Inference from Link-Tracing Network Samples

Krista J. Gile

University of Massachusetts, Amherst *

October 21, 2013

For more information:

- Krista J. Gile "Inference from Link-Tracing Network Samples: A Foundational Review," in preparation

*Research supported by NICHD grant 7R29HD034957 and NIDA grant 7R01DA012831, UW Networks Project (Martina Morris, PI) and by NSF grant DMS-0354131 (Mark Handcock, PI), NSF Grant 12-510 SES-1230081 with support from the National Agricultural Statistics Service.(Krista Gile PI), NIH Grant 1 R21 AG042737-01A1, Elena Erosheva PI, K.Gile Subcontract PI.

Hard-to-Reach Population Methods Research Group (HPMRG) (and Collaborators)

- Isabelle Beaudry, UMass Amherst
- Elena Erosheva, University of Washington
- Ian Fellows, FellStat
- Karen Fredriksen-Goldsen, University of Washington
- Krista J. Gile, UMass, Amherst
- Mark S. Handcock, UCLA
- Lisa G. Johnston, Tulane University, UCSF
- Corinne M. Mar, University of Washington
- Miles Ott, Carleton College
- Matt Salganik, Princeton University
- Amber Tomas, Mathematica Policy Research

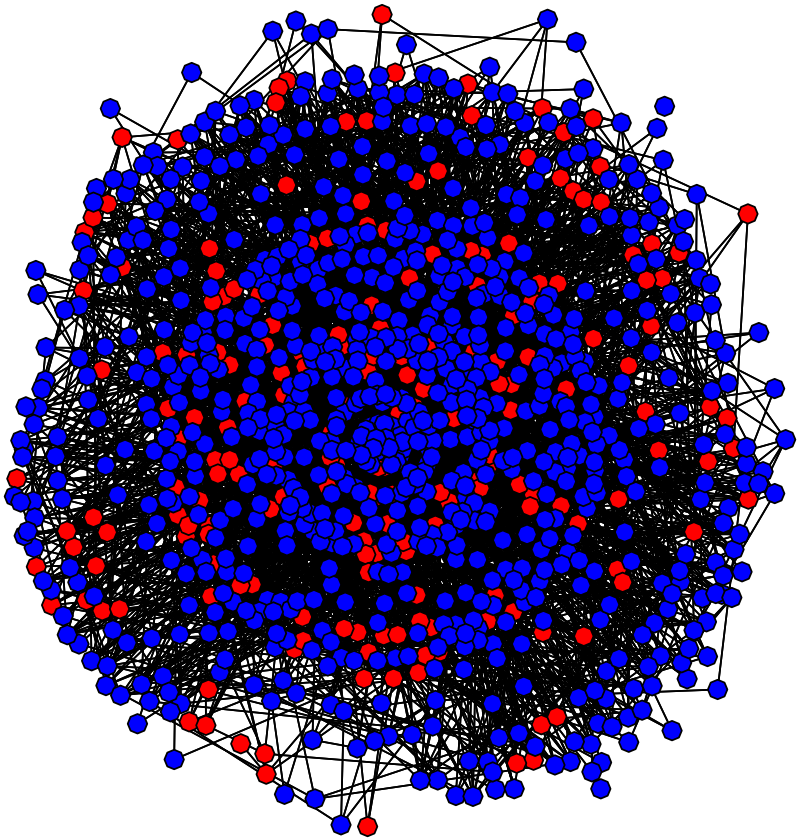
Outline

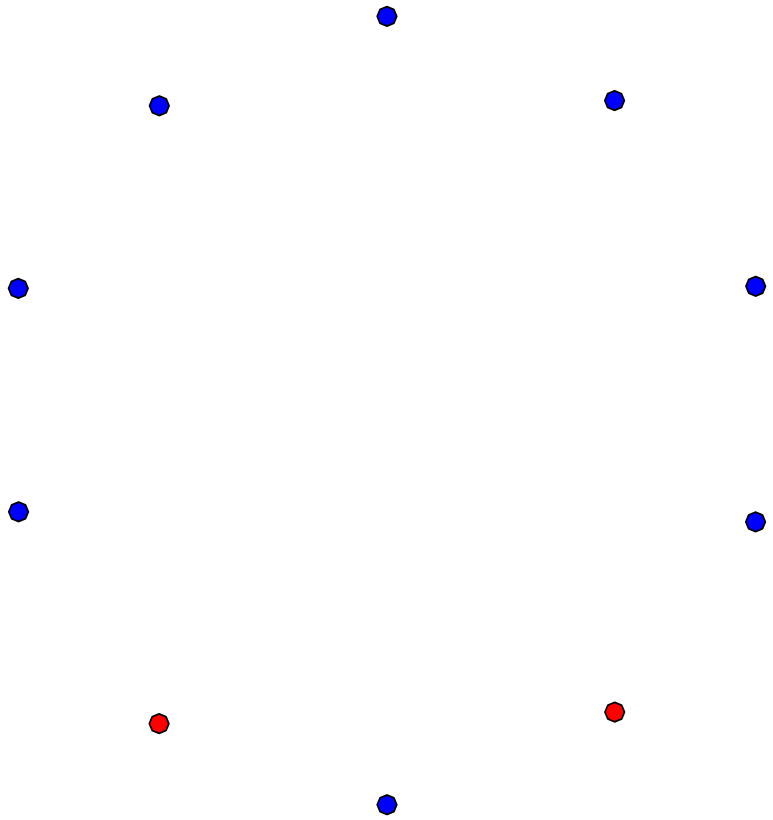
- Introduce link-tracing
- Link-tracing is (substantively) interesting
- Link-tracing is (statistically) interesting
- Comparison: Stratified Random Sample
- Challenge 1: Sampling depends on network
- Challenge 2: Unknown initial sample mechanism
- Challenge 3: Unknown population
- Example: Respondent-Driven Sampling
- Discussion

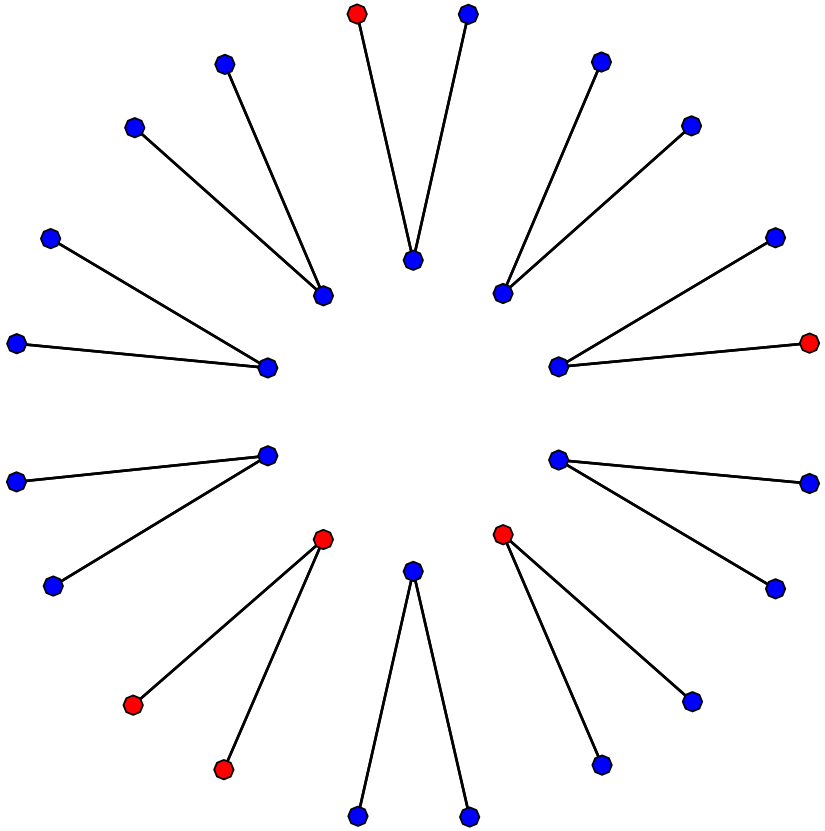
Link-Tracing Network Sampling

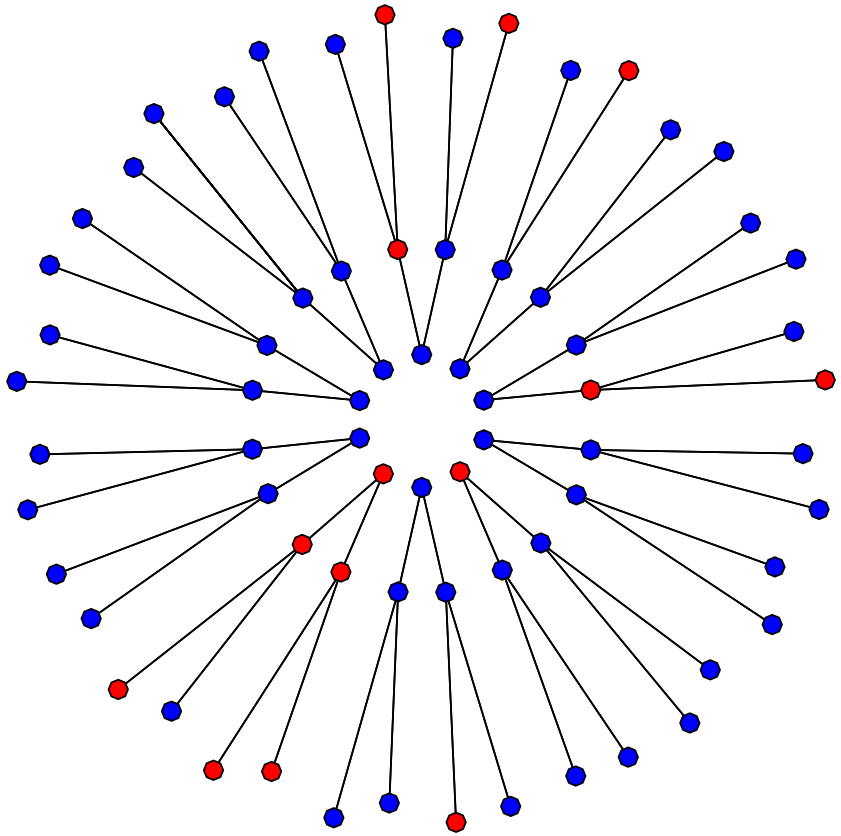
Link-Tracing Sampling:

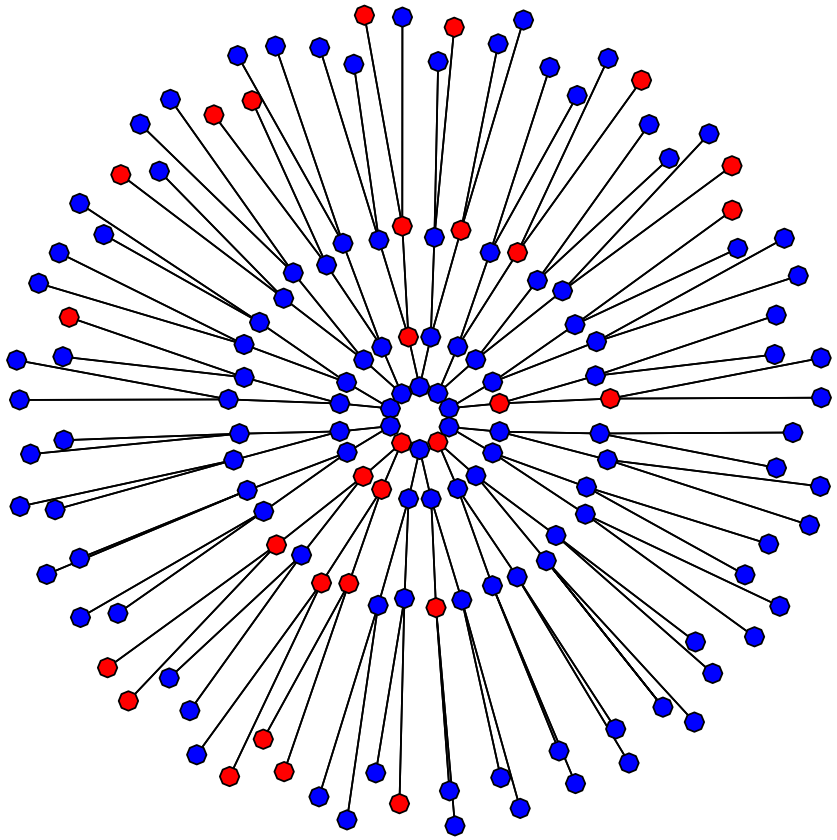
- A population of units (nodes) connected by a network
- A two-phase sampling process:
  1. And initial set of sampled units
  2. Subsequent units sampled from among network alters of current sample (traced links)

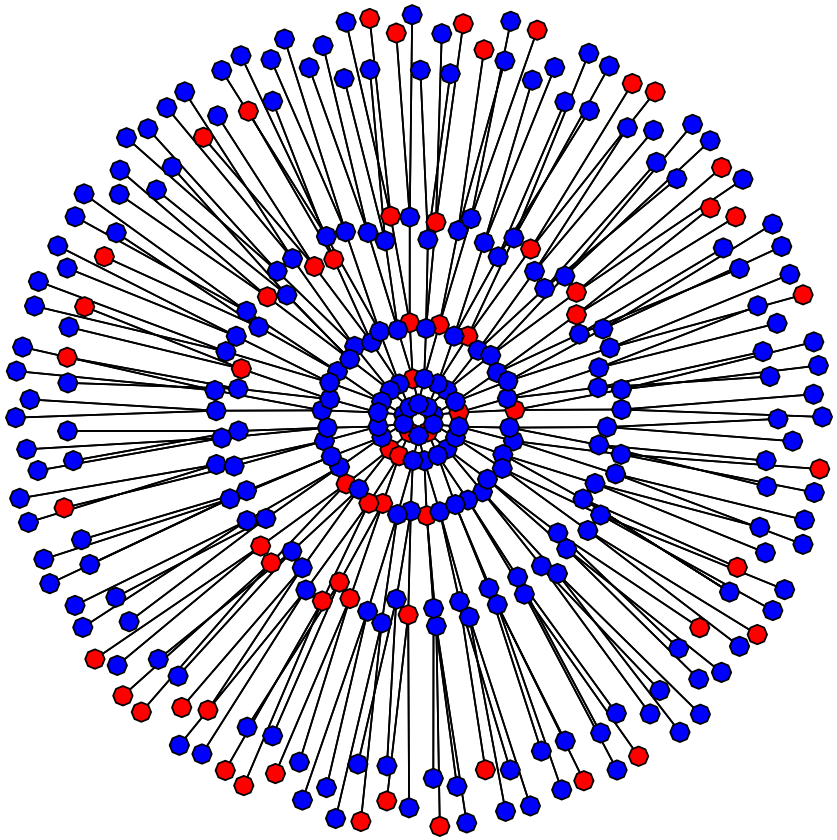


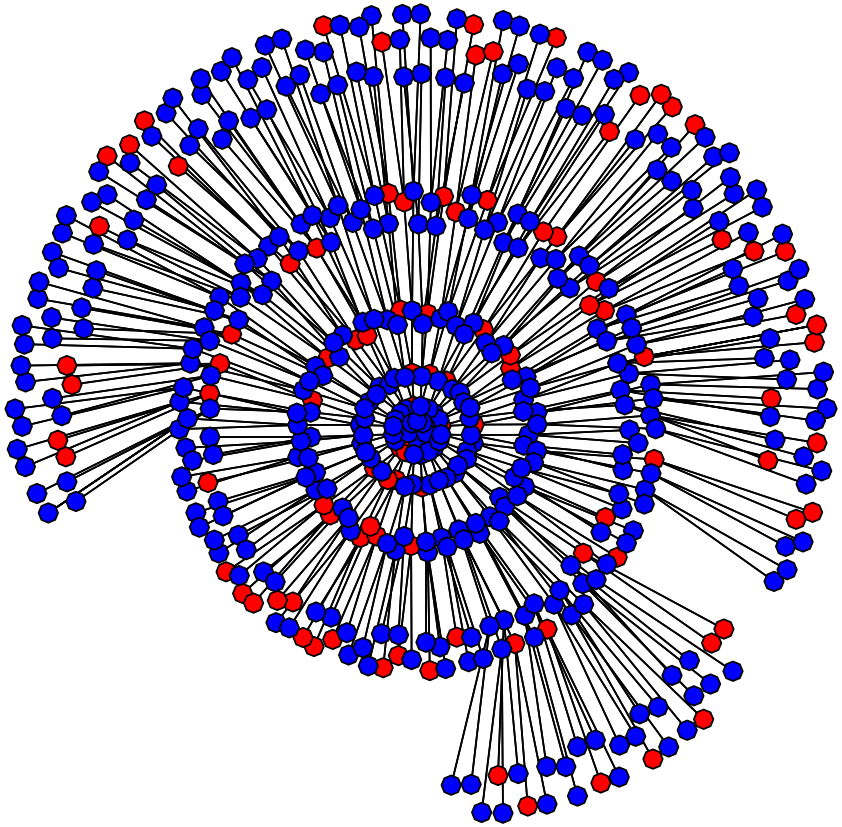


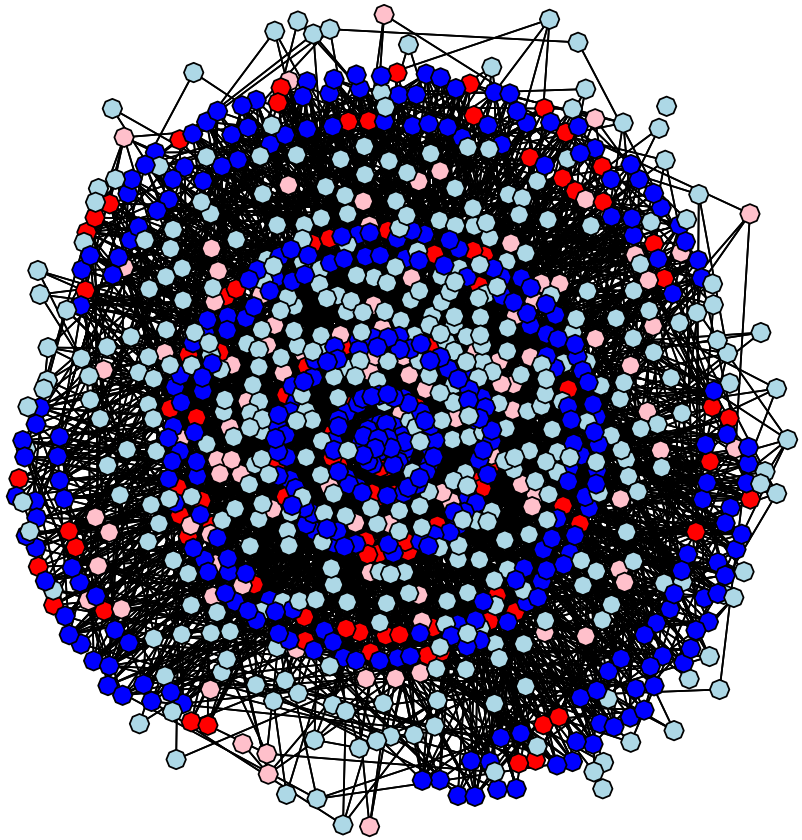




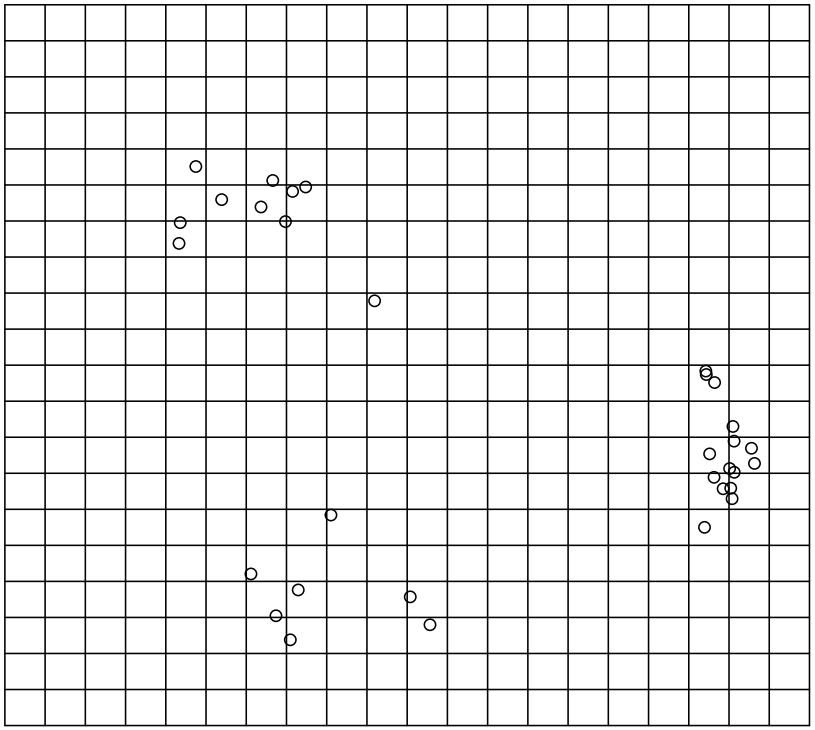






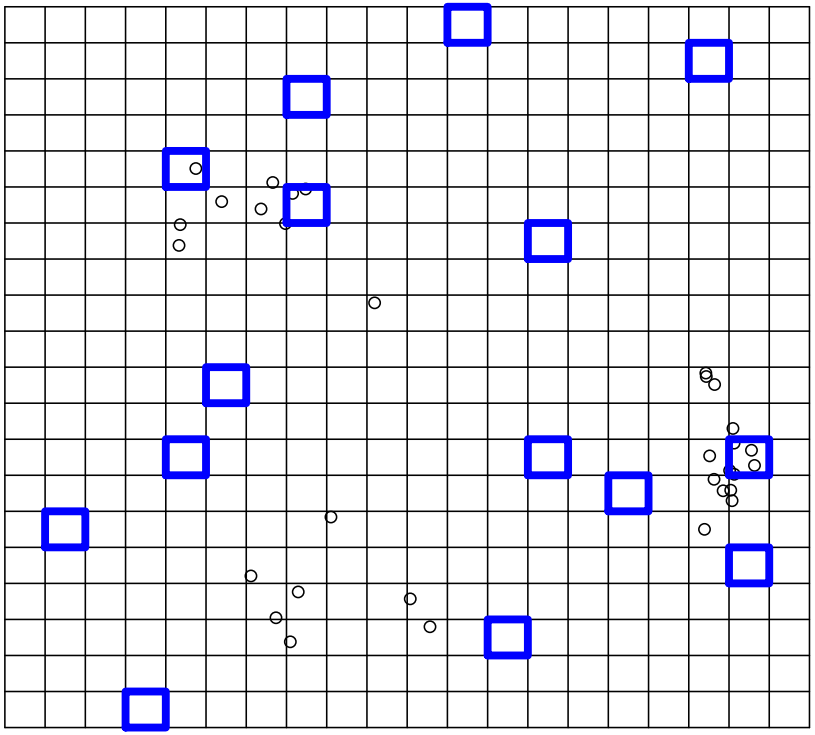


Link-tracing is (substantively) interesting



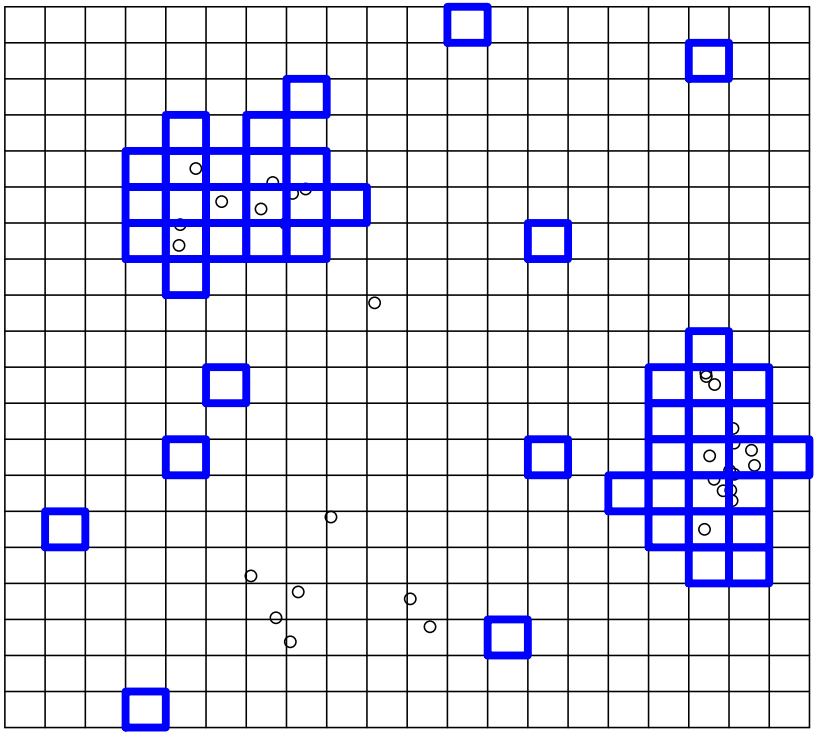
Thompson, S.K., 1990, "Adaptive Cluster Sampling," Journal of the American Statistical Association, 85, 1050-9.

Link-tracing is (substantively) interesting



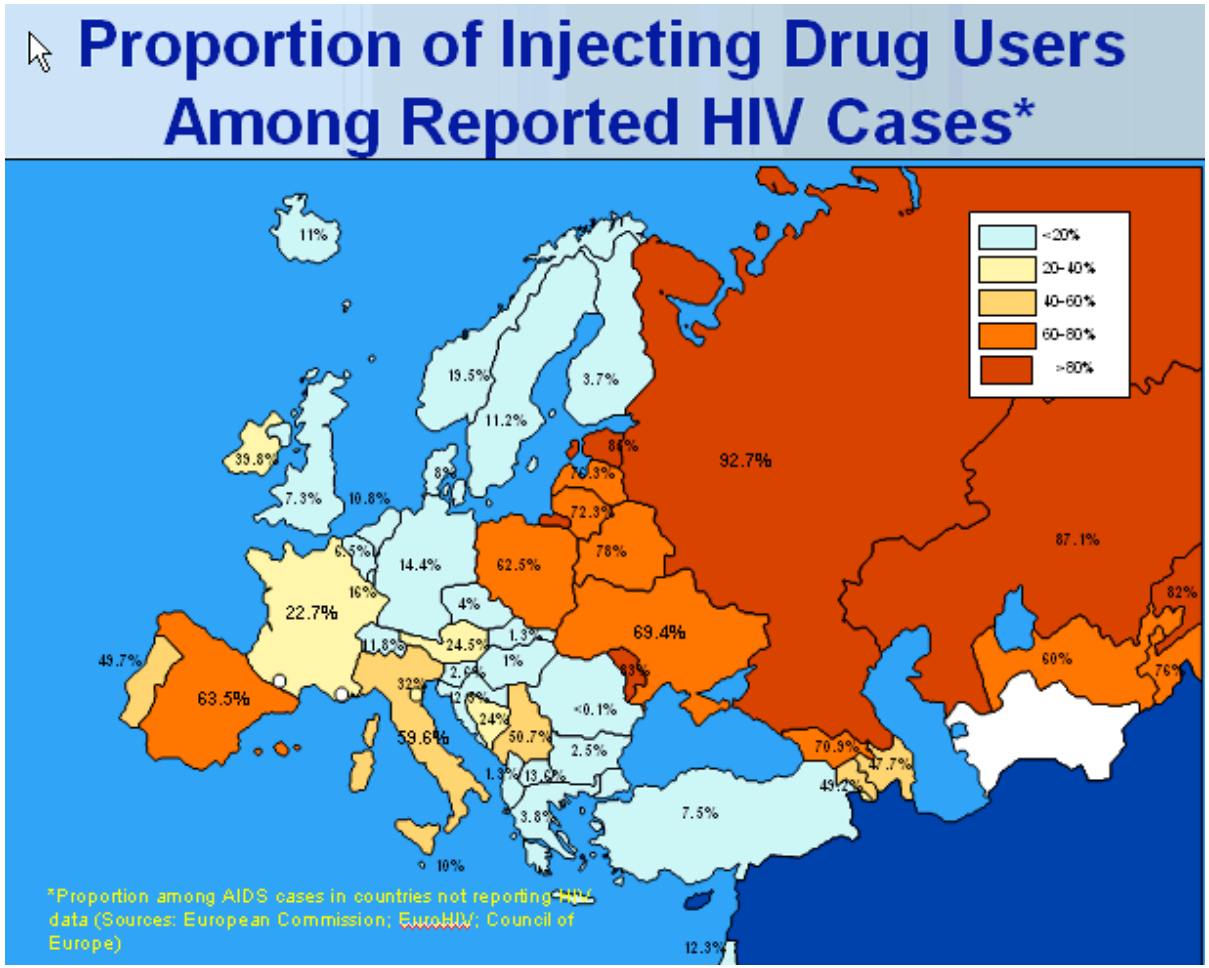
Thompson, S.K., 1990, "Adaptive Cluster Sampling," Journal of the American Statistical Association, 85, 1050-9.

Link-tracing is (substantively) interesting



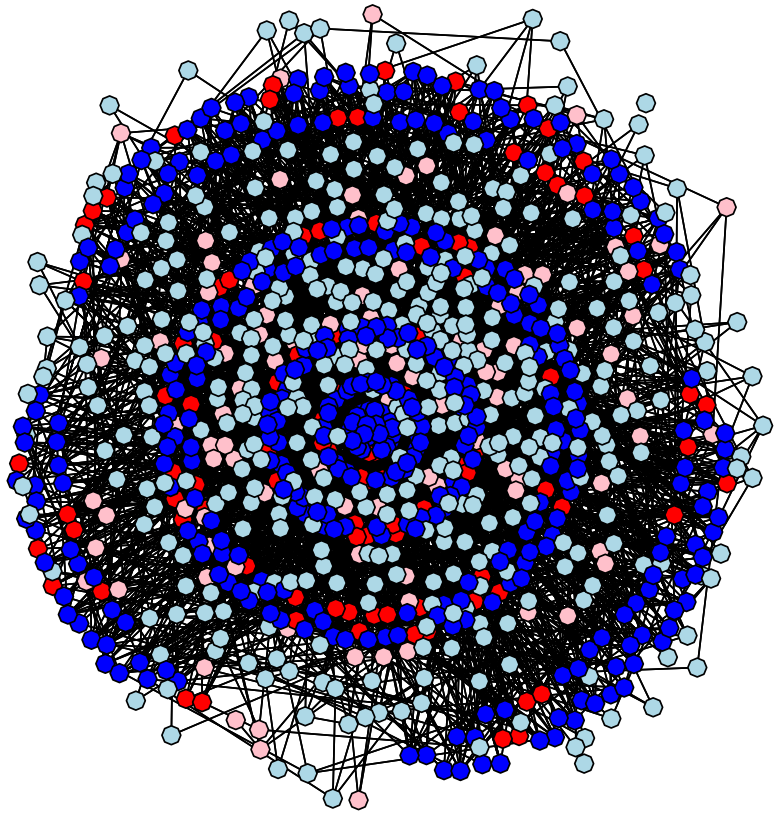
Thompson, S.K., 1990, "Adaptive Cluster Sampling," Journal of the American Statistical Association, 85, 1050-9.

Link-tracing is (substantively) interesting



From World Health Organization: `www.who.int`

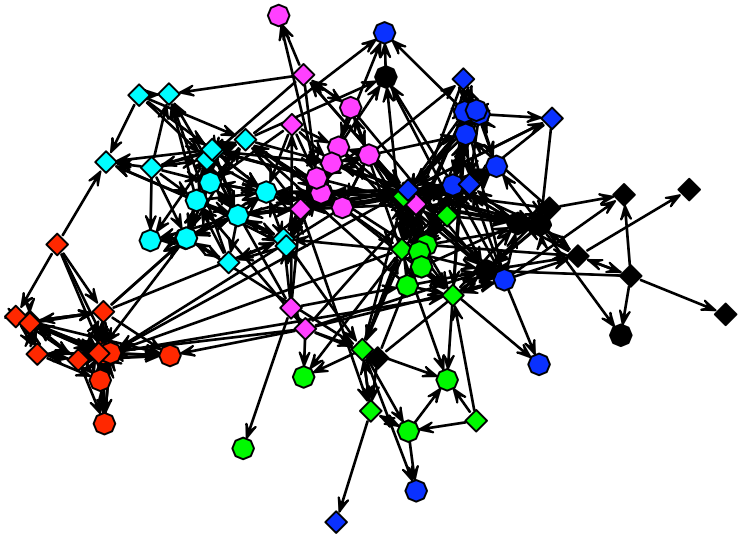
Link-tracing is (substantively) interesting



Link-tracing is (substantively) interesting

- Rare populations
- Stigmatized populations
- Internet: World Wide Web, Facebook

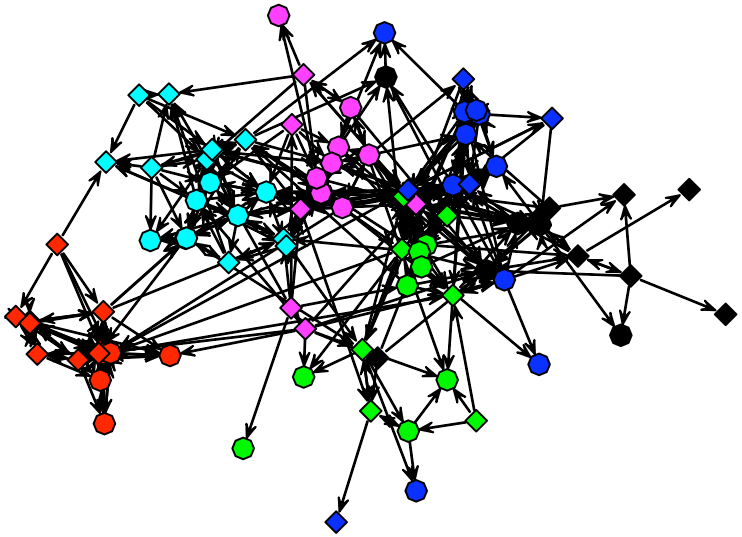
Networks are (statistically) interesting



Network:

- Two types units: nodes and relations. Relations between nodes.
- Nodal and dyadic characteristics (HIV status, amount of trade)
- Higher-order network features (e.g. triangles)
- Dependencies across types of units e.g.:
  - Connected nodes more similar
  - Dyads sharing a node more similar
- Conditioning: what is stochastic, what is fixed? (nodal characteristics, relations, degrees)

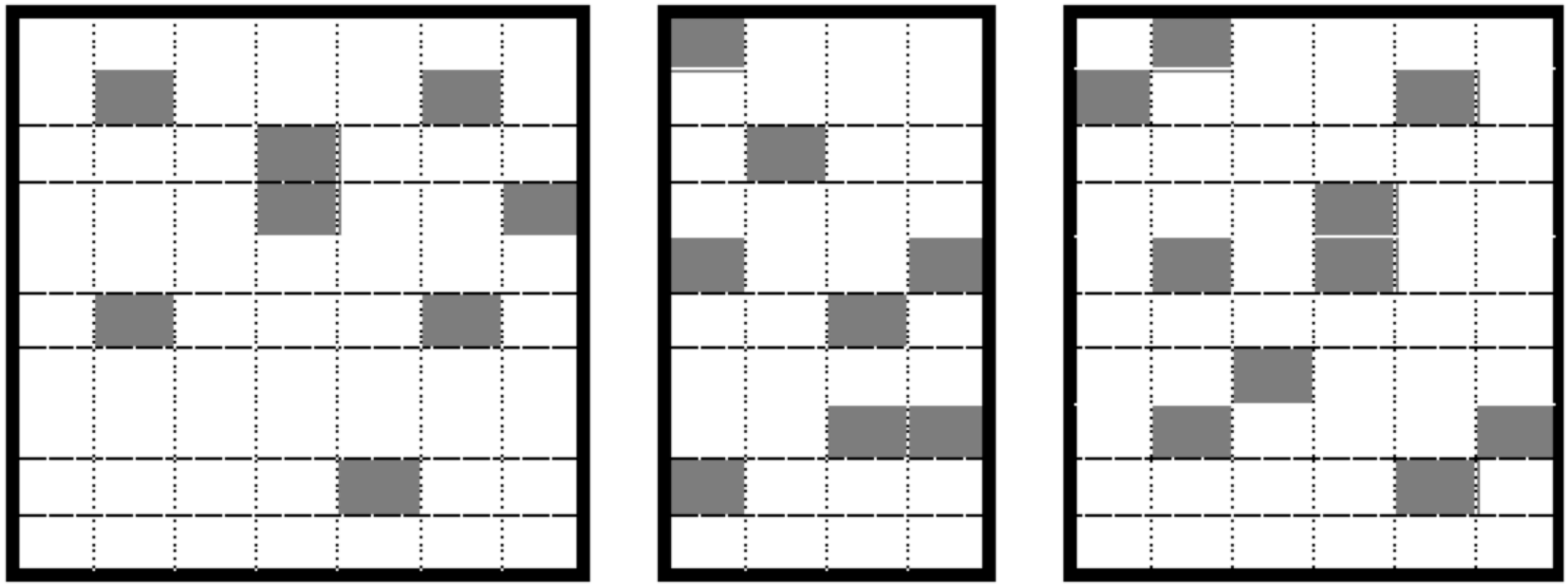
Link-tracing is (statistically) interesting



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



- 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. Similar approach for standard error.

Requires π_i for all sampled units.

Stratified Random Sample: Likelihood Inference

Observed Data: $X_{obs} = X_i : i \in 1 \ldots n$, AND $S_i, i \in 1 \ldots N$,
where $S_i = 1$ if unit i sampled, full population data X.

Assume a model:

$$X_i = \beta_0 + \beta_{k_i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

for k_i the strata of unit i. Parameter $\theta = \{\beta_0, \beta, \sigma^2\}$.
Inference based on:

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

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

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

Challenge 1: Sampling depends on network

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

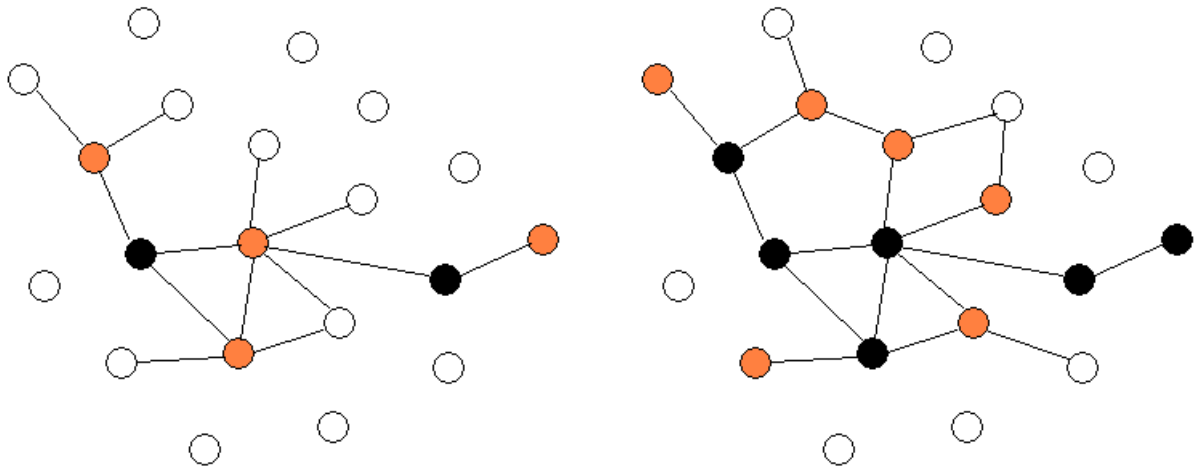
Sampling depends on network: design-based

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, in press.

Sampling depends on network: likelihood

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

If $P(S|Y,\theta) = P(S|Y_{obs})$ (MAR), then $L(\theta|X,S) \propto \sum_{Unobserved} P(Y|\theta)$. Consider Initial Sample S_0:

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

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

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.

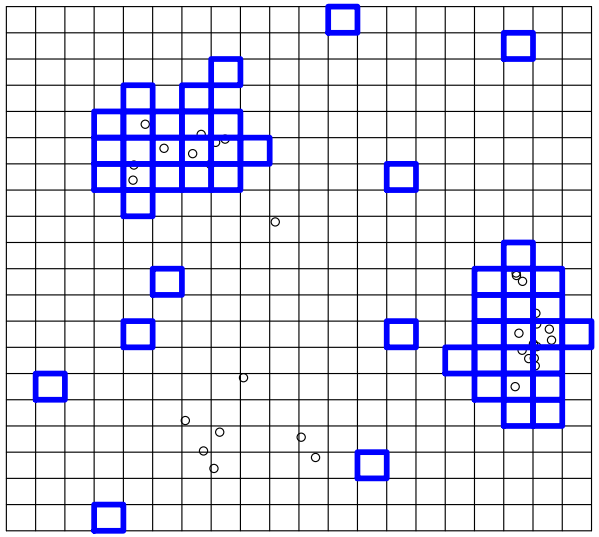
Thompson, S.K. and O. Frank, 2000, "Model-based estimation with link-tracking sampling designs." , Survey Methodology 26, 87-98.
Handcock, M.S. and K.J. Gile, 2010, "Modeling social networks from sampled data." , Annals of Applied Statistics.

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



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., 2011, "Improved Inference for Respondent-Driven Sampling Data with Application to HIV Prevalence Estimation," Journal of the American Statistical Association.

- 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.
Thompson, S.K., 2006, "Adaptive Web Sampling," Biometrics, 62, 1224-34.

Challenge 2: Unknown initial sample mechanism

Consider a hidden 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).

If $P(S|Y,\theta) = P(S|Y_{obs})$ (MAR), then $L(\theta|X,S) \propto \sum_{Unobserved} P(Y|\theta)$. Consider Initial Sample S_0:

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

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

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

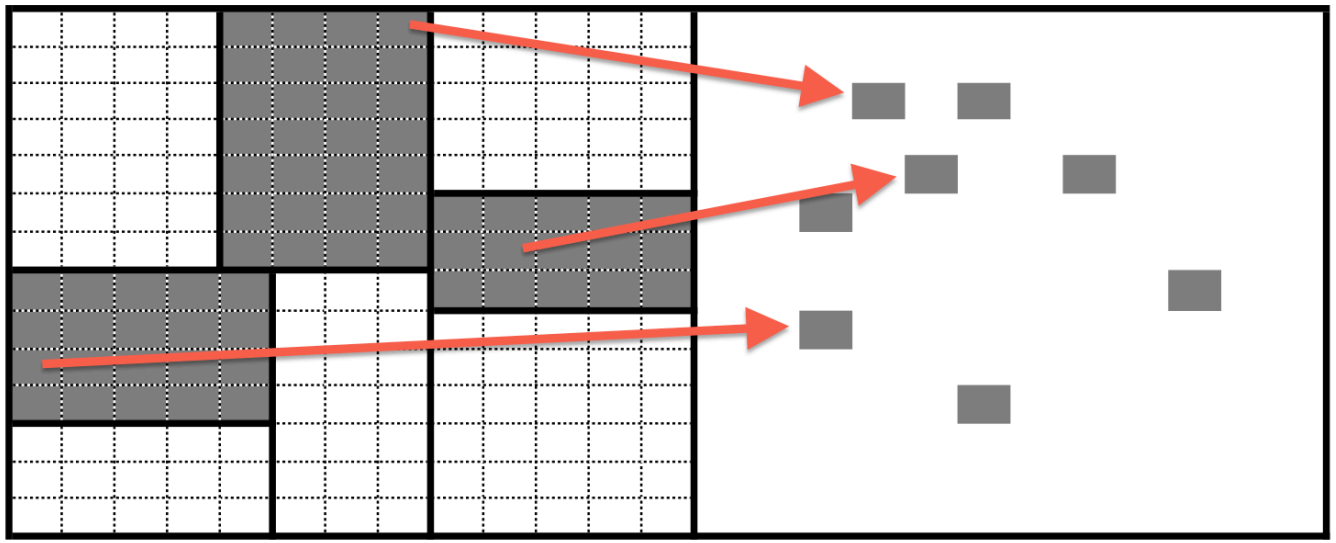
1. Assume initial sample well-behaved
2. Assume initial sample design partially known
3. Assume many waves of sampling decrease dependence on initial sample
4. Condition on initial sample

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



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

Gile, K.J., and M.S. Handcock, 2013, "Network Model-Assisted Inference from Respondent-Driven Sampling Data" , under revision, available on arXiv.

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

Challenge 3: Unknown population

Consider a hidden 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

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

- 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 $\implies$ π_i unknown for many i.

Unknown population: likelihood

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

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

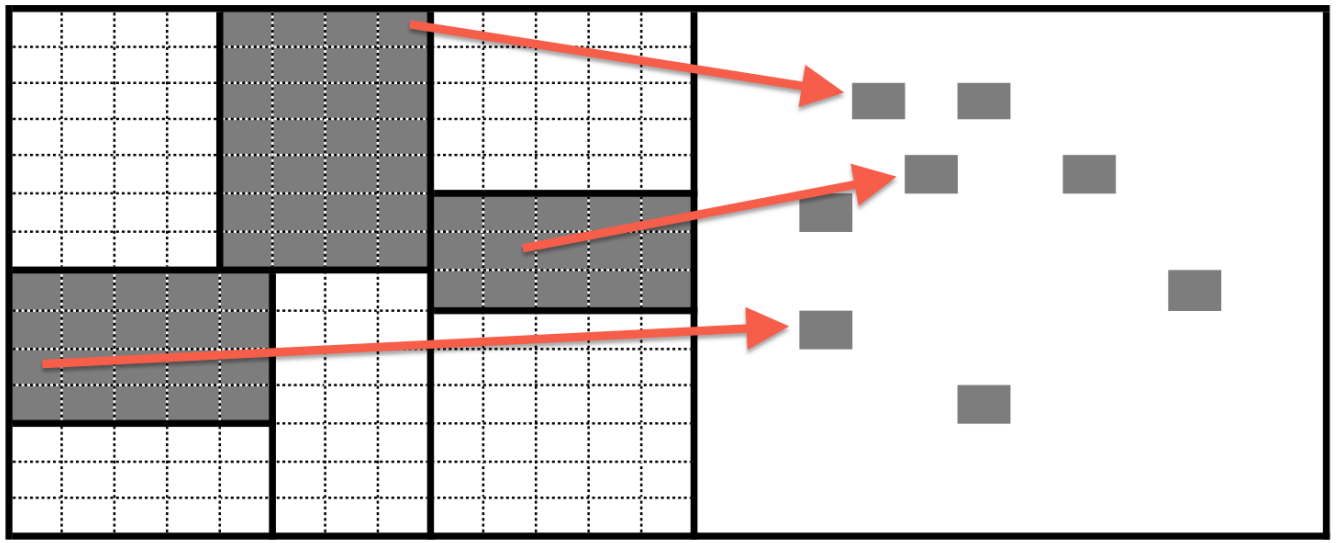
1. Assume N known
2. Estimate N
3. Ratio Estimator (design-based)
4. Condition on part of sample

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.

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, 2013, "Estimating Hidden Population Size using Respondent-Driven Sampling Data" , under revision

- 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., G.L. Robins, T.A.B. Snijders, and P. Wang 2013, "Conditional estimation of exponential random graph models from snowball sampling designs." , Journal of Mathematical Psychology, in press.

- 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.

Example: Respondent-Driven Sampling

- Hard-to-reach Populations - no conventional sampling frame
- Example: What proportion of injecting drug users in London are HIV positive?

Sampling:

- Begin with convenience sample.
- Respondents *drive* sample by passing coupons to contacts in social network
- Sample ends at desired sample size

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

Subject To:

- Sampling based on partially-observed networks
- Unknown initial sample mechanism
- Unknown population

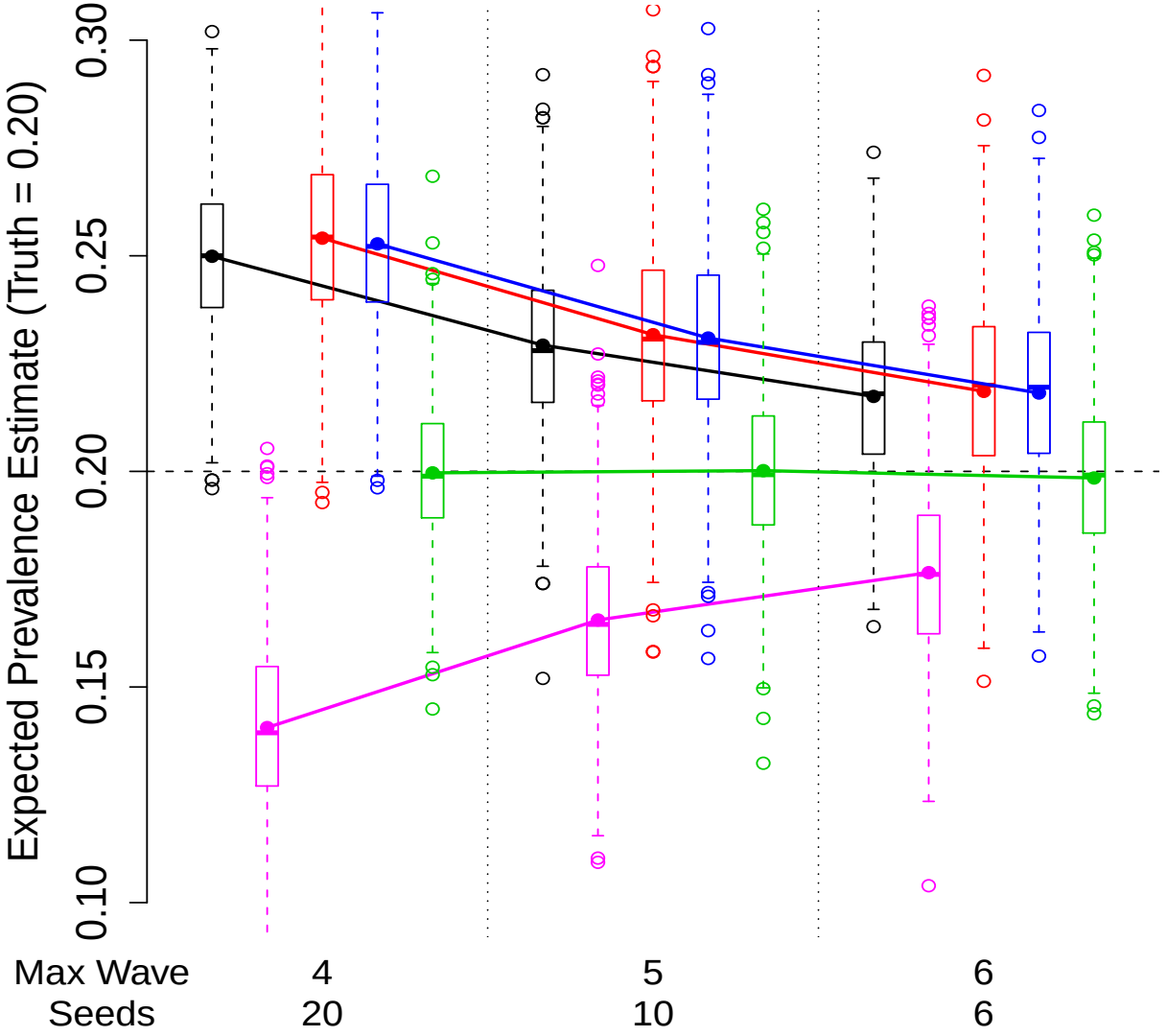
Approach: Use network model to correct for initial sample

Initial convenience sample, require knowledge of N

- Condition on initial non-probabilty sample
- Fit network model
- Estimate population proportions

Gile, K.J., and M.S. Handcock, 2013, "Network Model-Assisted Inference from Respondent-Driven Sampling Data" , under revision, available on arXiv.

All Infected Seeds, varying number of seeds, 50%



Other new approaches:

- Jointly model sampling and network:
  Fellows, I.E., and M.S. Handcock, 2013, "Analysis of Partially Observed Networks via Exponential family Random Network Models," under review, available on arXiv.
- Adjust for unequal sampling probabilities due to sampling aberrations:
  Ott, M.Q., K.J. Gile, et al. "Re-weighted Estimation for Respondent-Driven Sampling: Implications for Inference." In Preparation, 2013.

Discussion

- General Issues for Statistical Inference
  - Dependent data, plus data-dependent sampling
  - Not missing at random sampling (NMAR)
  - Implicit sampling frame
- Challenges for Link-Tracing Sampling
  - Sampling depends on (typically) partially-observed 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 is high-stakes application driving innovation.