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:


¹Research 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 \((D)\).

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

\[
= \sum_{Unobserved} P(D|Y, \delta) P(Y|\beta)
\]

• \(\beta\) is the model parameter
• \(\delta\) 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
Using link-tracing data to inform epidemiology

**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

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

\[ D = (S \cdot Z)^T + 1(S \cdot Z)^T - (S \cdot Z)(S \cdot Z)^T, \quad D_W = (S \cdot Z)^T \]

Record all uninfected individuals tested.
Contact Tracing Design 3: Contacts of Edge Units Sample

Figure 4: Design 3: Contacts of Edge Units Sample

\[ D = S1^T + 1S^T - SS^T, \quad DW = (S \cdot Z)1^T \]

Record relations of all individuals tested.
Contact Tracing Design 4: Full Contact Components Sample

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

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 $\beta$, 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:

\[
P(Z, Z_0, W | \tau, \eta, Y) = \eta Z_0^T T_1 (1 - \eta) N - Z_0^T_1 \tau^T W_1 (1 - \tau) Z^T Y (1 - Z) \prod_{i:Z_i=1} \mathbb{I}(RZ_0)_i \geq 1
\]

Where \( R \) is the reachability graph through transmitting arcs.

\( \eta \)  Probability of exogenous infection (from outside network)
\( \tau \)  Transmissibility (probability of transmission)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Meaning</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td>Sociomatrix of edges</td>
<td>( N \times N )</td>
</tr>
<tr>
<td>( Z )</td>
<td>Vector of infection</td>
<td>( N \times 1 )</td>
</tr>
<tr>
<td>( Z_0 )</td>
<td>Vector of exogenous infection</td>
<td>( N \times 1 )</td>
</tr>
<tr>
<td>( W )</td>
<td>Matrix of transmissions</td>
<td>( N \times 1 )</td>
</tr>
<tr>
<td>( Net )</td>
<td>Contact and Disease: ((Y, Z, Z_0, W))</td>
<td>( N \times N )</td>
</tr>
</tbody>
</table>
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
Using link-tracing data to inform epidemiology
Using link-tracing data to inform epidemiology
Using link-tracing data to inform epidemiology
Using link-tracing data to inform epidemiology
Using link-tracing data to inform epidemiology
Using link-tracing data to inform epidemiology
Using link-tracing data to inform epidemiology
Using link-tracing data to inform epidemiology
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
Using link-tracing data to inform epidemiology

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
Using link-tracing data to inform epidemiology

Simulation Results

Estimated Proportion Infected

0.10 0.15 0.20 0.25 0.30

Ideal Mean

SPPS Net
Using link-tracing data to inform epidemiology

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
Using link-tracing data to inform epidemiology

References

- **Missing Data and Sampling**

- **Modeling Social Network Data with Exponential-Family Random Graph Models**

- **Inference with Partially-Observed Network Data**

- **Other**

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

Thank you for your attention!