Using Link-Tracing Data to Inform Epidemiology

Krista J. Gile
Nuffield College, Oxford

joint work with Mark S. Handcock
University of Washington, Seattle

23 October, 2008

For details, see:

- Gile, K.J. (2008). Inference from Partially-Observed Network Data. PhD. Dissertation. University of Washington, Seattle.[1]

[1] Research supported by NICHD grant 7R29HD034957 and NIDA grant 7R01DA012831

Fitting Models to Partially Observed Social Network Data

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

$$P(Y_{obs}, D|\beta, \delta) = \sum_{Unobserved} P(Y, D|\beta, \delta)$$
$$= \sum_{Unobserved} P(D|Y, \delta) P(Y|\beta)$$

- β is the model parameter
- δ is the sampling parameter

If $P(D|Y, \delta) = P(D|Y_{obs}, \delta)$ (*adaptive sampling* or *missing at random*)

Then

$$P(Y_{obs}, D|\beta, \delta) = P(D|Y, \delta) \sum_{Unobserved} P(Y|\beta)$$

- Can find maximum likelihood estimates by summing over the possible values of unobserved, ignoring sampling
- Sample with Markov Chain Monte Carlo (MCMC)

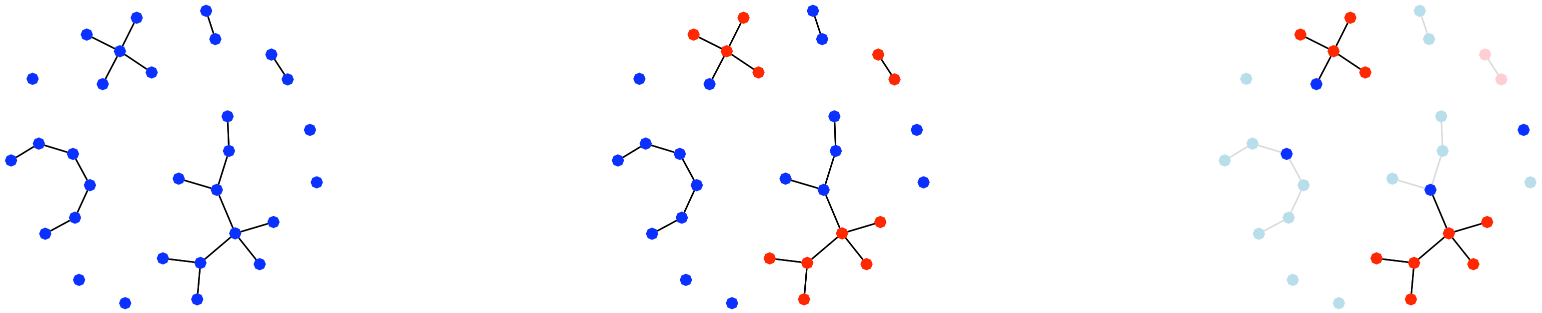
Contact Tracing

Reportable diseases reported to public health authorities. Partners of those infected reported, contacted, and tested.

- Reportable Diseases (King County, Washington - partial list)
  - AIDS, HIV
  - Chlamydia
  - Gonorrhea
  - Herpes
  - Syphilis
  - Measles
  - Rabies
  - Smallpox
  - Typhus
  - Yellow Fever
- Type of link-tracing design
- Traced from infected nodes only

Three Random Processes

Treat in layers: Contact Formation, Disease Propagation, Sampling Propagation

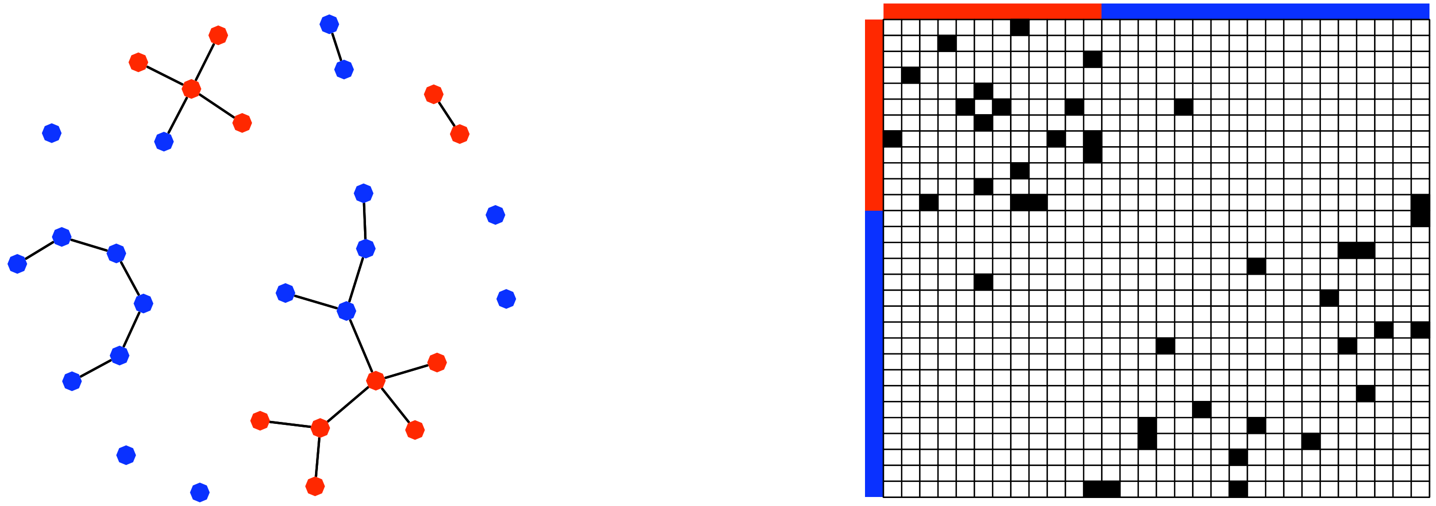


(a) Contacts

(b) Disease

(c) Sampling

Contact Tracing Sampling

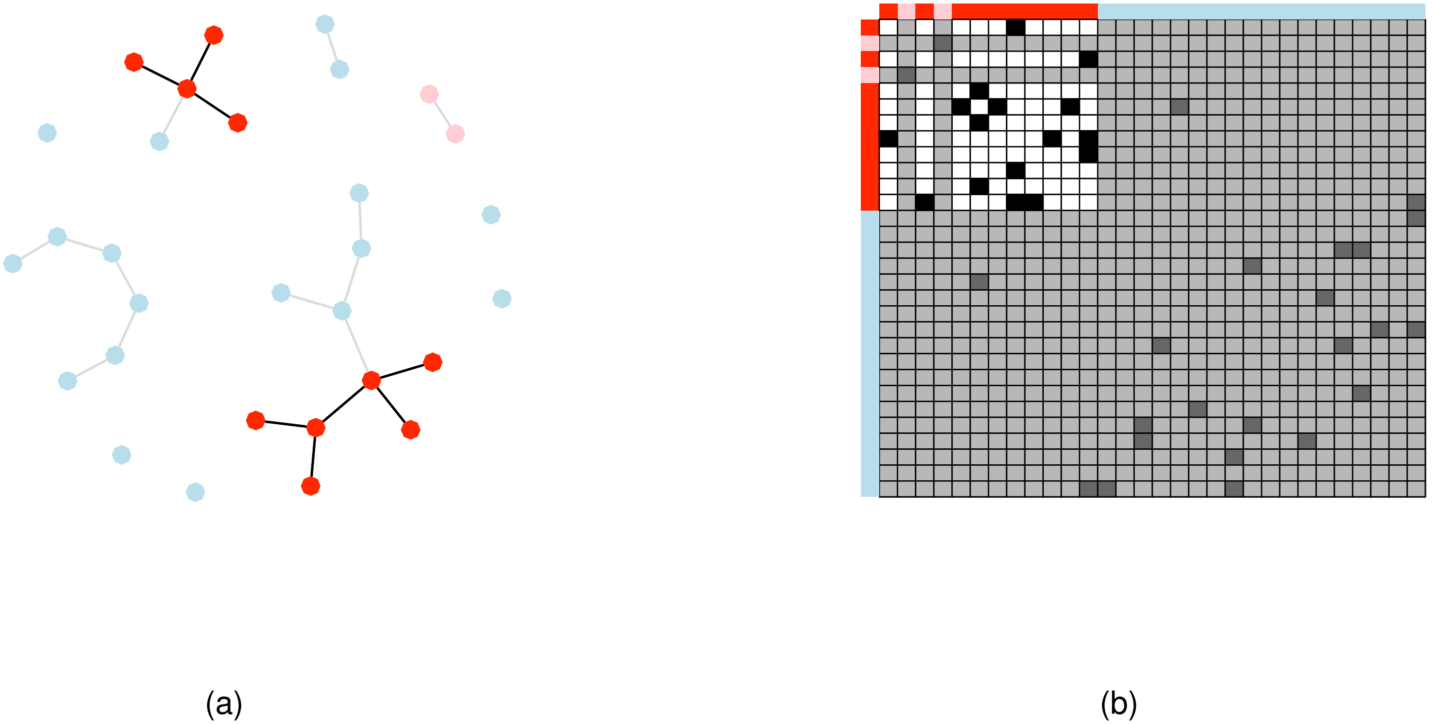


(d) Sociogram

(e) Sociomatrix

Figure 1: Full Network: Red Nodes Infected, Black squares are edges

Contact Tracing Design 1: Infected Only Sample



(a)

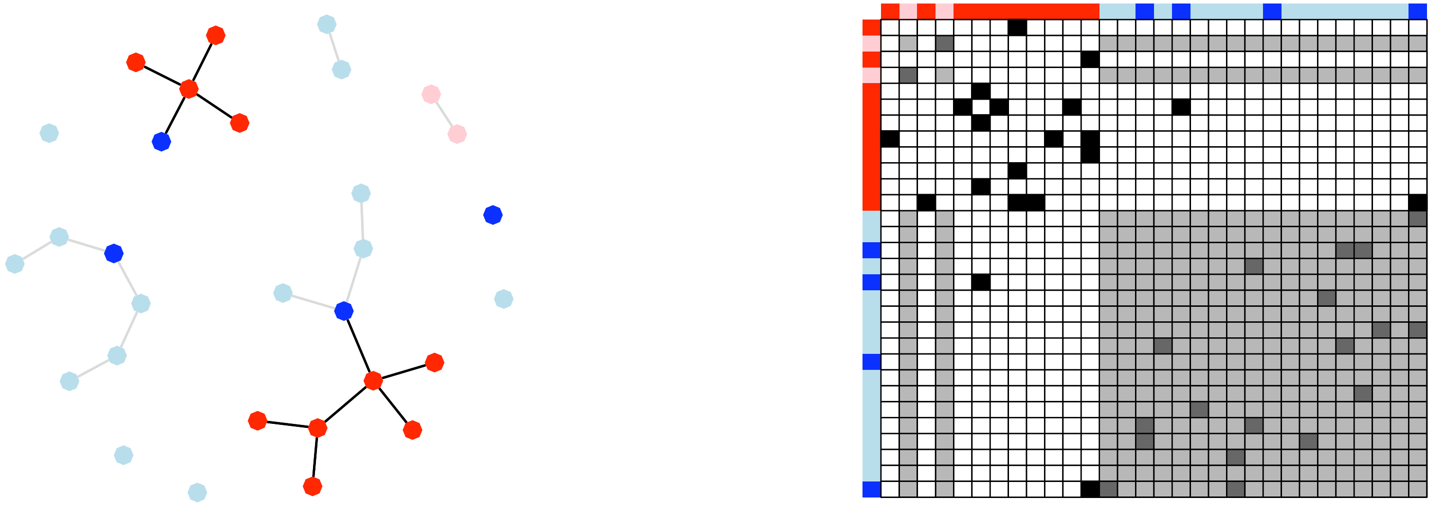
(b)

Figure 2: Design 1: Infected Only Sample

$$D = D_W = SS^T$$

Do not record any relations of uninfected individuals (as data currently exist).

Contact Tracing Design 2: Infected & Edge Units Sample



(a)

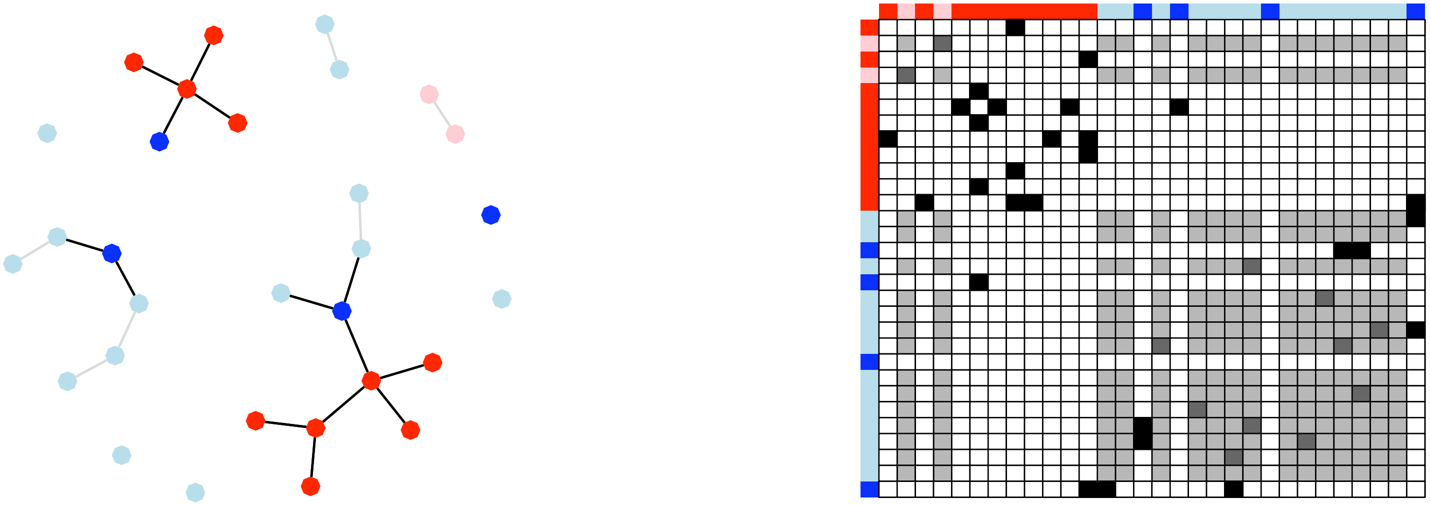
(b)

Figure 3: Design 2: Infected & Edge Units Sample

$$D = (S \cdot Z)1^T + 1(S \cdot Z)^T - (S \cdot Z)(S \cdot Z)^T, D_W = (S \cdot Z)1^T$$

Record all uninfected individuals tested.

Contact Tracing Design 3: Contacts of Edge Units Sample



(a)

(b)

Figure 4: Design 3: Contacts of Edge Units Sample

$$D = S1^T + 1S^T - SS^T, D_W = (S \cdot Z)1^T$$

Record relations of all individuals tested.

Contact Tracing Design 4: Full Contact Components Sample

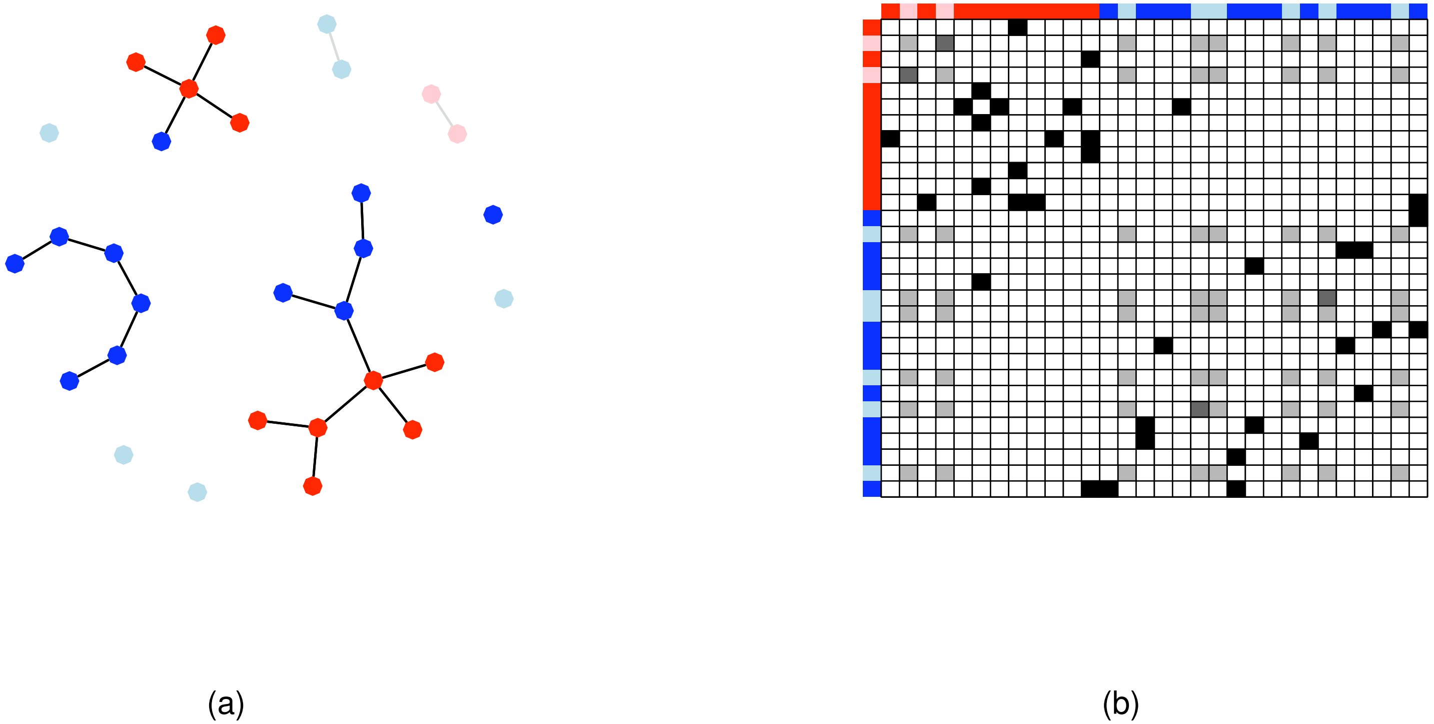


Figure 5: Design 4: Full Contact Components Sample

$$D = S1^T + 1S^T - SS^T, D_W = (S \cdot Z)1^T$$

Enroll any partners reported (most intrusive).

Epidemiological questions of interest

- What is the structure of possible disease-passing contacts in the population?
- What is the transmissibility of the disease?
- What is the epidemic potential in the population?

Contact Models

With parameters β, and covariate matrix X:

- Dyad Independent ERGM (logistic regression):

$$P(Y = y|X, \beta) = \prod_{i<j} \frac{\exp\{\beta^T X_{ij}\} Y_{ij}}{1 + \exp\{\beta^T X_{ij}\}}$$

- Inner Product Model:

$$P(Y = y|X, \beta) = \prod_{i<j} \frac{\exp\{\beta^T X_{ij} + \beta^* u_i u_j\} Y_{ij}}{1 + \exp\{\beta^T X_{ij} + \beta^* u_i u_j\}}$$

  Where u_i, u_j unobserved, assumed distributed $N(0, 1)$

- Dyad Dependent ERGM:

$$P(Y = y|X, \beta) = c^{-1} \exp\{\beta^T g(y, X)\}, c = \sum_w \exp\{\beta^T g(w, X)\}$$

  Where the normalizing constant is $c \equiv c(\beta)$ (sum over allowable graphs)

Modeling Disease Status Given Contact Structure

Disease Model:

$$P(Z, Z_0, W|\tau, \eta, Y) = \eta^{Z_0^T 1}(1-\eta)^{N-Z_0^T 1}\tau^{1^T W 1}(1-\tau)^{Z^T Y(1-Z)} \prod_{i:Z_i=1} \mathbb{I}_{(RZ_0)_i \geq 1)}$$

Where R is the reachability graph through transmitting arcs.

η	Probability of exogenous infection (from outside network)
τ	Transmissibility (probability of transmission)

Variable	Meaning	Dimension
Y	Sociomatrix of edges	$N \times N$
Z	Vector of infection	$N \times 1$
Z_0	Vector of exogenous infection	$N \times 1$
W	Matrix of transmissions	$N \times N$
Net	Contact and Disease: (Y, Z, Z_0, W)	

Discussion

Conclusions:

- Established a model-based frame for modeling contact and disease structure based on contact tracing data.
  - Estimate the structure of possible disease-passing contacts in the population
  - Estimate the transmissibility of the disease
  - Estimate the epidemic potential in the population

Limitations and Outstanding Questions:

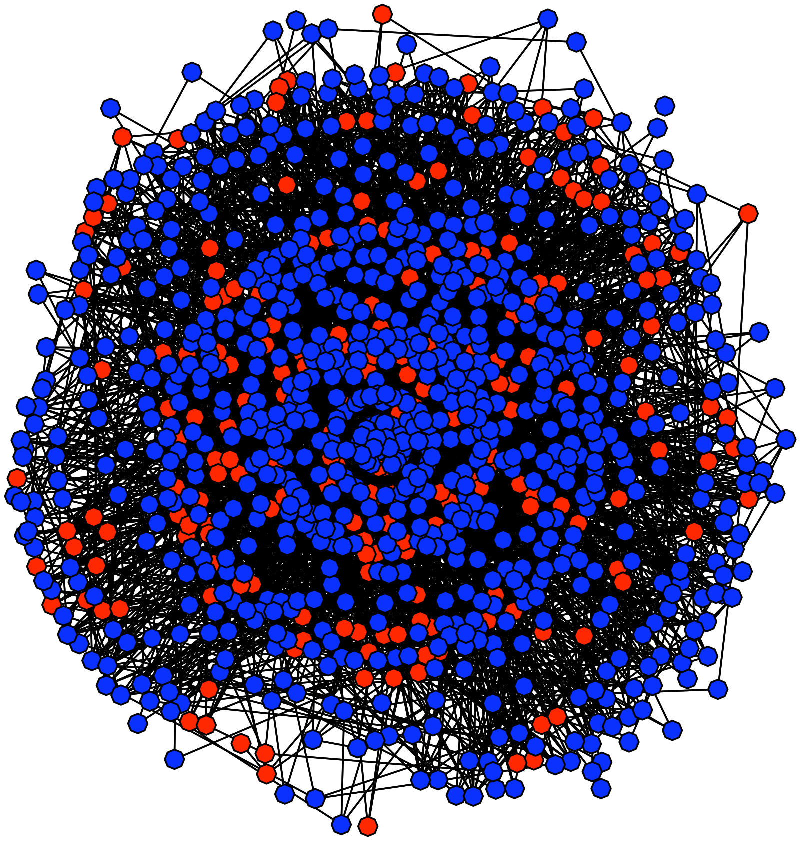
- Assumed MAR initial sample
  - *Is it possible to use auxiliary information to address NMAR?*
- Assumed known population size
  - *How often do we have a good estimate? Are there ways to estimate?*
- Ignored dynamics
  - *How critical is this limitation?*

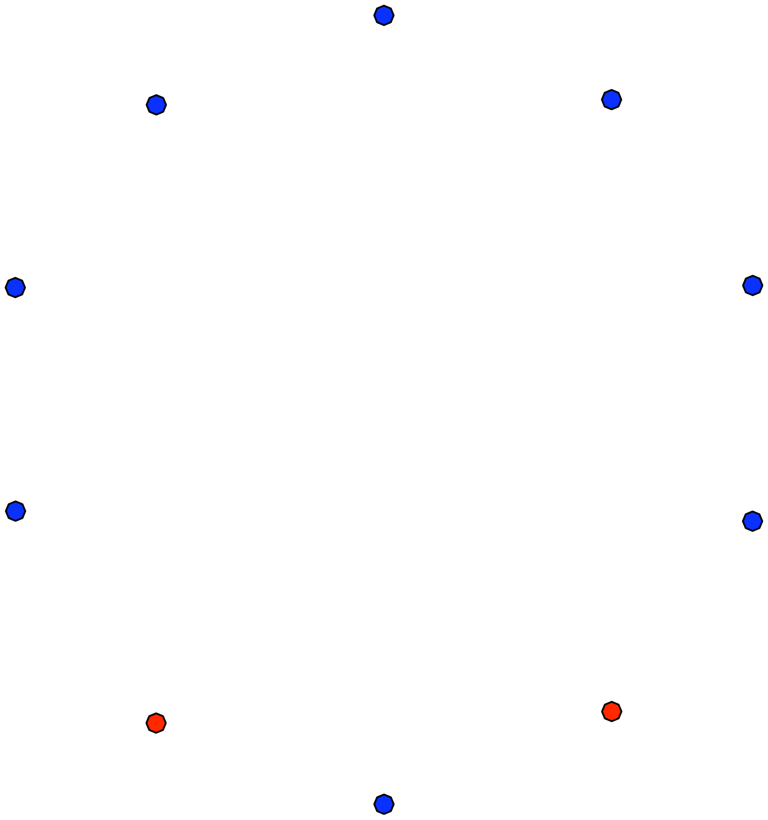
Respondent-Driven Sampling (RDS): Introduction

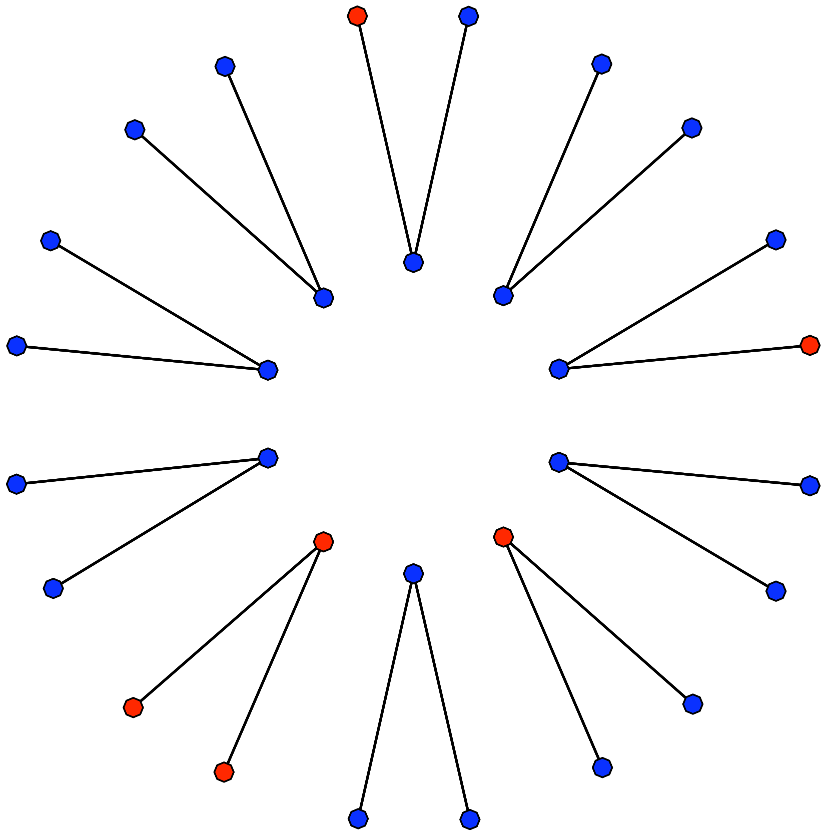
Example:
What proportion of Injection Drug Users in New York City are HIV positive?

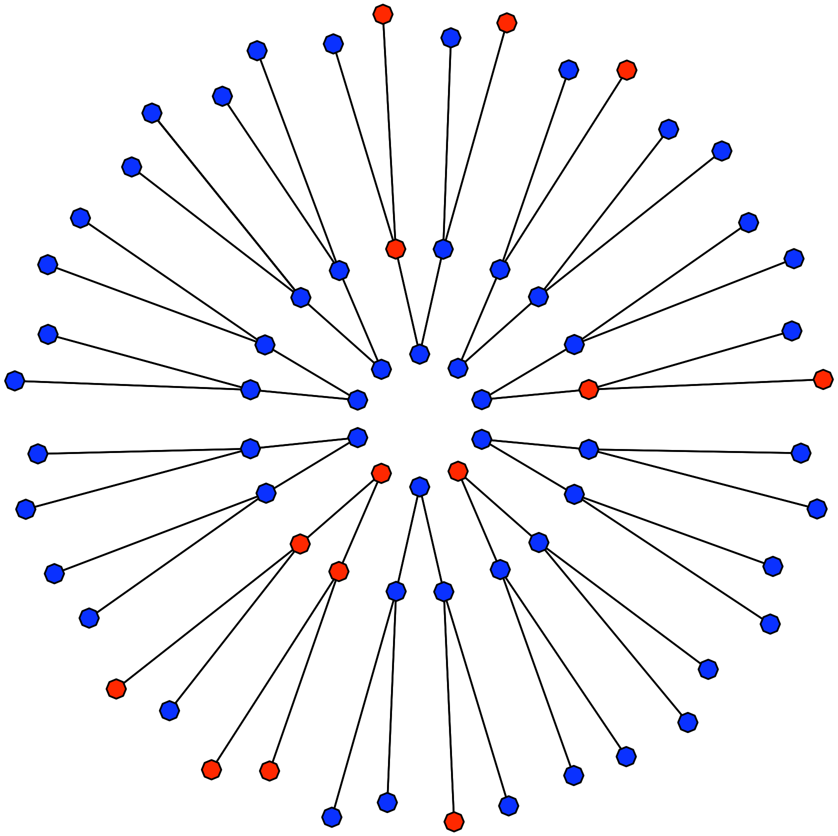
Hard-to-reach population

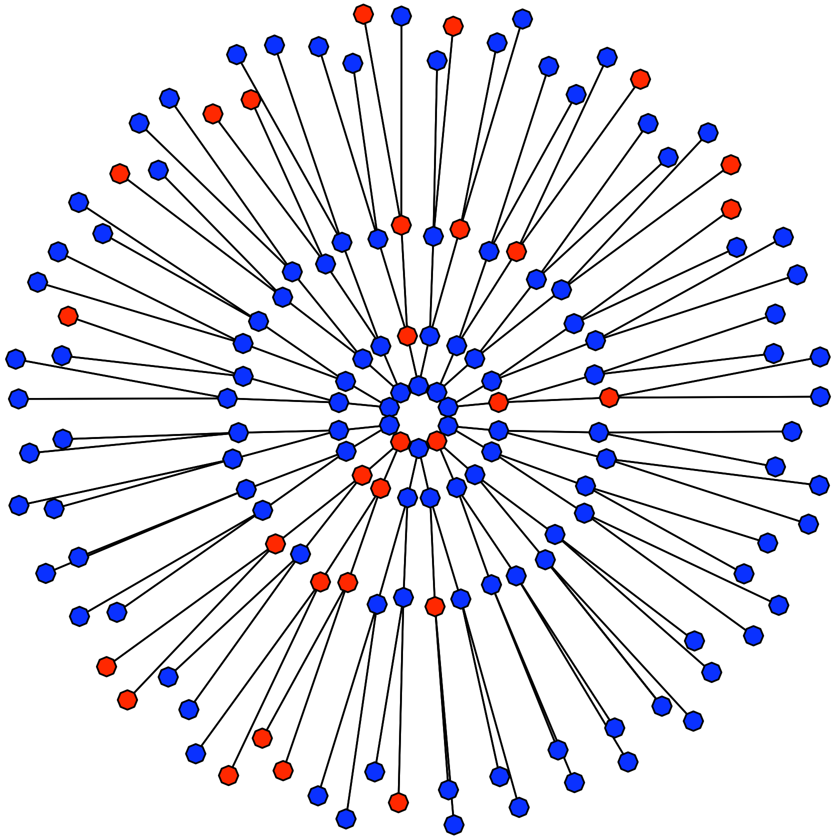
- Other Approaches:
  - Convenience samples of individuals (not probability sample)
  - Time-location samples (not probability sample of individuals)
  - Sample from larger existing sampling frame (too expensive)
- RDS: “Something like” probability sample

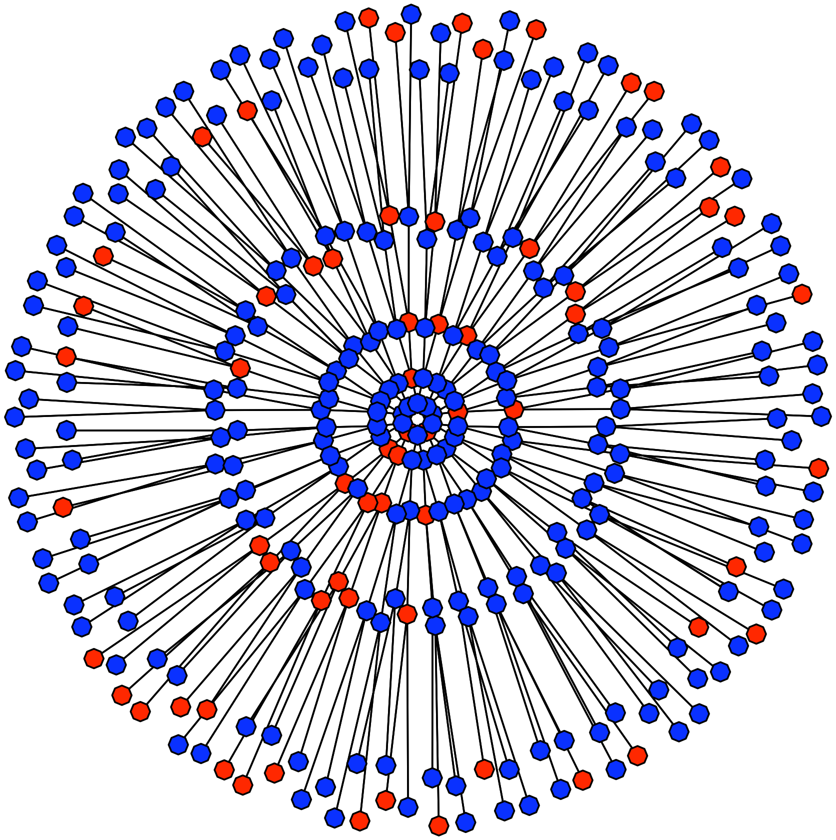


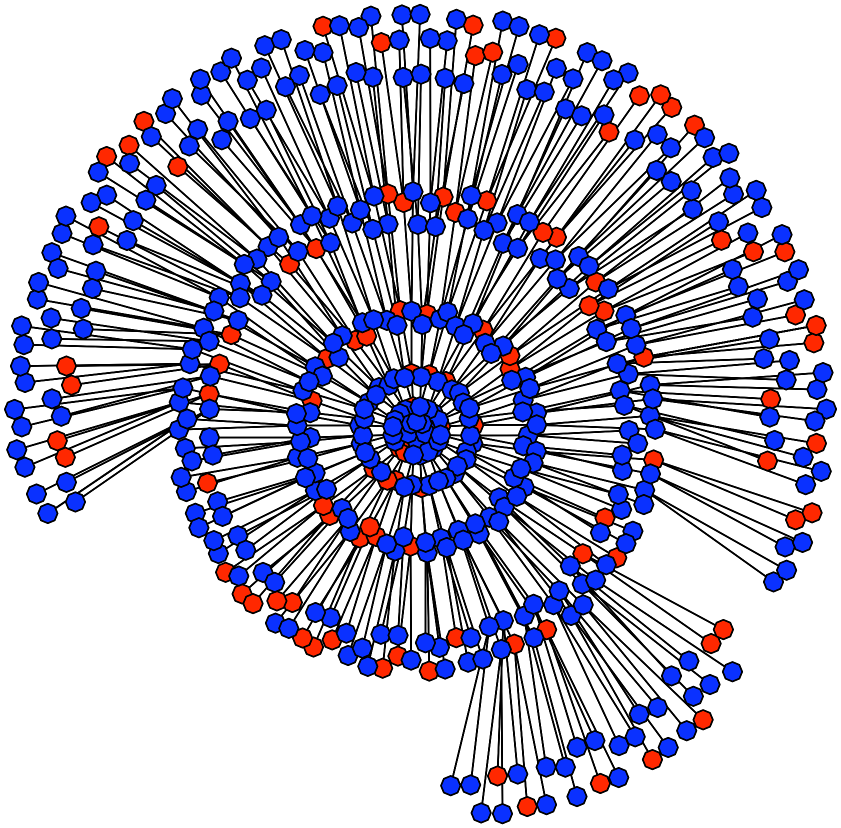


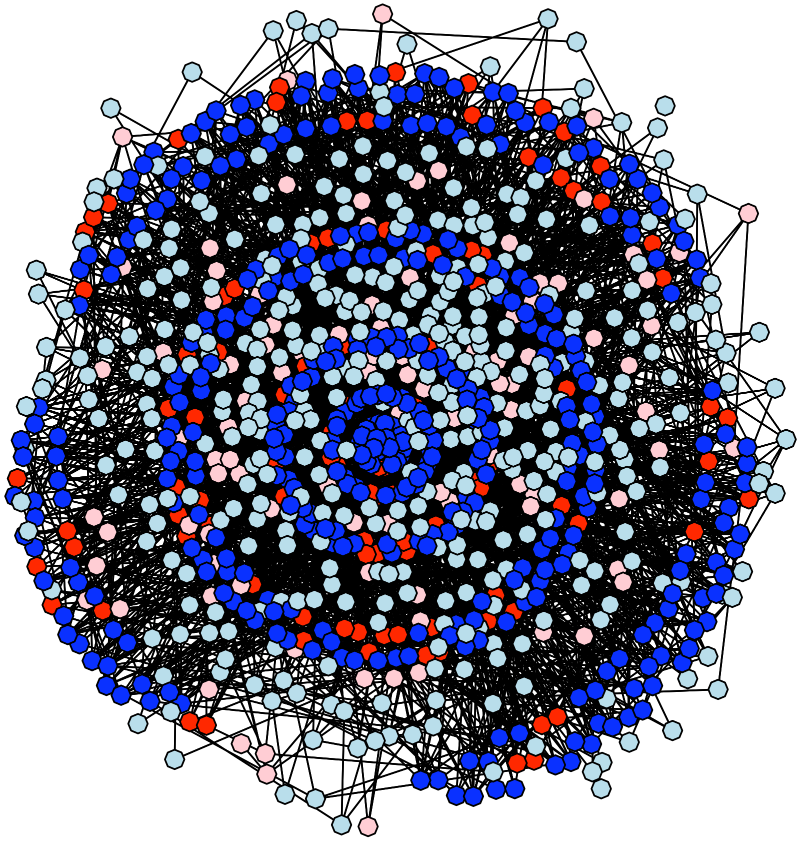












Sampling

Sampling:

- Begin with convenience sample of “seeds”
- Foster many waves of sampling to reduce dependence on convenience-sample seeds

- Good news: Large diverse samples in hard-to-reach populations!
- Bad news: Current inference problematic

Epidemiological questions of interest

- Characteristics of high-risk population
  - Proportion infected
  - Frequency of high-risk behaviors
- What is the structure of the social ties in the high-risk population
  - Note: network here is not strictly disease-contact

Structure of Analysis

Sample:

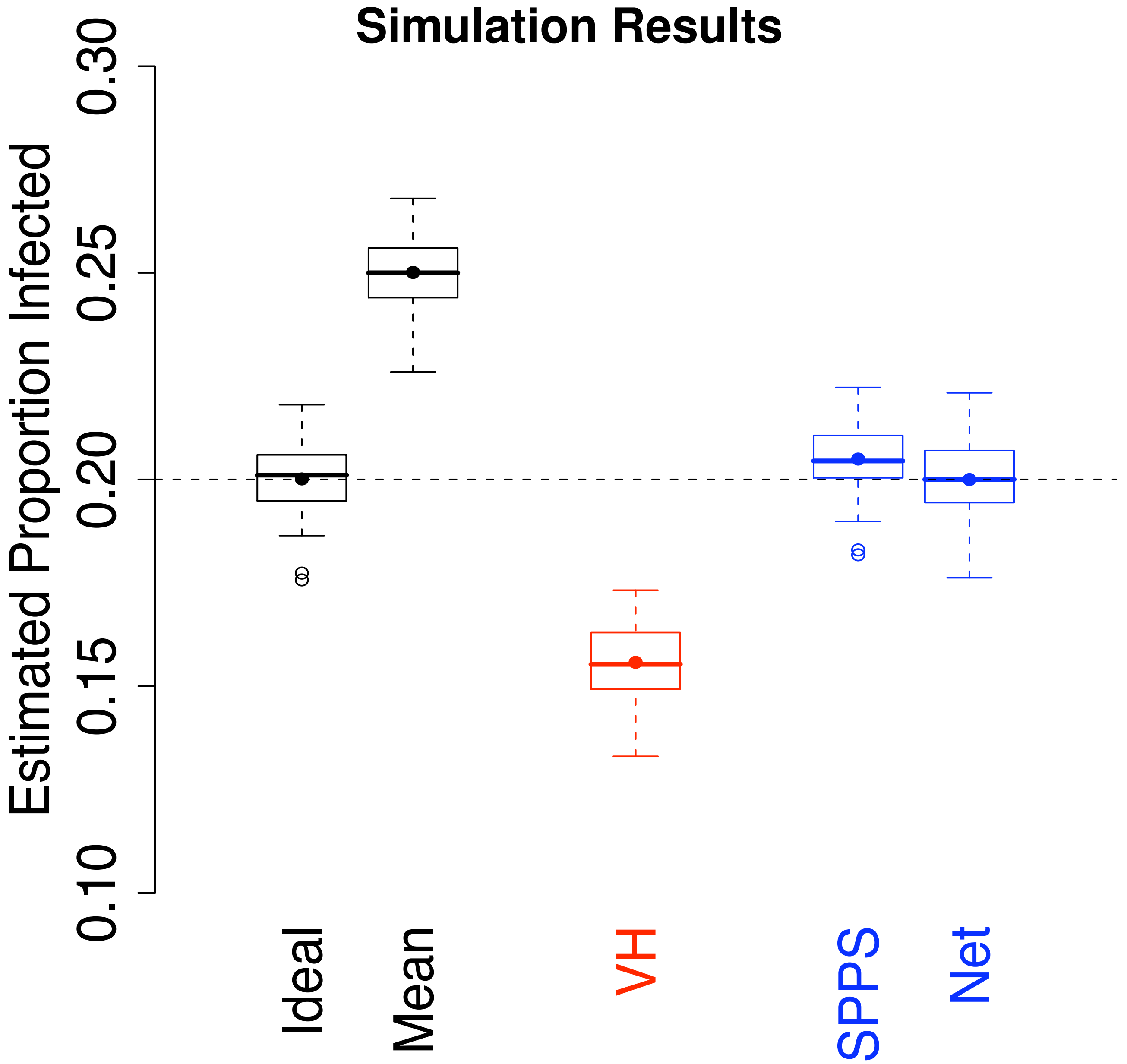
- Link-tracing sampling variant
- Ask number of contacts - but not who. Can’t identify alters.
- Network used as sampling tool

Existing Approach:

- Assume inclusion probability proportional to number of contacts (Volz and Heckathorn, 2008)
- Assume many waves of sampling remove bias of seed selection

Our work:

- Design-based (describe structure, not mechanism)
- Fit simple network model to observed data (model-assisted)
- Correct for biases due to network-based sampling, and observable irregularities


Simulation Results
Estimated Proportion Infected
0.10
0.15
0.20
0.25
0.30
Ideal
Mean
VH
SPPS
Net

Discussion

Conclusions:

- Can estimate nodal proportions of interest
  - Proportions infected
  - Frequencies of high-risk behaviors
- Network-Model estimator corrects for differential activity by infection status, unlike sample mean.
- Network-Model estimator uses appropriate sample weights for simulated high sample fraction, unlike sample mean or Volz-Heckathorn estimator.
- Network-Model estimator corrects for seed bias, unlike any existing method.

Limitations:

- Assume full network size known (subject of ongoing research)
- Can only correct for *observable* sampling biases
- Uncertainty may be quite high
- Computationally expensive

Discussion

- Network models can be applied to data from link-tracing samples to address scientific questions about the full population.
  - Contact Tracing
  - Respondent-Driven Sampling
- Some forms of additional information collected in the study can greatly improve possibilities for inference.
  - Edge unit information
  - Measurement of sampling biases
  - Any characteristics of unobserved population
- All models fit with Exponential-Family Random Graph Models using `statnet` R software.

Outstanding Issues:

- Unknown Network Size
- Boundary Specification Problem

References

- **Missing Data and Sampling**
  - Little, R. J.A. and D. B. Rubin, Second Edition (2002). *Statistical Analysis with Missing Data*, John Wiley and Sons, Hoboken, NJ.
  - Thompson, S.K., and G.A. Seber (1996). *Adaptive Sampling* John Wiley and Sons, Inc. New York.
- **Modeling Social Network Data with Exponential-Family Random Graph Models**
  - Handcock, M.S., D.R. Hunter, C.T. Butts, S.M. Goodreau, and M. Morris (2003) `statnet`: An R package for the Statistical Modeling of Social Networks. URL: `http://www.csde.washington.edu/statnet`.
  - Holland, P.W., and S. Leinhardt (1981), An exponential family of probability distributions for directed graphs, *Journal of the American Statistical Association*, **76**: 33-50.
  - Snijders, T.A.B., P.E. Pattison, G.L. Robins, and M.S. Handcock (2006). New specifications for exponential random graph models. *Sociological Methodology*, 99-153.
- **Inference with Partially-Observed Network Data**
  - Frank, O. (1971). *The Statistical Analysis of Networks* Chapman and Hall, London.
  - Frank, O., and T.A.B. Snijders (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, **10**: 53-67.
  - Gile, K.J. (2008). Inference from Partially-Observed Network Data. PhD. Dissertation. University of Washington, Seattle.
  - Gile, K. and M.S. Handcock (2006). Model-based Assessment of the Impact of Missing Data on Inference for Networks. Working paper, Center for Statistics and the Social Sciences, University of Washington.
  - Handcock, M.S., and K. Gile (2007). Modeling social networks with sampled data. Technical Report, Department of Statistics, University of Washington.
  - Thompson, S.K. and O. Frank (2000). Model-Based Estimation With Link-Tracing Sampling Designs. *Survey Methodology* , **26**: 87-98.
- **Other**
  - Harris, K. M., F. Florey, J. Tabor, P. S. Bearman, J. Jones, and R. J. Udry (2003). The National Longitudinal Study of Adolescent Health: Research design. Technical Report, Carolina Population Center, University of North Carolina at Chapel Hill.

E-mail: krista.gile@nuffield.ox.ac.uk

Thank you for your attention!