Modeling Networks from Partially-Observed Network Data

Mark S. Handcock
University of Washington

joint work with Krista J. Gile
Nuffield College, Oxford

MURI-UCI April 24, 2009

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.

Outline

- Network modeling from a statistical perspective
- Statistical Models for Social Networks
- Introduction of two social examples:
  - Friendships among school students
  - Collaborations within a law firm
- Statistical analysis of social networks
- Mechanisms for the partial observation of social networks
- Analysis of partially-observed social networks
- Missing Data Example: Friendships among school students
- Link-Tracing Sampling Example: Collaborations within a law firm
- Discussion

Network modeling from a statistical perspective

- Networks are widely used to represent data on relations between interacting actors or nodes.
- The study of social networks is multi-disciplinary
  - plethora of terminologies
  - varied objectives, multitude of frameworks
- Understanding the structure of social relations has been the focus of the social sciences
  - *social structure:* a system of social relations tying distinct social entities to one another
  - Interest in understanding how social structure form and evolve
- Attempt to represent the structure in social relations via networks
  - the data is conceptualized as a realization of a network model
- The data are of at least three forms:
  - individual-level information on the social entities
  - relational data on pairs of entities
  - population-level data

Deep literatures available

- Social networks community (Heider 1946; Frank 1972; Holland and Leinhardt 1981)
- Statistical Networks Community (Frank and Strauss 1986; Snijders 1997)
- Spatial Statistics Community (Besag 1974)
- Statistical Exponential Family Theory (Barndorff-Nielsen 1978)
- Graphical Modeling Community (Lauritzen and Spiegelhalter 1988, . . .)
- Machine Learning Community (Jordan, Jensen, Xing, . . .)
- Physics and Applied Math (Newman, Watts, . . .)
- Network Sampling (Frank 1971, Thompson and Seber 1996, Thompson 2002, . . .)

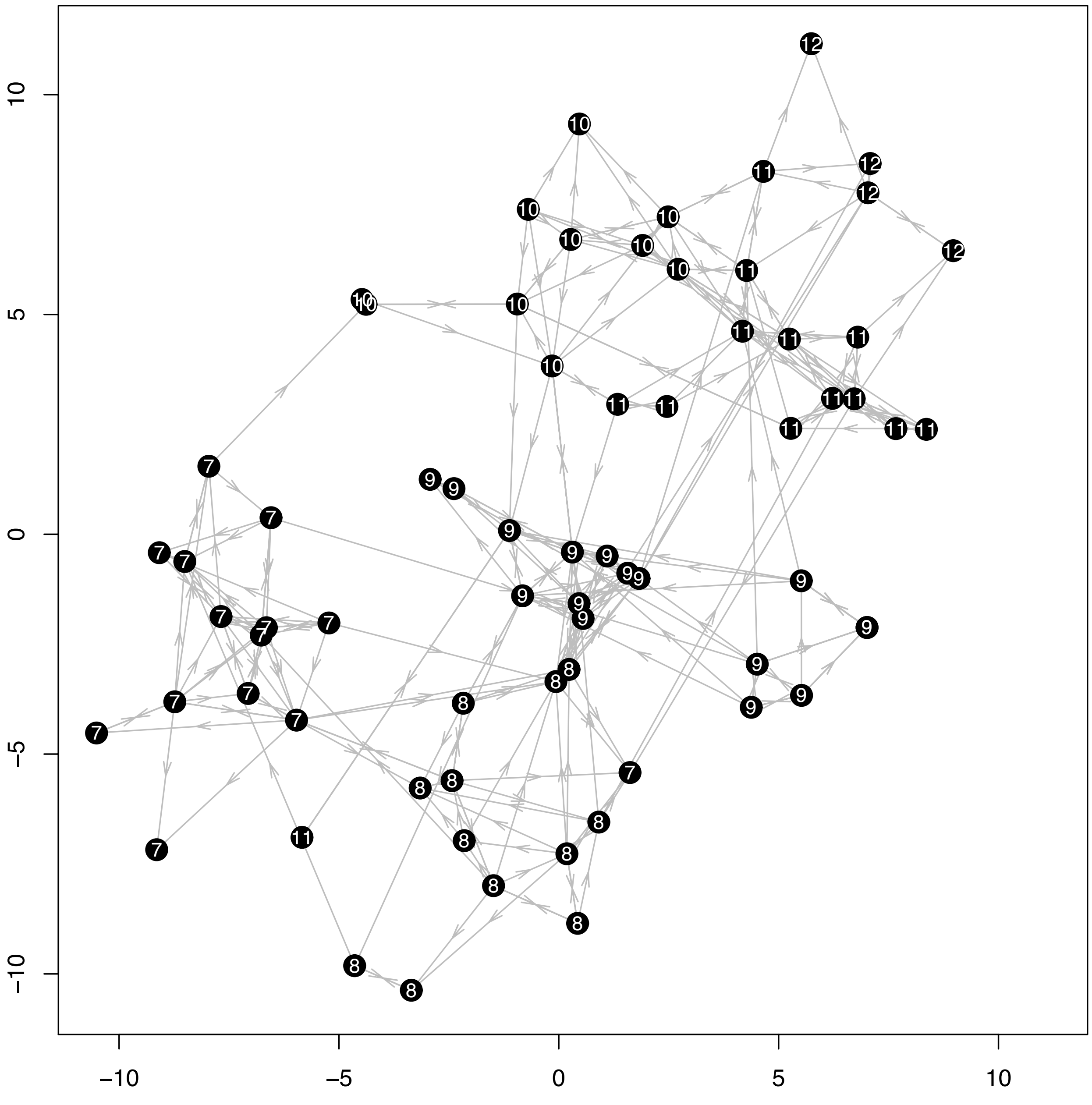
Examples of Friendship Relationships

- The National Longitudinal Study of Adolescent Health

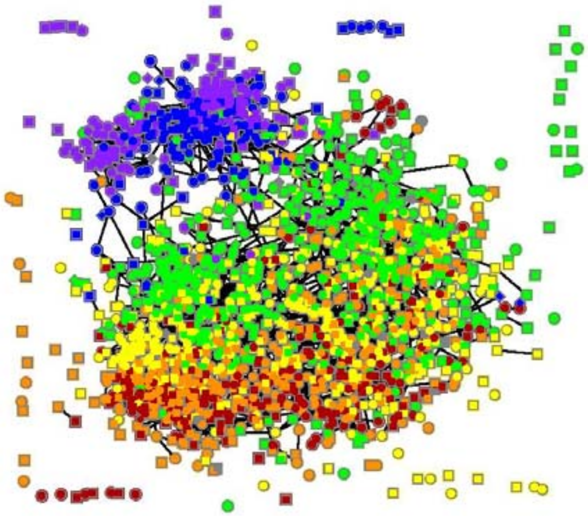
  $\Rightarrow$ `www.cpc.unc.edu/projects/addhealth`

  – “Add Health” is a school-based study of the health-related behaviors of adolescents in grades 7 to 12.

- Each nominated up to 5 boys and 5 girls as their friends
- 160 schools: Smallest has 69 adolescents in grades 7–12



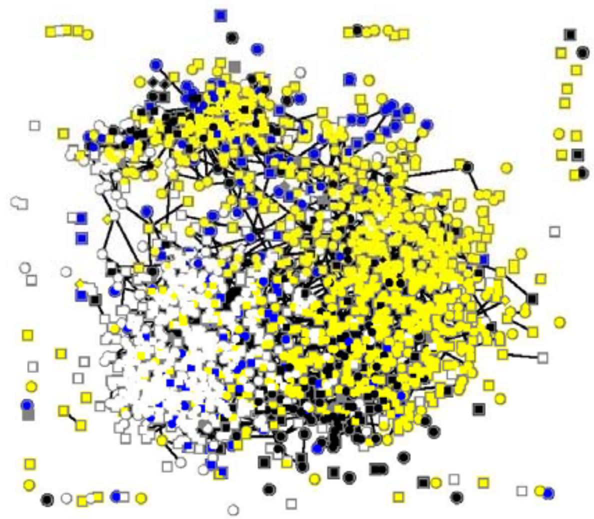
School Community Stratum 44
mutual friendships by Grade



2209 Students

- **Grade 7**
- **Grade 8**
- **Grade 9**
- **Grade 10**
- **Grade 11**
- **Grade 12**
- **Grade NA**

School Community Stratum 44
mutual friendships by Race



2209 Students

- **White** (non-Hispanic)
- **Black** (non-Hispanic)
- **Hispanic** (of any race)
- **Asian / Native Am / Other** (non-Hispanic)
- **Race NA**

Features of Many Social Networks

- *Mutuality* of ties
- *Individual heterogeneity* in the propensity to form ties
- *Homophily* by actor attributes

  $\Rightarrow$ Lazarsfeld and Merton, 1954; Freeman, 1996; McPherson et al., 2001

  - higher propensity to form ties between actors with similar attributes
    e.g., age, gender, geography, major, social-economic status
  - attributes may be observed or unobserved
- *Transitivity* of relationships
  - friends of friends have a higher propensity to be friends
- *Balance* of relationships $\Rightarrow$ Heider (1946)
  - people feel comfortable if they agree with others whom they like
- *Context* is important $\Rightarrow$ Simmel (1908)
  - triad, not the dyad, is the fundamental social unit

The Choice of Models depends on the objectives

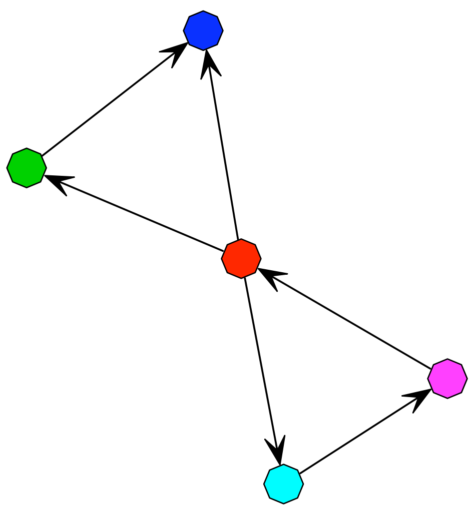
- Primary interest in the nature of relationships:
  - How the behavior of individuals depends on their location in the social network
  - How the qualities of the individuals influence the social structure
- Secondary interest is in how network structure influences processes that develop over a network
  - spread of HIV and other STDs
  - diffusion of technical innovations
  - spread of computer viruses
- Tertiary interest in the effect of *interventions* on network structure and processes that develop over a network

Perspectives to keep in mind

- Network-specific versus Population-process
  - *Network-specific*: interest focuses only on the actual network under study
  - *Population-process*: the network is part of a population of networks and the latter is the focus of interest
    - the network is conceptualized as a realization of a social process

(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

Statistical Models for Social Networks

Notation

A *social network* is defined as a set of n social “actors” and a social relationship between each pair of actors.

$$Y_{ij} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \\ 0 & \text{otherwise} \end{cases}$$

- call $Y \equiv [Y_{ij}]_{n \times n}$ a *sociomatrix*
  - a $N = n(n-1)$ binary array
- The basic problem of stochastic modeling is to specify a distribution for Y i.e., $P(Y = y)$

A Framework for Network Modeling

Let $\mathcal{Y}$ be the sample space of Y e.g. $\{0,1\}^N$

Any model-class for the multivariate distribution of Y
can be *parametrized* in the form:

$$P_\eta(Y = y) = \frac{\exp\{\eta \cdot g(y)\}}{\kappa(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

Besag (1974), Frank and Strauss (1986)

- $\eta \in \Lambda \subset R^q$ q-vector of parameters
- $g(y)$ q-vector of *network statistics*.
  $\Rightarrow$ $g(Y)$ are jointly sufficient for the model
- For a "saturated" model-class $q = |\mathcal{Y}| - 1$ e.g. $2^N - 1$
- $\kappa(\eta, \mathcal{Y})$ distribution normalizing constant

$$\kappa(\eta, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp\{\eta \cdot g(y)\}$$

Simple model-classes for social networks

Homogeneous Bernoulli graph (Erdős-Rényi model)

- Y_{ij} are independent and equally likely
  with log-odds $\eta =$ logit$[P_\eta(Y_{ij} = 1)]$

$$P_\eta(Y = y) = \frac{e^{\eta \sum_{i,j} y_{ij}}}{\kappa(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

  where $q = 1,\ g(y) = \sum_{i,j} y_{ij},\ \kappa(\eta, \mathcal{Y}) = [1 + \exp(\eta)]^N$
- homogeneity means it is unlikely to be proposed as a model for real phenomena

Dyad-independence models with attributes

- Y_{ij} are independent but depend on dyadic covariates $x_{k,ij}$

$$P_\eta(Y = y) = \frac{e^{\sum_{k=1}^{q} \eta_k g_k(y)}}{\kappa(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

$$g_k(y) = \sum_{i,j} x_{k,ij} y_{ij}, \quad k = 1, \ldots, q$$

$$\kappa(\eta, \mathcal{Y}) = \prod_{i,j} [1 + \exp(\sum_{k=1}^{q} \eta_k x_{k,ij})]$$

Of course,

$$\textit{logit}[P_\eta(Y_{ij} = 1)] = \sum_k \eta_k x_{k,ij}$$

Generative Theory for Network Structure

Actor Markov statistics $\Rightarrow$ Frank and Strauss (1986)

– motivated by notions of "symmetry" and "homogeneity"
– Y_{ij} in Y that do not share an actor are
conditionally independent given the rest of the network
$\Rightarrow$ analogous to nearest neighbor ideas in spatial modeling

- Degree distribution: $\mathrm{d}_k(y) =$ proportion of actors of degree k in y.
- k-star distribution: $\mathrm{s}_k(y) =$ proportion of k-stars in the graph y. (In particular, $\mathrm{s}_2 =$ proportion of edges that exist between pairs of actors.)
- triangles: $\mathrm{t}_1(y) =$ proportion of triads that from a complete sub-graph in y.

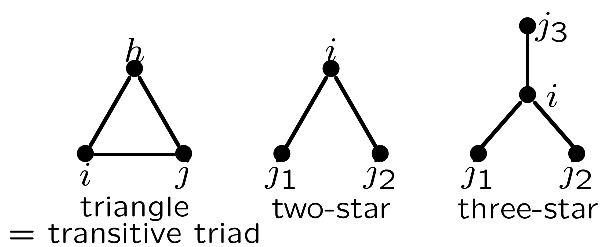
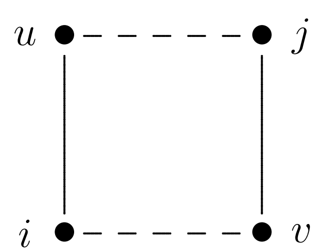


Figure 1: Some configurations for non-directed graphs

General mechanisms motivated by conditional independence

$\Rightarrow$ Pattison and Robins (2002), Butts (2005)
$\Rightarrow$ Snijders, Pattison, Robins and Handcock (2006)

– Y_{uj} and Y_{iv} in Y are conditionally independent given the rest of the network if they could not produce a cycle in the network



Partial conditional dependence when four-cycle is created

This produces features on configurations of the form:

- edgewise shared partner distribution: $\mathrm{ep}_k(y) =$
  proportion of edges between actors with exactly k shared partners
  $k = 0, 1, \ldots$

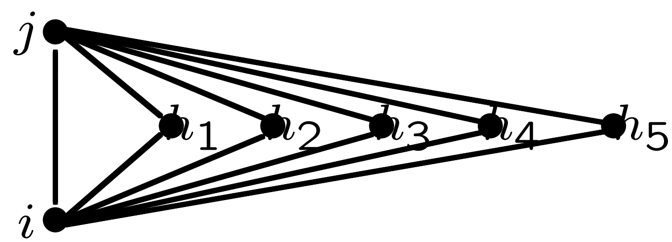


Figure 2: The actors in the non-directed (i, j) edge have 5 shared partners

- dyadwise shared partner distribution:
  $\mathrm{dp}_k(y) =$ proportion of dyads with exactly k shared partners
  $k = 0, 1, \ldots$

Structural Signatures

– identify social constructs or features
– based on intuitive notions or partial appeal to substantive theory

- Clusters of edges are often *transitive*:
  Recall $t_1(y)$ is the proportion of triangles amongst triads

$$t_1(y) = \frac{1}{\binom{g}{3}} \sum_{\{i,j,k\} \in \binom{g}{3}} y_{ij} y_{ik} y_{jk}$$

A closely related quantity is the
proportion of triangles amongst 2-stars

$$C(y) = \frac{3 \times t_1(y)}{s_2(y)}$$

mean clustering coefficient

Statistical Inference for η

Base inference on the loglikelihood function,

$$\ell(\eta) = \eta \cdot g(y_{\text{obs}}) - \log \kappa(\eta)$$

$$\kappa(\eta) = \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\eta \cdot g(z)\}$$

Approximating the loglikelihood

- Suppose $Y_1, Y_2, \ldots, Y_m \stackrel{\text{i}.i.d.}{\sim} P_{\eta_0}(Y = y)$ for some η_0.
- Using the LOLN, the difference in log-likelihoods is

$$\begin{aligned}
\ell(\eta) - \ell(\eta_0) &= \log \frac{\kappa(\eta_0)}{\kappa(\eta)} \\
&= \log \mathbf{E}_{\eta_0} \left(\exp \left\{ (\eta_0 - \eta) \cdot g(Y) \right\} \right) \\
&\approx \log \frac{1}{M} \sum_{i=1}^{M} \exp \left\{ (\eta_0 - \eta) \cdot (g(Y_i) - g(y_{\text{obs}})) \right\} \\
&\equiv \tilde{\ell}(\eta) - \tilde{\ell}(\eta_0).
\end{aligned}$$

- Simulate $Y_1, Y_2, \ldots, Y_m$ using a MCMC (Metropolis-Hastings) algorithm
  $\Rightarrow$ Snijders (2002); Handcock (2002).
- Approximate the MLE $\hat{\eta} = \text{argmax}_\eta \{\tilde{\ell}(\eta) - \tilde{\ell}(\eta_0)\}$ (MC-MLE)
  $\Rightarrow$ Geyer and Thompson (1992)
- Given a random sample of networks from P_{η_0}, we can thus approximate (and subsequently maximize) the loglikelihood shifted by a constant.

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} =

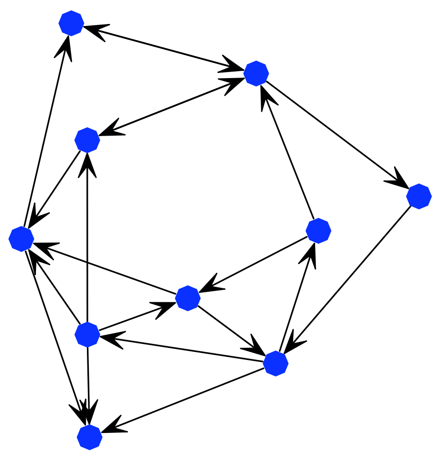
	A	B	C	D
A	-	?	?	?
B	?	-	?	?
C	0	0	-	0
D	1	1	1	-

D =

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

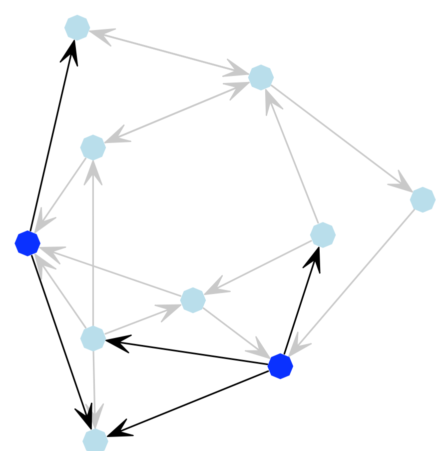
Partial Observation of Social Networks

- **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?"



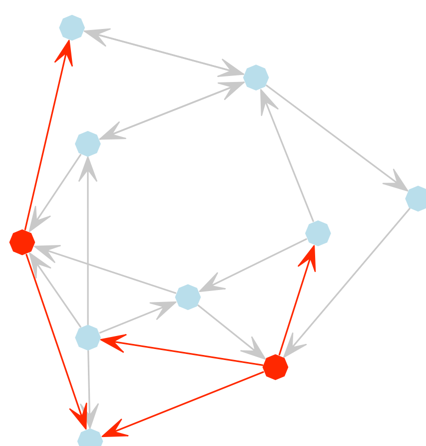
Partial Observation of Social Networks

- **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?"



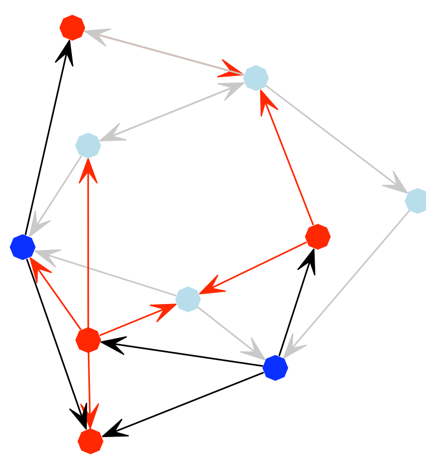
Partial Observation of Social Networks

- **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?"



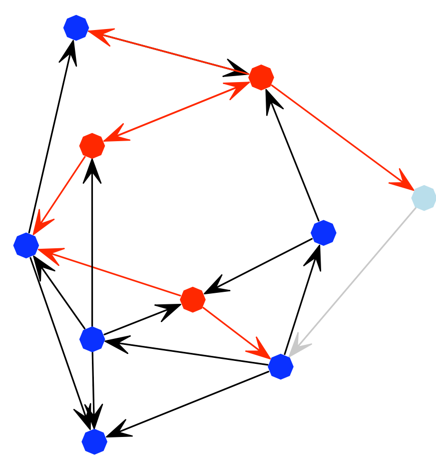
Partial Observation of Social Networks

- **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?"



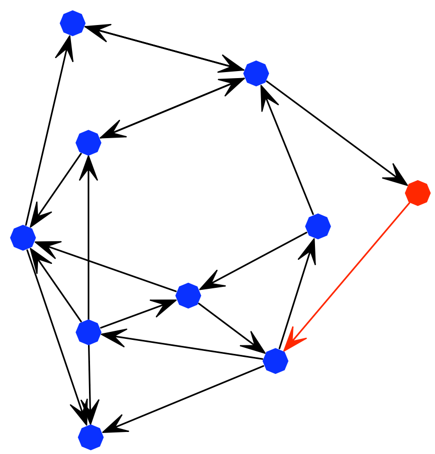
Partial Observation of Social Networks

- **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?"



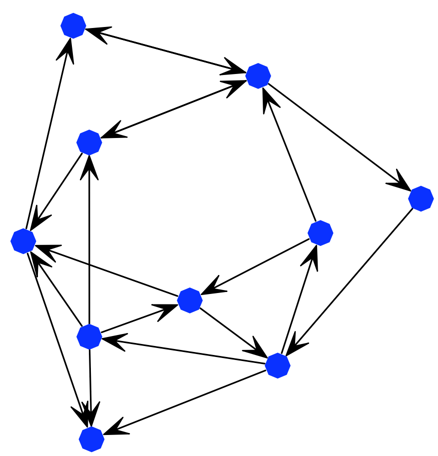
Partial Observation of Social Networks

- **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?"



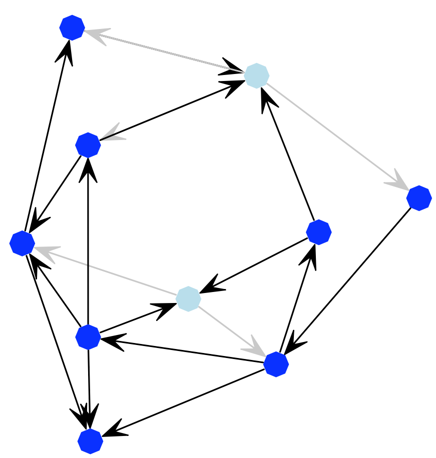
Partial Observation of Social Networks

- **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?"



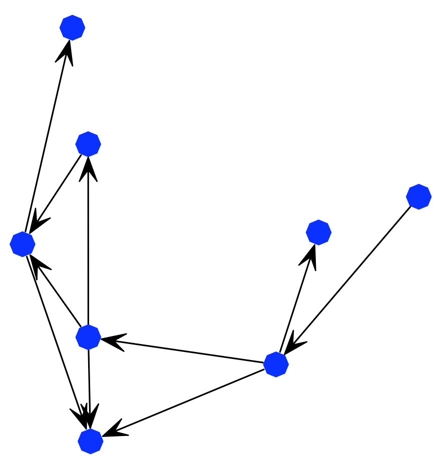
Partial Observation of Social Networks

- **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?"



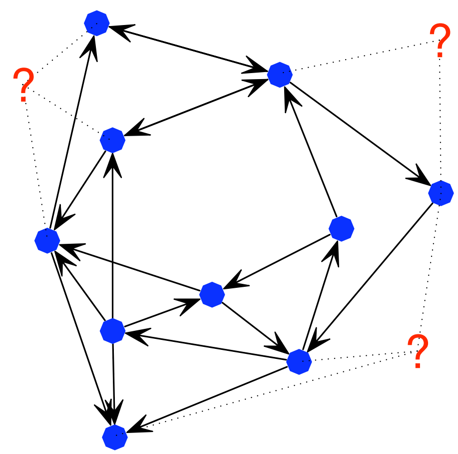
Partial Observation of Social Networks

- **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

Modeling with Missing and Sampled Data

- Most analysis ignores individuals with missing data
- Earlier work: assume and enforce reciprocity (Stork and Richards 1992)
- Treat respondents and non-respondents separately, pseudo-likelihood (Robins, Pattison, and Woolcock, 2004)
- Fit simple network model with non-observations (Thompson and Frank, 2000)
- This work: extend to full range of stochastic models; expand sophistication of model-checking

Design-based Inference for Describing Structure

- Example Scientific Questions:
  - What proportion of the social contacts of unemployed residents of London are with other unemployed residents?
  - What is the average donation size to each political candidate?
- Approach:
  - Make probability statements about the relations in the full network based on the observed part of the network
  - Weight each observation by the inverse of probability of being sampled
- Advantages:
  - Requires no assumptions about network structure
- Disadvantages:
  - Requires full knowledge of sampling mechanism, and sampling probabilities
  - Difficult to conduct complex analysis such as regression-type models

Social Network Modeling for Understanding Processes

- Example Scientific Questions:
  - Are men in a company more likely to collaborate with other men than with women?
  - Are countries more likely to trade with other countries with similar political structures?
- Approach:
  - Make probability statements about the social forces that could account for the network
  - Create complex regression-style model for relational information
- Advantages:
  - Flexible Models to answer complex questions
- Disadvantages:
  - Assumes chosen model form is accurate
  - Computationally expensive for complex models
  - Assume sampling is ”Missing at Random”
  - Initially, only fit to fully observed networks

Fitting Models to Networks with Incomplete Data

- Two types of units: nodes and relational structures
- Sampling typically on nodes, inference on relational structures

- Extend and adapt methods from survey sampling and missing data literature (Thompson and Seber, 1996, Little and Rubin, 2002)
- Extend former work on partially-observed network data (Frank, 1971, Frank and Snijders, 1994, Thompson and Frank, 2000)
- Novel Methods: Full range of stochastic models; expand model-checking (Handcock and Gile, 2007, Gile and Handcock, 2006)

- Key Point: require that statistical properties of unobserved relations do not depend on unobserved characteristics, given what was observed

Fitting Models to Partially Observed Social Network Data

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

$$\begin{aligned}\ell(\eta,\delta) &\equiv P(Y_{obs}=y_{obs},D|\eta,\delta)\\ &= \sum_{y_{unobs}} P(Y_{obs}=y_{obs},Y_{unobs}=y_{unobs},D|\eta,\delta)\\ &= \sum_{y_{unobs}} P(D|Y_{obs}=y_{obs},Y_{unobs}=y_{unobs},\delta)P_{\eta}(Y_{obs}=y_{obs},Y_{unobs}=y_{unobs})\end{aligned}$$

- η is the model parameter
- δ is the sampling parameter

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

Then

$$\begin{aligned}\ell(\eta,\delta) &\equiv P(Y_{obs}=y_{obs},D|\eta,\delta)\\ &= P(D|Y_{obs}=y_{obs},\delta)\sum_{y_{unobs}} P_{\eta}(Y_{obs}=y_{obs},Y_{unobs}=y_{unobs})\end{aligned}$$

Fitting Models to Partially Observed Social Network Data

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

$$\ell(\eta, \delta) \equiv P(Y_{obs} = y_{obs}, D|\eta, \delta)$$
$$= \sum_{y_{unobs}} P(Y_{obs} = y_{obs}, Y_{unobs} = y_{unobs}, D|\eta, \delta)$$
$$= \sum_{y_{unobs}} P(D|Y_{obs} = y_{obs}, Y_{unobs} = y_{unobs}, \delta) P_\eta(Y_{obs} = y_{obs}, Y_{unobs} = y_{unobs})$$

- η is the model parameter
- δ is the sampling parameter

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

Then

$$\ell(\eta, \delta) \equiv P(Y_{obs} = y_{obs}, D|\eta, \delta)$$
$$= P(D|Y_{obs} = y_{obs}, \delta) \sum_{y_{unobs}} P_\eta(Y_{obs} = y_{obs}, Y_{unobs} = y_{unobs})$$

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

When is Sampling MAR?

Examples of MAR Sampling:

- Individual sample, sample based on observed things like race, sex, and age that we know.
- Link-tracing sample starting with a MAR sample with follow-up based on observed relations with others in the sample, as well as things like race and sex and age.
- Link-tracing with probability proportional to number of partners is MAR!

Examples of NMAR (not missing at random) Sampling:

- Individual sample based on unobserved properties of non-respondents - like infection status or illicit activity.
- Link-tracing sample starting where links are followed dependent on unobserved properties of alters.

Application to ERGM

$$\ell(\eta, \delta) \equiv \ell(\delta)\ell(\eta)$$

$$\begin{aligned} \ell(\eta) &\equiv \sum_{y_{unobs}} P_\eta(Y_{obs} = y_{obs}, Y_{unobs} = y_{unobs}) \\ &= \sum_{y_{unobs}} \frac{\exp\{\eta \cdot g(y_{obs} + y_{unobs})\}}{\kappa(\eta, \mathcal{Y})} = \frac{\kappa(\eta, \mathcal{Y}|y_{obs})}{\kappa(\eta, \mathcal{Y})} \end{aligned}$$

where $\kappa(\eta, \mathcal{Y}|y_{obs}) = \sum_{y_{unobs}} \exp\{\eta \cdot g(y_{obs} + y_{unobs})\}$.

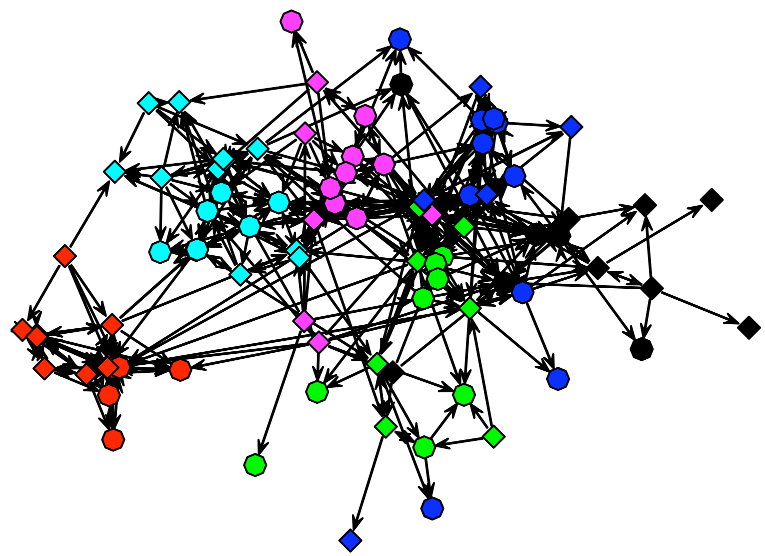
However

$$P_\eta(Y_{unobs} = y_{unobs}|Y_{obs} = y_{obs}) = \frac{\exp\{\eta \cdot g(y_{obs} + y_{unobs})\}}{\kappa(\eta, \mathcal{Y}|y_{obs})} \qquad y_{unobs} \in \mathcal{Y}(y_{obs})$$

where $\mathcal{Y}(y_{obs}) = \{y_{unobs} : y + y_{obs} \in \mathcal{Y}\}$
so estimate $\kappa(\eta, \mathcal{Y}|y_{obs})$ with same Markov Chain Monte Carlo (MCMC)

Example: Friendships in a School

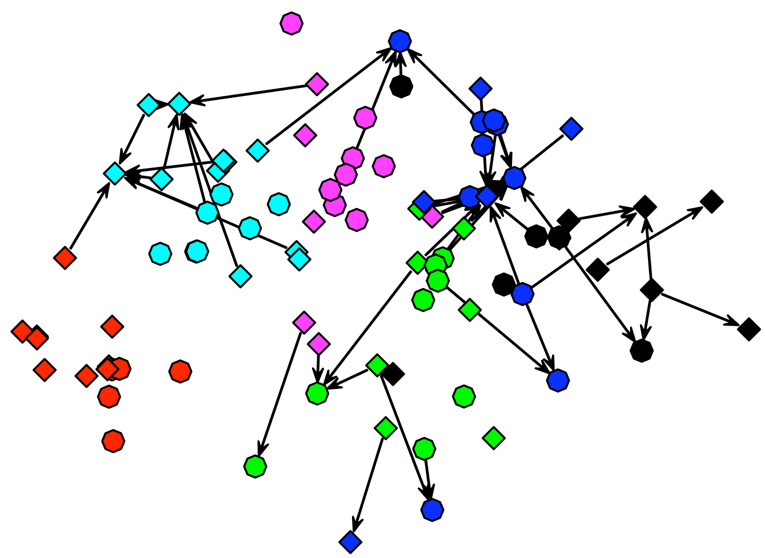
From the National Longitudinal Survey on Adolescent Health - Wave 1:



- 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

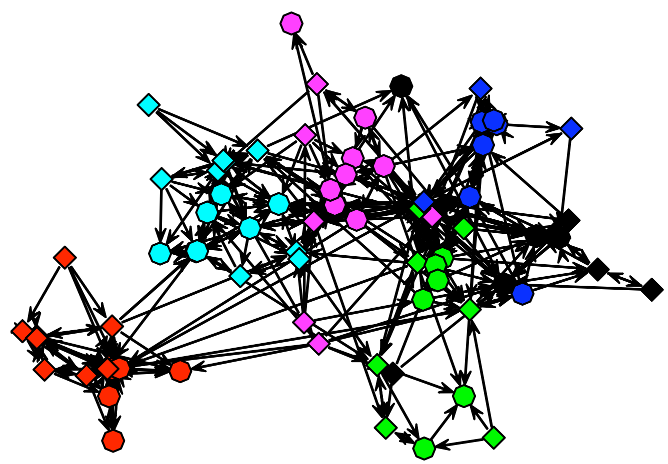
From the National Longitudinal Survey on Adolescent Health - Wave 1:



- 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

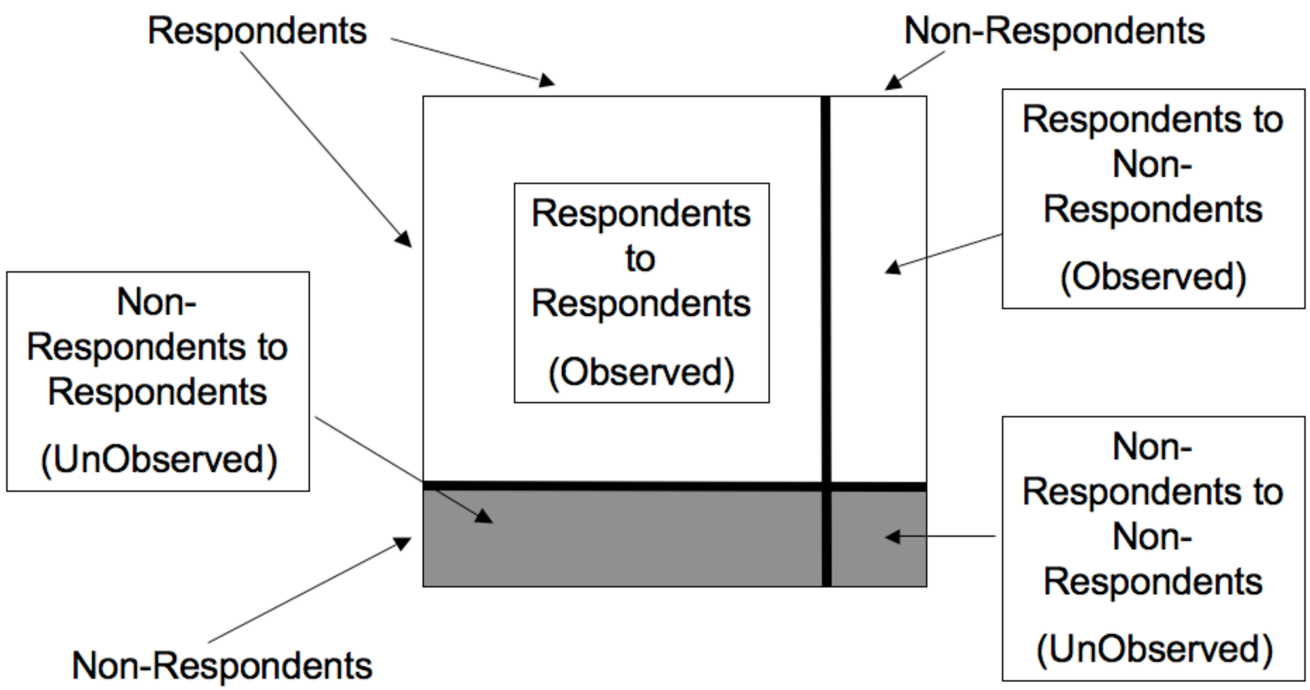
- **Scientific Question:** Do friendships form in an egalitarian or an hierarchical manner?
- **Methodological Question:** Can we fit a network model to a network with missing data? Is the fit different from that of just the observed data?

$P(D|Y, \delta) = P(D|y_{obs}, \delta)$ (*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

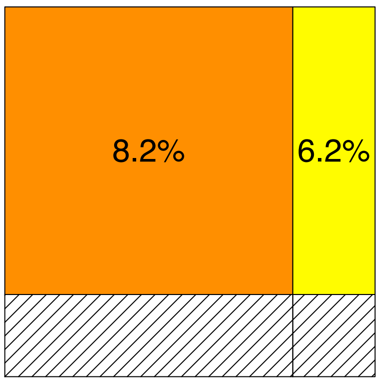
Fit an ERGM to the partially observed data, get coefficients like in logistic regression.

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** (hierarchical)
- Propensity for ties within sex and grade to be **cyclical** (egalitarian)
- **Isolation**: Propensity for students to receive no nominations

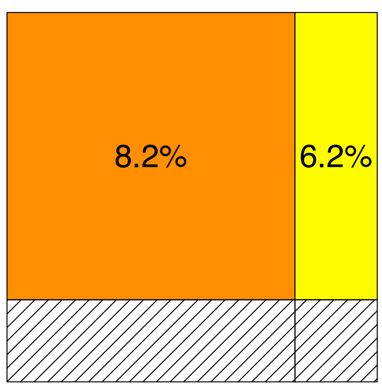
Percent of Possible Relations Realized

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

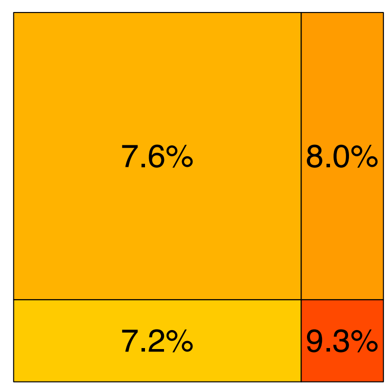


Goodness of Fit: Percent of Possible Relations Realized

	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



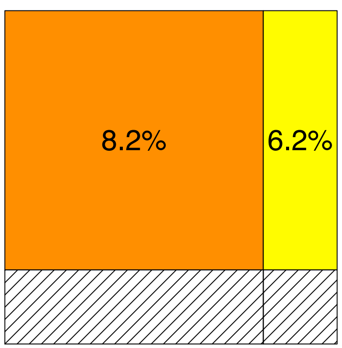
(a) Observed



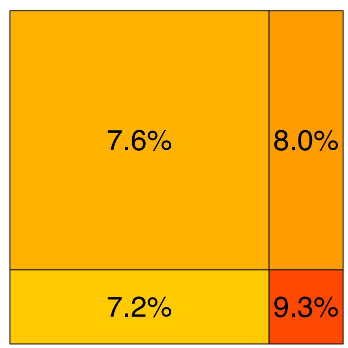
(b) Fit

Goodness of Fit: Percent of Possible Relations Realized

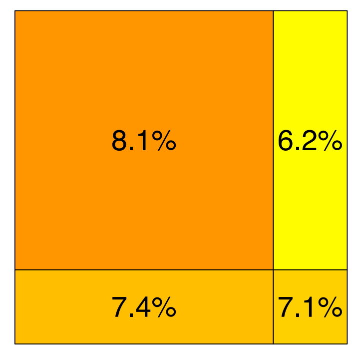
	Observed	Original	Diff. Popularity
Respondents to Respondents	8.2	7.6	8.1
Respondents to Non-Respondents	6.2	8.0	6.2
Non-Respondents to Respondents	-	7.2	7.4
Non-Respondents to Non-Respondents	-	9.3	7.1



(c) Observed



(d) Original



(e) 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^{***}

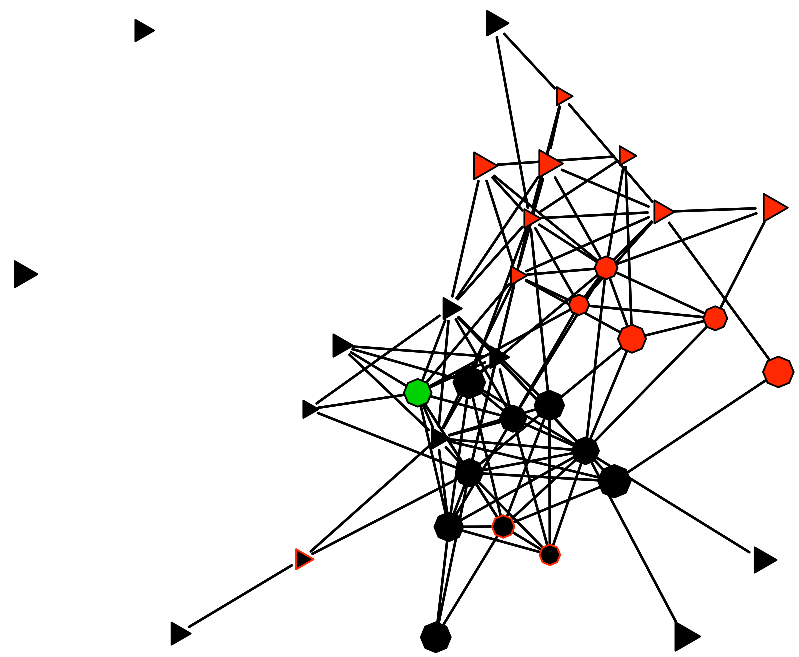
	coefficient	s.e.
Density	-1.138	0.19^{***}
Sex and Grade Factors		
Grade 8 Popularity	-0.178	0.14
Grade 9 Popularity	-0.420	0.16^{**}
Grade10 Popularity	-0.339	0.16^{*}
Grade 11 Popularity	0.256	0.19
Grade 12 Popularity	0.243	0.20
Male Popularity	0.779	0.17^{***}
Non-Resp Popularity	-0.322	0.10^{**}
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 hierarchical rather than egalitarian structure
- More students receive no nominations than we would expect at random.

Law Firm Collaboration Example

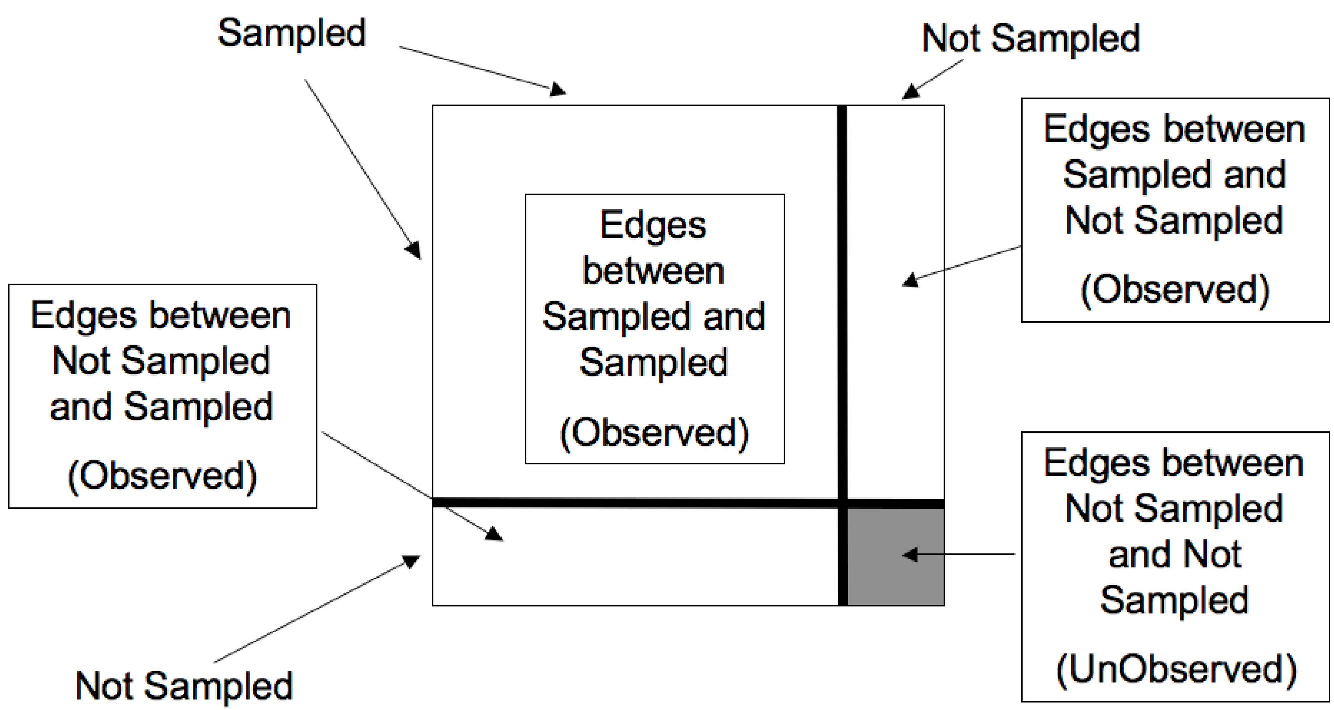
From the Emmanuel Lazega's study of a Corporate Law Firm:



- 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.

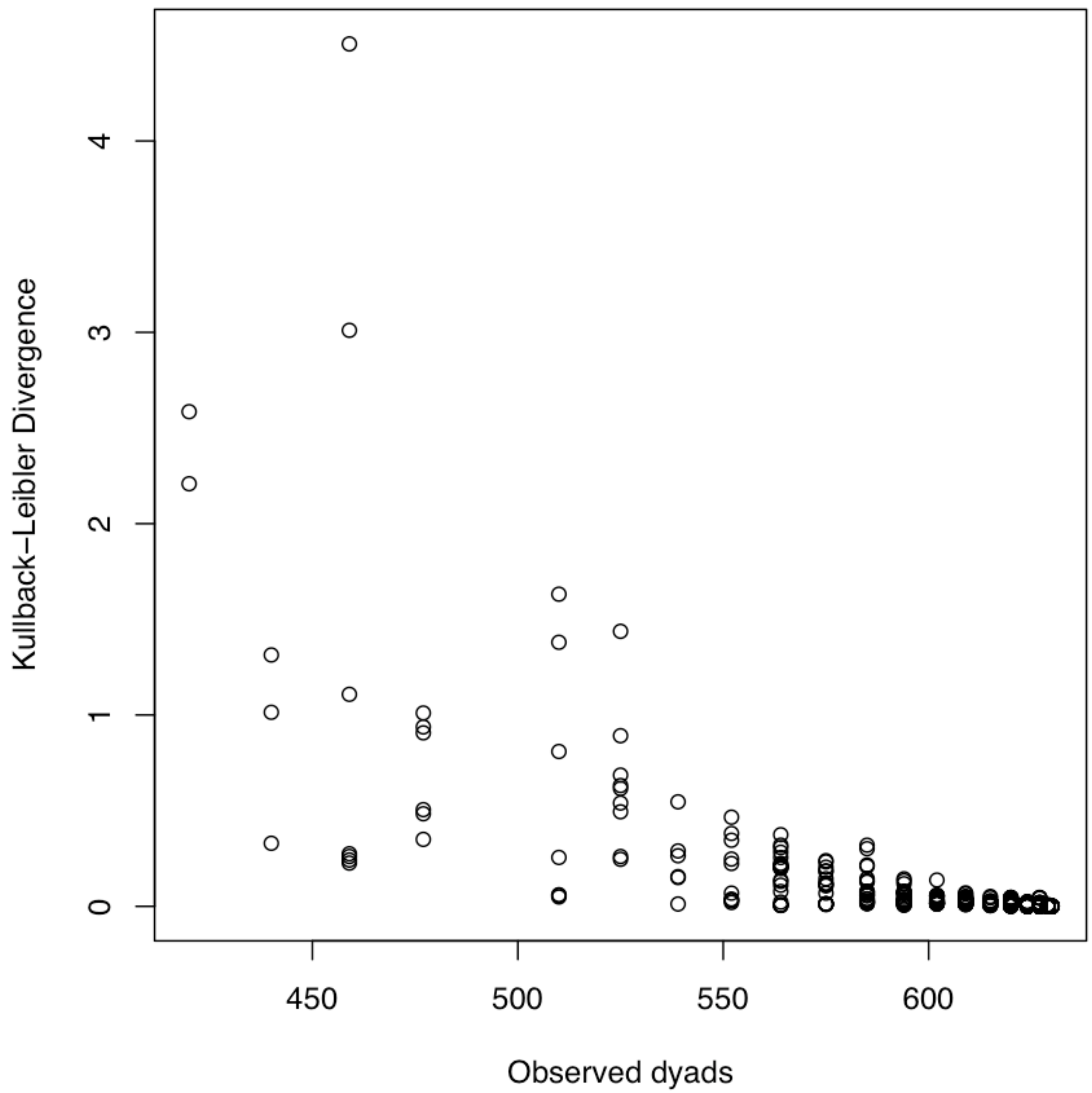
Performance of Parameter Estimates

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

Discussion

Missing Data, School Friendship Example:

- Challenge: Only part of network observed
- Fit model to all observed data
- Leverage information in sample
  - In-ties (and in-degrees)
  - Covariate information
- Limitations:
  - Assume full network size known
  - Requires identifiability of alters
  - Missing at Random data
- Implications for Study Design
  - Collect and keep data relating to non-respondents:
    - In-ties
    - Covariate information
    - Number of non-respondents
  - Likelihood inference is possible with missing data!

Discussion

Sampling, Law Firm Collaboration Example:

- Challenge: Observed data due to complicated link-tracing process
- Fit model to observed data
- Leverage information in sample
  - In-ties
  - Covariate information
- Link-tracing sample is Adaptive!
- Limitations
  - Assume full network size known
  - Requires identifiability of alters
  - Requires Missing at Random initial sample
- Implications for Study Design
  - Collect and keep data relating to non-respondents:
    - In-ties
    - Covariate information
    - Number of non-respondents
  - Likelihood inference is possible with link-tracing sample!

Discussion

- Network models can be applied to partially-observed network data to address scientific questions about the full network.
  - Missing Data (missing at random)
  - Sampled Data (egocentric or adaptive)
  - Do not need simple random sample to be representative
- Some forms of additional information collected in the study can greatly improve possibilities for inference.
  - If not *missing at random* or *adaptive*, can use extra information to improve inference
  - Measurement of sampling biases
  - Any characteristics of unobserved units
- All models fit with an Exponential-Family Random Graph Model using `statnet` R software.