Respondent-Driven Sampling Estimation and the National HIV Behavioral Surveillance System

Michael "Trey" Spiller

Division of HIV/AIDS Prevention

Centers for Disease Control and Prevention

Atlanta, GA, United States

NSF-Census Research Network Meeting

September 11, 2014

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

The Challenge

- Sample “hard-to-reach” or “hidden” populations
  - Rare
  - Actively hide membership
- Data needed
  - Public health monitoring
  - Populations relevant for other statistical agencies
    - Homeless, undocumented residents
- Standard methods will not work
  - No sampling frame
  - Difficult to contact population members

CDC's National HIV Behavioral Surveillance (NHBS) System

- Monitor HIV risk and prevention behaviors and HIV prevalence
- Ongoing data collection began in 2003
- Cities with high AIDS burden
- Standard protocol
- NHBS conducted among:
  - Men who have sex with men (MSM)
  - Injection drug users (IDU)
  - Heterosexuals at increased risk of HIV infection
- Annual rotating cycles

MSM	IDU	HET	MSM	IDU	HET	MSM	IDU	HET
2003-04	2005	2006-07	2008	2009	2010	2011	2012	2013

Respondent-Driven Sampling (RDS)

- Link-tracing sampling method
  - Modifications to standard link-tracing approaches
- Used in hundreds of studies since 1997, including surveys of populations most at risk for acquiring HIV
- NHBS uses RDS to produce estimates of:
  - HIV infection
  - Sharing syringes
  - Condomless sex
  - Other topics related to HIV risk and prevention

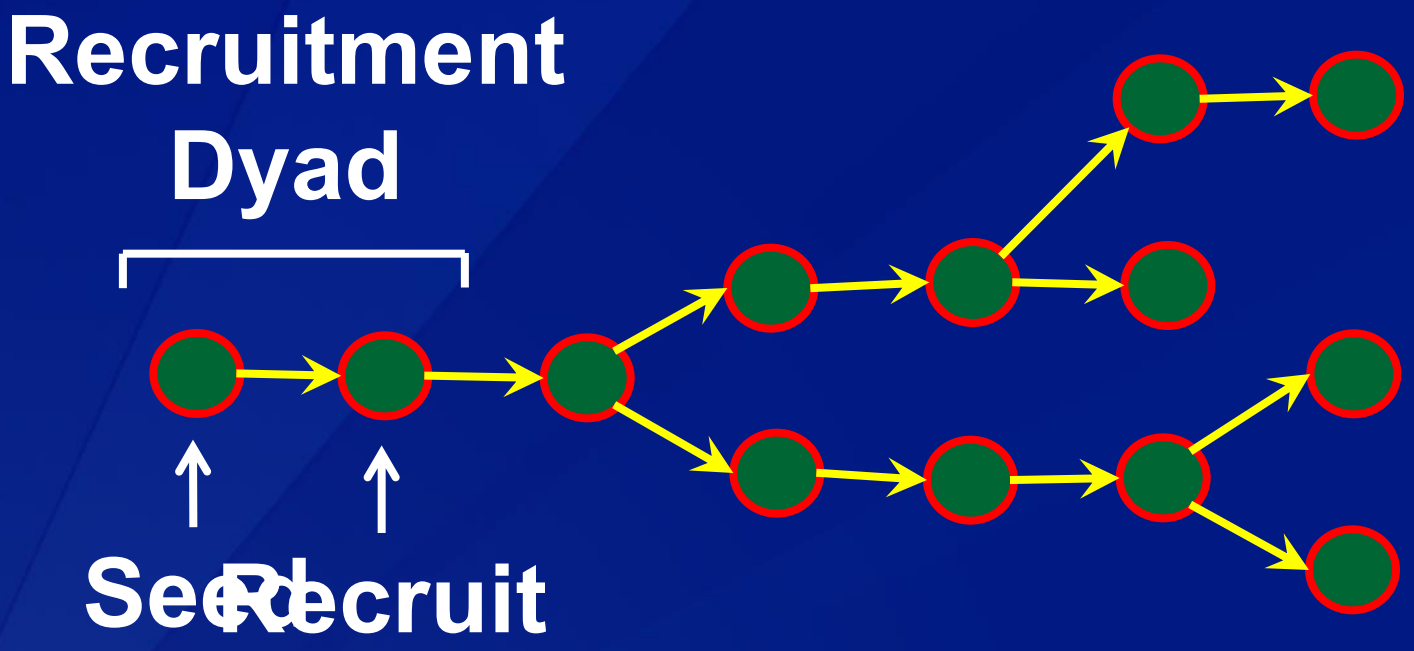
Snowball Sampling versus RDS

- **Snowball**
  - Participants report contacts' information
  - Researchers recruit participants
- **RDS**
  - Participants recruit each other
- **Advantages of RDS**
  - Fosters population member trust in survey
  - Researchers don't have to go to unsafe locations
- **Disadvantages of RDS**
  - Researcher has less control over sampling
  - Researcher has less information about sampling

RDS Implementation

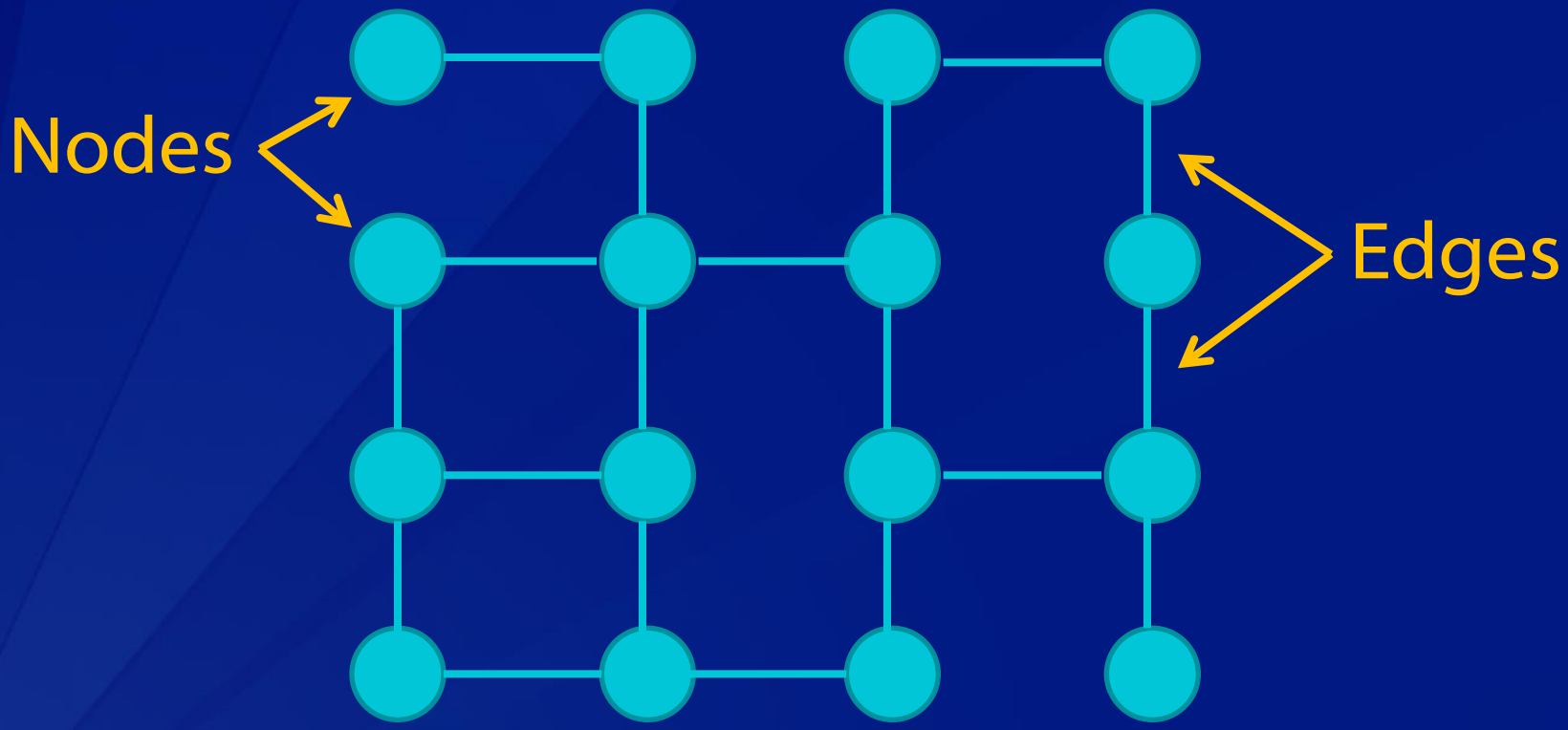
- Small number of population members (typically 3-10) purposively selected
- Interviewed at a field site and given a small number of uniquely numbered coupons
- Invite other population members they know to participate by giving them a coupon
- Those people are interviewed and given coupons, and so on, until the total sample size is reached

RDS Recruitment



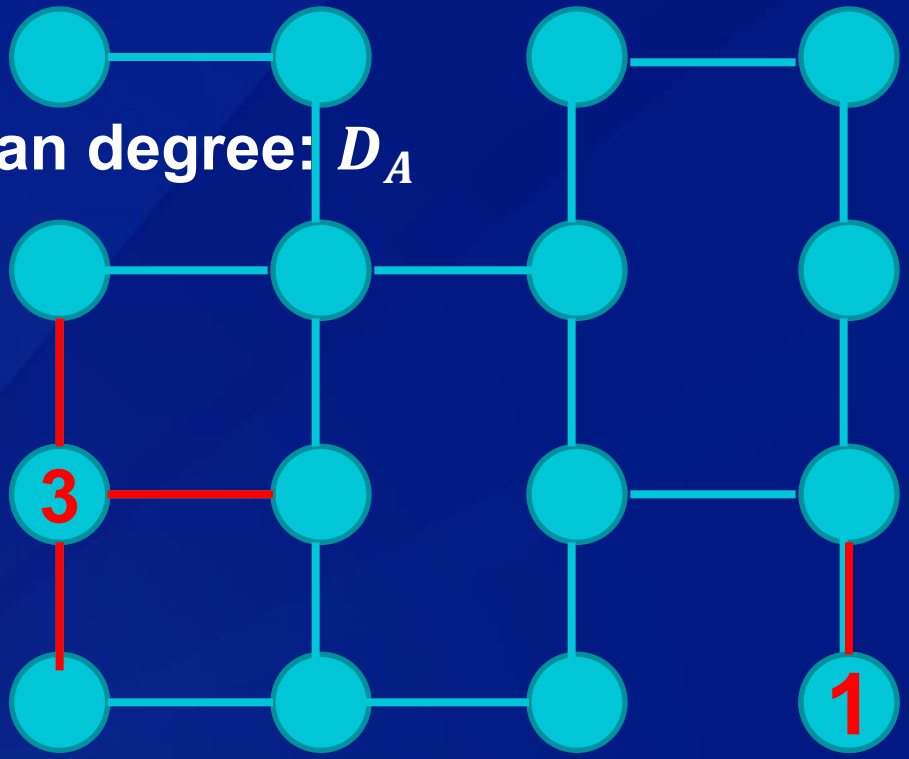
Recruitment Tree Waves

Network



Degree

- Individual node's degree: d_i
- Group mean degree: D_A



Estimation Challenges

- **Selection probabilities dependent on unobserved structure of network**
- **Sampling informative and unamenable**
- **Specific challenges**
  - Seeds
  - Number of waves
  - Sampling without replacement
    - Edge depletion

RDS Estimators

- **RDS as Markov process**
  - RDS-I
    - Estimates from edges radiating from each group
    - Addresses non-random selection of seeds
  - RDS-II
    - Estimates directly from Markov model
- **Successive Sampling**

Salganik and Heckathorn 2004; Volz and Heckathorn 2008

RDS-I Estimation (1)

- Consider two groups, A and B

- Number of edges radiating from members of group A

$$R_A = \sum_{i \in A} d_i = N_A \cdot D_A$$

Salganik and Heckathorn 2004

RDS-I Estimation (2)

- Probability of a cross-group edge radiating from each group:

$$C_{A,B} = \frac{T_{AB}}{R_A} \qquad \& \qquad C_{B,A} = \frac{T_{BA}}{R_B}$$

- Assumption: all ties in the network are reciprocal

$$T_{AB} = T_{BA}$$

$$N_A \cdot D_A \cdot C_{A,B} = N_B \cdot D_B \cdot C_{B,A}$$

Salganik and Heckathorn 2004

RDS-I Estimation (3)

- Divide through by N

$$\frac{N_A}{N} \cdot D_A \cdot C_{A,B} = \frac{N_B}{N} \cdot D_B \cdot C_{B,A}$$

- Proportional group sizes

$$P_A \cdot D_A \cdot C_{A,B} = P_B \cdot D_B \cdot C_{B,A}$$

$$P_A + P_B = 1$$

Salganik and Heckathorn 2004

RDS-I Estimation (4)

$$P_A = \frac{D_B \cdot C_{B,A}}{D_A \cdot C_{A,B} + D_B \cdot C_{B,A}}$$

Salganik and Heckathorn 2004

Estimating Mean Degree (D_A)

- **Probability proportional to degree (PPD)**
  - More friends = more people who could recruit you
- **Self-reported degree measure**
  - "How many people in New York City do you know who inject and whom you have seen in the past 30 days? Please include the person who gave you the coupon."
- **Assume that error in self-reported degrees is proportional to degree, not similar in the magnitude of absolute error across degrees**

Estimating Mean Degree (2)

- **Hansen-Hurwitz based estimator**
  - Harmonic mean

$$\widehat{D}_A = \frac{n_A}{\sum_1^{n_A} \frac{1}{d_i}}$$

Salganik and Heckathorn 2004

Estimating Mean Degrees: Assumptions

- Network is connected
- Sampling is with replacement
- Each participant is given one coupon
- Recruitment is uniformly at random
- Seeds selected with PPD

Salganik and Heckathorn 2004

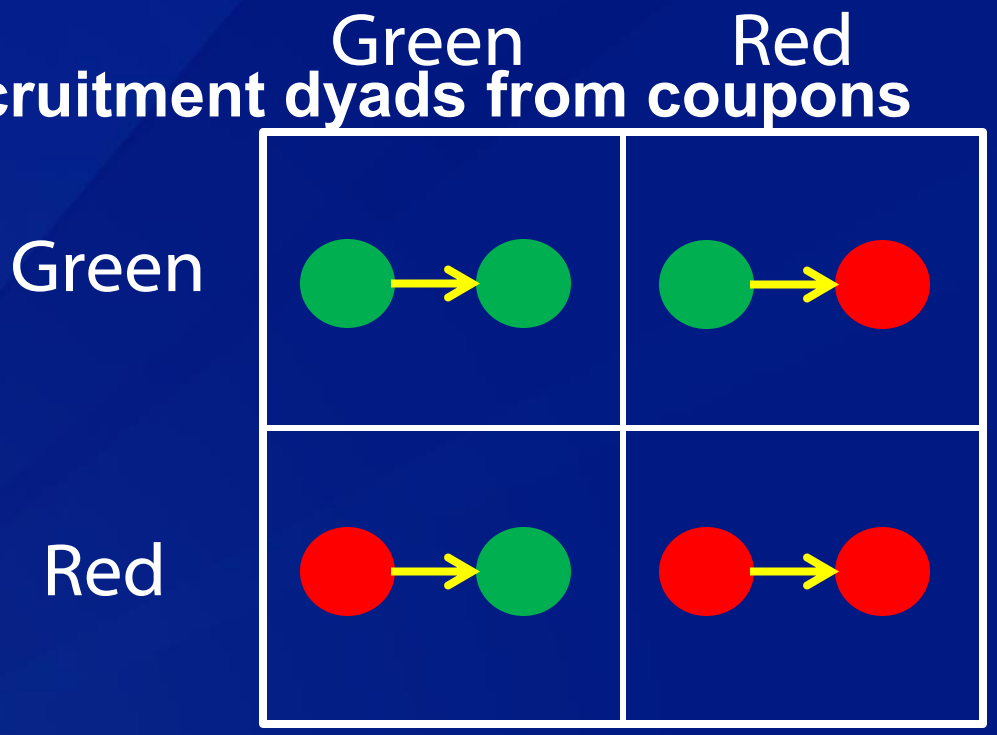
RDS-I: Addressing Seed Bias

- **Seeds are biased sample**
- **First-order Markov chain**
  - Seed bias negligible after enough steps
  - Few coupons = many waves
- **Chain state space is nodes**
  - Random walk on the network

Salganik and Heckathorn 2004

Estimating Cross-Group Edges (C_A)

- Two groups = four combinations
- Classify recruitment dyads from coupons



Recruitment Matrix

Recruiter \ Recruit	Green	Red
Green	15	20
Red	10	5

Transition Matrix (1)

Recruiter \ Recruit	Green	Red
Green	.375	.625
Red	.667	.333

Transition Matrix (2)

Recruiter \ Recruit	A	B
A	.375	.625
B	.667	.333

RDS-I Estimator

$$\widehat{P}_A = \frac{\widehat{D}_B \cdot \widehat{C}_{B,A}}{\widehat{D}_A \cdot \widehat{C}_{A,B} + \widehat{D}_B \cdot \widehat{C}_{B,A}}$$

Salganik and Heckathorn 2004

Recruitment and Demographic Adjustment

- **Estimator of cross-group ties assumes members of each group make the same average number of recruitments**
  - Random structure = no problem
- **Real networks have non-random structure**
  - If structure is related to estimand, transition probability estimates biased
- **Demographic adjustment**
  - Equilibrium of transition matrix
  - Multiply equilibrium transition probabilities by total number recruitments in the sample

Heckathorn 2002

Data Smoothing (1)

- Two groups

$$P_A \cdot D_A \cdot C_{A,B} = P_B \cdot D_B \cdot C_{B,A}$$

$$P_A + P_B = 1$$

- Three groups

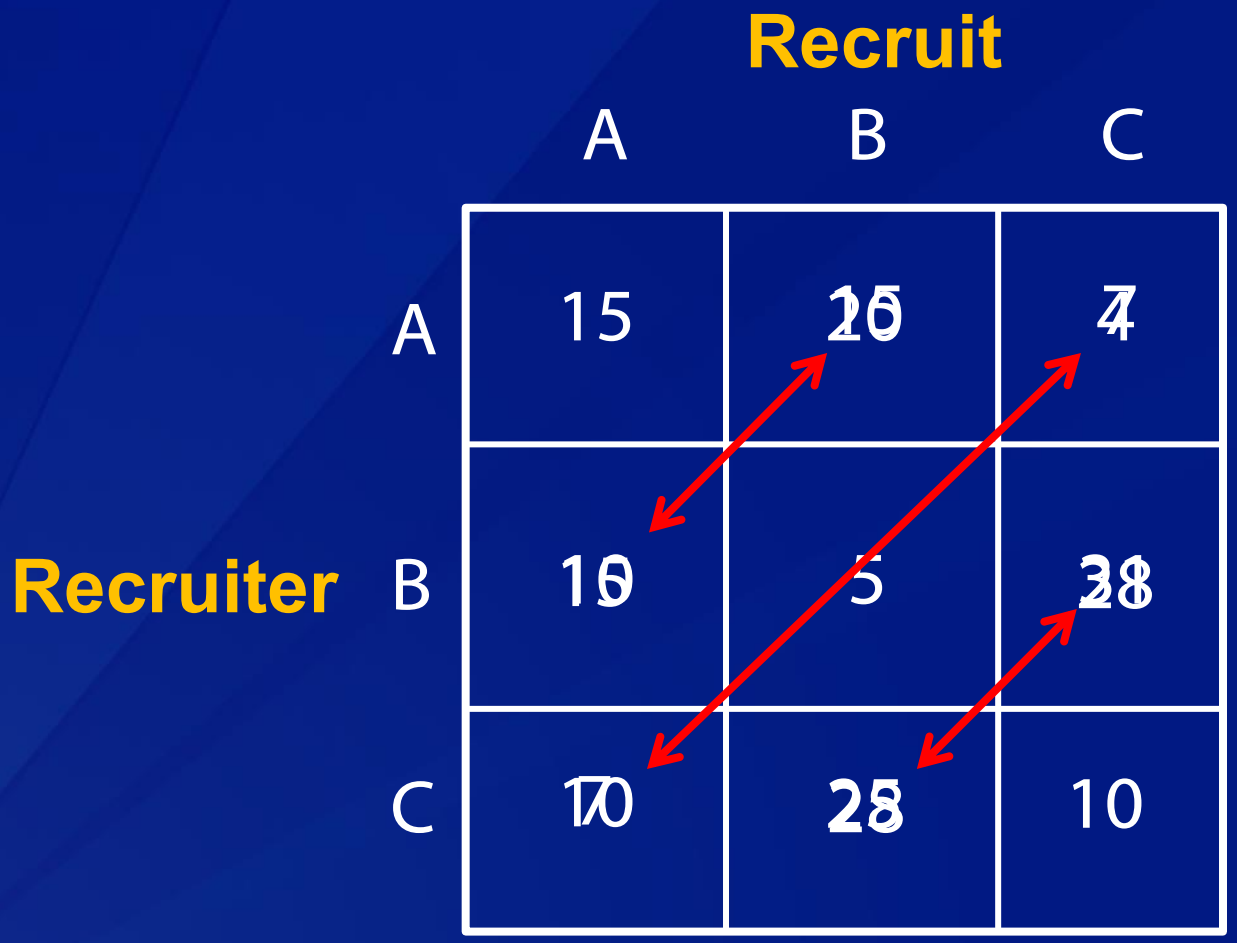
$$P_A \cdot D_A \cdot C_{A,B} = P_B \cdot D_B \cdot C_{B,A}$$

$$P_A \cdot D_A \cdot C_{A,C} = P_C \cdot D_C \cdot C_{C,A}$$

$$P_B \cdot D_B \cdot C_{B,C} = P_C \cdot D_C \cdot C_{C,B}$$

$$P_A + P_B + P_C = 1$$

Data Smoothing (2)



NHBS and RDS Estimation

- **NHBS currently uses RDS-I**
- **Recruitment efficiency bias**
  - Network structure related to estimands
  - Different average numbers of recruitments by groups
- **Real-world estimation details addressed**
  - Missing data for estimands
  - Missing degree data
  - Reported degrees of 0
  - Lost coupon data
- **Software**

RDS-II

- **RDS-II linked RDS estimation directly to standard complex sampling estimators**
- **Similar to RDS-I**
  - Identical estimates in some situations
- **Markov chain on nodes**
  - Random walk on network

Volz and Heckathorn 2008

RDS-II Estimation (1)

- Probability proportional to degree

- Horvitz-Thompson estimator
- Generalized Horvitz-Thompson estimator
  - More flexible than RDS-I

$$\widehat{P}_A = \frac{1}{N}\sum_{i=1}^{N} S_i \cdot \frac{A_i}{d_i}$$

$$\widehat{P}_A = \frac{\sum_{i=1}^{N} S_i \cdot \frac{A_i}{d_i}}{\sum_{i=1}^{N} S_i \cdot \frac{1}{d_i}}$$

Volz and Heckathorn 2008; Gile and Handcock 2010

RDS-II Estimation (2)

- Alternative representation: adjusting the sample proportion

$$\widehat{P}_A = \left(\frac{n_A}{n}\right) \cdot \left(\frac{\widehat{D}}{\widehat{D}_A}\right)$$

Volz and Heckathorn 2008

RDS-I and RDS-II Assumptions

- **RDS-II assumptions equivalent to RDS-I**
  - Does not relax single recruit assumption
- **RDS-II estimates similar to RDS-I unless:**
  - Some groups recruit more than others
  - Network has meaningful structure addressed by RDS-I
- **RDS-II directly tied to standard sampling estimation literature**
- **RDS-I used in majority of published RDS studies**

Volz and Heckathorn 2008 ; Gile and Handcock 2010

Successive Sampling Estimator (1)

- **RDS without replacement**
- **Nodes with large degree sampled earlier**
  - Variance of degree distribution shrinks
  - Variance of selection probabilities shrinks
- **RDS-II - random walk over sampled network**
- **Consider random walks over all networks with same degree distribution as network being sampled**
  - With replacement ≈ RDS-II
  - Without replacement ≠ RDS-II

Gile 2011

Successive Sampling Estimator (2)

- **For known population size N**
- **Iteratively estimate via simulation:**
  - Population degree distribution and mapping of nodal degree to selection probability
  - Mapping is a function of the order of sequence of sampled degrees
- **Use estimated selection probabilities in generalized Horvitz-Thompson estimator**

Gile 2011

Successive Sampling Estimator (3)

- Large sampling fraction

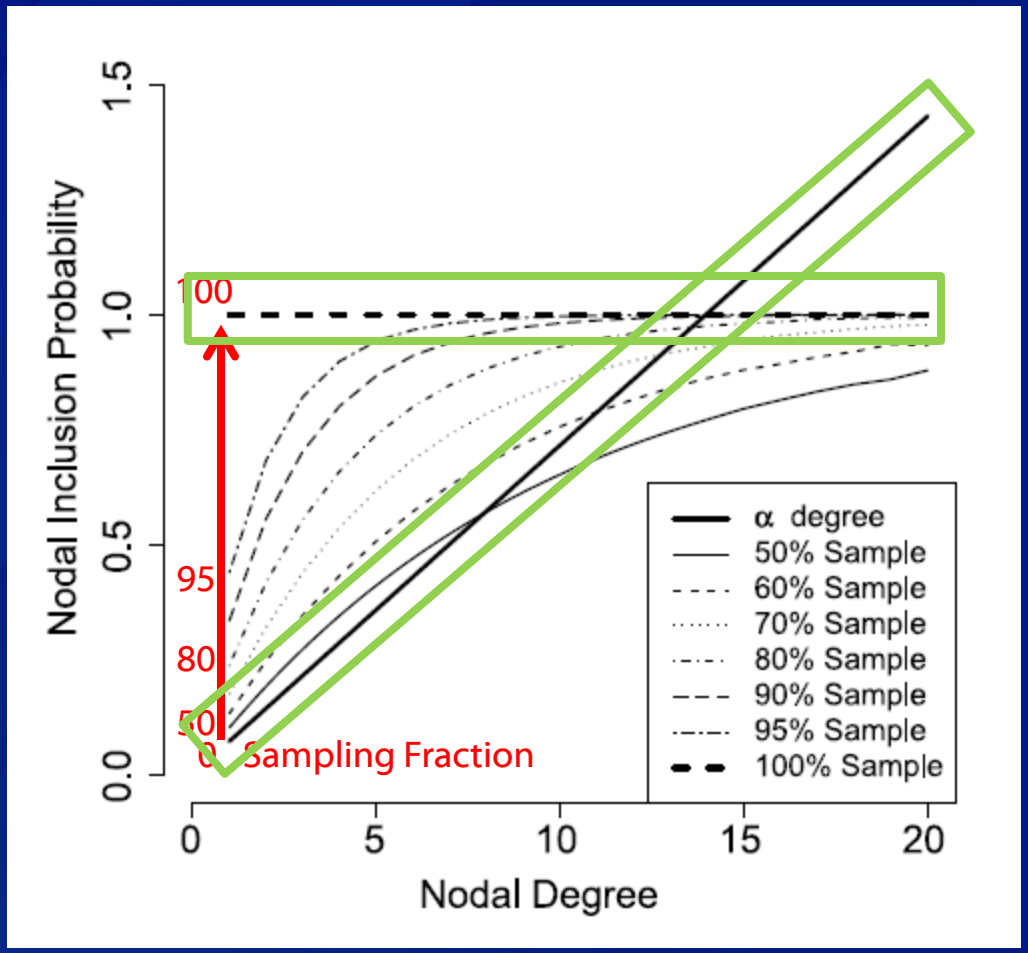
$$\lim_{sf \to 1} \widehat{P}_A^{SS} = \frac{n_A}{n}$$

- Small sampling fraction

$$\lim_{sf \to 0} \widehat{P}_A^{SS} = \widehat{P}_A^{RDS-II}$$

Gile 2011

Successive Sampling Estimator (4)



Gile 2011

Estimator Assumptions

Estimator	Network Assumptions	Sampling Assumptions
All three	Network Connected	Many sample waves
	Edges reciprocal	Degree accurately measured
	Structure weak enough	Random recruitment

Estimator	Network Assumptions	Sampling Assumptions
RDS-I RDS-II		Sampling with replacement OR Sampling fraction small enough
		Single, non-branching chain

Estimator	Network Assumptions	Sampling Assumptions
Successive Sampling (SS)	Known population size	Initial sample unbiased

Gile 2011

Estimators in Progress

- Model-assisted
- Edges not reciprocal
- Infection over network
- Information about unrecruited friends
- Fully Bayesian

Gile and Handcock 2011; Lu et al. 2012; Malmros et al. 2013; Berchenko et al. 2012; Crawford 2014; Lu 2012; Lunagomez and Airoldi 2014

Variance Estimation

- **Closed form for RDS-II**
  - Not widely used
  - Few comparisons to others

- **All others are bootstrap variants**
  - Salganik bootstrap – Markov chain on the transition matrix with samples from sample degree distributions
  - Successive sampling - PPD without replacement draws from model of degree distribution
  - Model assisted – simulated RDS on synthetic networks

Volz and Heckathorn 2008; Salganik 2006; Gile 2011; Gile and Handcock 2011

Unresolved Questions

- **Most effort on creating point estimators**
- **Non-simulation assessments of estimators and assumptions less common**
  - Ground truth data difficult to gather
  - A few projects
- **Variance estimation**
- **Multivariable modeling**

Conclusion

- Information needed about hidden populations
- Estimation challenging; requires strong assumptions
- Estimation literature highly active
- We at NHBS look forward to your contributions to unresolved questions!

Acknowledgements

- **NHBS sites and participants**
- **Behavioral Surveillance Team**

Gabriela Paz-Bailey
Dita Broz
Winston Abara
Johnathan Cook
Laura Cooley
Melissa Cribbin
Paul Denning
Alicia Edwards
Teresa Finlayson
Kathy Hageman
Kristen Hess
Brooke Hoots
Wade Ivy
Binh Le
Rashunda Lewis
Stacey Mason
Lina Nerlander
Katie Salo
Catlainn Sionean
Amanda Smith
Justin Smith
Cyprian Wejnert
Mingjing Xia

Thank You!

Michael W. Spiller, PhD

Epidemiologist
Division of HIV/AIDS Prevention
Behavioral and Clinical Surveillance Branch
Centers for Disease Control and Prevention

MSpiller@cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.