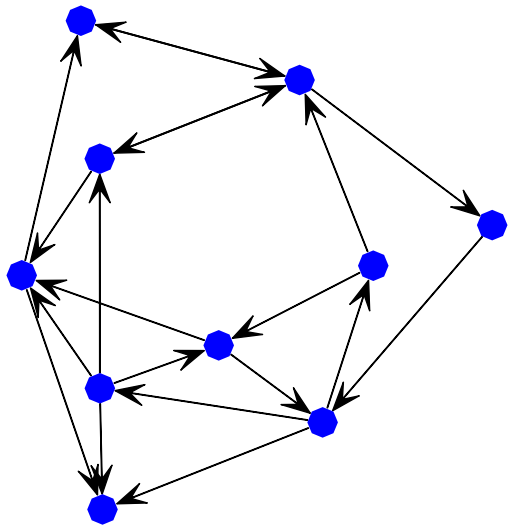


Figure: Graphical representation of the recruitment tree for the sampling of PWID. The nodes are the respondents and the wave number increases as you go down the page. The node color indicates the geographic neighborhood.

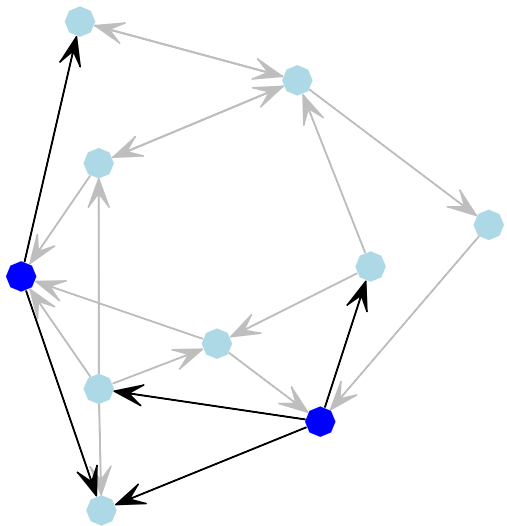
Partial Observation of Social Networks

- **Sampling Design:** Choose which part of an local community to observe: "Ask 10% of people about who have a big influence on them"



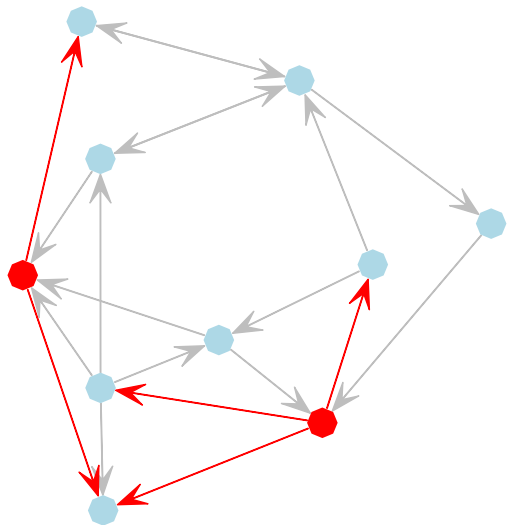
Partial Observation of Social Networks

- *Sampling Design:* Choose which part of an local community to observe: "Ask 10% of people who have a big influence on them"
  - *Egocentric*



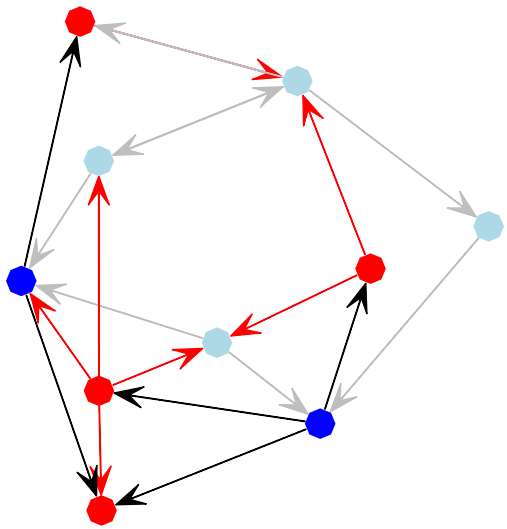
Partial Observation of Social Networks

- *Sampling Design:* Choose which part of an local community to observe: "Ask 10% of people who have a big influence on them"
  - Egocentric
  - *Adaptive*



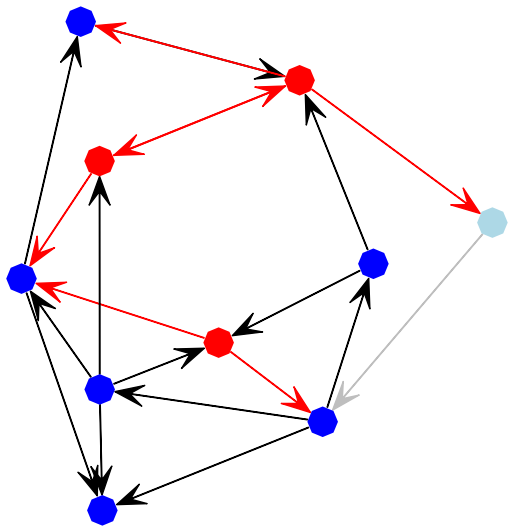
Partial Observation of Social Networks

- *Sampling Design:* Choose which part of an local community to observe: "Ask 10% of people who have a big influence on them"
  - Egocentric
  - *Adaptive*



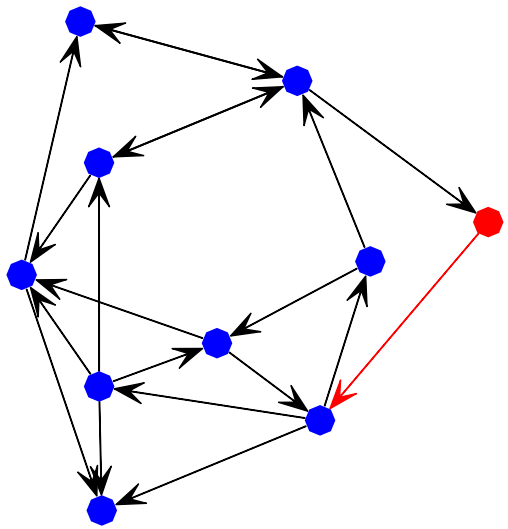
Partial Observation of Social Networks

- *Sampling Design:* Choose which part of an local community to observe: "Ask 10% of people who have a big influence on them"
  - Egocentric
  - *Adaptive*



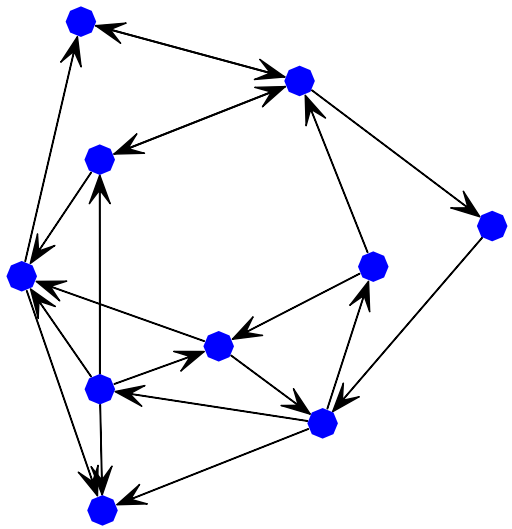
Partial Observation of Social Networks

- *Sampling Design:* Choose which part of an local community to observe: "Ask 10% of people who have a big influence on them"
  - Egocentric
  - *Adaptive*



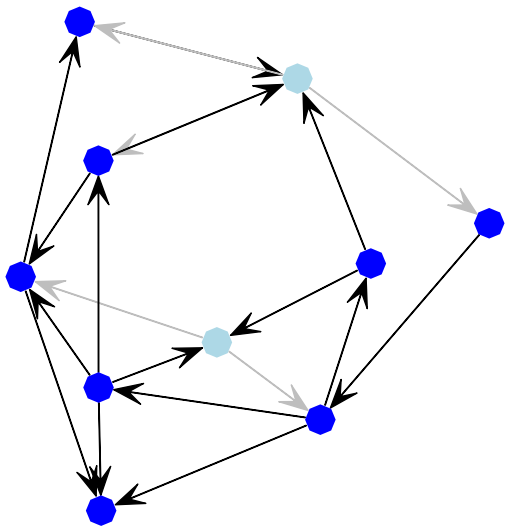
Partial Observation of Social Networks

- *Sampling Design:* Choose which part of an local community to observe: "Ask 10% of people who have a big influence on them"
  - Egocentric
  - *Adaptive*



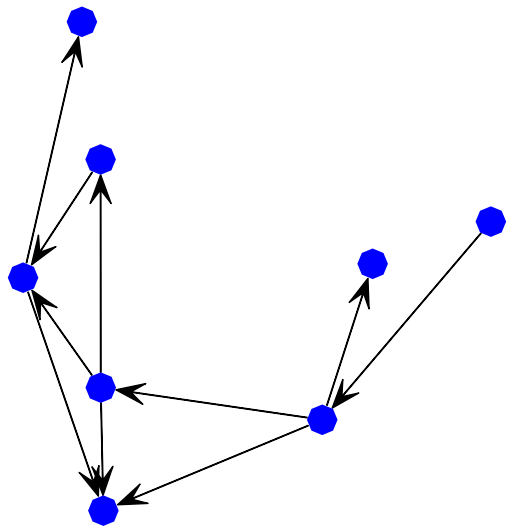
Partial Observation of Social Networks

- **Sampling Design:** Choose which part of an local community to observe:
  "Ask 10% of people who have a big influence on them"
  - Egocentric
  - Adaptive
- *Out-of-design Missing Data:*
  "Try to survey the whole community, but someone is unavailable"



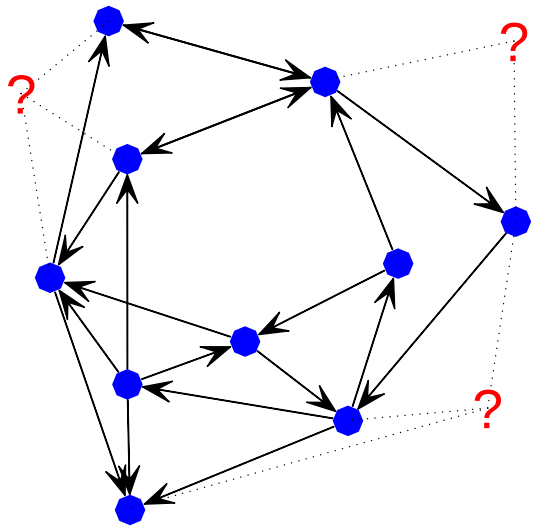
Partial Observation of Social Networks

- **Sampling Design:** Choose which part of an local community to observe:
  "Ask 10% of people who have a big influence on them"
  - Egocentric
  - Adaptive
- *Out-of-design Missing Data:*
  "Try to survey the whole community, but someone is unavailable"



Partial Observation of Social Networks

- **Sampling Design:** Choose which part of an local community to observe:
  "Ask 10% of people who have a big influence on them"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole community, but someone is unavailable"
- *Boundary Specification Problem:*
  "Should an overseas family member be considered a part of the community?"



Adaptive Network Sampling

Suppose:

- The population is 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 the sample by the researchers sampling those tied to those already in the sample.

Concerns:

- Seed Dependence: final sample depends on sampling mechanism of seeds
- Confidentiality: some populations prefer to stay "hidden"
- Estimation: The sample and sampling probabilities depend on the unknown network

Adaptive Network Sampling

Strengths:

- Exploits information in the network of relationships
- Network structure used to improve the design
- Increases the range of possible designs
- Adjusts for discovered features in the population
- Leads to increased efficiency of sampling

Adaptive Network Sampling

Strengths:

- Exploits information in the network of relationships
- Network structure used to improve the design
- Increases the range of possible designs
- Adjusts for discovered features in the population
- Leads to increased efficiency of sampling

Issues:

- Seed Dependence: final sample depends on sampling mechanism of seeds
- Privacy: some populations prefer to stay "hidden"
- Link-tracing can be challenging: confidentiality, logistics
- Estimation: The sample and sampling probabilities depend on the unknown network

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	Yes	Yes	Yes	Yes
One-Wave	Yes	No	No	No
k—Wave, $1 < k < \infty$	No	No	No	No
Saturated	Yes	No	No	No

(Unconditional) sampling probabilities unknown 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.

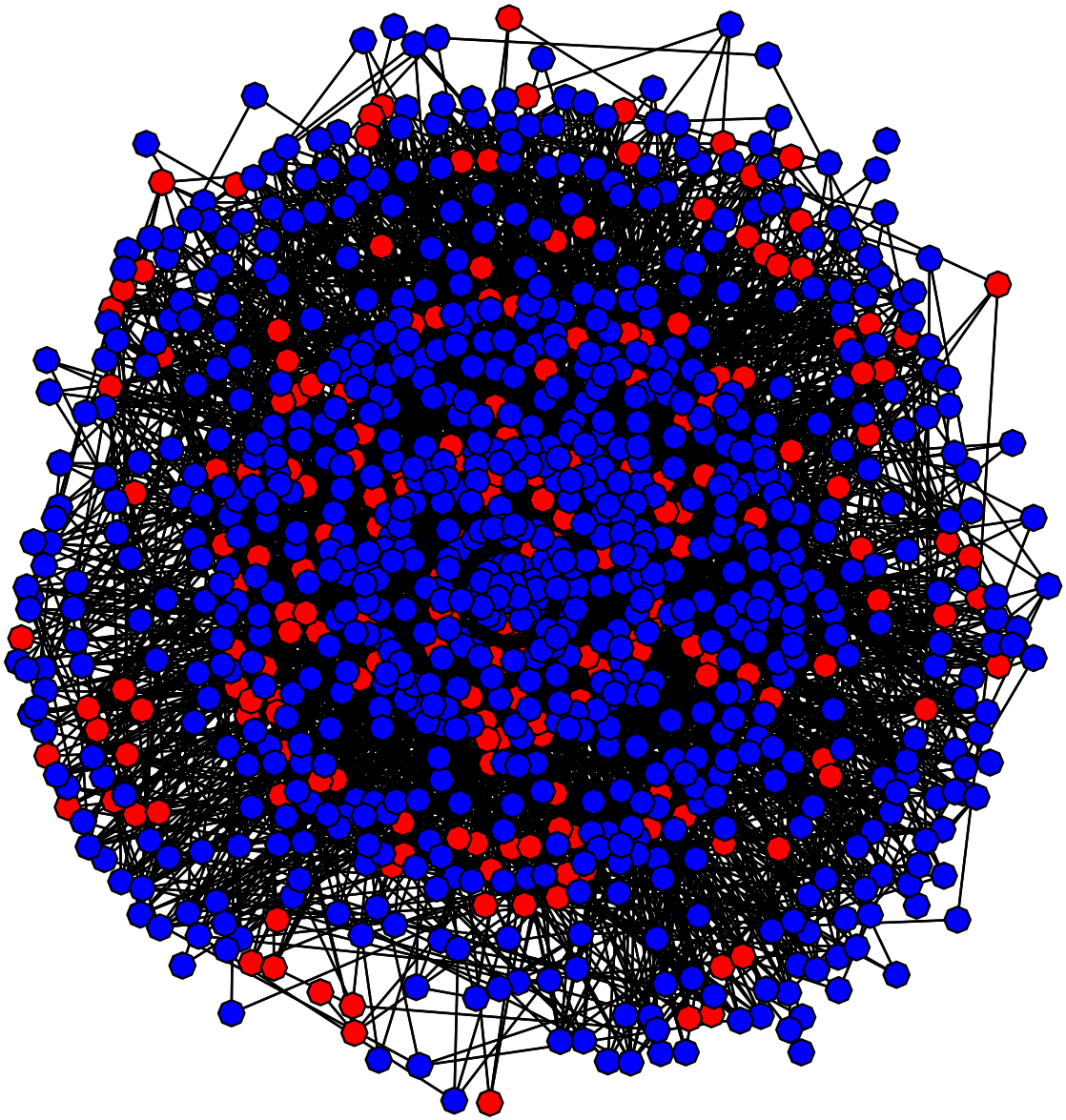
A peculiar case: Respondent-Driven Sampling

- *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
- *Privacy*: ***respondent-driven:*** respondents distribute uniquely identified coupons. no names.
- *Link-tracing*: none by researchers, done by respondents.
- *Estimation*: Challenging to get valid estimates
- Effective at obtaining large varied samples in many populations.
- Widely used: over 100 studies, in over 30 countries. Often HIV-risk populations.

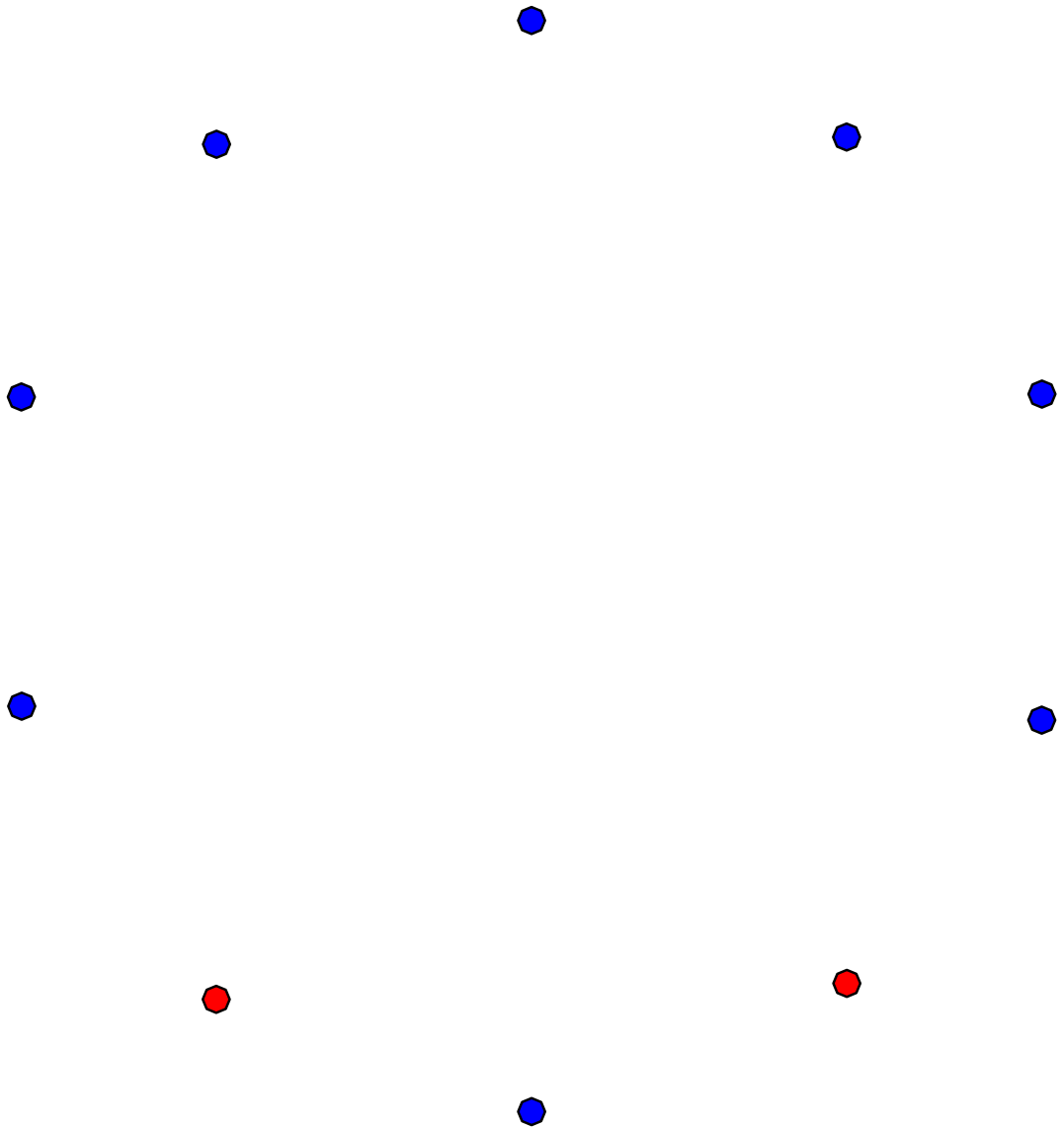
Heckathorn, D.D., "Respondent-driven sampling: A new approach to the study of hidden populations." Social Problems, 1997.

Handcock, M.S. and K.J. Gile, "On the Concept of Snowball Sampling." Sociological Methodology, 2011.

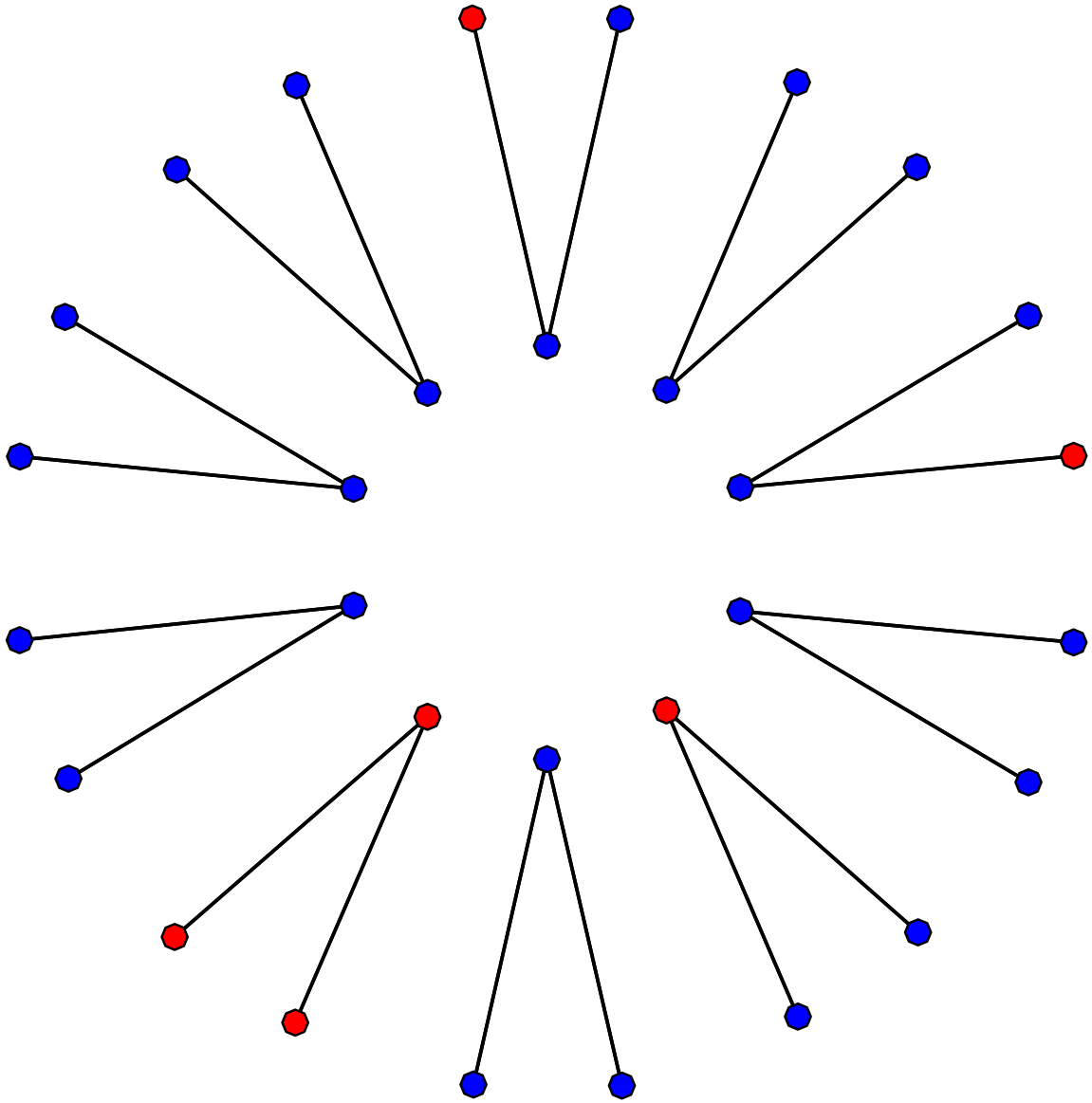
Stylized population



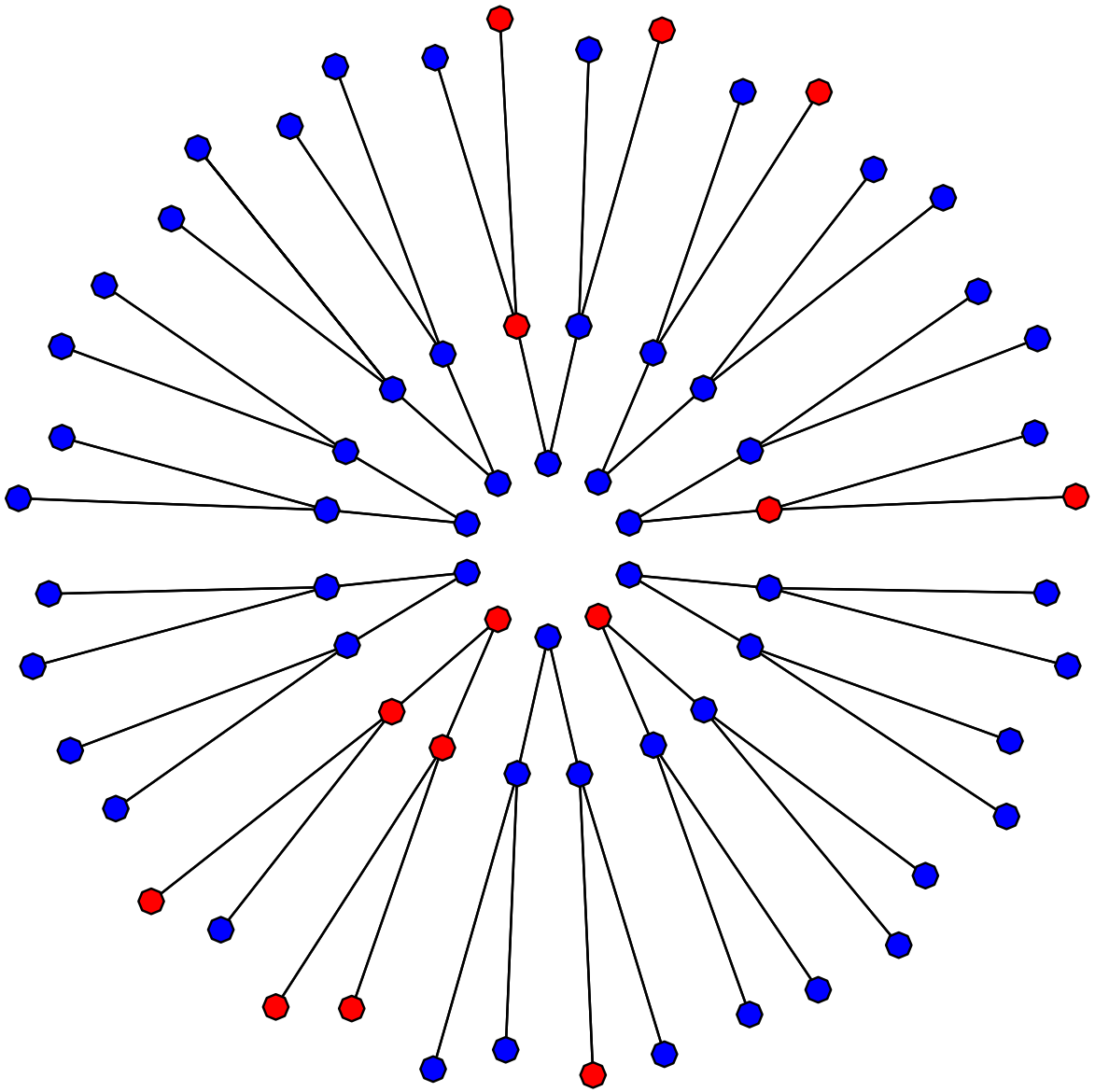
Start with seeds …



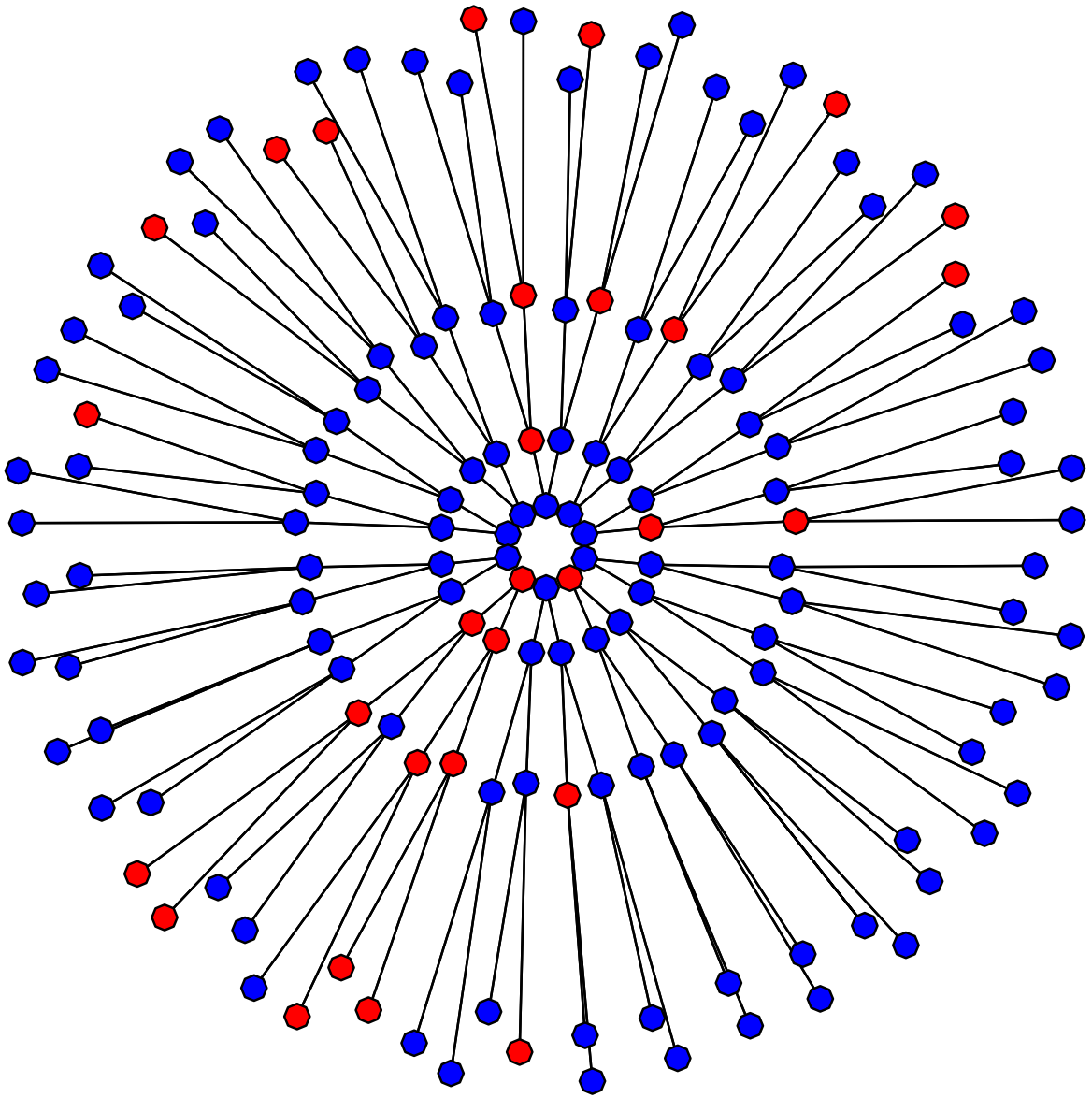
Seeds recruit the first wave …



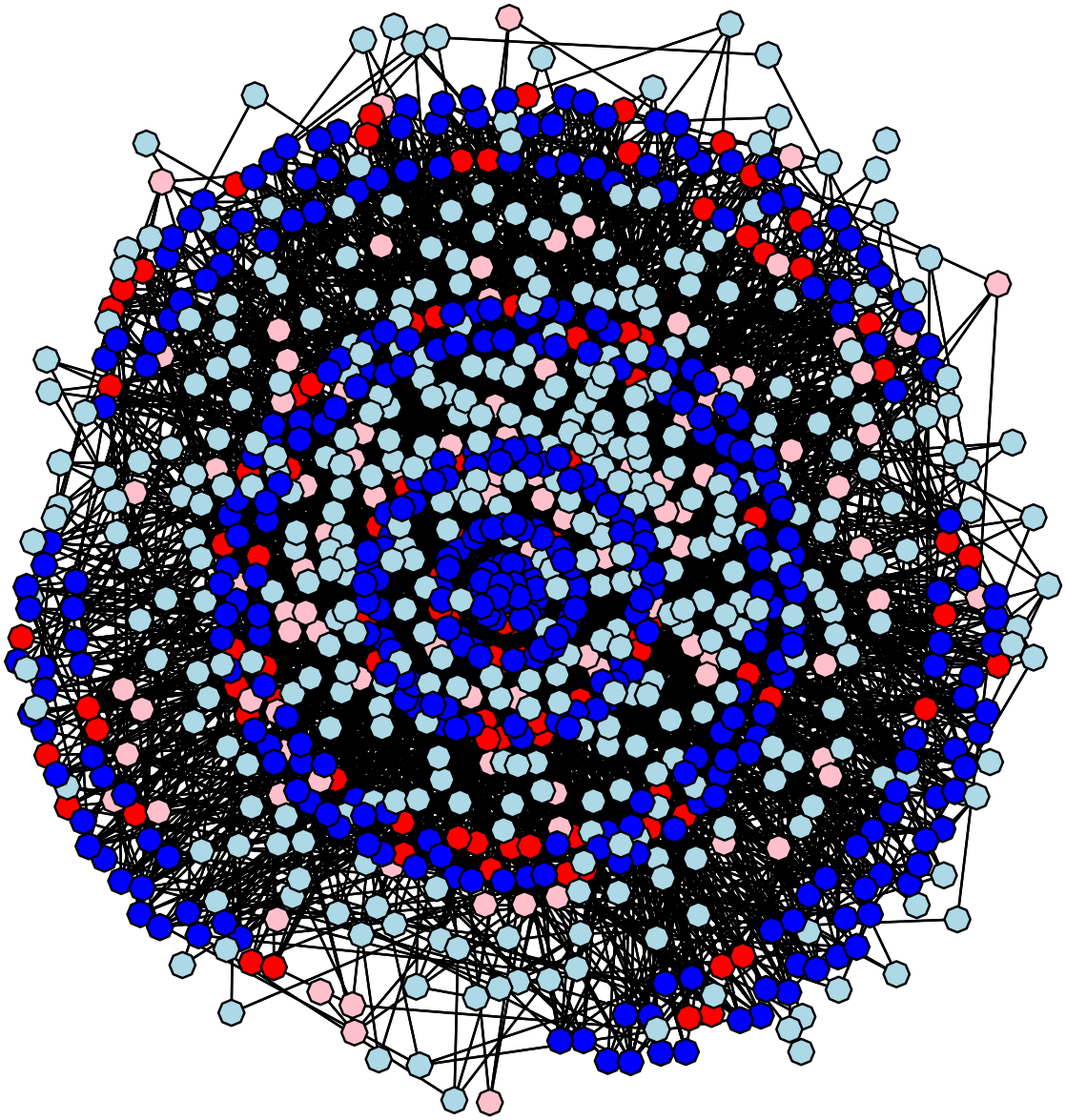
The first wave recruit the second wave …

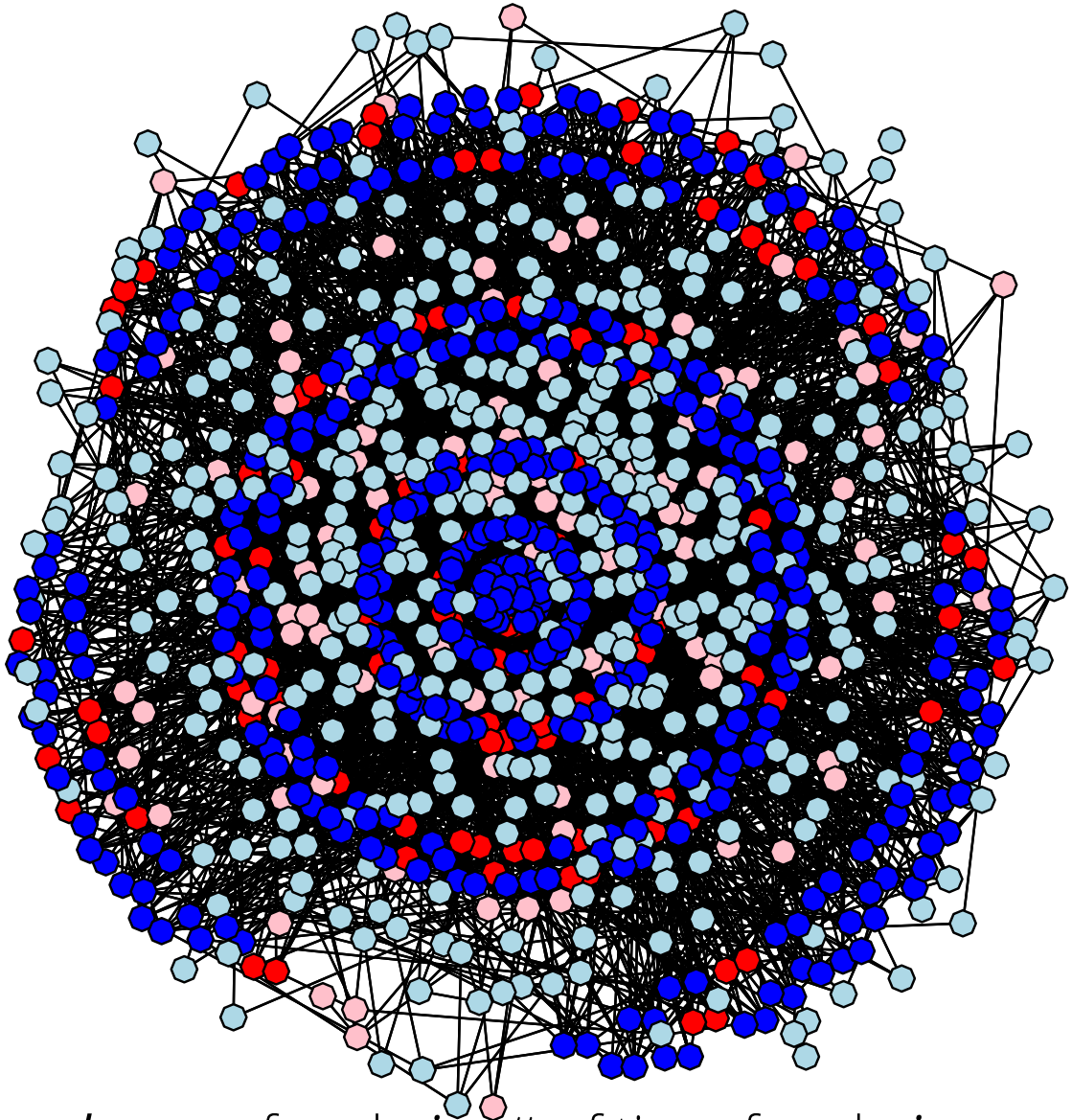


and so on …

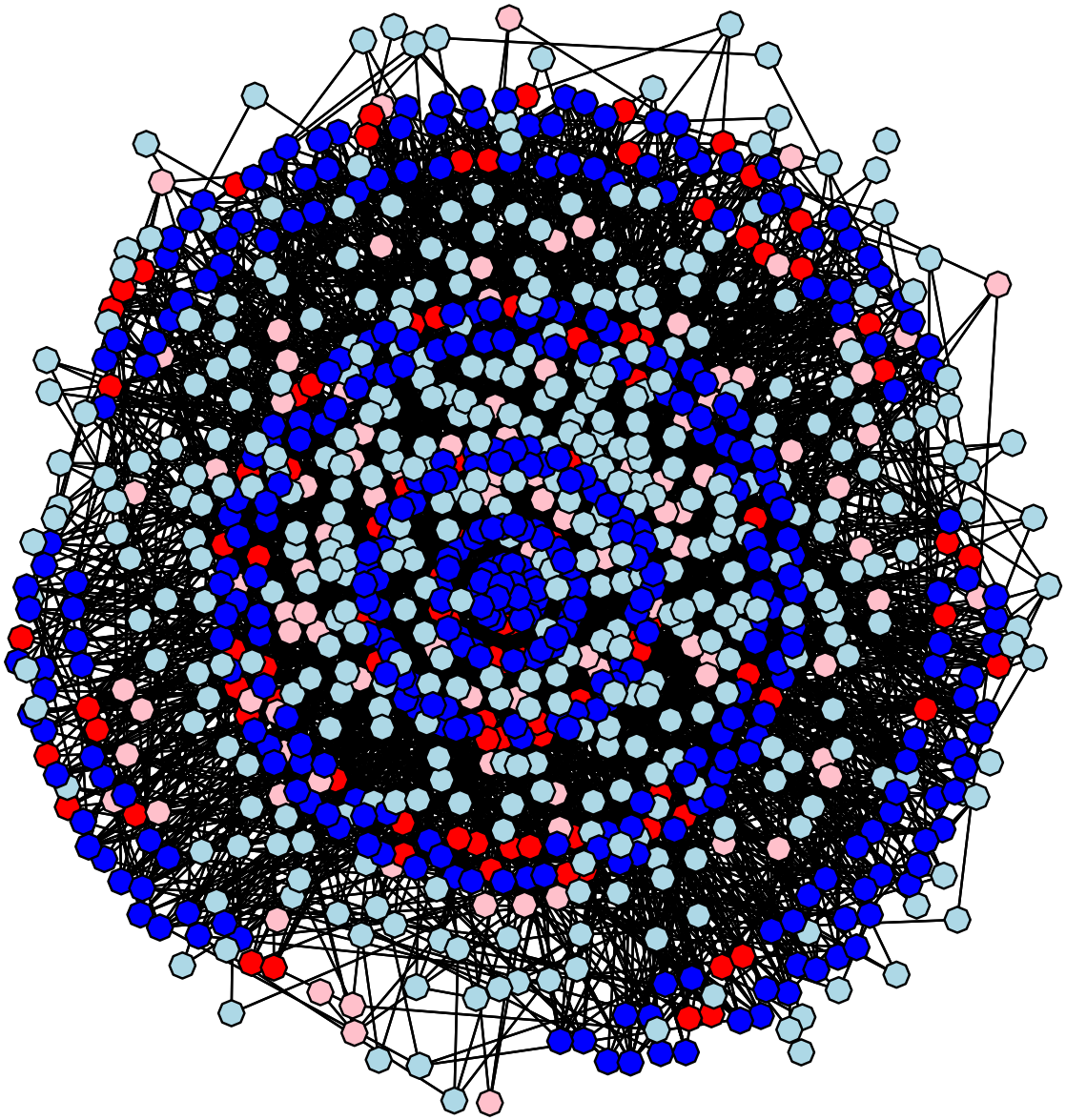


At the end (the unsampled are shaded)





degree of node ***i*** = # of ties of node ***i***



Classic Design-Based Inference: Generalized Horvitz-Thompson Estimators

- Goal: Estimate the population mean of y:

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

where

$$y_i = \begin{cases} 1 & i \text{ “positive"} \\ 0 & i \text{ “negative".} \end{cases}$$

- Hajek Estimator:

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

where

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

- The key point: Estimator requires $\pi_i = P(S_i = 1) \quad \forall\, i : S_i = 1$

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.

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

$$\hat{\mu}_{\mathrm{VH}} = \frac{\sum_i s_i \frac{y_i}{d_i}}{\sum_i s_i \frac{1}{d_i}}$$

where d_i = degree of node i, s_i sample indicator, y_i quantity of interest.

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

Design of a Simulation Study of the approximation

Simulate Populations

- Number of people: 1000, 835, 715, 625, 555, or 525 nodes
- 20% of the population are "positive"

Simulate Social Networks over those Populations

(from ERGM, using `ergm/statnet`)

- Mean degree 7
- Homophily on Positivity: $R = \frac{P(\text{positive to negative tie})}{P(\text{negative to positive tie})} = 5$ (or other)
- Differential Activity: $w = \frac{\text{mean degree for positives}}{\text{mean degree for negatives}} = 1$ (or other)

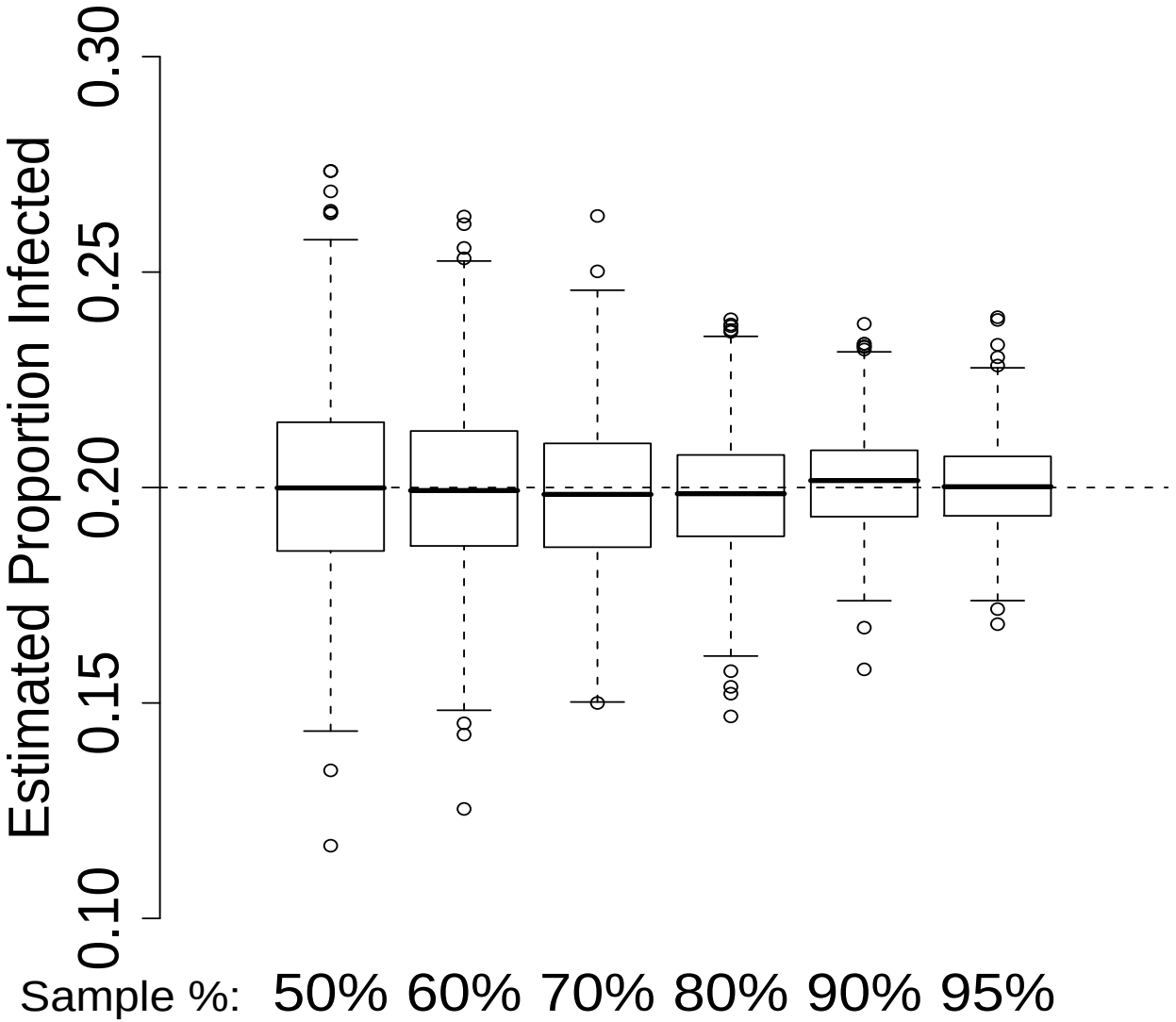
Simulate Respondent-Driven Samples from the Populations

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

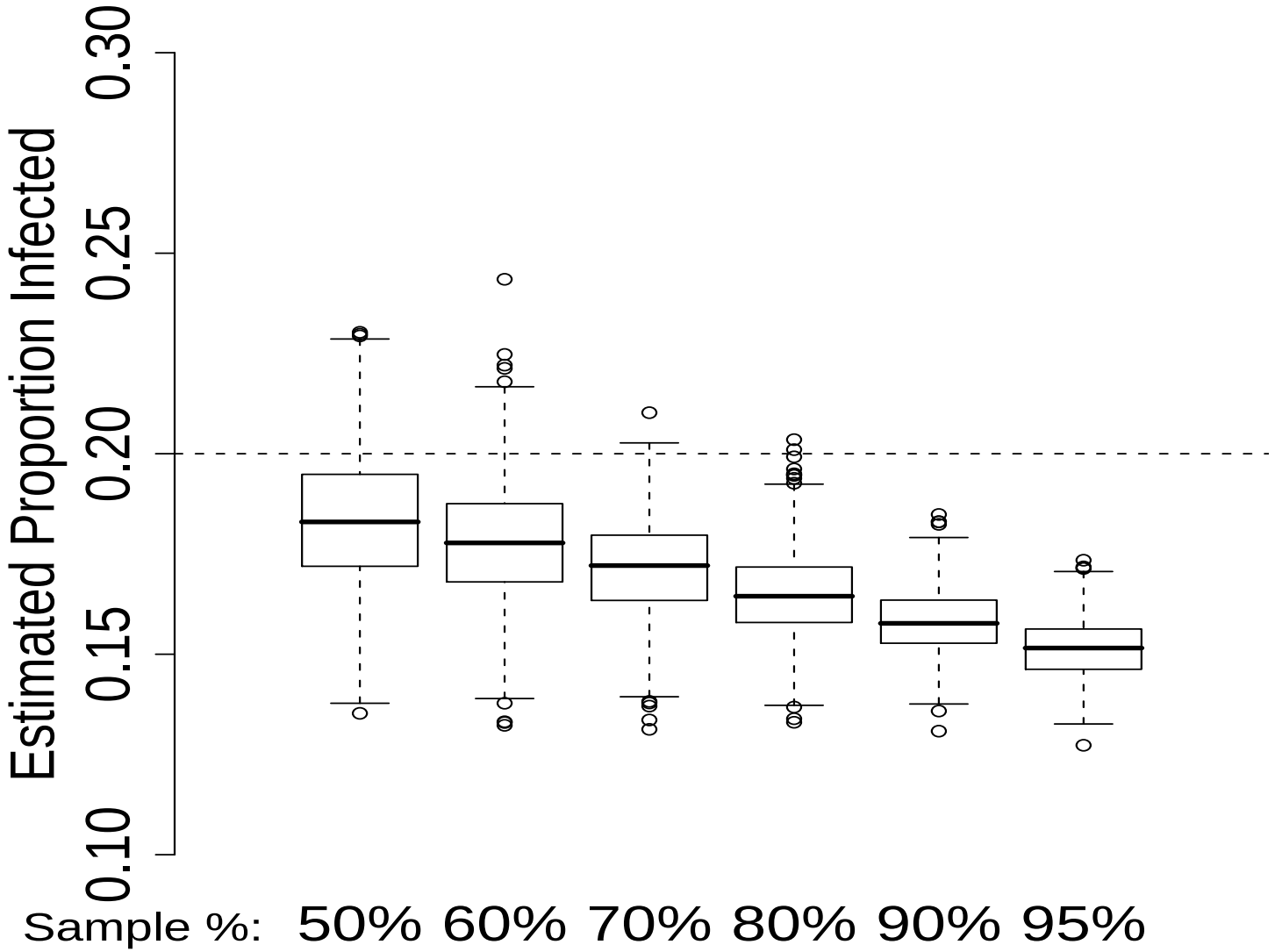
Repeat 1000 times

Blue parameters varied in study.

Volz-Heckathorn, when no differential activity (w=1)



When the positives are more active (w=1.5)

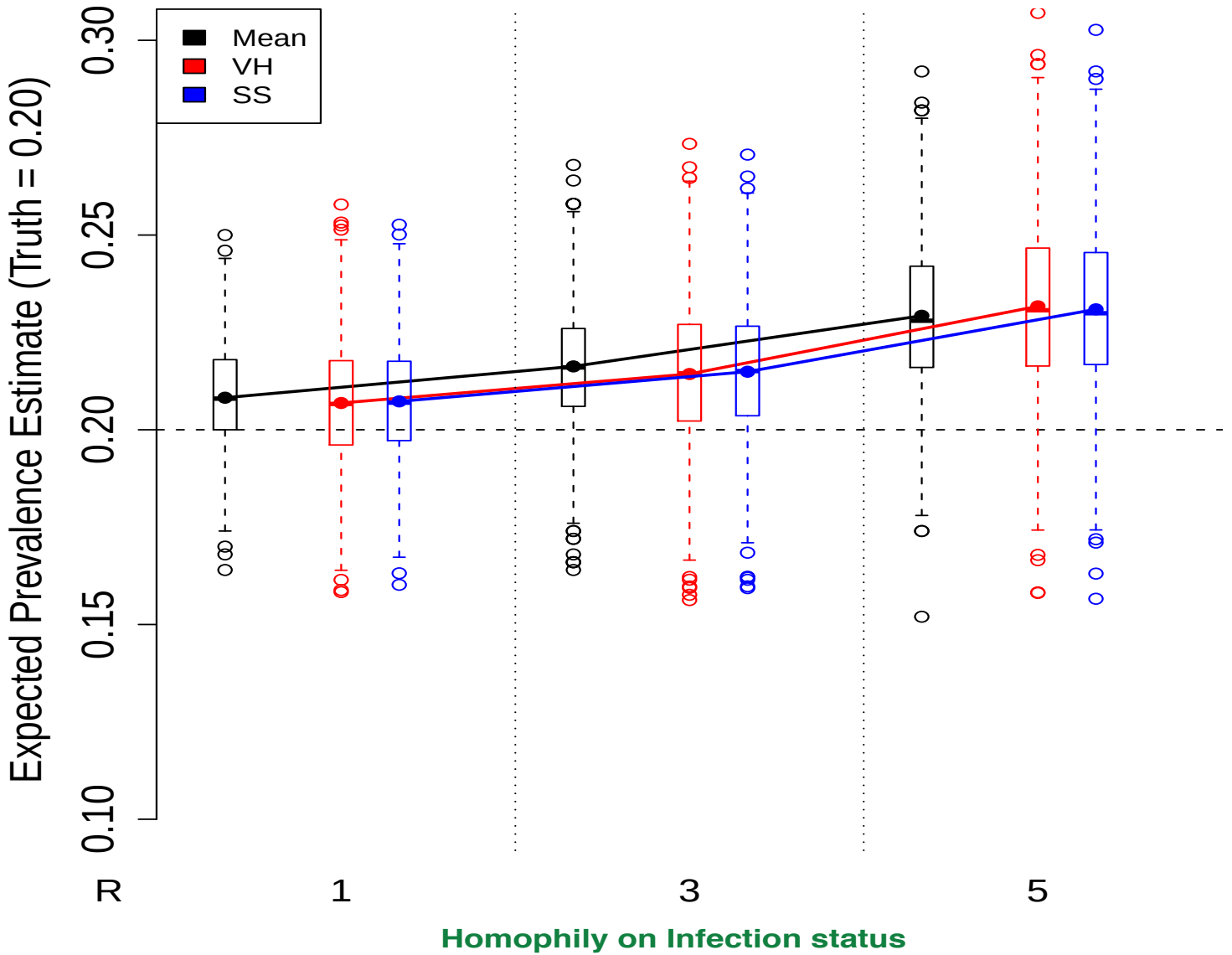


Improvements to RDS design-based estimators

The key is the modeling of the sampling process

- Salganik and Heckathorn (2004): simple Markov Chain model over classes. Struggles with Seed bias and finite population, good on homophily
- Volz and Heckathorn (2008): Markov Chain model over people. Seed bias, finite population, differential activity, homophily
- Gile (2008, 2011): Develops a model based on the successive sampling of people in time.
  Adjusts for without-replacement and finite population effects
- Fellows (2018) introduced the homophily configuration graph (HCG) estimator that has the good features of the SH and SS estimators.

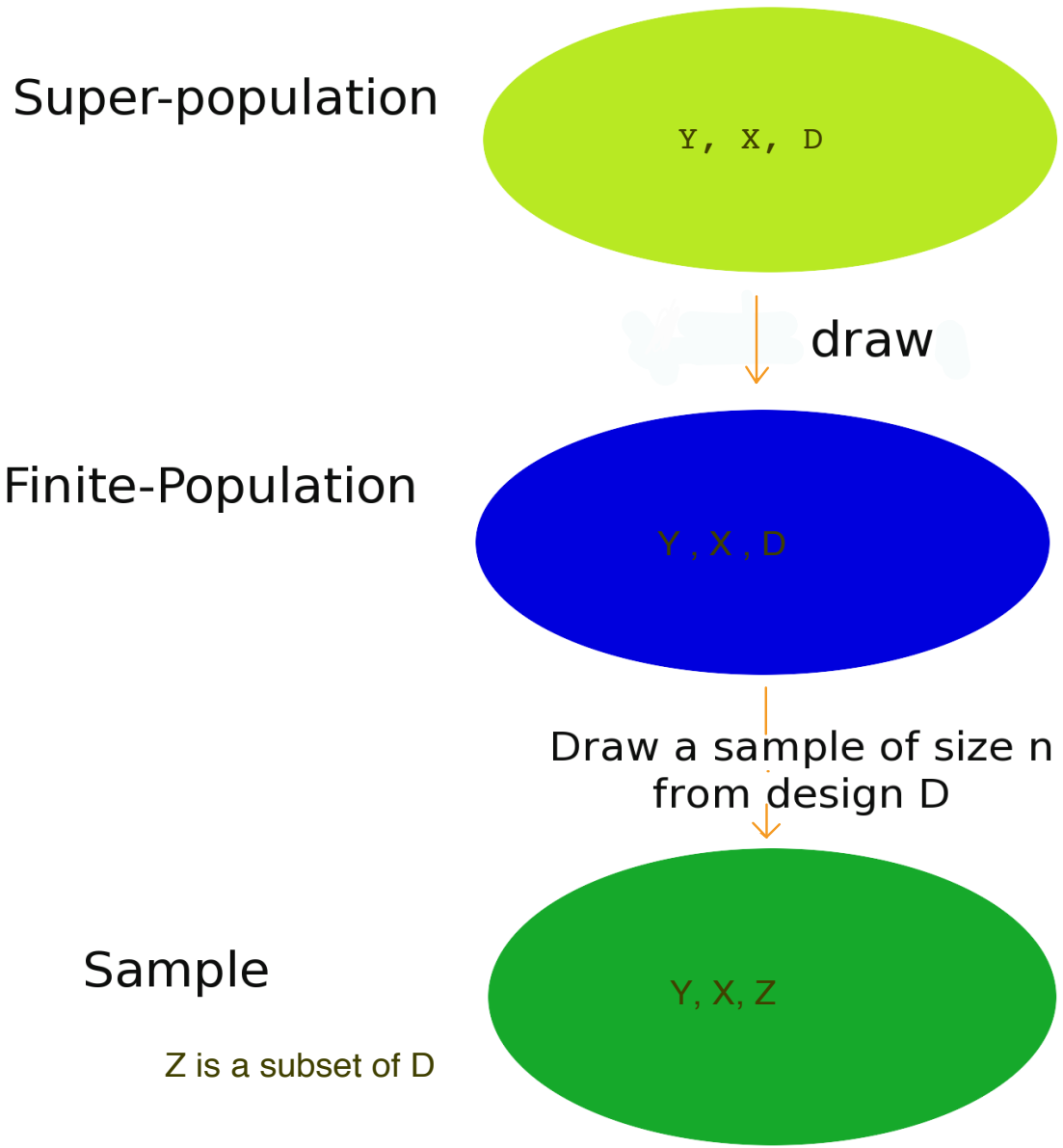
When all the seeds are positives, homophily causes bias



Doing better: Broader Perspectives

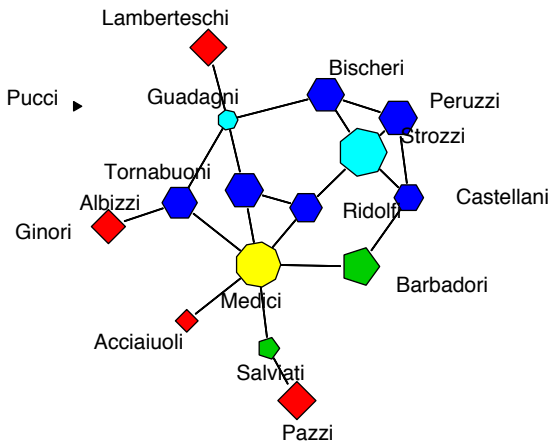
- Network-specific versus Population-process
  - Network-specific: interest focuses only on the actual network under study
  - Super-Population-process: the network is part of a population of networks, but interest is in the specific network
    - the network is conceptualized as a realization of a social process
  - Population-process: Interest focuses on the super-population process and the sampled network is thought of as data
- The choice of models depends on the objectives
  - The complexity of most network processes precludes complete modeling
  - We choose those aspects of the network we represent and model well

Conceptual Sampling Design Framework



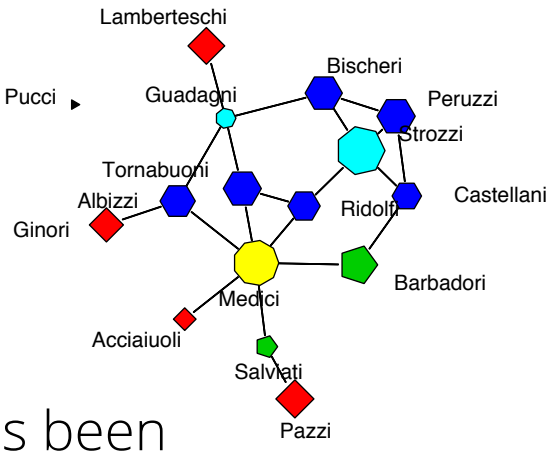
Network representations of the population

- Networks are widely used to represent data on relations between interacting entities or nodes.



Network representations of the population

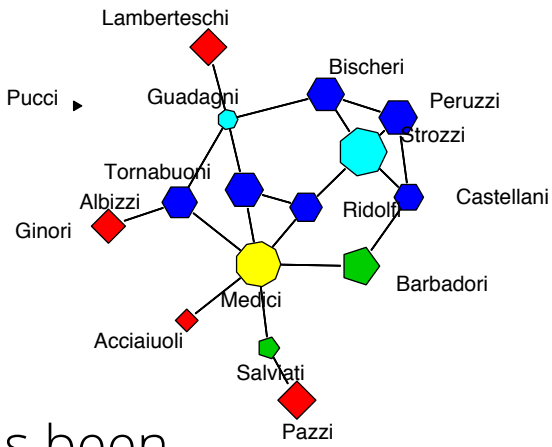
- Networks are widely used to represent data on relations between interacting entities or nodes.



- The study of social networks is multi-disciplinary
  - plethora of terminologies
  - varied objectives, multitude of frameworks
- Understanding the structure of social relations has been the focus of the social sciences
- Attempt to represent the structure in social relations via networks

Network representations of the population

- Networks are widely used to represent data on relations between interacting entities or nodes.



- The study of social networks is multi-disciplinary
  - plethora of terminologies
  - varied objectives, multitude of frameworks
- Understanding the structure of social relations has been the focus of the social sciences
- Attempt to represent the structure in social relations via networks
- *Network representations can be helpful in many settings*

Statistical Models for Social Networks?

Notation
A *social network* is defined as a set of n social "actors", a social relationship between each pair of actors, and a set of variables on those actors/pairs.

$$Y_{ij} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \\ 0 & \text{otherwise} \end{cases}$$

- call $Y \equiv [Y_{ij}]_{n \times n}$ a *graph*
  - a $N = n(n-1)$ binary array

Statistical Models for Social Networks?

Notation
A *social network* is defined as a set of n social "actors", a social relationship between each pair of actors, and a set of variables on those actors/pairs.

$$Y_{ij} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \\ 0 & \text{otherwise} \end{cases}$$

- call $Y \equiv [Y_{ij}]_{n \times n}$ a *graph*
  - a $N = n(n-1)$ binary array
- X be $n \times q$ matrix of actor and dyadic covariates
- call (Y, X) a *network*
- The basic problem of stochastic modeling is to specify a distribution for X, Y i.e., $P(Y = y, X = x)$

Features of Many Social Networks

- *Mutuality* of ties
- *Individual heterogeneity* in the propensity to form ties
- *Homophily* by actor attributes

  ⇒ Lazarsfeld and Merton, 1954; Freeman, 1996; McPherson et al., 2001

  - higher propensity to form ties between actors with similar attributes
    e.g., age, gender, geography, major, social-economic status
  - attributes may be observed or unobserved
- *Transitivity* of relationships
  - friends of friends have a higher propensity to be friends

The ERGM Framework for Graph Modeling

Let $\mathcal{Y}$ be the sample space of Y e.g. $\{0,1\}^N$.

- $g(y), y \in \mathcal{Y}$ d-vector of *graph statistics*
- represent graph features of interest (e.g., density, transitivity)
- desire $g(Y)$ to be jointly sufficient for the model

The ERGM Framework for Graph Modeling

Let $q(y)$ be a probability mass function over $\mathcal{Y}$.
Recall the maximum entropy motivation for exponential-families:

$$\underset{q}{\text{maximize}} \quad \sum_{y} q(y) \log(q(y))$$

subject to

$$E_q[g_i(Y)] = \mu_i, \qquad \forall\, i \in \{1, \ldots, d\}$$

Leads to:

$$P_\eta(Y = y) = \frac{\exp\{\eta \cdot g(y)\}}{c(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

$$E_\eta[g(Y)] = \mu$$

The ERGM Framework for Network Modeling

Let $\mathcal{Y}$ be the sample space of Y e.g. $\{0,1\}^N$
and $\mathcal{X}$ be the sample space of X.
Model the multivariate distribution of Y given X via:

$$P_\eta(Y = y|X = x) = \frac{\exp\{\eta \cdot g(y|x)\}}{c(\eta, x, \mathcal{Y})} \qquad y \in \mathcal{Y}, \;\; x \in \mathcal{X}$$

Frank and Strauss (1986)

- $\eta \in \Lambda \subset R^d$ d-vector of parameters
- $g(y|x)$ d-vector of *graph statistics*.
  $\Rightarrow$ $g(Y|x)$ are jointly sufficient for the model
- $c(\eta, x, \mathcal{Y})$ distribution normalizing constant

$$c(\eta, x, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp\{\eta \cdot g(y|x)\}$$

Fitting Models to Partially Observed Social Network Data

- Focus on the joint distribution of $Z = (Y, X)$.
- Types of data: Observed relations, nodal and dyadic variables ($z_{obs} = (y_{obs}, w_{obs})$), and indicators of relations and covariates
- $Z = (Z_{obs}, Z_{unobs})$

$$
\begin{aligned}
L(\boldsymbol{\eta}, \psi) &\equiv P(Z_{obs} = z_{obs}, D|\boldsymbol{\eta}, \psi) \\
&= \sum_{Z_{unobs}} P(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}, D|\boldsymbol{\eta}, \psi) \\
&= \sum_{Z_{unobs}} P(D|Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}, \psi) P_{\boldsymbol{\eta}}(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}) \\
&= \sum_{Z_{unobs}} P(D|Z, \psi) \times P_{\boldsymbol{\eta}}(Z = z)
\end{aligned}
$$

sampling design $\times$ network model

- $\boldsymbol{\eta}$ is the network model parameter ("super population")
- ψ is the sampling parameter

Adaptive Sampling Designs

- A sampling design adaptive if:

$$P(D = d|Z_{obs}, Z_{mis}, \psi) = P(D = d|Z_{obs}, \psi) \qquad \forall z \in \mathcal{Z}.$$

that is, it uses information collected during the survey to direct subsequent sampling, but the sampling design depends only on the observed data.

- adaptive sampling designs satisfy a "*missing at random*" condition from Rubin (1976) in the context of missing data.
- **Result:** standard network sampling designs such as conventional, adaptive web, and multi-wave link-tracing sampling designs are adaptive
  $\Rightarrow$ Thompson and Frank (2000), Handcock and Gile (2007).

When is sampling non-adaptive?

- Individual sample based on unobserved properties of non-respondents - like infection status or illicit activity.
- Link-tracing sample starting where links are followed dependent on unobserved properties of alters.

Adaptive Sampling Designs and their Amenable Models

Definition: Consider a sampling design governed by parameter $\psi \in \Psi$ and a stochastic network model $P_{\boldsymbol{\eta}}(Z = z)$ governed by parameter $\boldsymbol{\eta} \in \Xi$. We call the sampling design amenable to the model if the sampling design is adaptive and the parameters ψ and $\boldsymbol{\eta}$ are distinct.

Result: If the sampling design is amenable to the model the likelihood for $\boldsymbol{\eta}$ and ψ is

$$L[\boldsymbol{\eta}, \psi | Z_{obs} = z_{obs}, D = d] \propto L[\psi | D = d, Z_{obs} = z_{obs}] L[\boldsymbol{\eta} | Z_{obs} = z_{obs}]$$

sampling design likelihood $\times$ face-value likelihood

$$L[\psi | D = d, Z_{obs} = z_{obs}] = P(D | Z_{obs} = z_{obs}, \psi)$$

$$L[\boldsymbol{\eta} | Z_{obs} = z_{obs}] = \sum_{z_{unobs}} P_{\boldsymbol{\eta}}(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs})$$

Adaptive Sampling Designs and their Amenable Models

Result: If the sampling design is ***not*** amenable to the model the likelihood for $\boldsymbol{\eta}$ and ψ is

$$L(\boldsymbol{\eta}, \psi) = \sum_{z_{unobs}} P(D|Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs}, \psi) P_{\boldsymbol{\eta}}(Z_{obs} = z_{obs}, Z_{unobs} = z_{unobs})$$

and the design will need to be represented.

Clearly $P(D|Z, \psi)$ can be modeled when it is unknown.

Doing better: A Network Model-Based Estimator

- Fit a network model to observed data (ERGM, using **`statnet`** R package)
- Estimate sampling probabilities based on network model, and weight sample appropriately
- Can estimate conditional on seed selection, to reduce bias induced by seed selection.

Network Model-Assisted Estimator

- Interested in sampling probabilities $\pi_i = P_y(S_i = 1)$.
- Should reflect:
  - Nodal degree d_i
  - Sample fraction
  - Seed selection
  - Homophily and branching structure of sampling
- This is very difficult to do without known the underlying social network y
- So we develop a "super-population" representation for y with the purpose of "assisting" the design-based inference

Network Model-Assisted Estimator

- Approach: Retain design-based framework, but estimate the unknown finite-population sampling probabilities $\pi_i(y) = \mathbf{E}(S_i|Y = y)$.

Idea:

1. For given network y, can compute

$$\pi_i(y) = \mathbf{E}(S_i|Y = y)$$

2. Estimate π_i via

$$\hat{\pi}_i = \sum_{y_{unobs}} \pi_i(y) P_\eta(Y = y|Y_{obs} = y_{obs})$$

3. We do not know η, so we estimate it from the data.

Example: A simple model for sociality in a network

Let Y be an undirected network ($y_{ij} = y_{ji}$). The β model is:

$$P_\eta(Y = y) = \frac{\exp\{\sum_i \eta_i y_{i+}\}}{\kappa(\eta)}$$

where $y_{i+} = \sum_j y_{ij}$.

– η_i *sociality* of node i
– $\{y_{1+}, y_{2+}, \ldots, y_{n+}\}$ is referred to as the ***degree sequence*** of the network.

Extensive development of conditional models

- Classes of $g(y|x)$ (Generative Theory, Structural signatures)
- Inference on the loglikelihood function,

$$\ell(\eta|y_{\text{obs}}; x_{\text{obs}}) = \eta \cdot g(y_{\text{obs}}|x_{\text{obs}}) - \log c(\eta|x_{\text{obs}})$$

$$c(\eta|x_{\text{obs}}) = \sum_{z \in \mathcal{Y}} \exp\{\eta \cdot g(z|x_{\text{obs}})\}$$

- For computational reasons, approximate the likelihood via Markov Chain Monte Carlo (MCMC)

Returning to RDS: Fitting the Model

Use networks statistics $g(y)$:

- $\{y_{1+}, y_{2+}, \ldots, y_{n+}\}$, the ***degree sequence*** of the network.
- the number of ties between those positive and those negative.

Problem: Requires (unknown) networked population statistics $g(y)$.
Solution: Use design-based estimators

$$\hat{g}(\eta) = \sum_{i=1}^{N} \frac{\mathbf{s}_i \tilde{g}(y_{obs})}{\hat{\pi}_i}$$

where $\tilde{g}(y_{obs})$ are corresponding sample statistics.

Problem: This, in turn, requires sampling probabilities.
Solution: Novel iterative algorithm to find self-consistent solution.

Model-Assisted Estimator: Algorithm

- Goal: Estimate sampling probabilities (π_i).
- A function of homophily (η), and population of degrees and infection $\mathbf{N}$.
- Initiate via $\hat{\pi}_i$ estimated by simple rule.
- Iterate the following steps:
  - Estimate $\hat{\boldsymbol{g}}(\eta)$ using $\hat{\pi}_i$.
  - Find corresponding model parameter η (`ergm` R package)
  - Simulate M networks, and samples from networks. Estimate $\hat{\pi}_i$ by simulation.
- Use the resulting estimated probabilities, $\hat{\pi}_i$, to form weighted estimator.

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

Standard Error Estimation

Population Bootstrap:

- Simulate *M* populations
  - Conditional on z_{obs}
  - Use model parameter η
- Simulate adaptive network samples over each population
- Compute MA estimates. Average estimates over M populations
- Results:
  - Performs well when statistics are homophily, degree distribution and differential activity (*w*)
  - Computationally expensive

Krista J. Gile and Mark S. Handcock, "Network Model-Assisted Inference from Respondent-Driven Sampling Data." JRSS, A, 178, 3, 619-639, 2015.

Simulation Study
comparing design-based to model-assisted estimators

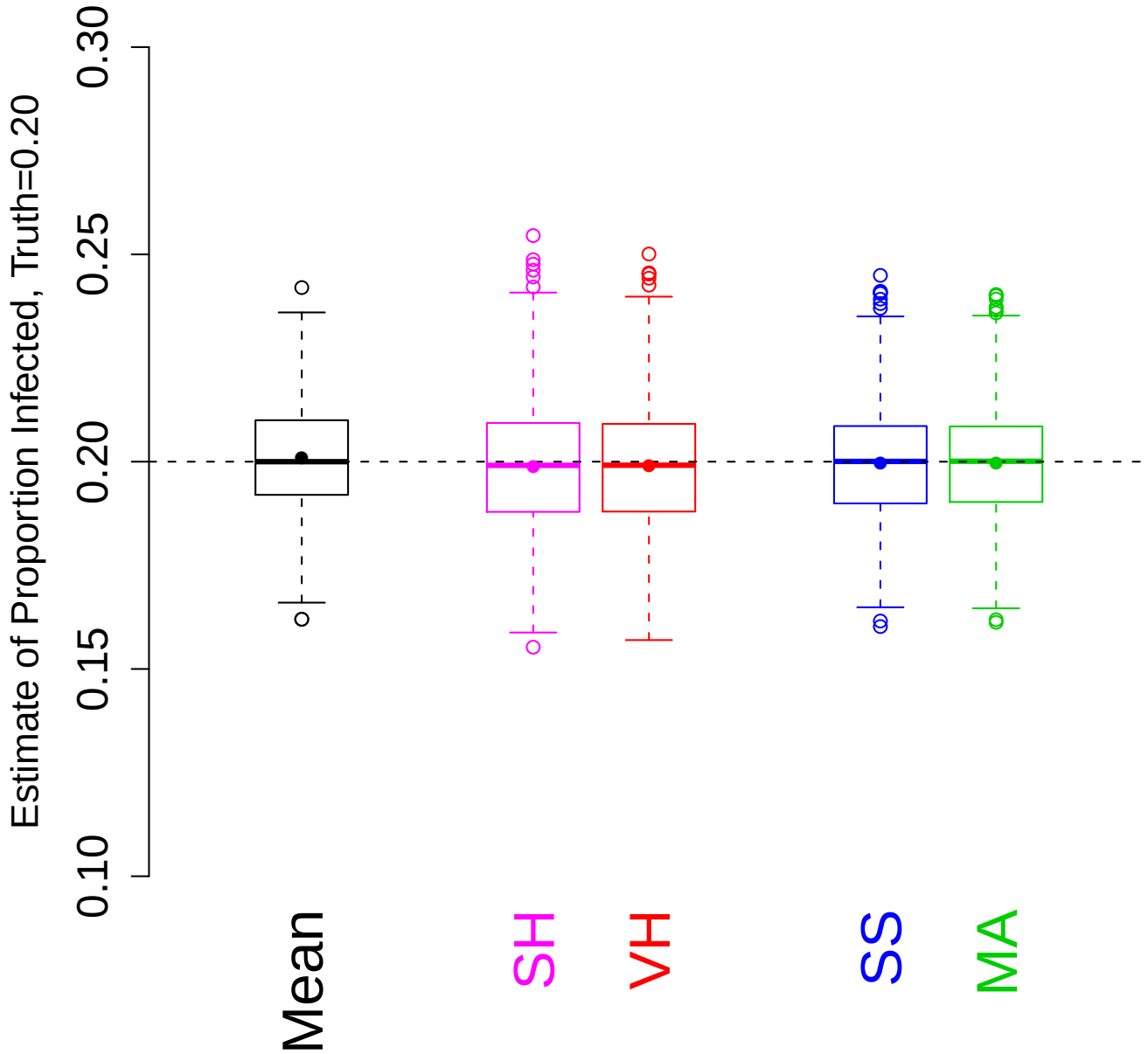
Critical Questions:

- Does Model-Assisted estimator perform well for differential activity and large sample fraction?
- Does Model-Assisted estimator correct for seed bias?
- What about unknown population size and network structure?

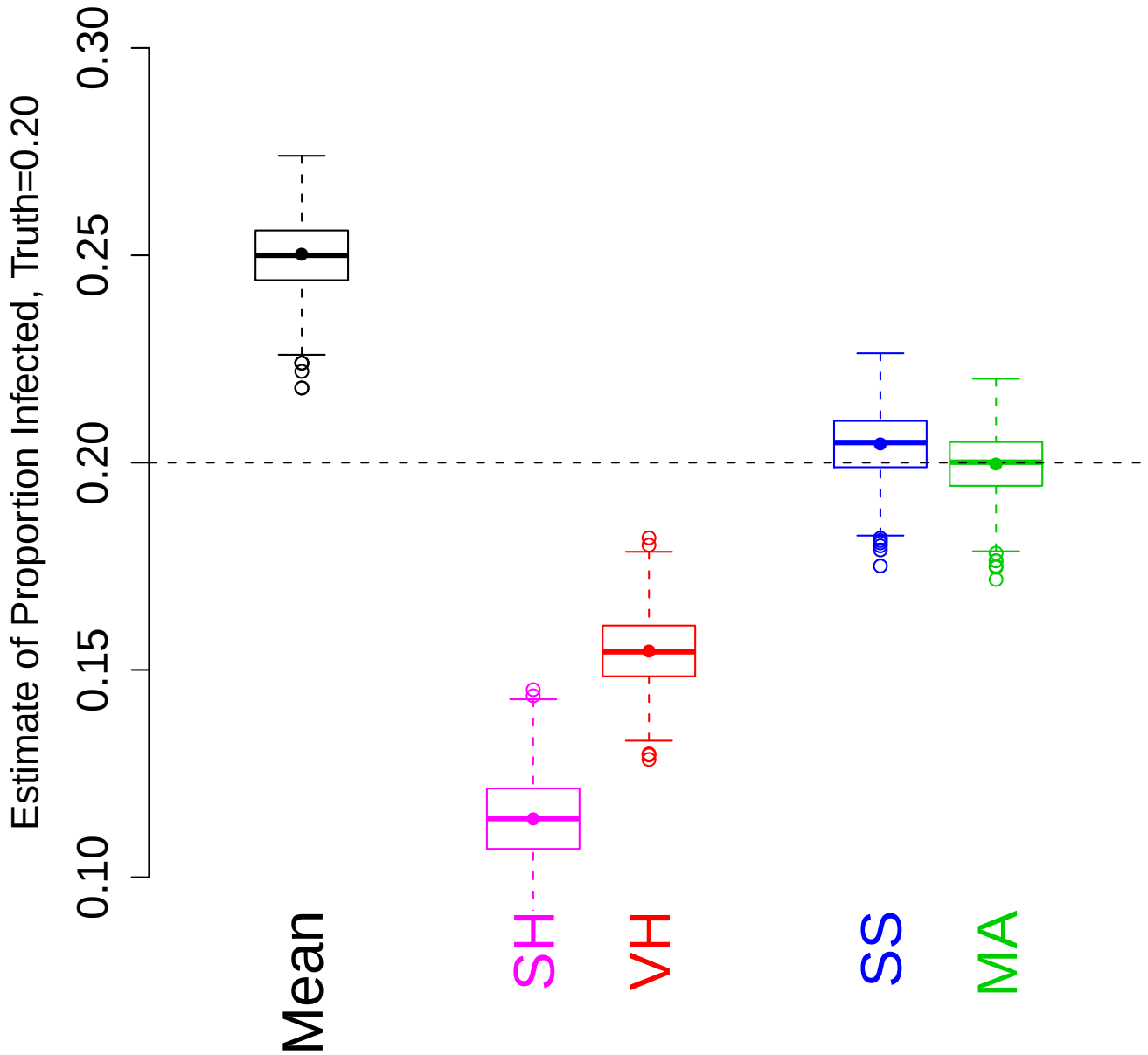
Comparison with design-based estimators:

- Mean: Naive Sample Mean
- SH: Salganik-Heckathorn: based on MME of number of cross-relations
- VH: Volz-Heckathorn Estimator
- SS: Gile's sequential sampling (SS) estimator
- MA: Network Model-Assisted Estimator

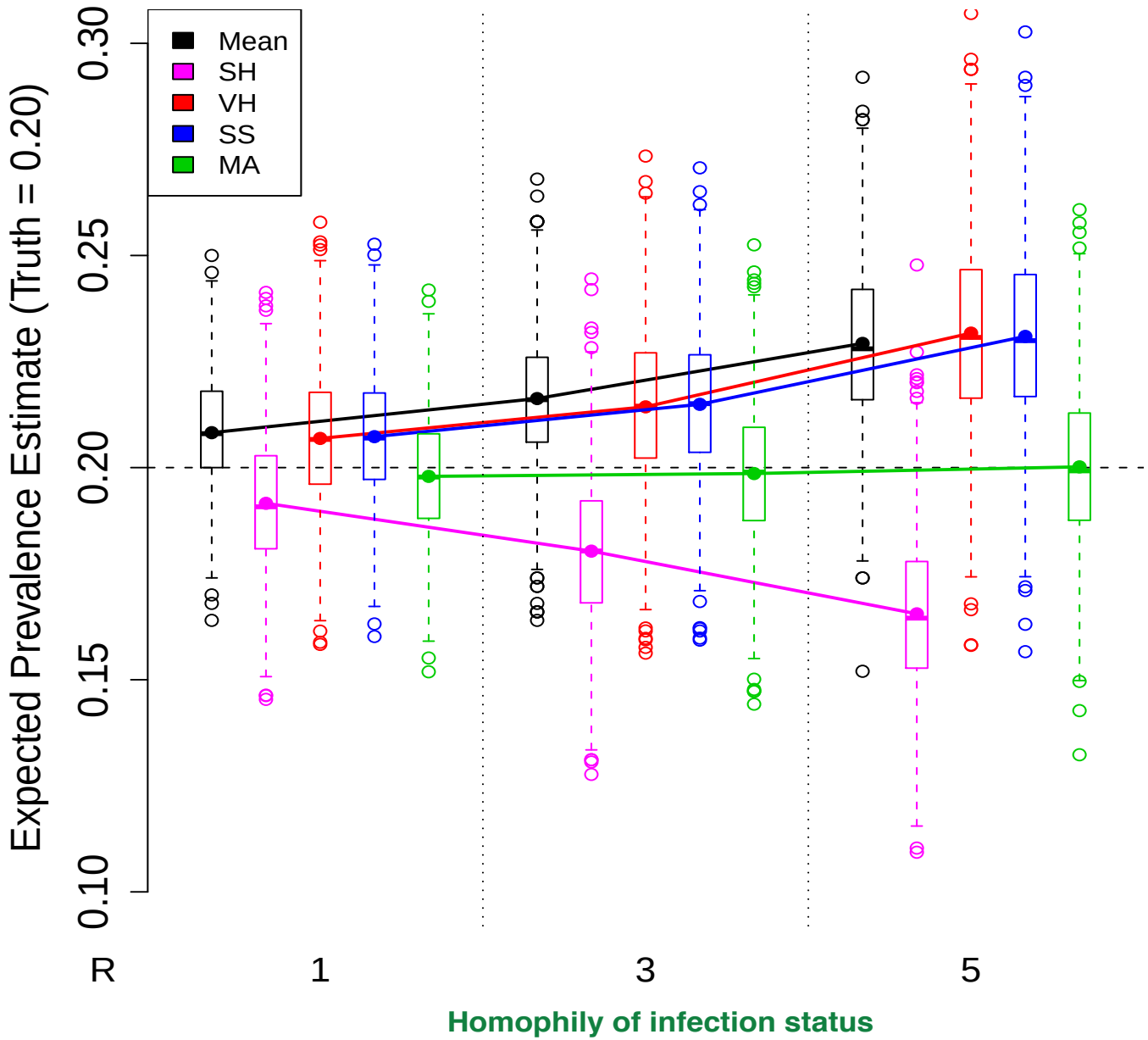
For a 50% Sample, No homophily, Random Seeds



For a 70% Sample, positives more active and homophily, Positive Seeds



When all Seeds are positive, varying Homophily



More and deeper studies

Journal of Survey Statistics and Methodology (2018) ***6****, 23–45*

EVALUATING VARIANCE ESTIMATORS FOR RESPONDENT-DRIVEN SAMPLING

MICHAEL W. SPILLER*
KRISTA J. GILE
MARK S. HANDCOCK
CORINNE M. MAR
CYPRIAN WEJNERT

Spiller *et. al* 2018

- The first systematic evaluation of the different RDS variance estimators
- Evaluation based on statistical performance on realistic but simulated populations
- Compare over simulated populations close to those of interest to the CDC
- Based on the CDC's National HIV Behavioral Surveillance system (NHBS)
  - NHBS sampled persons-who-inject-drugs (PWID) in 20 U.S. cities in both 2009 and 2012
  - Different surveys for heterosexuals, MSM and PWID, ongoing, 5 rounds since 2003.
  - a standardized protocol is used
  - 40 populations are simulated using information from the $2 \times 20 = 40$
- Primary focus is on estimates of confidence intervals (i.e., *coverage*)
- Confidence intervals are the primary RDS estimates

Confidence Interval Estimation

- The key is the modeling of the sampling process, as we have dependent data.
- Most successful methods use a form of *bootstrapping* the recruitment chains.
  - Salganik bootstrap: resamples dyads from the infection mixing matrix
  - Volz and Heckathorn estimator: Uses the Salganik bootstrap
  - Gile's SS: builds a population and resamples that using SS.

Gile's SS bootstrap:

Population Bootstrap:

- Simulate the networked population
  - Estimate the infection status by degree distribution
  - Estimate infection mixing matrix by infection status
- Simulate without-replacement sampling
  - Choose recruit infection status according to mixing matrix
  - Choose recruit degree by successive sampling
  - Update available population and mixing matrix
- Compute SS Estimates
- Results:
  - Performs well across differential activity and sample fraction
  - Performs well with homophily
  - Unreliable when seeds biased.

Outline of the study:

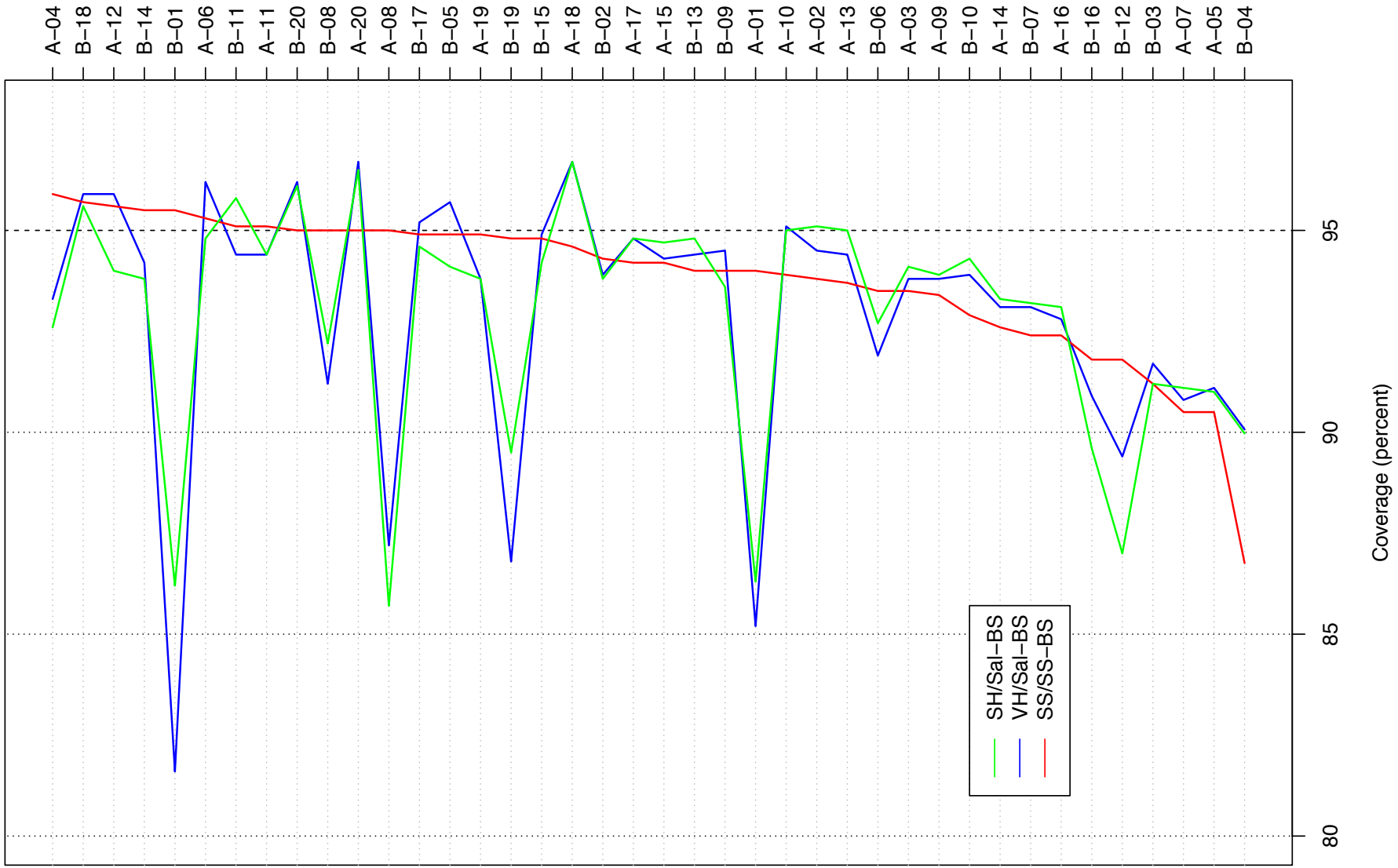
The first systematic evaluation of the different RDS CI estimators

- Compare VH, SH and SS bootstrap sampling procedures
- Compare two different CI computation procedures
  - percentile: Use quantiles of the bootstrap sample
  - studentized: Use the standard deviation of the bootstrap samples
- Compare over simulated populations close to those of interest to the CDC
  - The CDC's National HIV Behavioral Surveillance system (NHBS)
  - NHBS sampled persons-who-inject-drugs (PWID) in 20 U.S. cities in both 2009 and 2012
  - a standardized protocol is used
  - 40 populations are simulated using information from the $2 \times 20 = 40$ surveys.

Details of the simulated populations:

- four characteristics matched:
  - infection prevalence
  - network homophily on infection status
  - mean degree of population members
  - differential activity (DA) of the population by infection status
- For each population, simulated 1000 networks using ERGM with a population size of 10000 PWID.
- For each simulated population, approximated the RDS using the known design characteristics
- Estimated the various CI for each simulated RDS and compared to known truth.

The coverage rate of the CI estimators:



The coverage rate of the CI estimators:

Table 2: 95% Confidence Interval (CI) coverage rate for four RDS point and variance estimator pairs by bootstrap CI method

Point Estimator	Variance Estimator	Boostrap CI Method	Mean	Standard Deviation	Median	Range	Percentage of CIs with coverage between 93 and 97 inclusive*
Sample mean	SRS variance	N/A	67.4	23.8	74.9	[14, 96]	5.0
Salganik- Heckathorn	Salganik	Percentile	87.0	12.8	91.9	[41, 96]	40.0
Salganik- Heckathorn	Salganik	Studentized bootstrap	93.0	2.8	93.9	[86, 97]	67.5
Volz-Heckathorn	Salganik	Percentile	87.0	12.8	91.8	[41, 96]	42.5
Volz-Heckathorn	Salganik	Studentized bootstrap	92.9	3.2	93.9	[82, 97]	67.5
Successive Sampling	Successive Sampling	Percentile	94.1	1.8	94.6	[87, 97]	80.0
Successive Sampling	Successive Sampling	Studentized bootstrap	93.8	1.8	94.2	[87, 96]	75.0

* The percentage of CIs with coverage between 93% and 97% is presented as a summary measure of the percentage of CIs with acceptable coverage rates for a given estimator pair and bootstrap CI method.

Studies based on with and with-out replacement methods

- We compared RDS simulated when:
  - respondents are able to be re-sampled / re-surveyed "*with replacement*"
  - respondents are surveyed at most once "*with-out replacement*"
- Estimators based on with-out replacement outperformed those the used with replacement
- Studies based on with replacement give unrealistic and erroneous results

Estimation of Design Effects

- The *Design Effect*: a measure of the variability of an estimator relative to a SRS from the population.

$$\text{Design Effect} = \frac{\text{Variance of the RDS estimator}}{\text{Variance of the RDS estimator from a hypothetical SRS with the same sample size}}$$

- Typical design effects for (non-RDS) complex surveys RDS are between 1.5 and 2
- Some prior studies claim design effects in RDS surveys are much larger
- Our study shows RDS design effects are in the range of other complex survey designs.

The design-effects of the sampling:

Table 3: Design effects for four RDS point estimators for 40 sets of RDS simulations

Point Estimator (sampling method)	Range	Median	Mean	Standard Deviation
Sample mean (without replacement)	[0.71, 2.46]	1.28	1.35	0.46
Salganik-Heckathorn (without replacement)	[0.79, 90.35]	1.61	7.1	18.73
Volz-Heckathorn (without replacement)	[0.77, 5.76]	1.59	1.81	0.88
Successive Sampling (without replacement)	[0.78, 5.61]	1.56	1.79	0.86
Volz-Heckathorn (with replacement)*	[1.01, 7.97]	2.34	2.77	1.46

* Point estimator and sampling method used in Goel and Salganik 2010

Estimation of Effective Sample Size

- A much better measure for RDS surveys is the *effective sample size*:

$$\text{effective sample size} = \frac{\text{sample size}}{\text{Design Effect}}$$

  This is the number of observations in a comparable SRS.
- For example, an effective sample size of 50 is small, even in $n = 100$.

Conclusions: Methodological

- Coverage of nominally 95% RDS CI are usually above 90%
- Suggests that accurate RDS CI estimation is feasible
- The SS/SS-BS combination performed the best
  - the SH and VH CI estimators are poor when differential activity is low and homophily is high
- Fellows (2018) introduced the homophily configuration graph (HCG) estimator. It is model assisted and based on the homophily configuration graph model. It has the good features of the SH and SS estimators. It appears to be the best estimator.
- Improvements for cases with extreme low prevalence possible using alternative CI types
  - the combined Agresti-Coull and the bootstrap-t interval of Mantalos and Zografos (2008)
- Both CI coverage rates and design-effects are often acceptable but not perfect.

Conclusions: Broader

- Focus on the *effective sample size* rather than *design effect* or *sample size*.
- CI are a lower bound on the actual uncertainty
- The studies suggest that they are close to the actual uncertainty if the sampling is executed well.
- The availability of estimates and sensitivity methods in *user-friendly software* is an essential research contribution

Strengths and weaknesses of Respondent-Driven Sampling

Strengths:

- Effective at obtaining large varied samples in many populations.
- Can be used in situations where a sampling frame does not exist.
- Unlike other link-tracing methods, does not require initial probability sample.
- Widely used: over 150 studies, in over 30 countries. Often populations at high risk for HIV.

Weaknesses:

- Still subject to many assumptions, especially data quality
- The degree to which it can be considered a probability sample depends on the quality of the implementation and network characteristics.
- Requires case-specific sensitivity analysis to justify its validity.

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