Modeling Networks from Partially-Observed Network Data

Mark S. Handcock
University of Washington, Seattle

joint work with Krista J. Gile
Nuffield College, Oxford

23 October, 2008

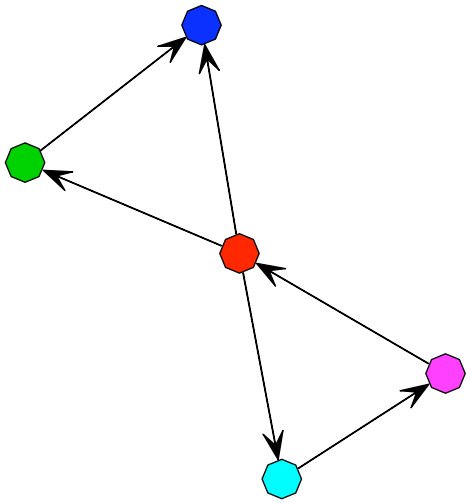
For details, see:

- 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 grant 7R01DA012831

Social Networks

- Social Network: Tool to formally represent and quantify relational social structure.
- Relations can include: sexual partnerships, needle sharing, 'knowing someone'
- Represent mathematically as a sociomatrix, Y, where
  $Y_{ij} =$ the value of the relationship from i to j



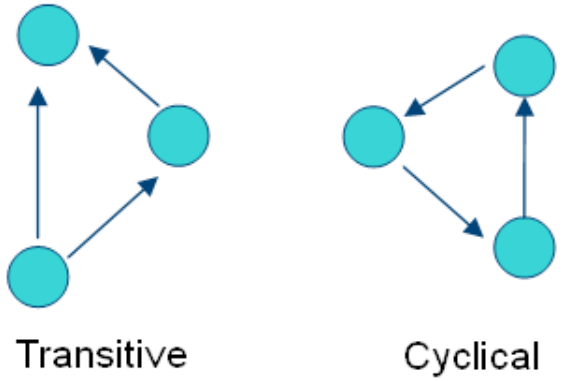
(a) Sociogram

	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

Exponential-Family Random Graph Models (ERGMs or p*)

- Independence models rarely capture relational structure



Exponential-family Random Graph Model (ERGM):

(Holland and Leinhardt (1981), Snijders et al, (2006),...)

$$P_\beta(Y = y) = c(\beta)e^{\beta_1 g_1(y)+\beta_2 g_2(y)+\ldots}$$

- $g(y)$ represent features of the social process
- $c(\beta)$ is the normalizing 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	-

$Y_{obs} =$

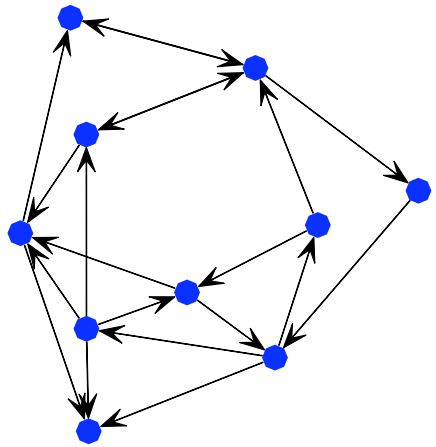
	A	B	C	D
A	-	?	?	?
B	?	-	?	?
C	0	0	-	0
D	1	1	1	-

$D =$

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

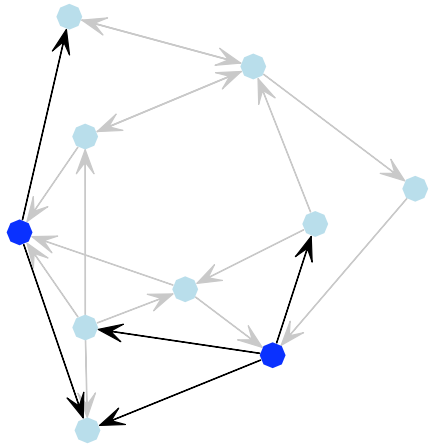
Partial Observation of Social Networks

- **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?"



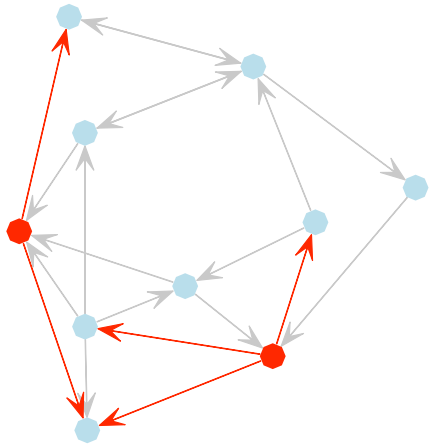
Partial Observation of Social Networks

- **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?"



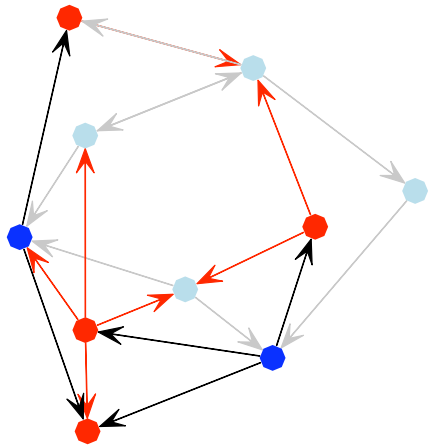
Partial Observation of Social Networks

- **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?"



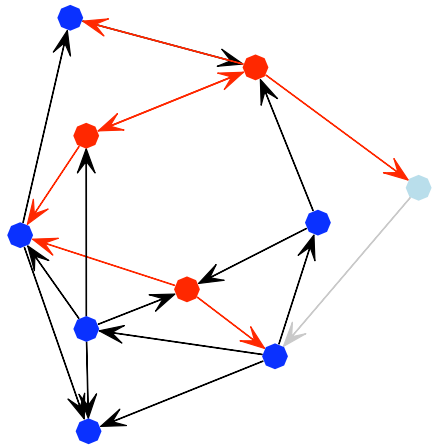
Partial Observation of Social Networks

- **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?"



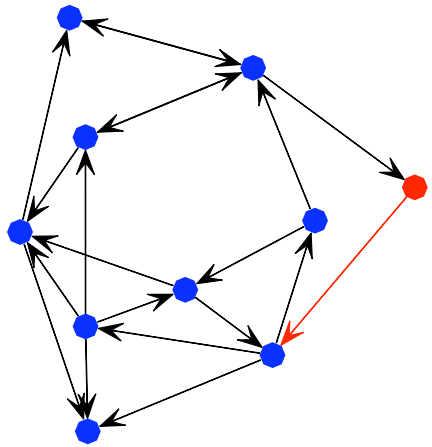
Partial Observation of Social Networks

- **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?"



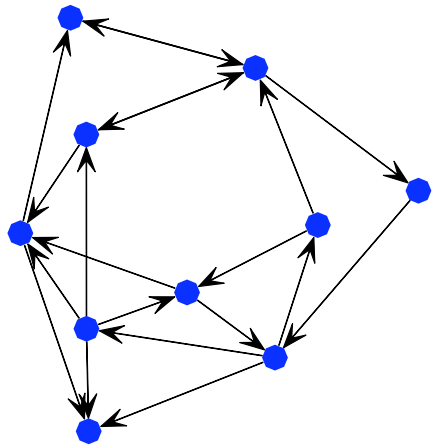
Partial Observation of Social Networks

- **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?"



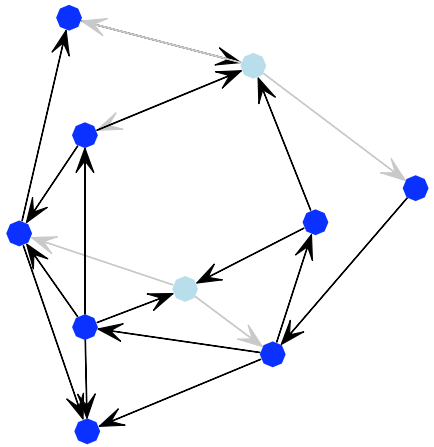
Partial Observation of Social Networks

- **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?"



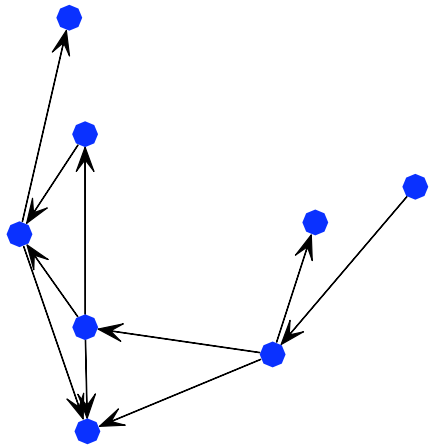
Partial Observation of Social Networks

- **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?"



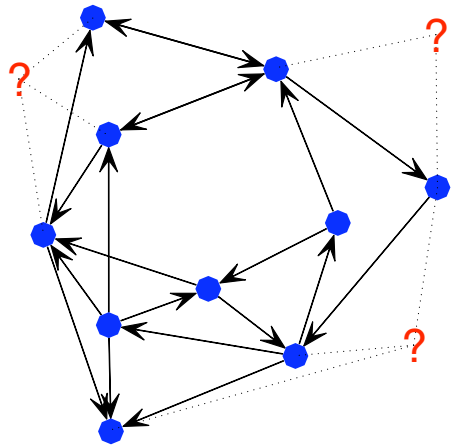
Partial Observation of Social Networks

- **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?"



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 compute likelihood by summing over the possible values of unobserved, ignoring sampling
- In practice, use 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*)

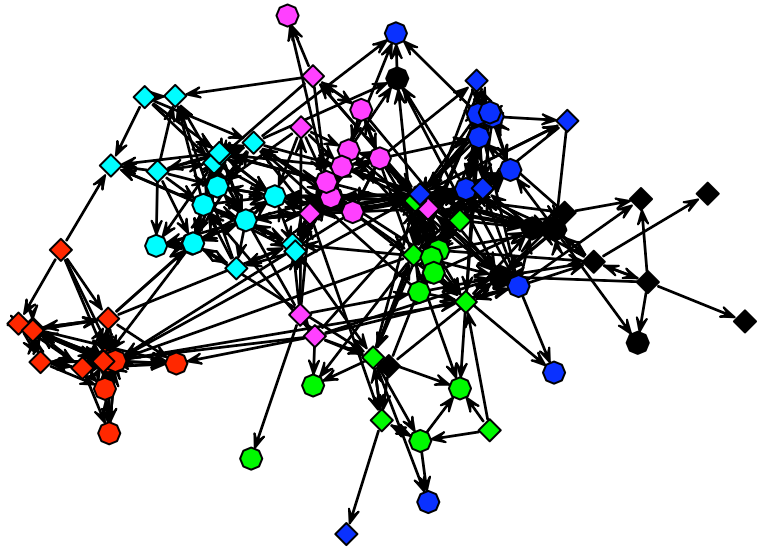
Then

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

- Can compute likelihood by summing over the possible values of unobserved, ignoring sampling
- In practice, use Markov Chain Monte Carlo (MCMC)

Friendships in a School

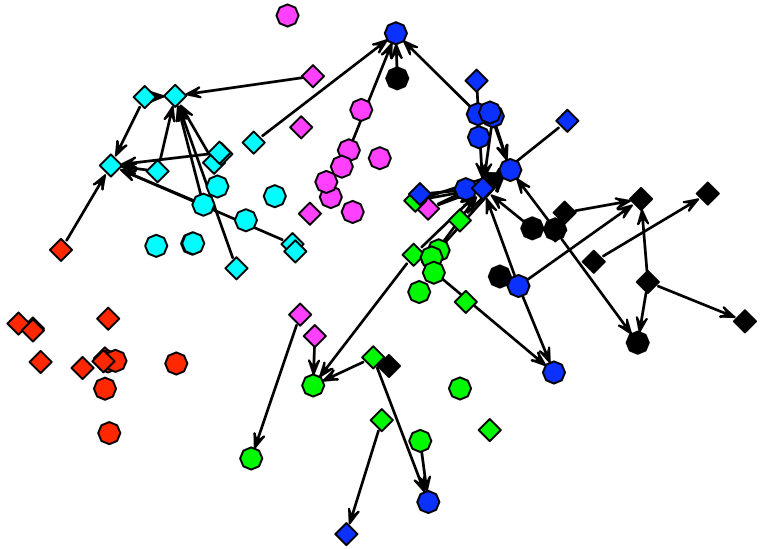
From the National Longitudinal Survey on Adolescent Health - Wave 1:



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

Friendships in a School

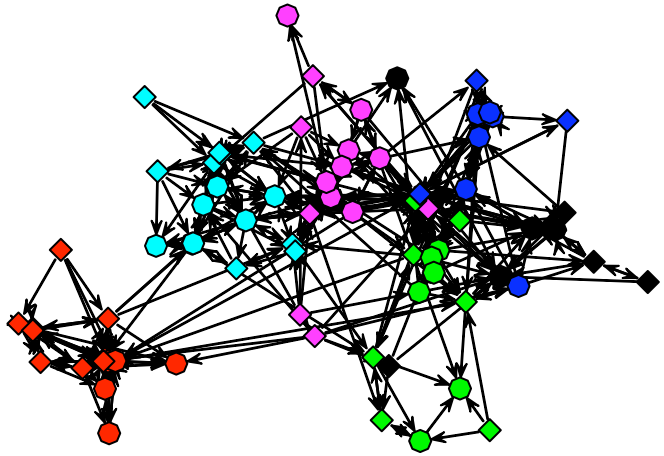
From the National Longitudinal Survey on Adolescent Health - Wave 1:



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

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.

Friendships in a School

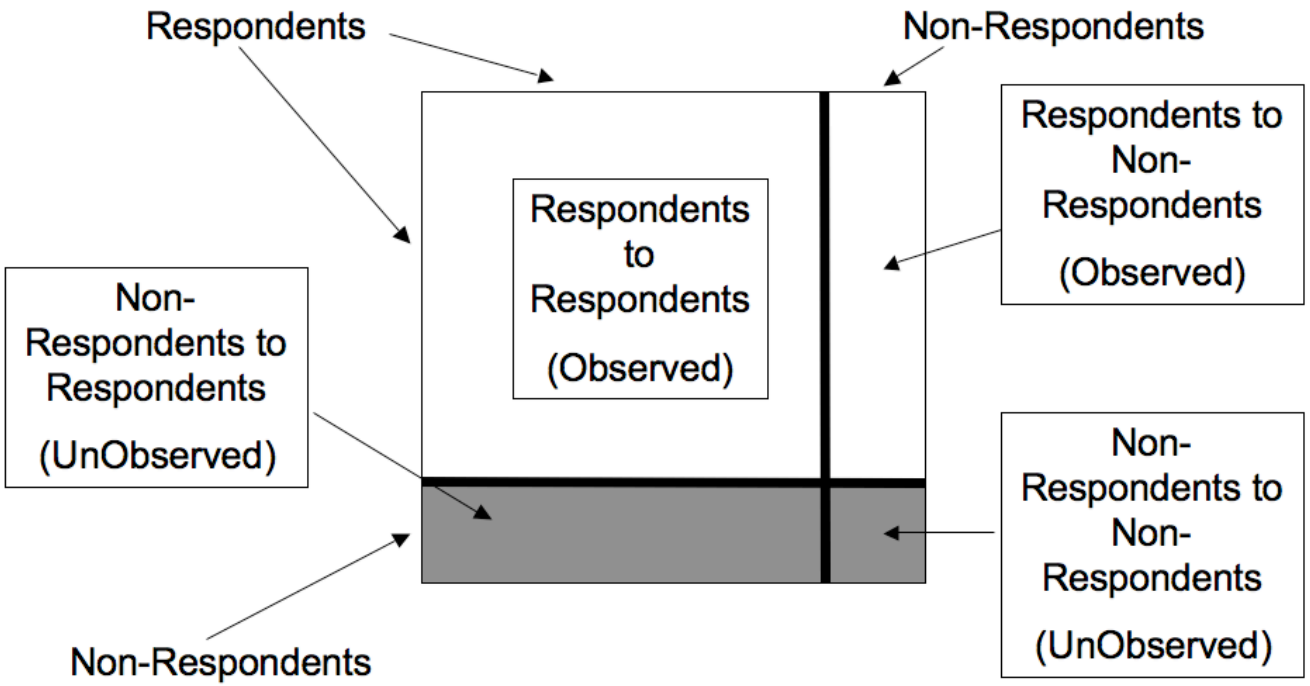
- **Methodological Question:** Can we fit a network model to a network with missing data? Is the fit different from that of just the observed data?

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



	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***

	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***

Link-Tracing is Adaptive!

- 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.

Discussion

- Likelihood inference is possible with missing data!
- 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.

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: handcock@stat.washington.edu

Thank you for your attention!