Analysis of Cross-Sectional Networks with Missing and Sampled Data

Krista J. Gile
Nuffield College, Oxford

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

MURI Kickoff Meeting, November 18, 2008

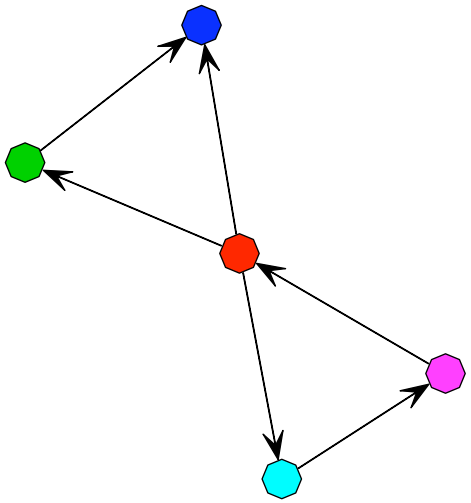
For details, see:

- Gile, K. and Handcock, M.S. (2006). Model-based Assessment of the Impact of Missing Data on Inference for Networks. Working Paper #66, Center for Statistics and the Social Sciences, University of Washington. (http://www.csss.washington.edu)[1]
- Handcock, M.S., and Gile, K.J. (2007). Modeling social networks with sampled data. Technical Report #523, Department of Statistics, University of Washington. (http://www.stat.washington.edu)
- Gile, K.J. (2008). Inference from Partially-Observed Network Data. PhD. Dissertation. University of Washington, Seattle.

[1]Research supported by NICHD grant 7R29HD034957 and NIDA 7R01DA012831, and ONR award N00014-08-1-1015.

(Cross-Sectional) Social Networks

- Social Network: Tool to formally represent and quantify relational social structure.
- Relations can include: friendships, workplace collaborations, international trade
- Represent mathematically as a sociomatrix, Y, where
  $Y_{ij} =$ the value of the relationship from i to j



(a) Sociogram

	0	1	1	1	0
	0	0	1	0	0
	0	0	0	0	0
	0	0	0	0	1
	1	0	0	0	0

(b) Sociomatrix

Partially-Observed Social Network Data

Some portion of the social network is often unobserved.

$Y =$

	A	B	C	D
A	-	1	0	0
B	0	-	1	1
C	0	0	-	0
D	1	1	1	-

$Y_{obs} =$

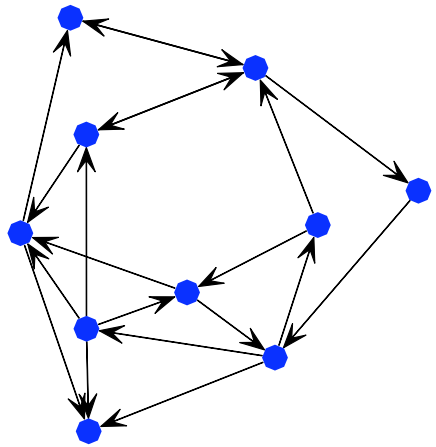
	A	B	C	D
A	-	?	?	?
B	?	-	?	?
C	0	0	-	0
D	1	1	1	-

$D =$

	A	B	C	D
A	-	0	0	0
B	0	-	0	0
C	1	1	-	1
D	1	1	1	-

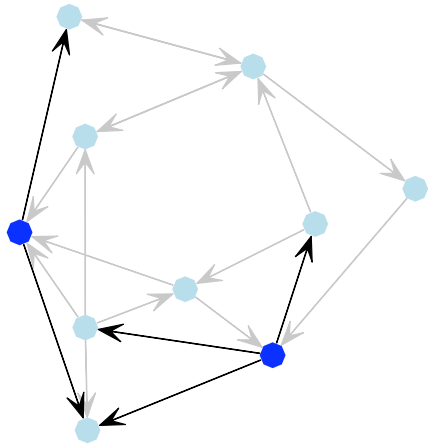
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



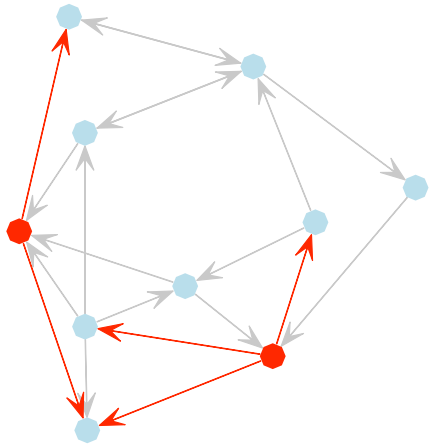
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



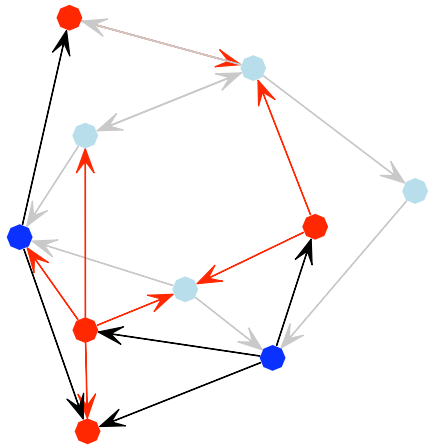
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



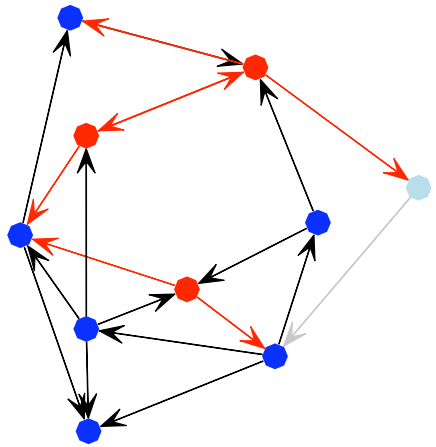
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



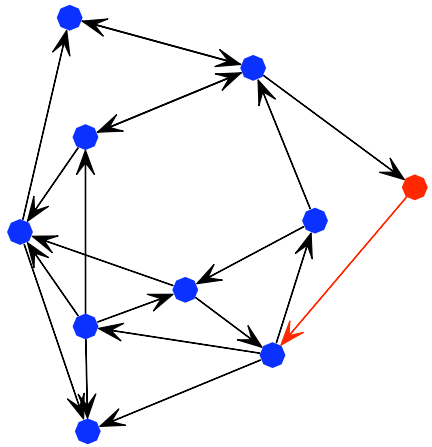
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



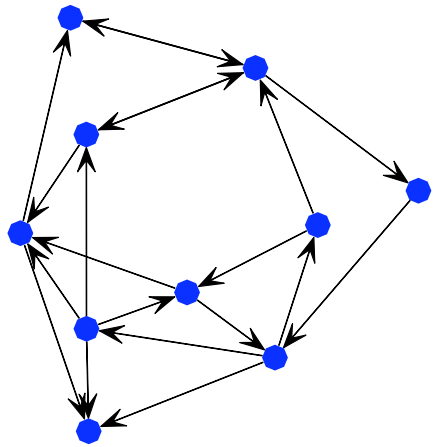
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



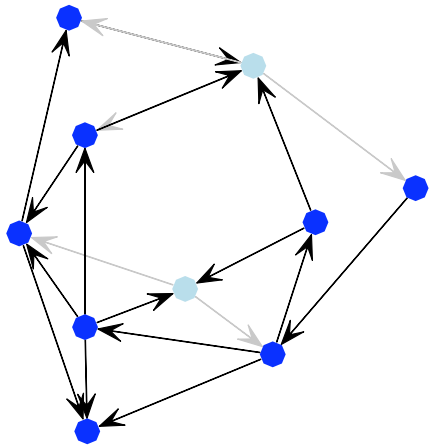
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



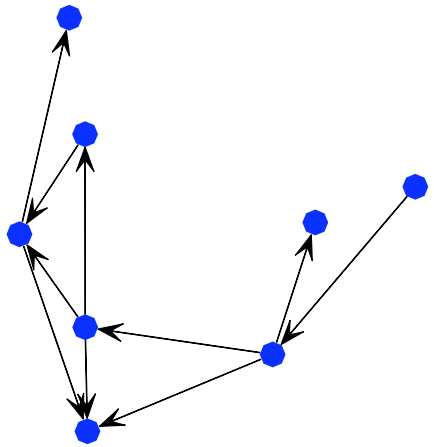
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



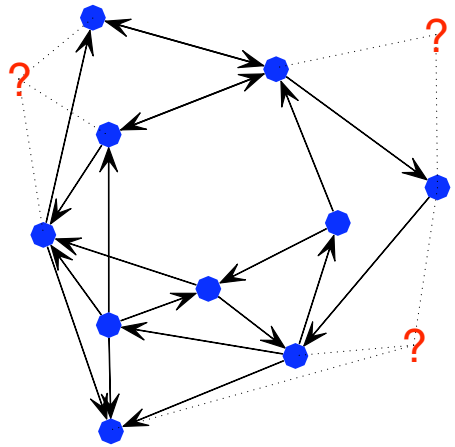
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



Frameworks for Statistical Analysis

	Describe Structure	Describe Mechanism
Fully Observed Data	Description	Modeling (Statistical)
Partially Observed Data	Design-Based Inference	Likelihood Inference

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)

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*)

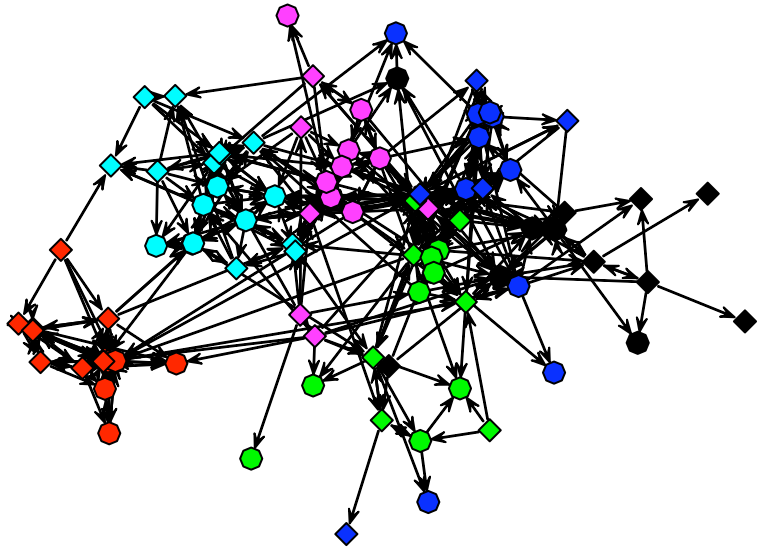
Then

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

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

Example: Friendships in a School

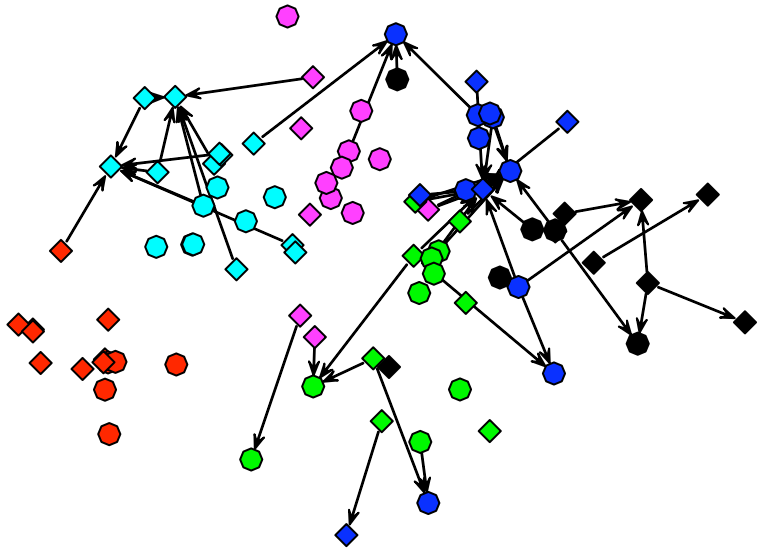
From the National Longitudinal Survey on Adolescent Health - Wave 1:



- Each student asked to nominate up to 5 male and 5 female friends
- Sex and Grade available for 89 students, 70 students reported friendships.

Example: Friendships in a School

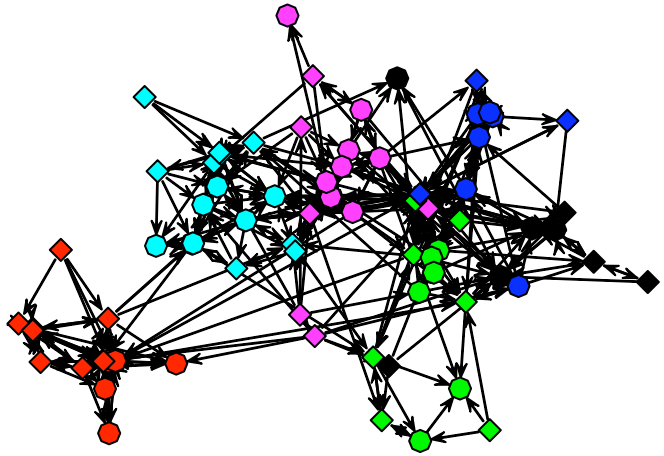
From the National Longitudinal Survey on Adolescent Health - Wave 1:



- Each student asked to nominate up to 5 male and 5 female friends
- Sex and Grade available for 89 students, 70 students reported friendships.

Example: Friendships in a School

From the National Longitudinal Survey on Adolescent Health - Wave 1:



- Each student asked to nominate up to 5 male and 5 female friends
- Sex and Grade available for 89 students, 70 students reported friendships.

Example: Friendships in a School

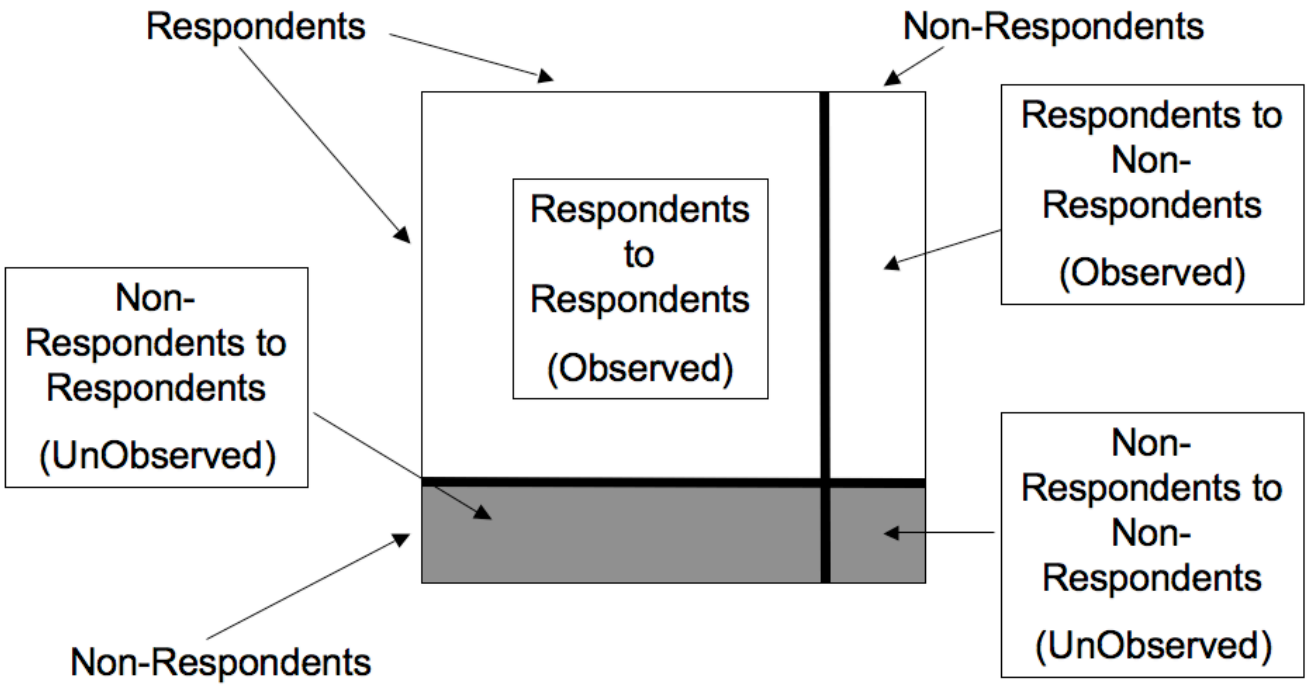
- **Scientific Question:** Do friendships form in an egalitarian or an hierarchal manner?
- **Methodological Question:** Can we fit a network model to a network with missing data? Is the fit different from that of just the observed data?

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

Does observed status depend on unobserved characteristics?

Structure of Data

- Up to 5 female friends and up to 5 male friends
- 89 students in school
- 70 completed friendship nominations portion of survey



Example: Friendships in a School

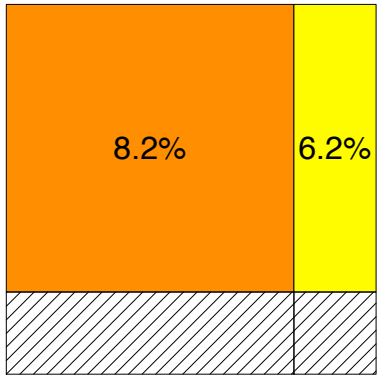
Fit an ERGM to the partially observed data, get coefficients like in logistic regression.

Terms in the model:

- **Density**: Overall rate of ties
- **Reciprocity**: Do students tend to reciprocate nominations?
- **Popularity by Grade**: Do students in different grades receive different rates of ties?
- **Popularity by Sex**: Do boys and girls receive different rates of ties?
- **Age:Sex Mixing**: Rates of ties between older and younger boys and girls
- Propensity for ties within sex and grade to be **transitive** (hierarchal)
- Propensity for ties within sex and grade to be **cyclical** (egalitarian)
- **Isolation**: Propensity for students to receive no nominations

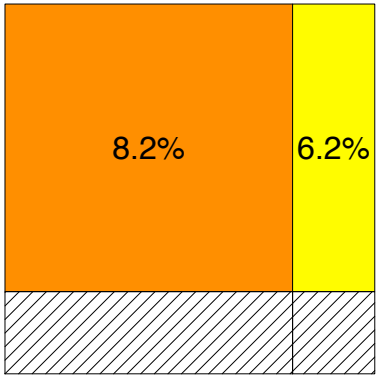
Percent of Possible Relations Realized

	Observed
Respondents to Respondents	8.2
Respondents to Non-Respondents	6.2
Non-Respondents to Respondents	-
Non-Respondents to Non-Respondents	-

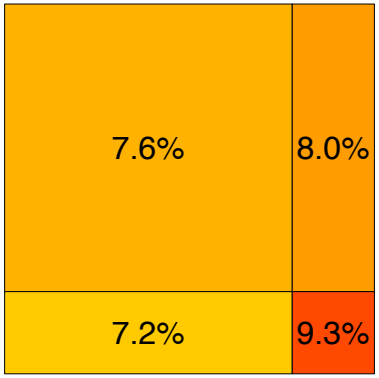


Goodness of Fit: Percent of Possible Relations Realized

	Observed	Fit
Respondents to Respondents	8.2	7.6
Respondents to Non-Respondents	6.2	8.0
Non-Respondents to Respondents	-	7.2
Non-Respondents to Non-Respondents	-	9.3



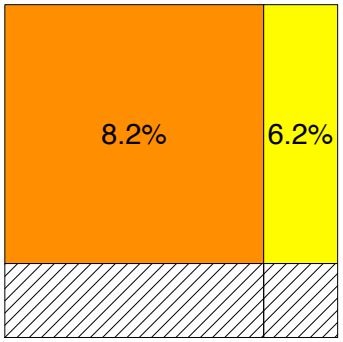
(c) Observed



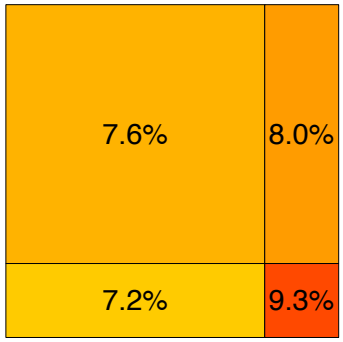
(d) Fit

Goodness of Fit: Percent of Possible Relations Realized

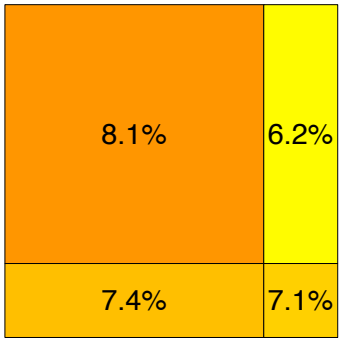
	Observed	Original	Diff. Popularity
Respondents to Respondents	8.2	7.6	8.1
Respondents to Non-Respondents	6.2	8.0	6.2
Non-Respondents to Respondents	-	7.2	7.4
Non-Respondents to Non-Respondents	-	9.3	7.1



(e) Observed



(f) Original



(g) Differential Popularity

	coefficient	s.e.
Density	−1.138	0.19***
Sex and Grade Factors		
Grade 8 Popularity	−0.178	0.14
Grade 9 Popularity	−0.420	0.16**
Grade10 Popularity	−0.339	0.16*
Grade 11 Popularity	0.256	0.19
Grade 12 Popularity	0.243	0.20
Male Popularity	0.779	0.17***
Non-Resp Popularity	−0.322	0.10**
Sex and Grade Mixing		
Girl to Same Grade Boy	0.308	0.23
Boy to Same Grade Girl	−0.453	0.23*
Girl to Older Girl	−1.406	0.16***
Girl to Younger Girl	−1.873	0.21***
Girl to Older Boy	−1.412	0.14***
Girl to Younger Boy	−2.129	0.24***
Boy to Older Boy	−1.444	0.16***
Boy to Younger Boy	−2.788	0.35***
Boy to Older Girl	−1.017	0.14***
Boy to Younger Girl	−1.660	0.18***
Mutuality	3.290	0.22***
Transitivity		
Transitive Same Sex and Grade	0.844	0.04***
Cyclical Same Sex and Grade	−1.965	0.16***
Isolation	5.331	0.64***

	coefficient	s.e.
Density	−1.138	0.19***
Sex and Grade Factors		
Grade 8 Popularity	−0.178	0.14
Grade 9 Popularity	−0.420	0.16**
Grade10 Popularity	−0.339	0.16*
Grade 11 Popularity	0.256	0.19
Grade 12 Popularity	0.243	0.20
Male Popularity	0.779	0.17***
Non-Resp Popularity	−0.322	0.10**
Sex and Grade Mixing		
Girl to Same Grade Boy	0.308	0.23
Boy to Same Grade Girl	−0.453	0.23*
Girl to Older Girl	−1.406	0.16***
Girl to Younger Girl	−1.873	0.21***
Girl to Older Boy	−1.412	0.14***
Girl to Younger Boy	−2.129	0.24***
Boy to Older Boy	−1.444	0.16***
Boy to Younger Boy	−2.788	0.35***
Boy to Older Girl	−1.017	0.14***
Boy to Younger Girl	−1.660	0.18***
Mutuality	3.290	0.22***
Transitivity		
Transitive Same Sex and Grade	0.844	0.04***
Cyclical Same Sex and Grade	−1.965	0.16***
Isolation	5.331	0.64***

	coefficient	s.e.
Density	−1.138	0.19***
Sex and Grade Factors		
Grade 8 Popularity	−0.178	0.14
Grade 9 Popularity	−0.420	0.16**
Grade10 Popularity	−0.339	0.16*
Grade 11 Popularity	0.256	0.19
Grade 12 Popularity	0.243	0.20
Male Popularity	0.779	0.17***
Non-Resp Popularity	−0.322	0.10**
Sex and Grade Mixing		
Girl to Same Grade Boy	0.308	0.23
Boy to Same Grade Girl	−0.453	0.23*
Girl to Older Girl	−1.406	0.16***
Girl to Younger Girl	−1.873	0.21***
Girl to Older Boy	−1.412	0.14***
Girl to Younger Boy	−2.129	0.24***
Boy to Older Boy	−1.444	0.16***
Boy to Younger Boy	−2.788	0.35***
Boy to Older Girl	−1.017	0.14***
Boy to Younger Girl	−1.660	0.18***
Mutuality	3.290	0.22***
Transitivity		
Transitive Same Sex and Grade	0.844	0.04***
Cyclical Same Sex and Grade	−1.965	0.16***
Isolation	5.331	0.64***

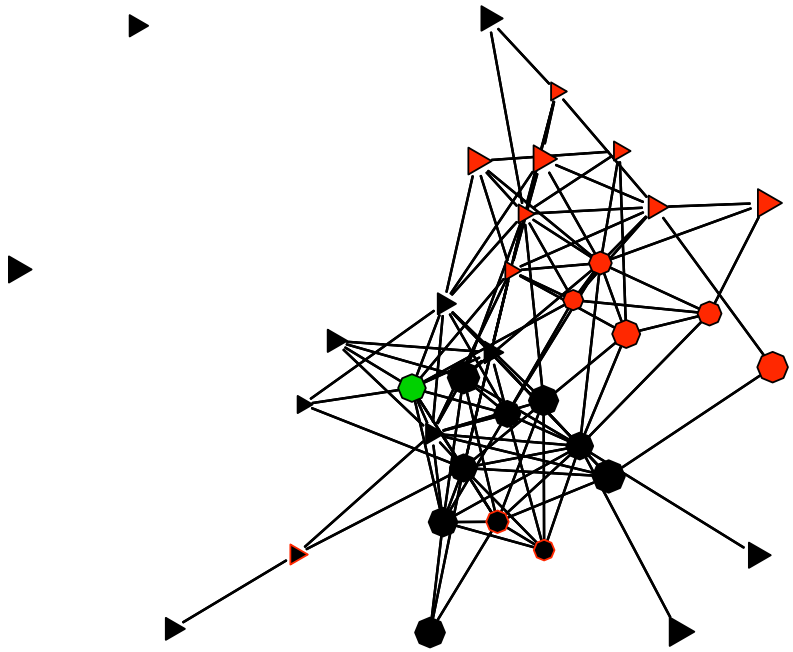
	coefficient	s.e.
Density	−1.138	0.19***
Sex and Grade Factors		
Grade 8 Popularity	−0.178	0.14
Grade 9 Popularity	−0.420	0.16**
Grade10 Popularity	−0.339	0.16*
Grade 11 Popularity	0.256	0.19
Grade 12 Popularity	0.243	0.20
Male Popularity	0.779	0.17***
Non-Resp Popularity	−0.322	0.10**
Sex and Grade Mixing		
Girl to Same Grade Boy	0.308	0.23
Boy to Same Grade Girl	−0.453	0.23*
Girl to Older Girl	−1.406	0.16***
Girl to Younger Girl	−1.873	0.21***
Girl to Older Boy	−1.412	0.14***
Girl to Younger Boy	−2.129	0.24***
Boy to Older Boy	−1.444	0.16***
Boy to Younger Boy	−2.788	0.35***
Boy to Older Girl	−1.017	0.14***
Boy to Younger Girl	−1.660	0.18***
Mutuality	3.290	0.22***
Transitivity		
Transitive Same Sex and Grade	0.844	0.04***
Cyclical Same Sex and Grade	−1.965	0.16***
Isolation	5.331	0.64***

Conclusions, School Friendships Example

- Nominations are reciprocated at a higher rate than random
- Males receive nominations from other males at a higher rate than females from females
- Nominations within grade are more likely than outside grade
- Nominations of older students are more likely than younger students
- Nominations within sex and grade are more consistent with a hierarchal rather than egalitarian structure
- More students receive no nominations than we would expect at random.

Law Firm Collaboration Example

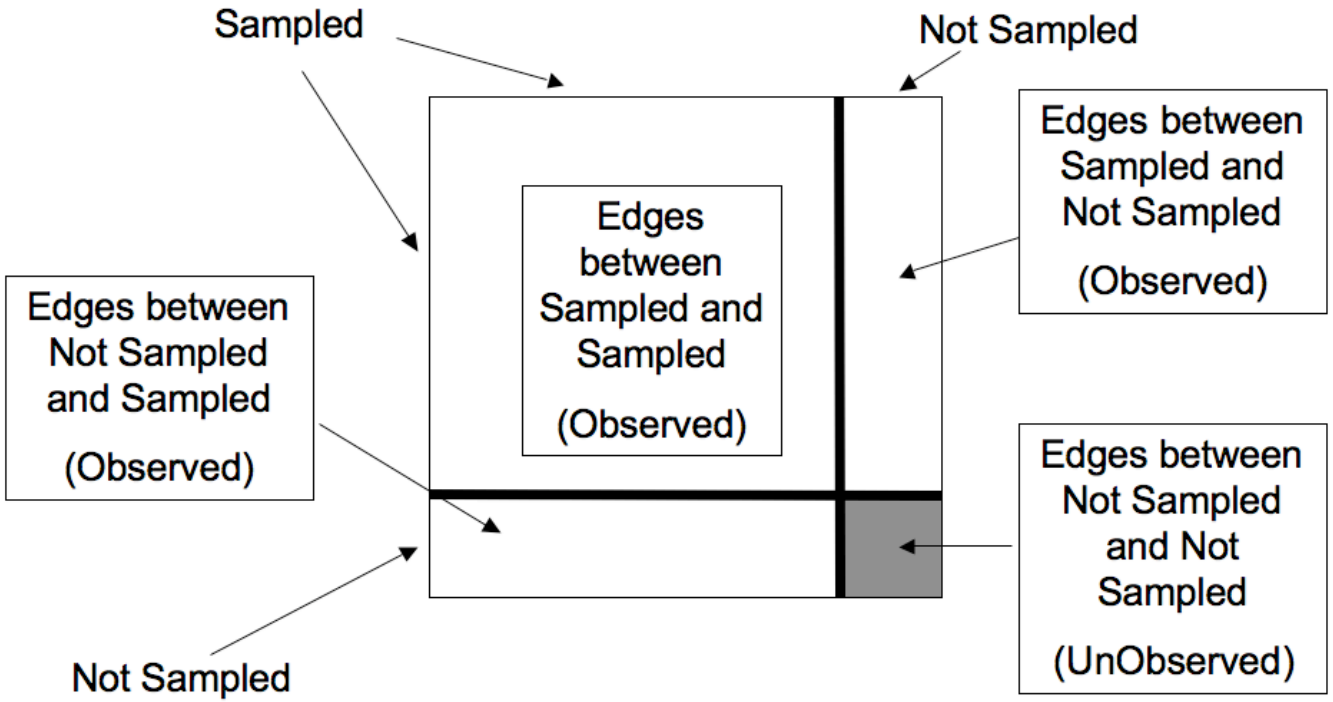
From the Emmanuel Lazega's study of a Corporate Law Firm:



- Each partner asked to identify the others with whom (s)he collaborated.
- Seniority, Sex, Practice (corporate or litigation) and Office (3 locations) available for all 36 partners.
- Simulated sampling: Start with 2 partners and include all their collaborators, as well as all collaborators of their collaborators.

Structure of Data

- 36 partners total, each reported all their collaborations
- Simulated samples: each begins with 2 seeds, samples 2 waves
- Between 2 (once) and 36 (3 times) partners sampled among 630 possible samples



Law Firm Collaboration Example

- **Scientific Question:** Do collaborations happen more often within the same practice, controlling for location and clustering?
- **Methodological Question:** Can we fit a network model to a network sampled by link-tracing?

$P(D|Y,\delta) = P(D|Y_{obs},\delta)$ (*adaptive sampling*)

Does observed status depend on unobserved quantities?

$$P(D|Y,\delta) = P(seeds)P(D|Y,\delta,seeds) = P(seeds)P(D|Y_{obs},\delta,seeds)$$

So if initial sample missing at random, link-tracing adaptive.

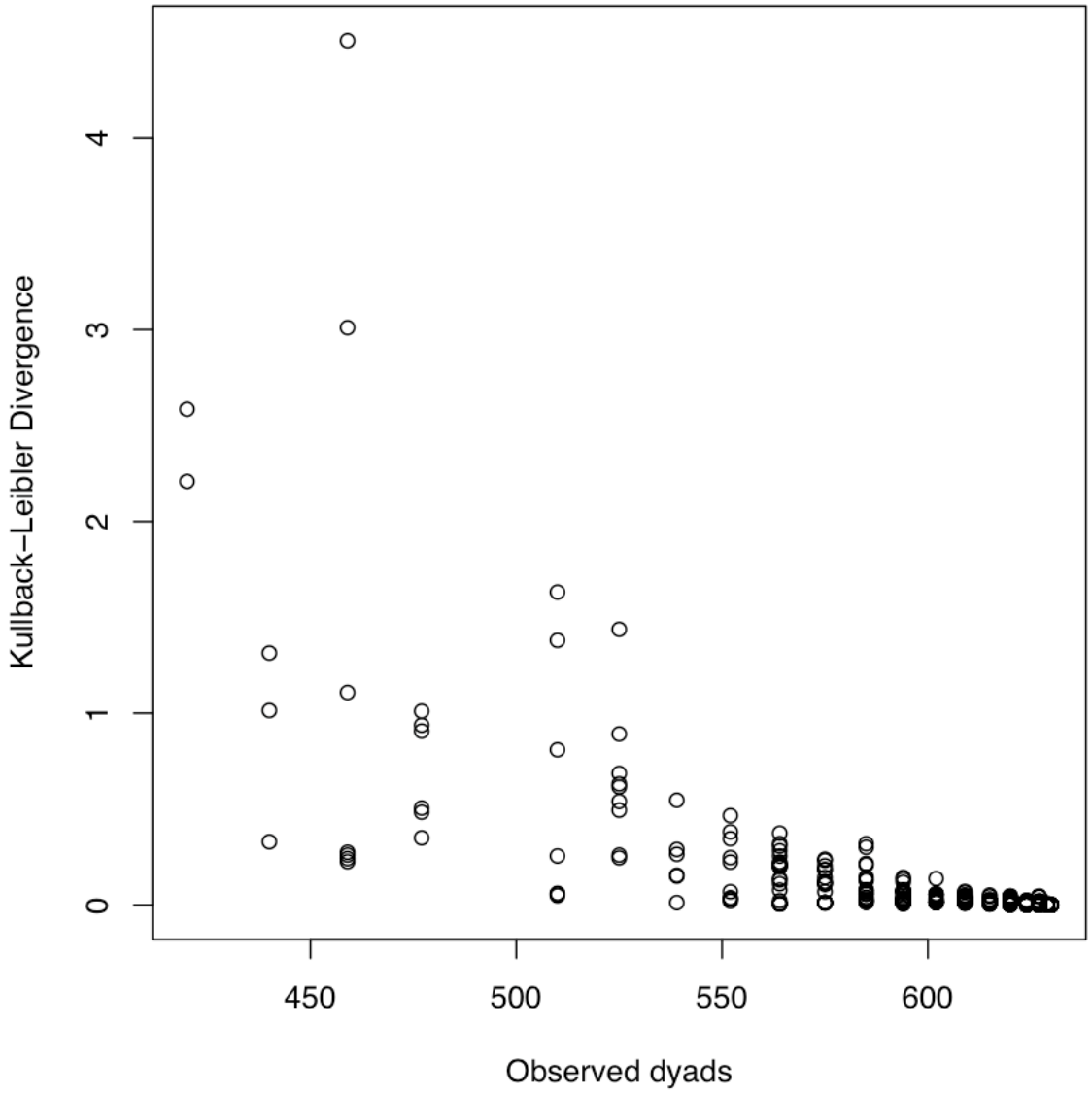
Performance of Parameter Estimates

parameter	complete data value	s.e.	bias (%)	RMSE (%)	efficiency loss (%)
Structural					
Density	−6.51	0.57	0.2	1.2	1.7
GWESP	0.90	0.15	0.8	3.7	5.1
Nodal					
Seniority	0.85	0.24	0.3	3.1	1.3
Practice	0.41	0.12	0.4	5.3	3.5
Homophily					
Practice	0.76	0.19	0.8	4.3	2.9
Gender	0.70	0.25	0.9	4.7	1.7
Office	1.15	0.19	0.7	2.9	2.8

Performance of Parameter Estimates

parameter	complete data value	s.e.	bias (%)	RMSE (%)	efficiency loss (%)
Structural					
Density	−6.51	0.57	0.2	1.2	1.7
GWESP	0.90	0.15	0.8	3.7	5.1
Nodal					
Seniority	0.85	0.24	0.3	3.1	1.3
Practice	0.41	0.12	0.4	5.3	3.5
Homophily					
Practice	0.76	0.19	0.8	4.3	2.9
Gender	0.70	0.25	0.9	4.7	1.7
Office	1.15	0.19	0.7	2.9	2.8

Model Fits: Kullback-Leibler divergence from Truth

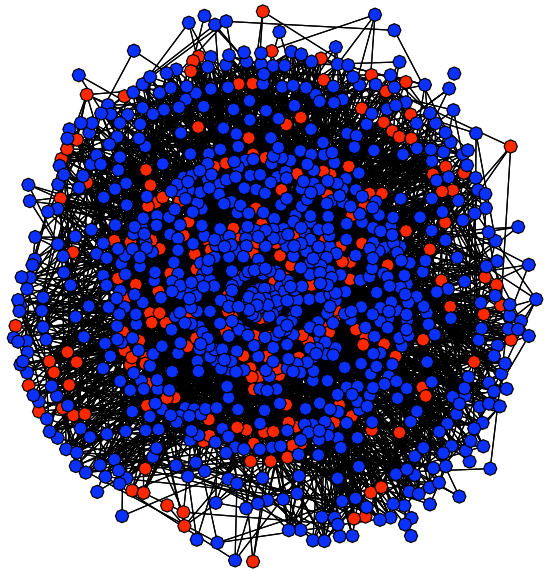


Conclusions, Law Firm Collaborations Example

- Collaborations clustered more than at random
- Senior lawyers collaborate more than junior lawyers
- Corporate lawyers collaborate more than litigation lawyers
- Collaboration more likely between same-sex pairs
- Collaboration more likely between same-office pairs
- Collaboration more likely between same-practice pairs

Injection Drug User Example

Simulations corresponding to U.S. Center for Disease Control study of Injection Drug Users in New York City



- Injection Drug Users (IDU) asked how many IDUs they know, and HIV tested.
- Given 2 coupons to pass to other IDUs they know, to invite them to join study.
- Sampling continues until sample size 500 (from simulated population 715).

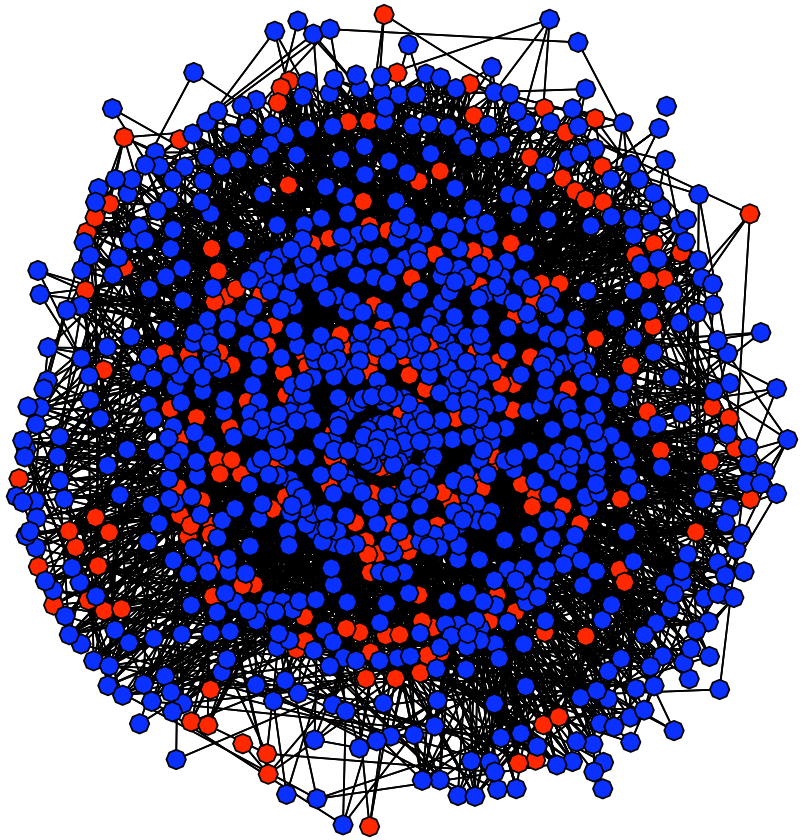
Injection Drug User Example

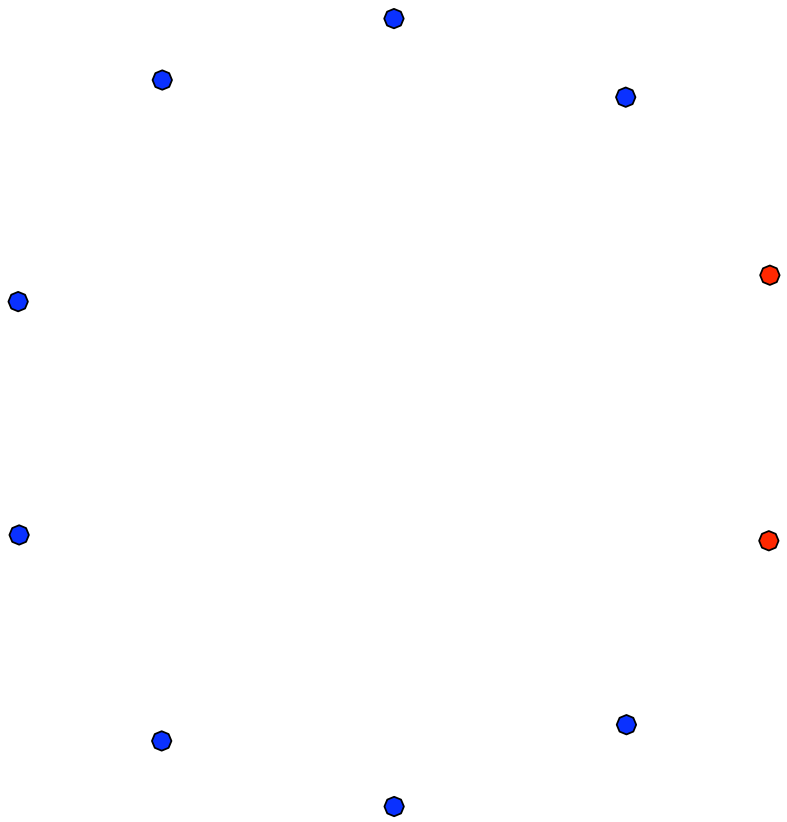
- **Scientific Question:** What proportion of Injection Drug Users in New York City are HIV positive?
- **Methodological Question:** Can we estimate population proportions from samples starting at a convenience sample of seeds?

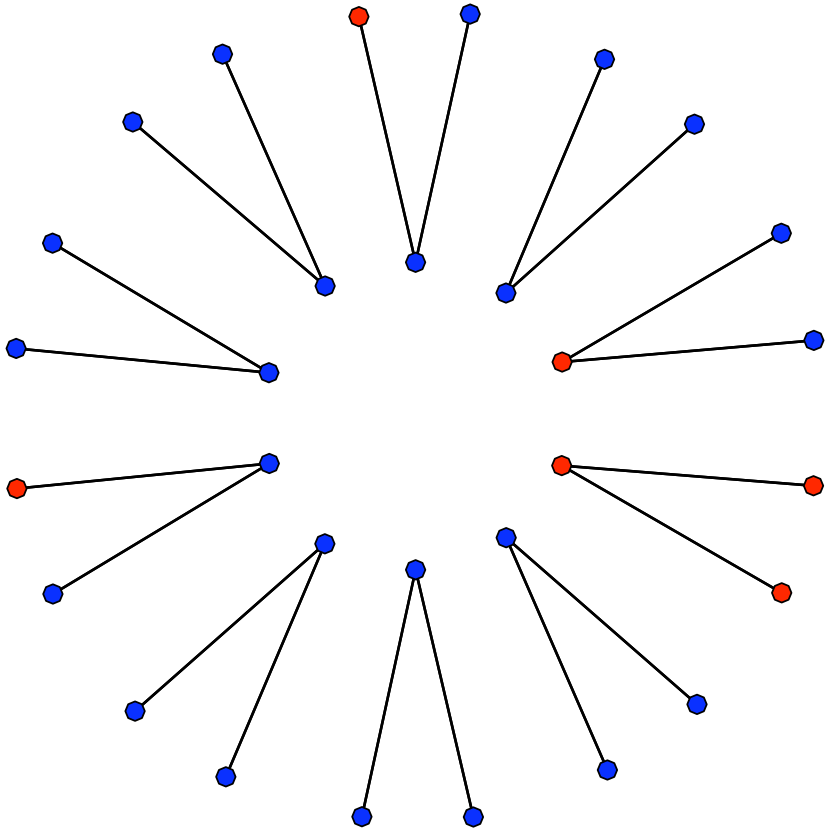
$P(D|Y,\delta) = P(D|Y_{obs},\delta)$ (*adaptive sampling*)

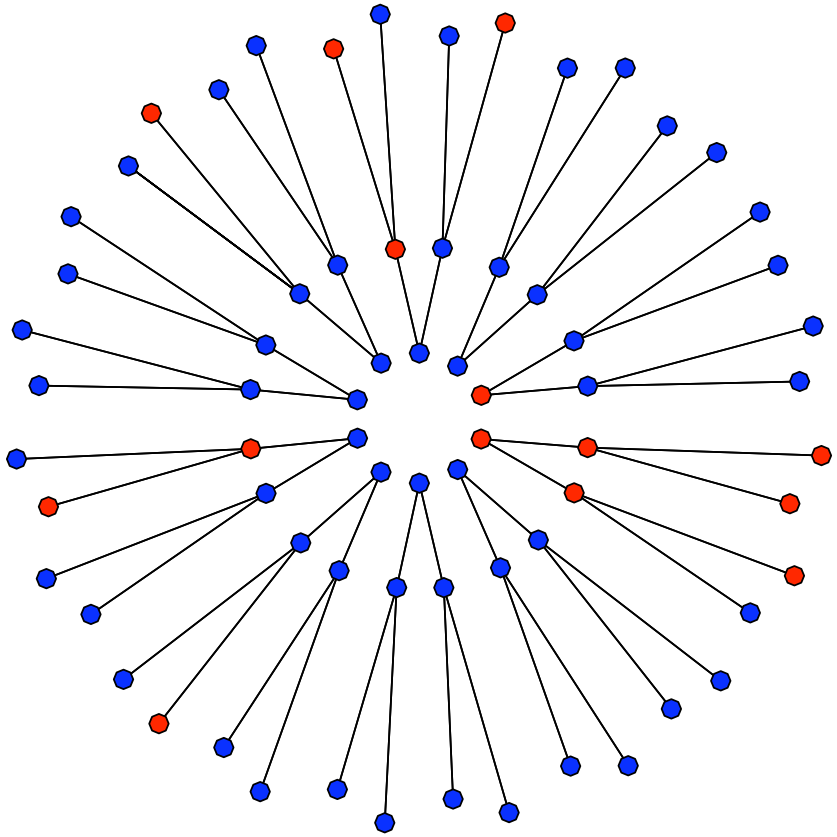
$$P(D|Y,\delta) = P(seeds|Y,\delta)P(D|Y,\delta,seeds) = P(seeds)P(D|Y_{obs},\delta,seeds)$$

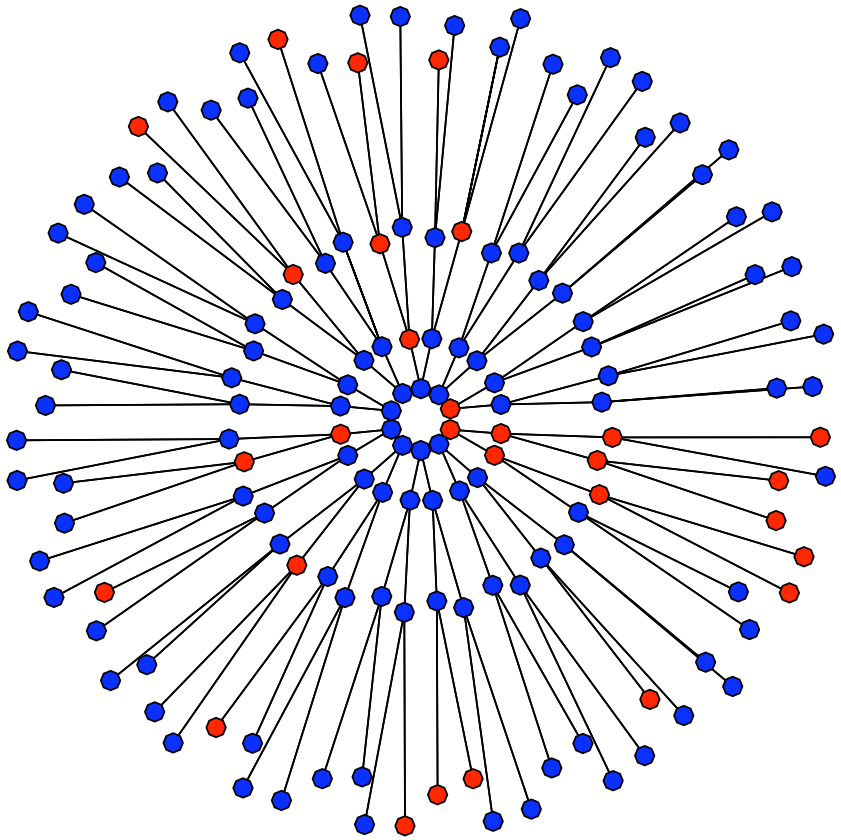
but typically $P(seeds|Y,\delta) \neq P(seeds|Y_{obs},\delta)$

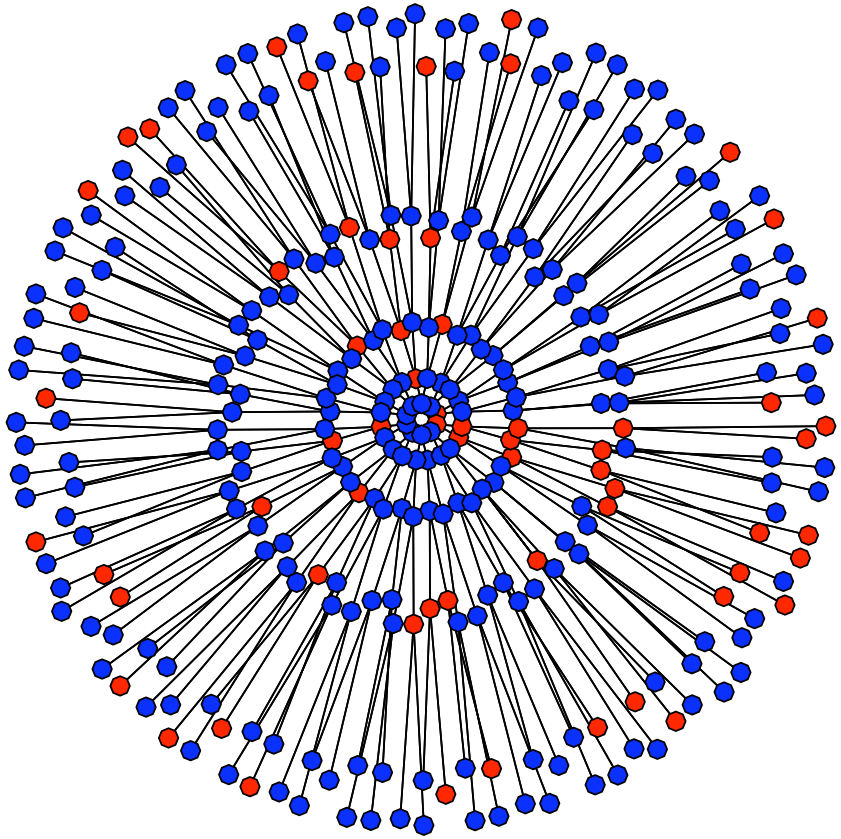


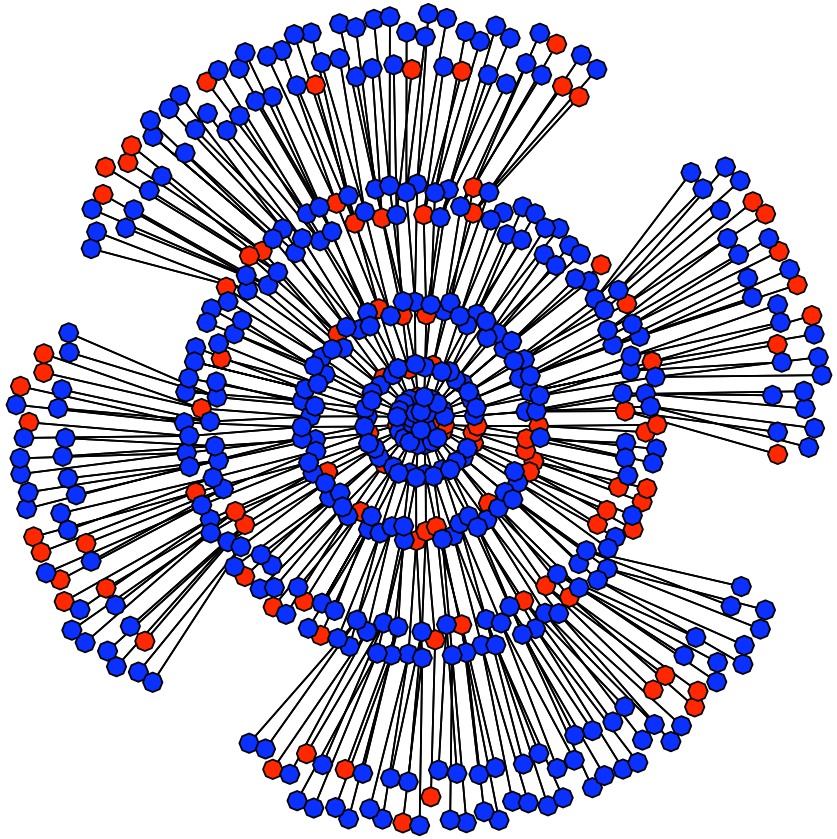


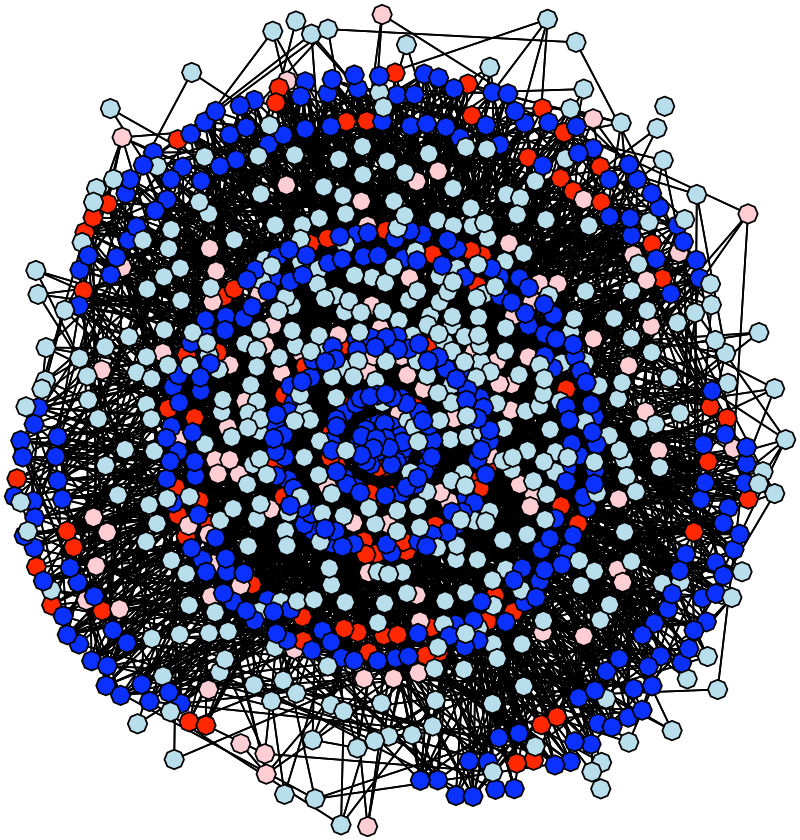












Structure of Analysis

Sample:

- Link-tracing sampling variant - *Respondent-Driven Sampling*
- Ask number of contacts - but not who. Can't identify alters. No matrix.
- 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

Generalized Horvitz-Thompson Estimator

- Goal: Estimate proportion "infected" :

$$\mu = \frac{1}{N}\sum_{j=1}^{N} z_i$$

where

$$z_i = \begin{cases} 1 & \text{node } i \text{ infected} \\ 0 & \text{node } i \text{ uninfected.} \end{cases}$$

- Generalized Horvitz-Thompson Estimator:

$$\hat{\mu} = \frac{\sum_{i:S_i=1} \frac{1}{\pi_i} z_i}{\sum_{i:S_i=1} \frac{1}{\pi_i}}$$

where

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

Key Point: Requires $\pi_i \, \forall \, i : S_i = 1$