Social Networks and Epidemiology

Missing data in social network analysis

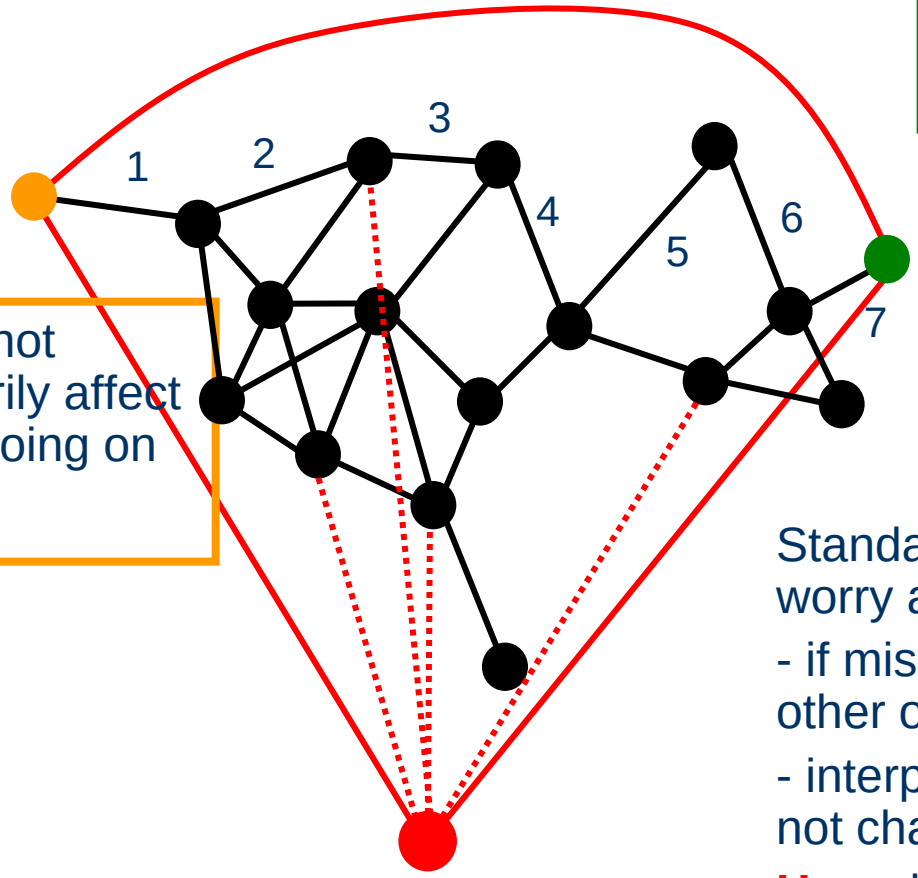
Johan Koskinen*

The 8th Asia-Pacific Complex Systems Conference, July 2-5, 2007, Surfers Paradise, QLD
*Department of Psychology, School of Behavioural Science, The University of Melbourne,
research sponsored by the DSTO

ARC Centre for Complex Systems
School of ITEE | The University of Queensland | ST LUCIA QLD 4069 | AUSTRALIA

T: +61 7 3365 1003 | F: +61 7 3365 1533 | E: outreach@accs.edu.au

Intuitive understanding of distance in networks



what goes on here…

… does not necessarily affect what is going on here

Standard quantitative methods only worry about:

- if missing observation differs from other observations

- interpretation of other observations do not change

Here, because of dynamics, self organisation, other observations may be affected **even** if missing observations similar to observed observations

THE UNIVERSITY OF MELBOURNE

Outline

What is a network - The “boundary problem”

Concepts and problems of missingness (standard cases)

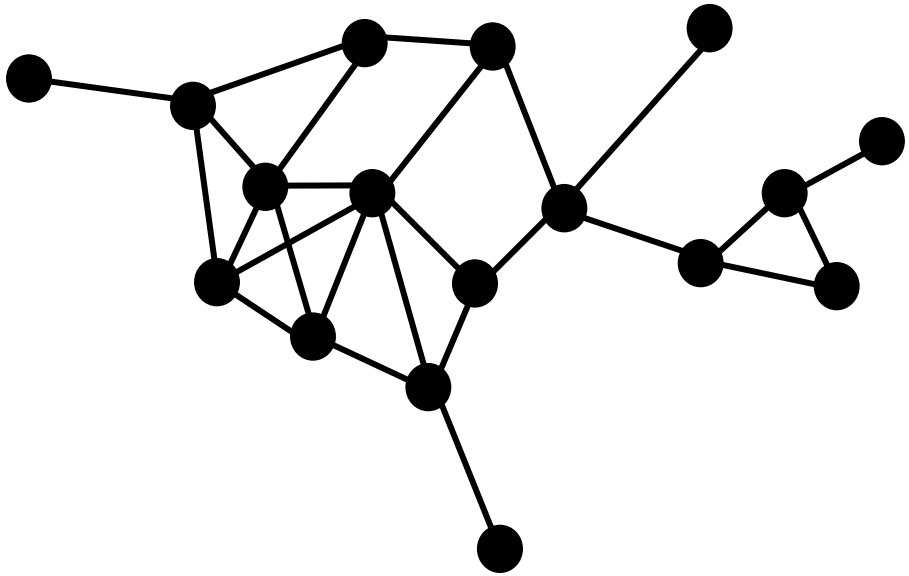
- Missing edges and non-respondents
- Pooling different views of the network (unobserved network I)
- Missing nodes
  - Sampling in/on networks
- Vertex doppelgangers - Node mapping between networks
- Misc.
  - Underlying patterns (unobserved network II)
  - Snapshots in time

Case: an approach for missing edges

What is a network

Ontological problem

Epistemological problem



The boundary problem (Laumann et al., 1983)

Nominalist approach

- group membership
- actor attributes
- relations
- events

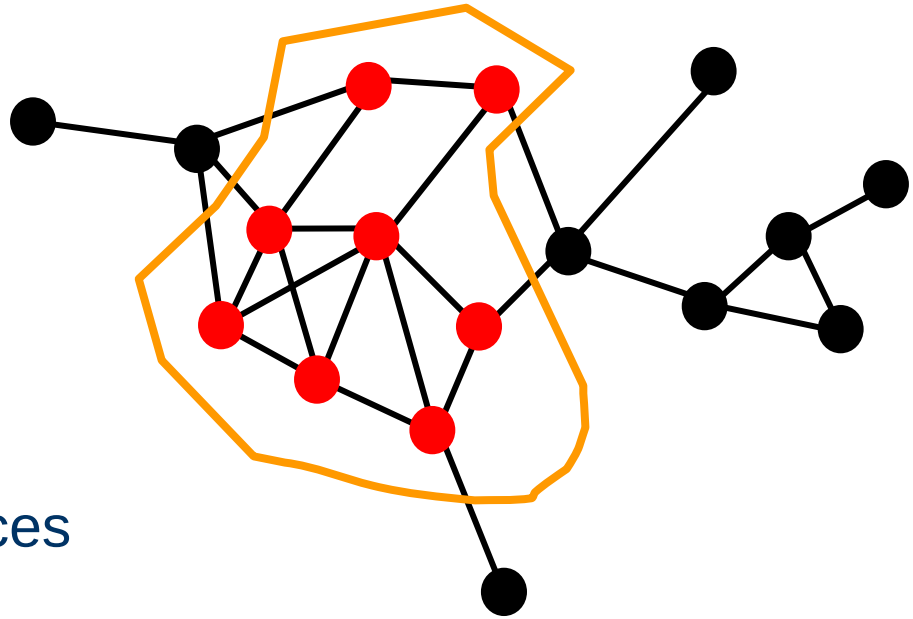
Same type of tie crossing

- e.g. vertiex set ≈ school, choices outside school
- What if (s.o.t.) most important ties are to others?

Different type of tie crossing

-Tie crossing boundary ties into process of study-tie (c.p. exchange; measured advice nw contingent on unmeasured friendship nw)

- Overlap of settings



The boundary problem (Laumann et al., 1983)

Realist approach

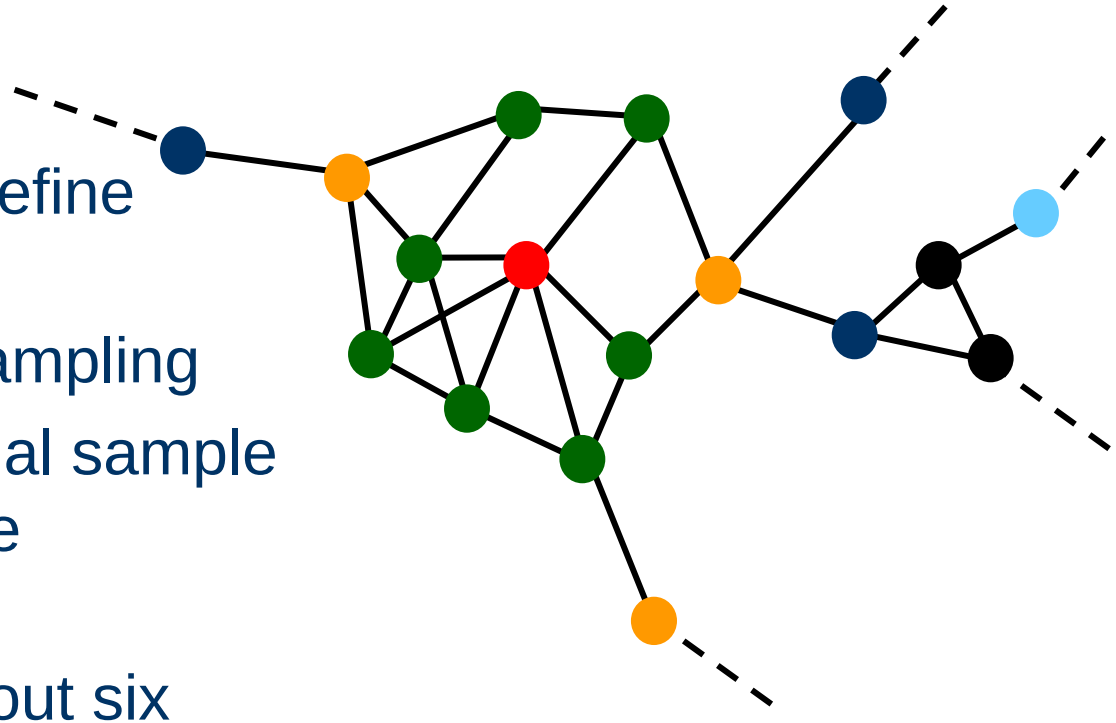
Let actors themselves define network boundaries

- essentially snowball sampling

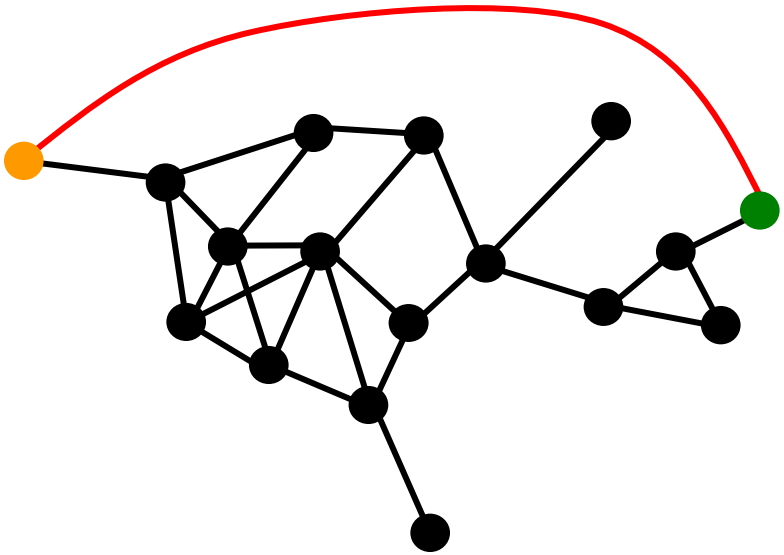
- how do we choose initial sample (if we need that to define "population")

- what if they're right about six degrees…

- where do we stop without violating the realist approach?



Concepts and problems: Missing edges and indicators



Missing "edges"

- fixed (maximum) choice
  - by design
  - by fatigue/response bias
  - observational inadequacy

Missing edge indicators

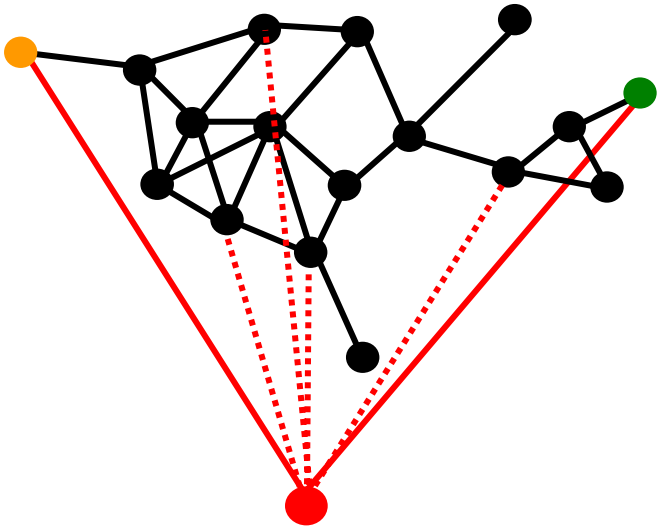
(Note: distinct from erroneously reported ties/non-ties)

- badly defined "group" for toft
- by fatigue/response bias
- observational inadequacy

When inferring structural properties:

possible to estimate missing observations and structural parameter simultaneously (Koskinen, 2007)

Missing edges and non-respondents



Non-respondents

- special case of missing edge indicators
- potentially different/specific missing data mechanism

Infering structural properties

- for digraphs: “complement” data (Stork & Richards, 1992)
- treat missing nodes as “special” (Robins et al., 2004)
- treat as missing edge indicators (Gile & Handcock, 2006; Koskinen, 2007)

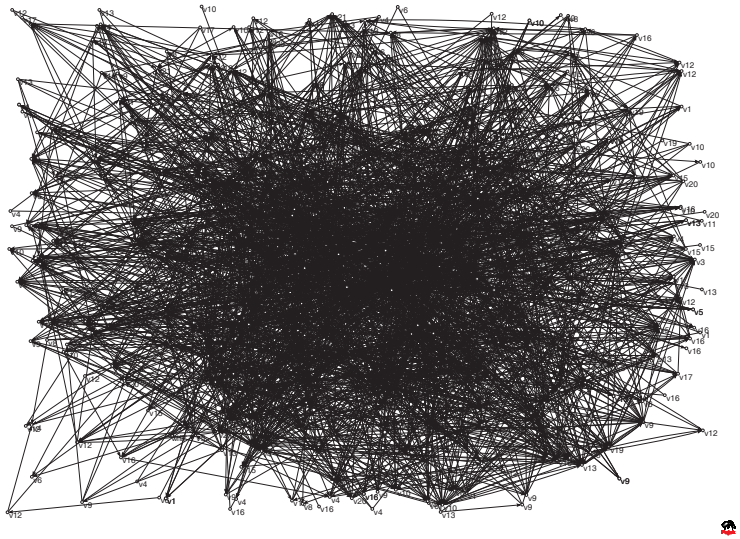
How much can we take away?

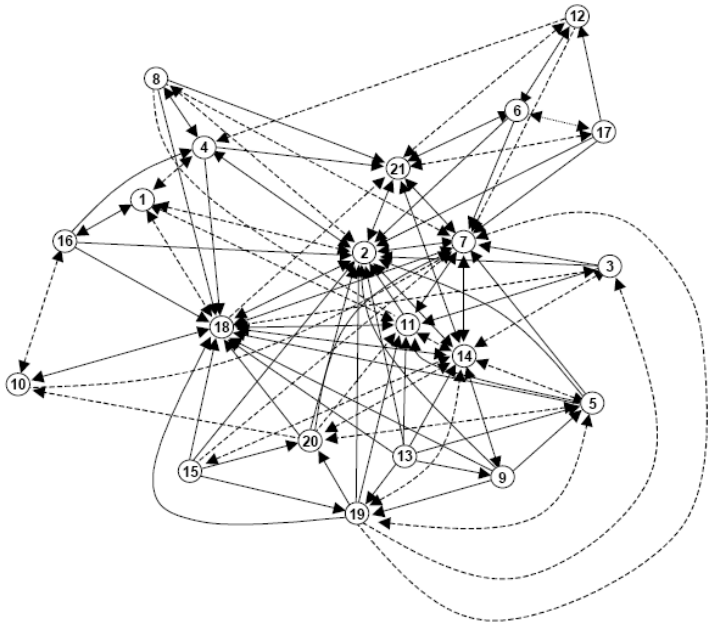
Effect on indices etc see Kossinets (2006); Costenbader & Valente (2003); & Huisman (2007)

Pooling different views of the network (unobserved network I)

Cognitive Social Structures (CSS) (Krackhardt, 1987)

Let each actor give his or her view of the entire network – n versions





What is missing? The "true" network:

- sender receiver agree on tie present
- at least one of s/r reports tie
- let majority decide
- use some paramteric model (e.g. Butts, 2003)

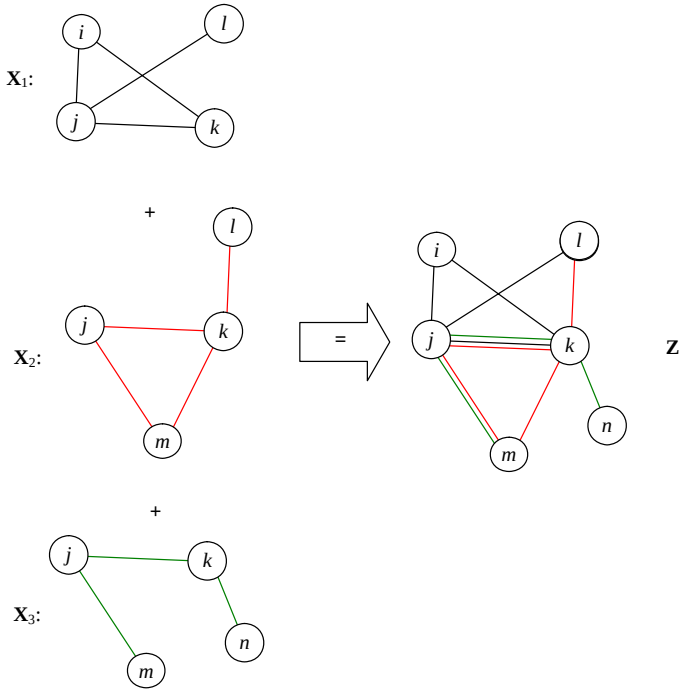
Pooling different views of the network (unobserved network I)

Partial Cognitive Social Structures (PCSS)

- What if each actor/observer give his or her view of part of the network – *K* versions of *K* different parts
- Different possibly conflicting sources of information

Can we pool these using CSS?

- applying CSS straight (not so clever; Koskinen et al., 2002)
- incorporating some mechanism for what parts are reported on (clever; Butts, 2007)
- Work needed on distribution for $\mathbf{Z}$ (prior or latent variable)



Missing nodes

Known who is missing?

Can we treat the missing node(s) as non-respondent(s)?

Unknown who, how many and where missing nodes are?

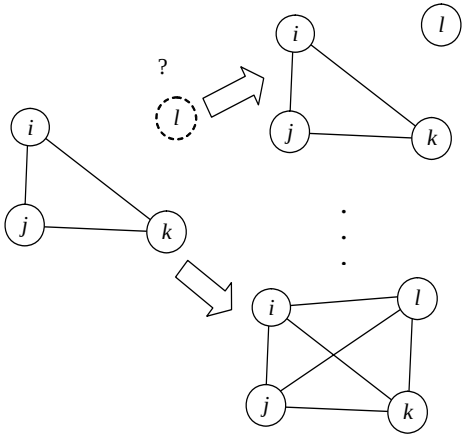
Multiple ERGM

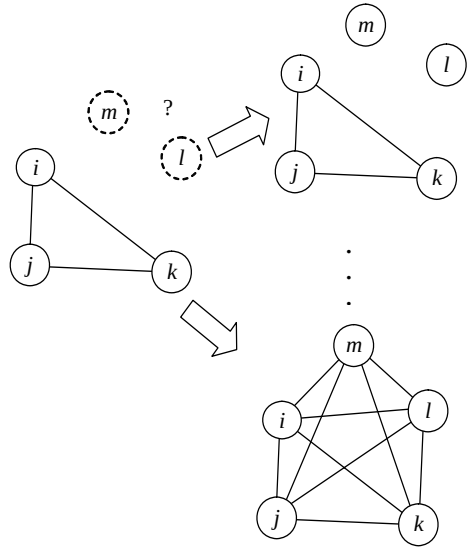
- can we infer structural properties under different hypothetical scenarios for what is missing?

- do inferred structural properties inform us where what is missing is?

Missing by design

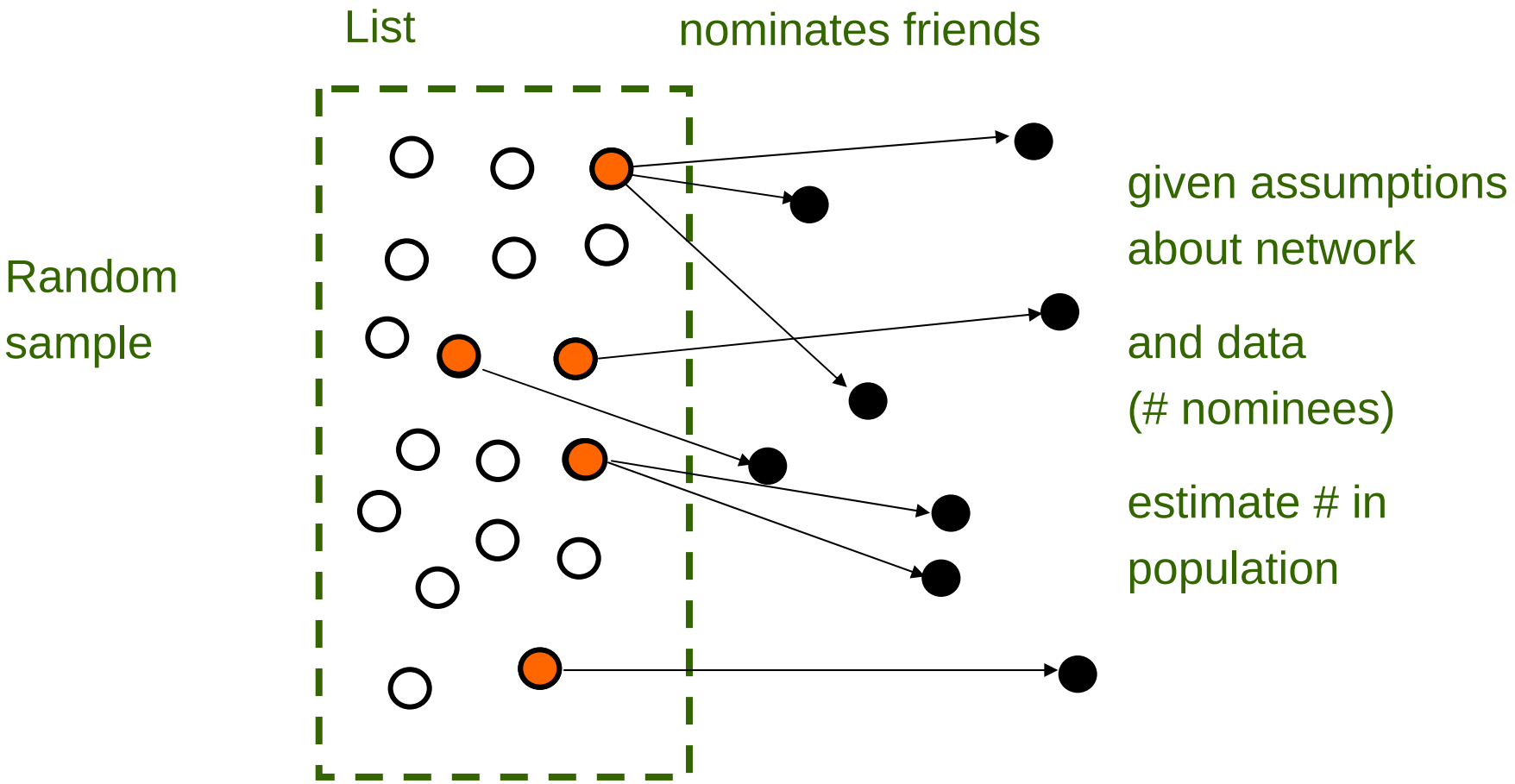
Can we make do with the sample we got? Snowball…





Sampling in/on networks – networks as auxiliary variables

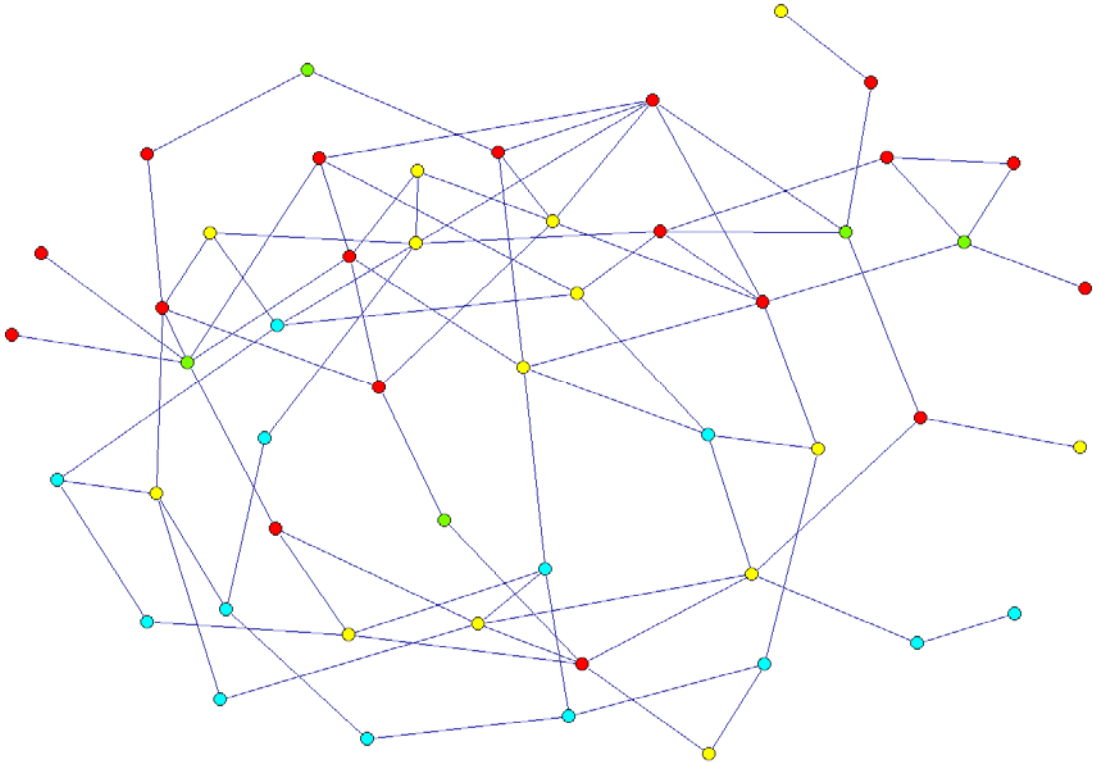
Estimating the size of a population (Frank & Snijders, 1994)

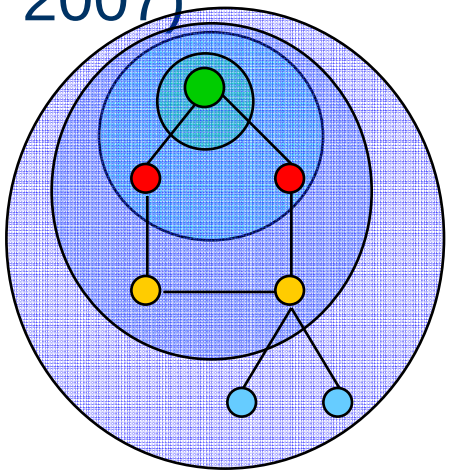


Sampling in/on networks – inferring networks

Results for how sample properties relate to populations under srs (e.g. Karlberg, 1998)

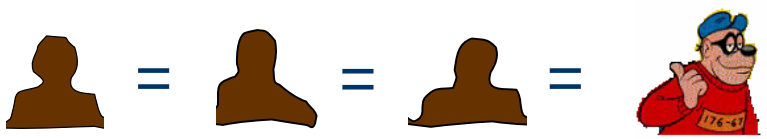
Using snowball sampling to infer structural dynamics in population (Pattison et al., 2007; Handcock, 2007)

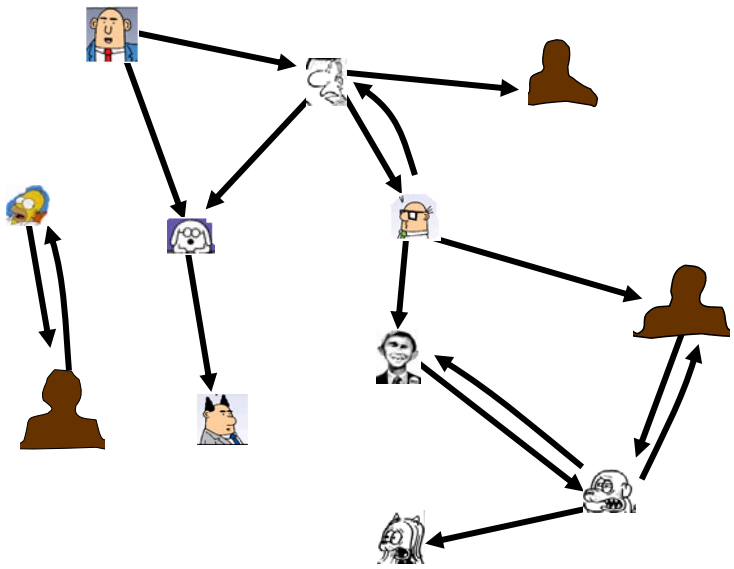


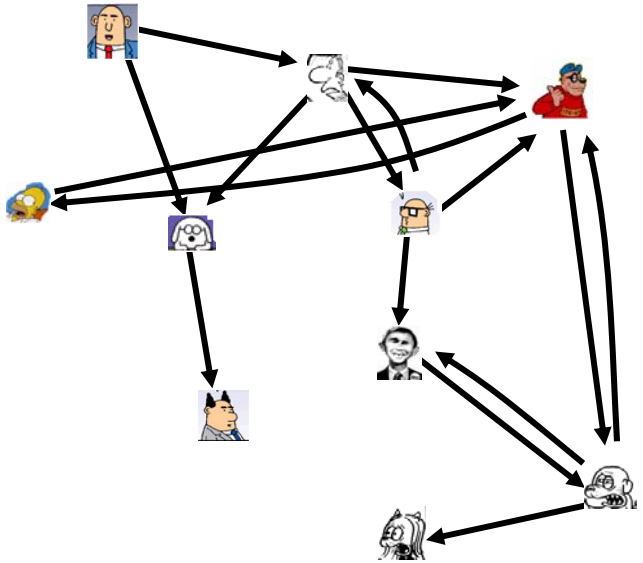


Vertex doppelgangers - Node mapping between networks

Suppose you have email exchange but some people have multiple email addresses

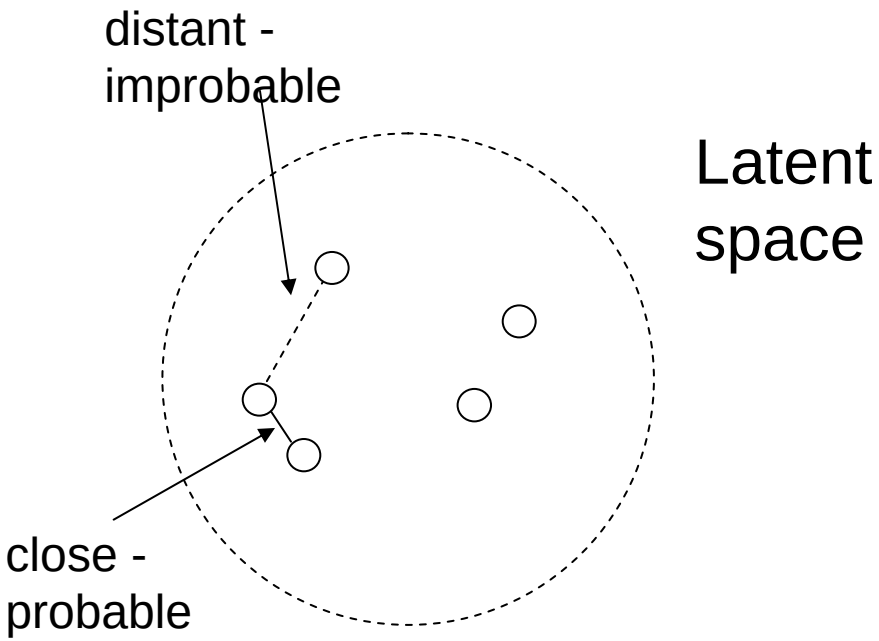






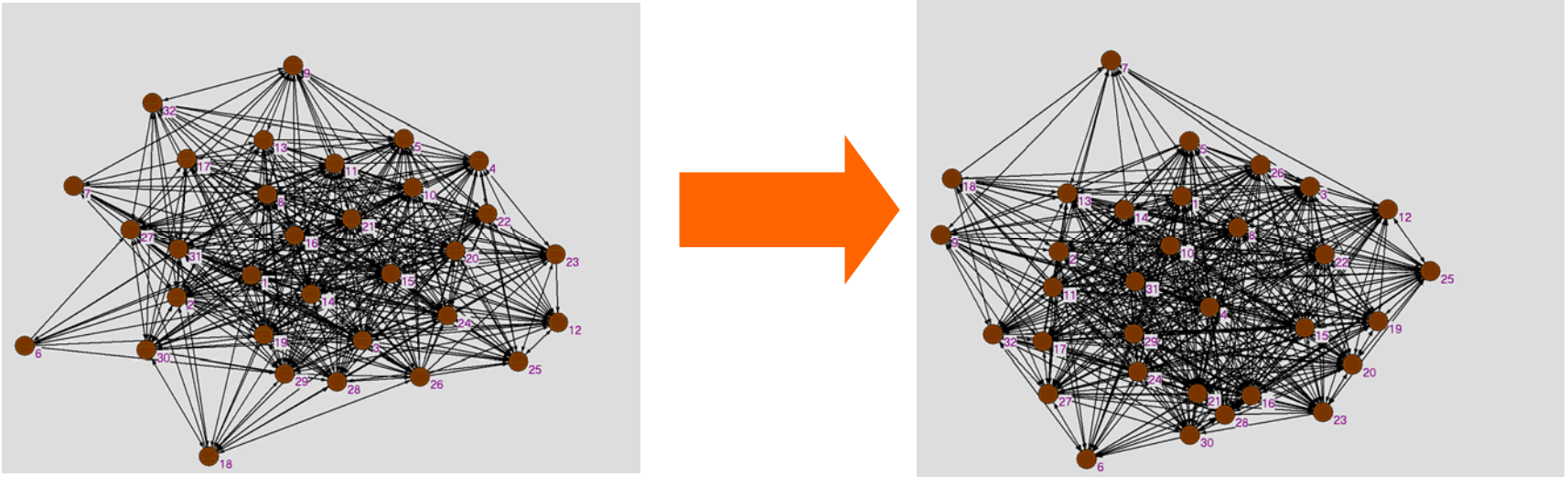
Underlying patterns (unobserved network II)

What if observed/measured interaction reflects some underlying social space (Schweinberger & Snijders, 2003; Hoff, Raftery, & Handcock, 2002)



Snappshots in time

Assume that we have two observations of the same network at different points in time



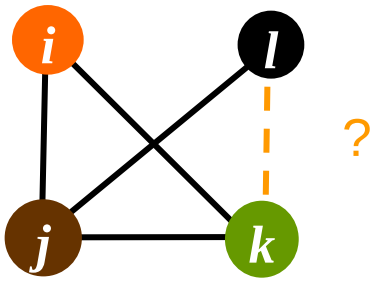
Typically the two observations differ

there has been change

the order in which changes took place and reversed changes is missing…

Case: an approach for missing edges

The aim

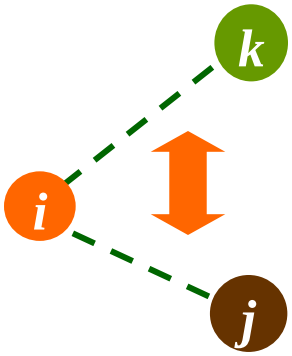


Fit exponential random graph (ERGM/p*) model to data using

- Bayesian inference
- Assuming dyads are missing at random

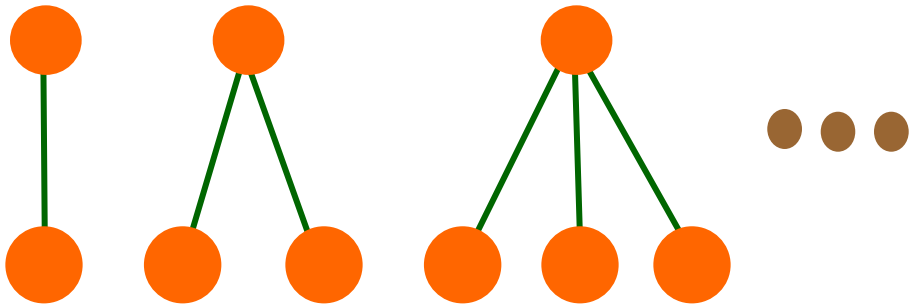
ERGM (p*): I. Markov Graphs

(Frank and Strauss, 1986)

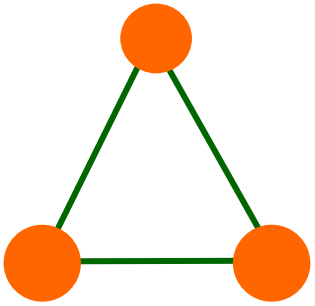


two edge indicators $\{i,j\}$ and $\{i',k\}$ are conditionally dependent if $\{i,j\} \cap \{i',k\} \neq \varnothing$

Hammersley-Clifford theorem (with homogeneity restrictions) imply that a Markov Graph has as sufficient statistics (counts of):



degree distribution; preferential attachment, etc



friends meet through friends; clustering; etc

ERGM (p*): II. Extensions Markov Graphs

Attributes of actors and dyads to capture selection effects, homophily, etc

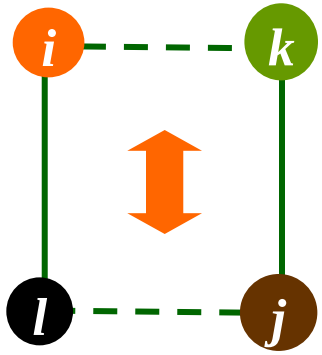
counts:

$$\log \frac{\Pr(i \text{ — } k \text{ given the rest})}{\Pr(i \quad k \text{ given the rest})} \approx \sigma_1 \Delta\# + \sigma_2 \Delta\# \cdots + \tau \Delta\#$$

p* and the pseudo likelihood (almost logistic regression… ?)

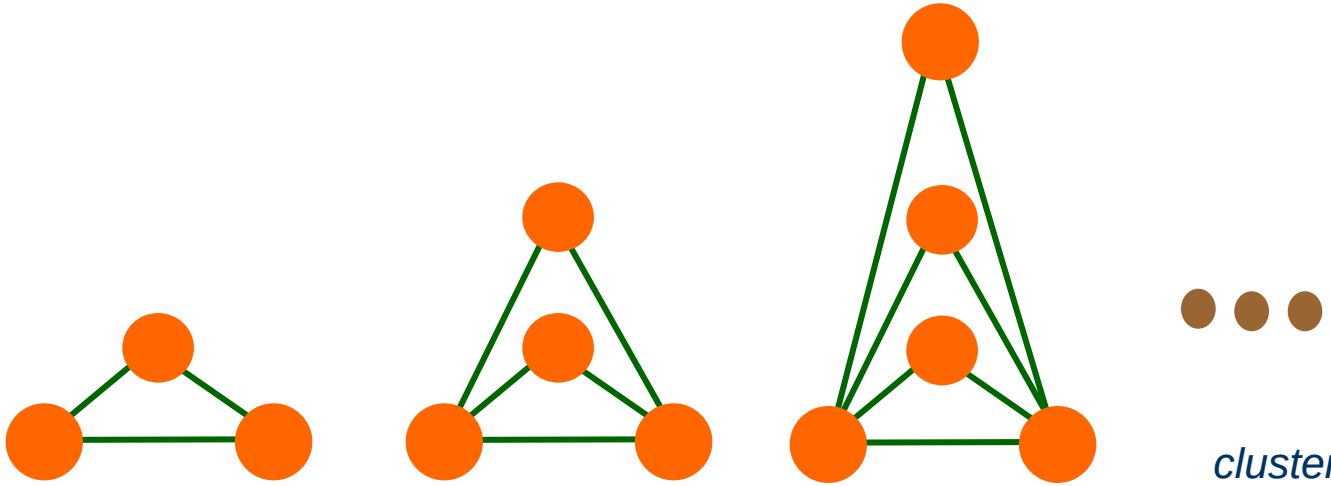
ERGM (p*): III. Extensions Markov Graphs

(Snijders, Pattison, Robins and Handcock, 2006)



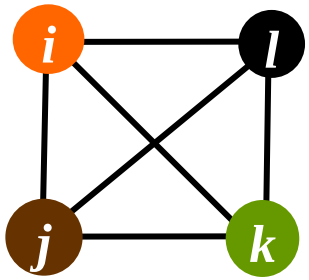
two edge indicators $\{i,k\}$ and $\{l,j\}$ are conditionally dependent if $\{i,l\}$, $\{l,j\} \in E$

Hammersley-Clifford theorem (with homogeneity restrictions) imply that sufficient statistics are (counts of; in addition to Markov):



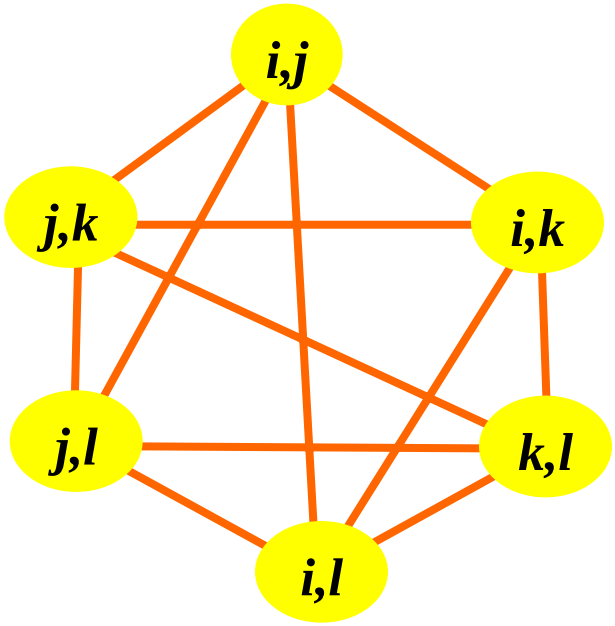
clustered regions

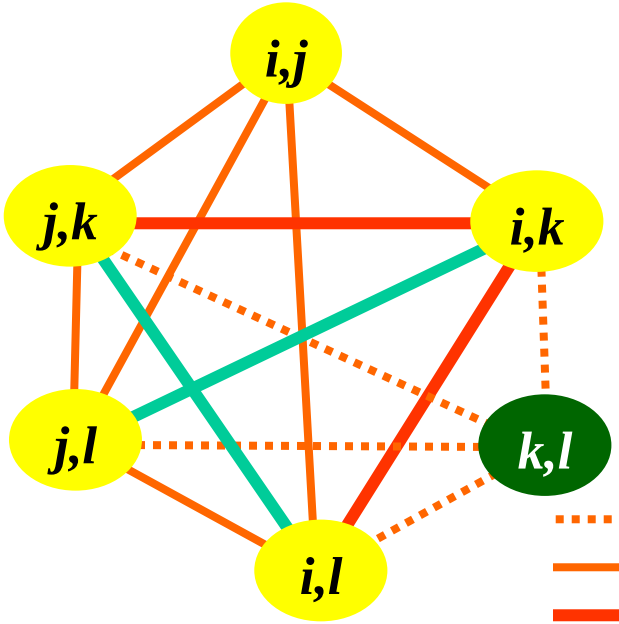
Why important – Markov Graphs



$\{i,j\}$ & $\{k,l\}$ cond. dep. if

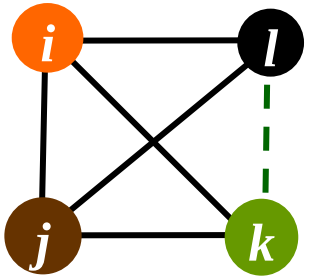
$\{i,j\} \cap \{k,l\} \neq \emptyset$





remove $\{k,l\}$,

join parents



?

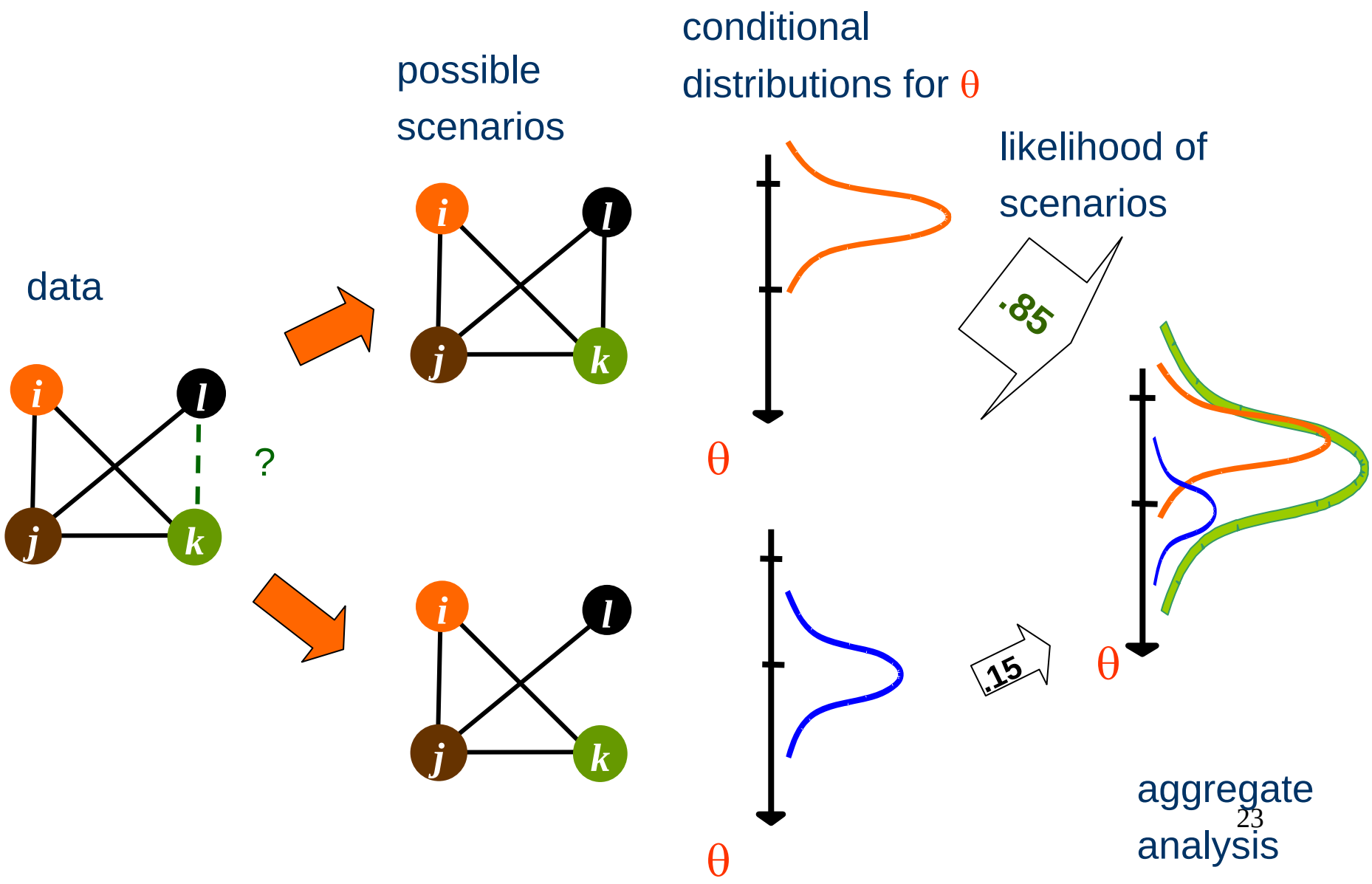
removed

old

con. parents

new

Why important – Markov Graphs



Why important – Markov Graphs

Exponential random graph model - **the model:**

$$p(x \mid \theta) = \exp\left\{ \sum_k \theta_k s_k - \psi(\theta) \right\}$$

normalising constant:

$$\psi(\theta) = \log\left(\sum_{\text{all graphs}} \left\{ \exp \sum_k \theta_k s_k \right\} \right)$$

Missing data: partition data x into observed part u and missing v

Posterior for any given realisation $x = (u, v)$

$$\pi(\theta \mid x) = \frac{\exp\left\{ \sum_k \theta_k s_k - \psi(\theta) \right\} \pi(\theta)}{\int \exp\left\{ \sum_k \theta_k s_k - \psi(\theta) \right\} \pi(\theta)\, \mathrm{d}\theta} \propto \exp\left\{ \sum_k \theta_k s_k - \psi(\theta) \right\} \pi(\theta)$$

Fitting ERGM with missing edges using Metropolis-Hastings

We use **LISA** (Koskinen, 2007) to draw θ from $\pi(\theta|\, x)$ for any $x=(u,v)$

$$\pi(\theta \mid x) \propto \exp\left\{\sum_k \theta_k s_k - \psi(\theta)\right\}\pi(\theta)$$

our "best guess" given x — **θ**

iterate

$$\pi(v \mid \theta, u) = \frac{p(u,v \mid \theta)\pi(\theta)}{\sum_{\text{all config } v} p(u,v \mid \theta)\pi(\theta)}$$

our "best guess" given x_{-i} & θ — **v**

For each v_i in $v_1, \ldots, v_m$

$$\pi(v_i \mid \theta, u, v_1, \mathrm{K}, v_{i-1}, v_{i+1}, \mathrm{K}, v_m) = \left[1+\exp(\theta^{\mathrm{T}}\delta_i x)\right]^{-1}$$

Change in s as element v_i change to 1- v_i

Fitting ERGM with missing edges – Lazega (2001) Lawyers

Collaboration network among 36 lawyers in a New England law firm (Lazega, 2001)

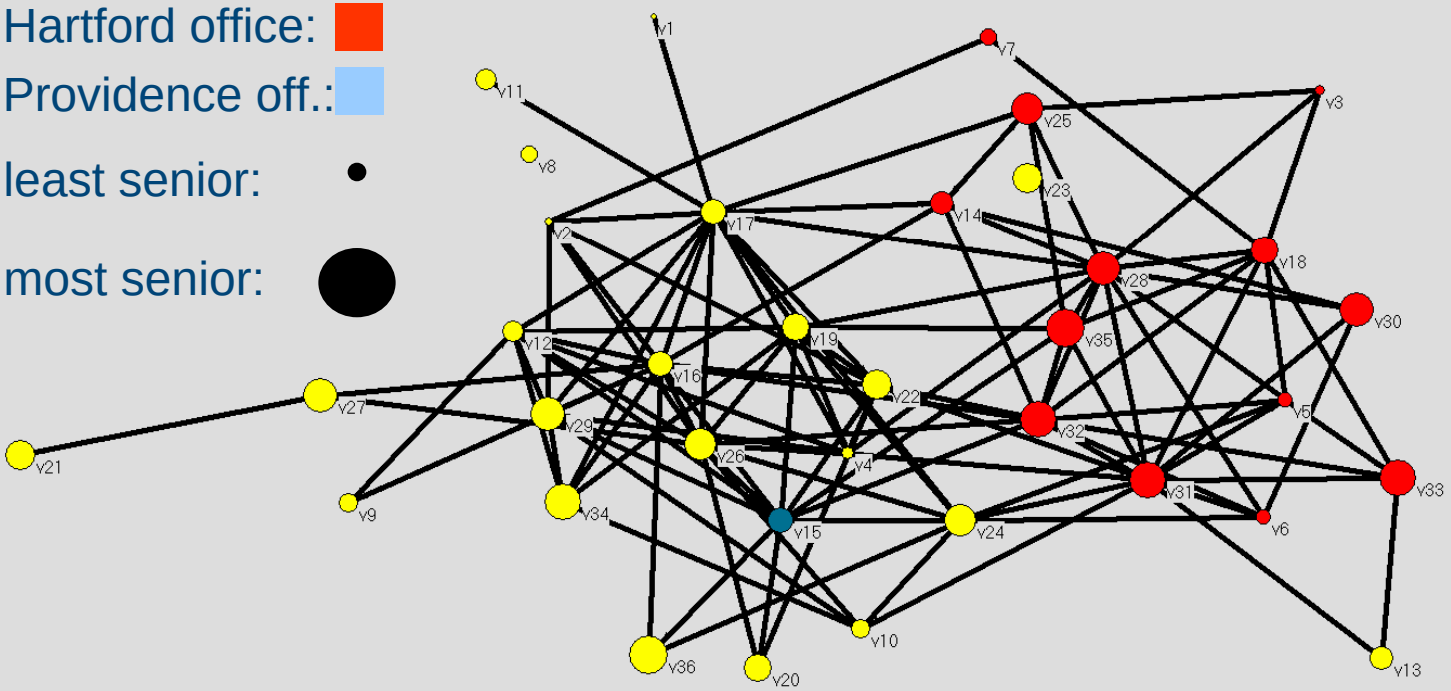
Boston office:

Hartford office:

Providence off..:

least senior:

most senior:



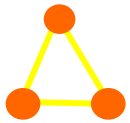
Fitting (C) ERGM with missing edges – Lazega Lawyers

Collaboration network among 36 lawyers in a New England law firm

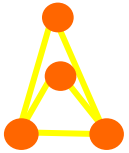
Model specification according to Snijders et al. (2006) and Hunter and Handcock (2006)

(b_i = 1, if i corporate, 0 litigation)

t_1 :



t_2 :



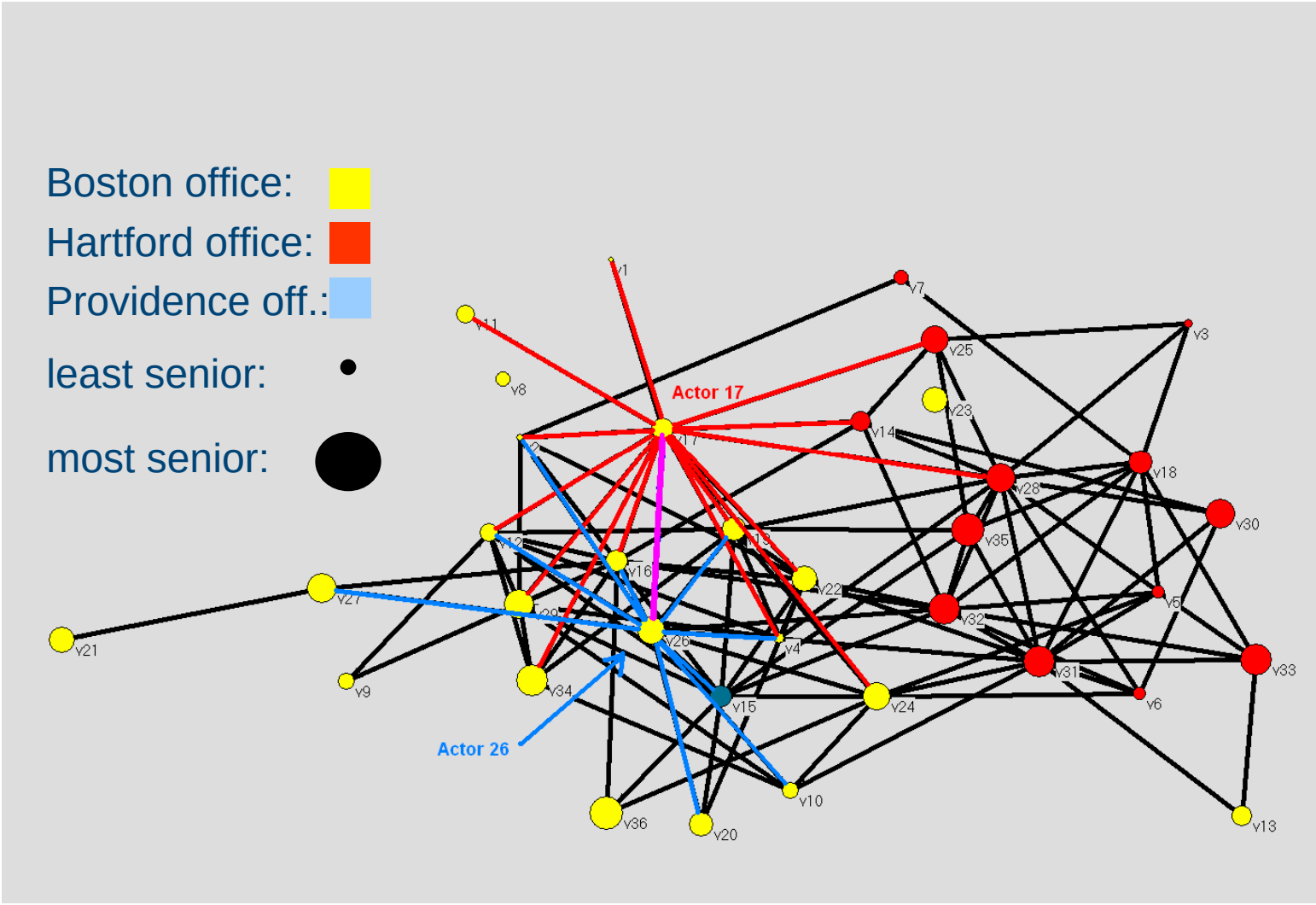
etc.

	Edges:	$\sum x_{ij}$
Main effect	Seniority:	$\sum x_{ij}(a_i + a_j)$
	Practice:	$\sum x_{ij}(b_i + b_j)$
Homophily	Practice:	$\sum x_{ij}\mathbf{1}(b_i = b_j)$
	Sex:	$\sum x_{ij}\mathbf{1}(c_i = c_j)$
	Office:	$\sum x_{ij}\mathbf{1}(d_i = d_j)$
GWESP:		$3t_1(x) - \frac{t_2(x)}{\lambda^1} + \Lambda + (-1)^{n-3}\frac{t_{n-2}(x)}{\lambda^{n-3}}$

with $\theta_8 = \log(\lambda)$

Fitting ERGM to Lazega Lawyers – an experiment

Now, let us pretend observations $X_{1,2}$ and $X_{17,26}$ are missing



Fitting ERGM to Lazega Lawyers – an experiment

Now, let us pretend observations ✠$_{1,2}$ and ✠$_{17,26}$ are missing

3-block M-H, circle through these 3 steps a large number of times

prior: $\pi(\theta)$

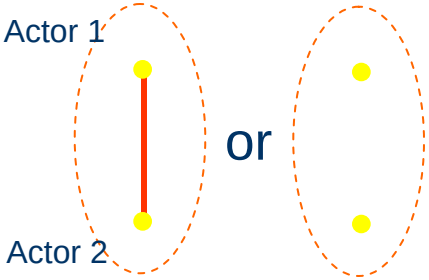
draw θ from

$\pi(\theta \mid u, x_{1,2}, x_{17,26})$

$\theta_1, \theta_2, \ldots, \theta_p$ given the rest

draw $x_{1,2}$ from

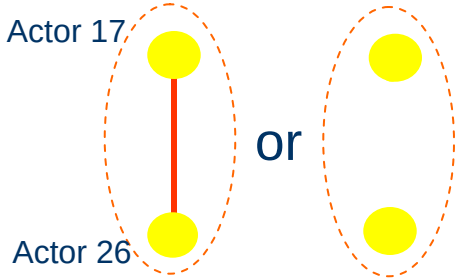
$\pi(x_{1,2} \mid u, \theta, x_{17,26})$



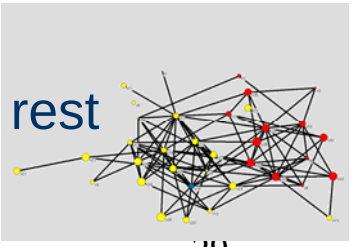
given the rest

draw $x_{17,26}$ from

$\pi(x_{17,26} \mid u, \theta, x_{1,2})$

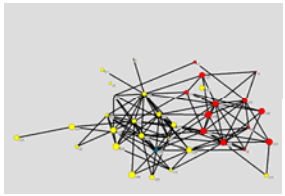


given the rest



Lazega Lawyers – an experiment: Posterior predictive distributions

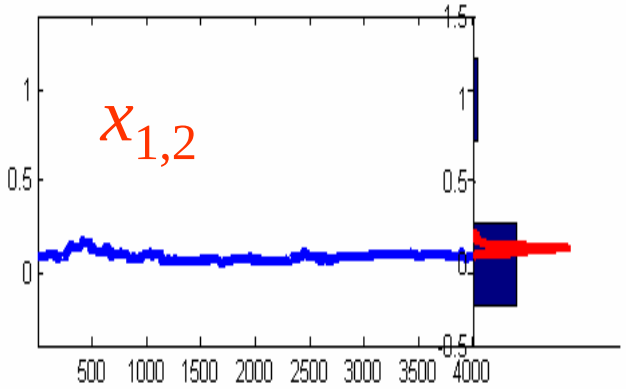
given *only*:

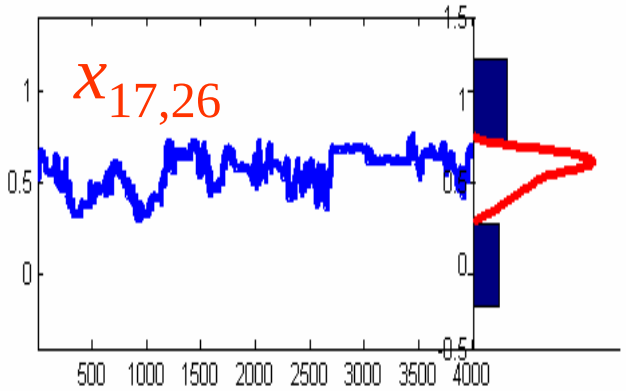


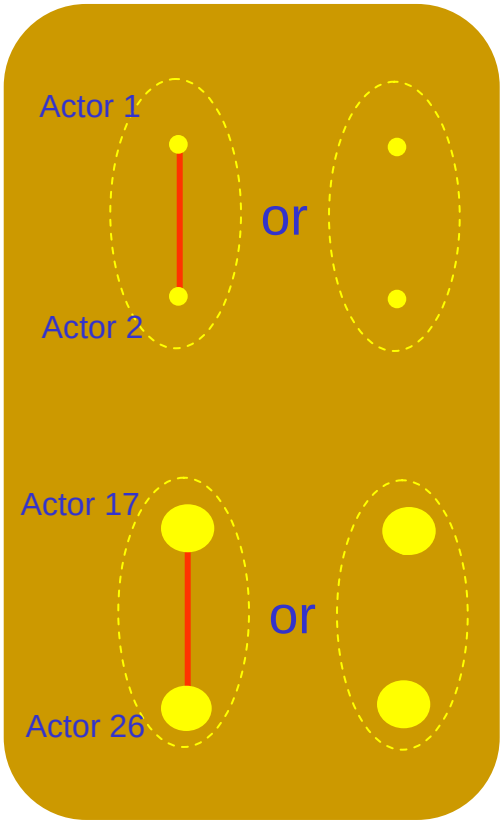
$\pi(v_i|\text{rest})$

hist $\pi(v_i|\text{rest})$

$\pi(v_i|\text{u})$

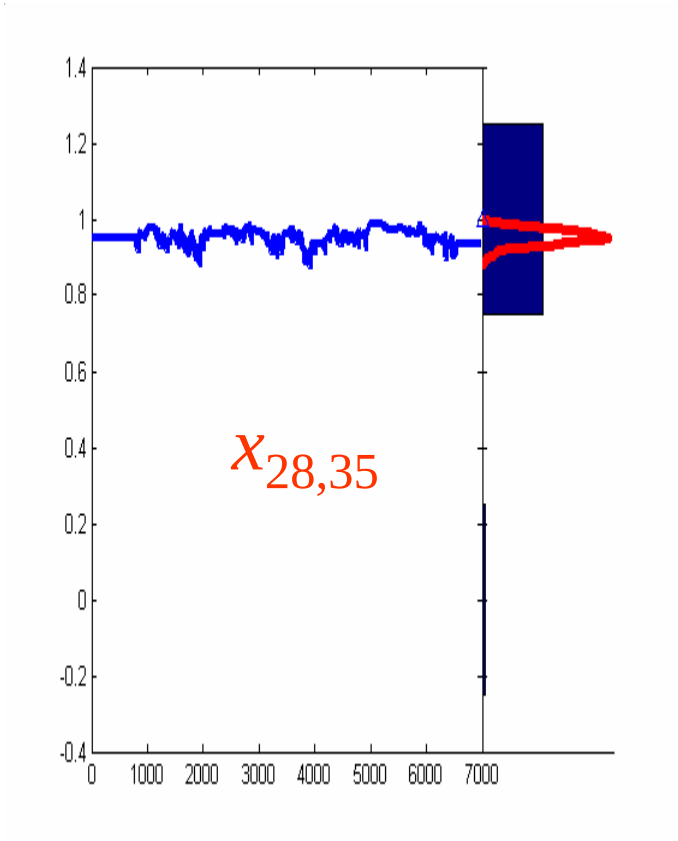


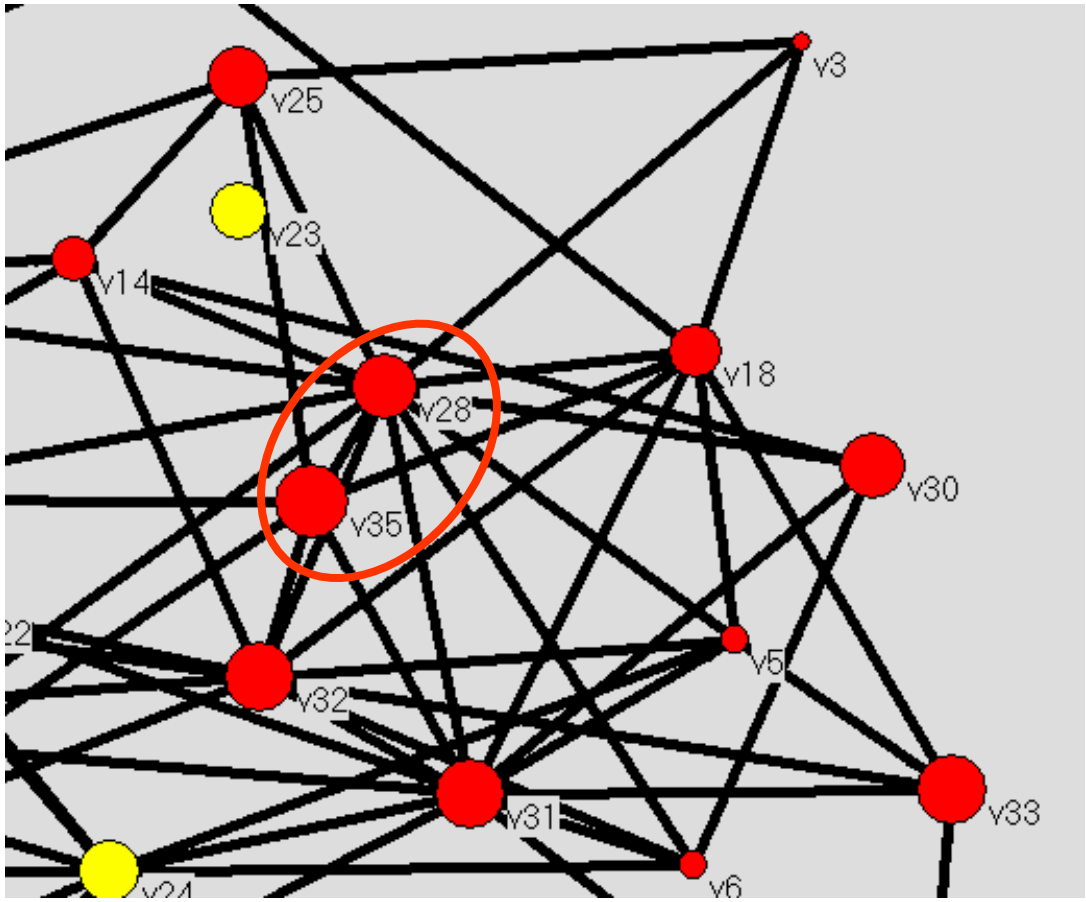




Lazega Lawyers – an experiment: Posterior predictive distributions

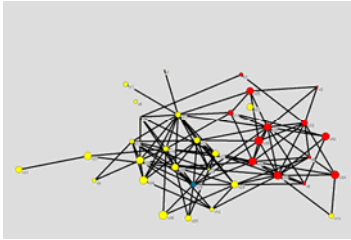
({28,35} one of 10 dyads removed at random)





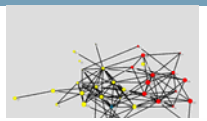
ERGM, Lazega Lawyers – more experiments

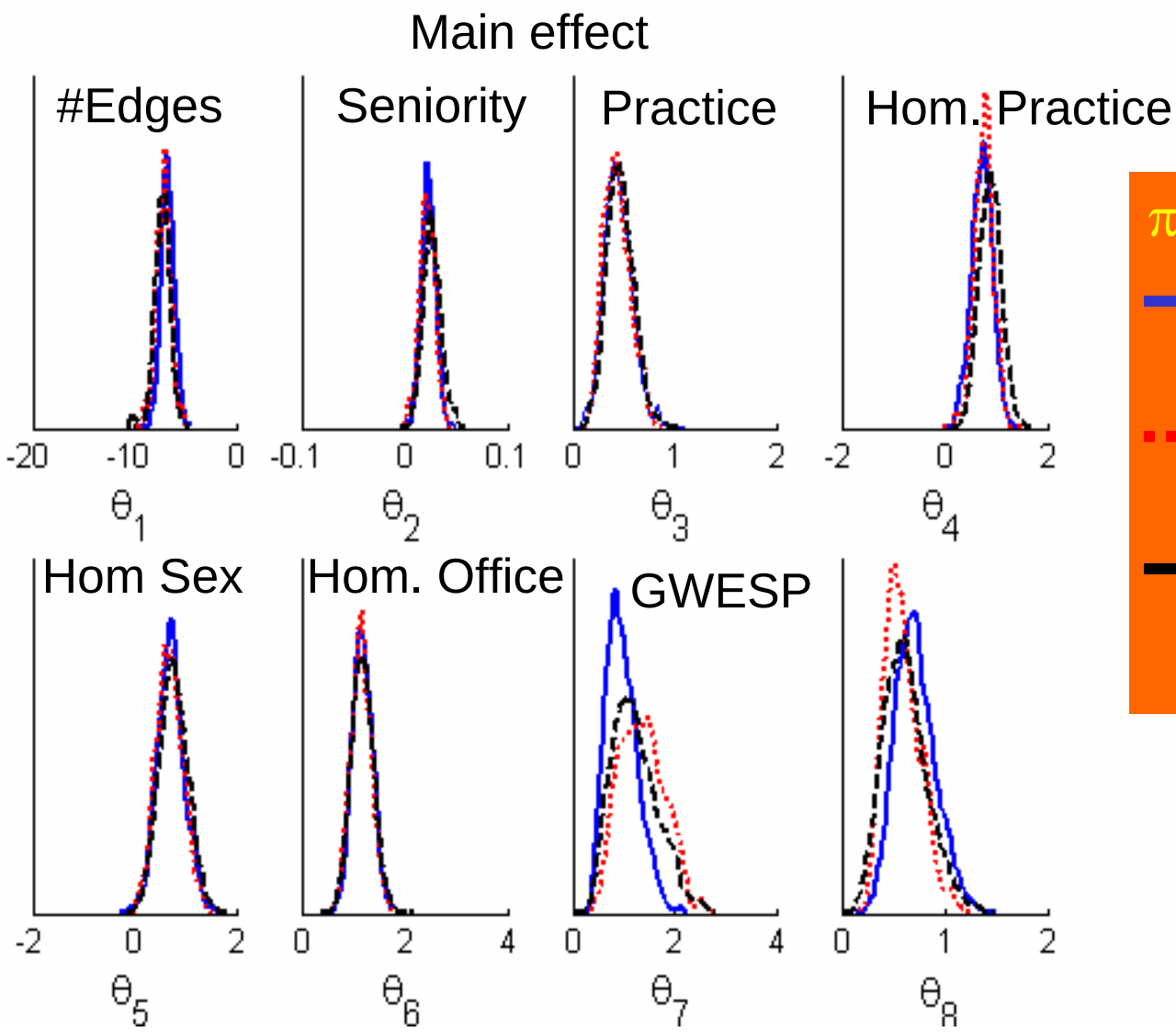
- Missing data scenarios
- Removed set of dyads
- A: ∅
- B:{1,2},{17,26}
- C: {1}×*N* \{1}, {2}×*N* \{2} (remove respondents 1 and 2)
- D: {1}×*N* \{1}, {2}×*N* \{2}, {3}×*N* \{3}, {4}×*N* \{4} (remove respondents 1, 2, 3 and 4, ca 20% of dyads)



Lazega Lawyers – more experiments: Posterior distributions

given **only** the observed part:

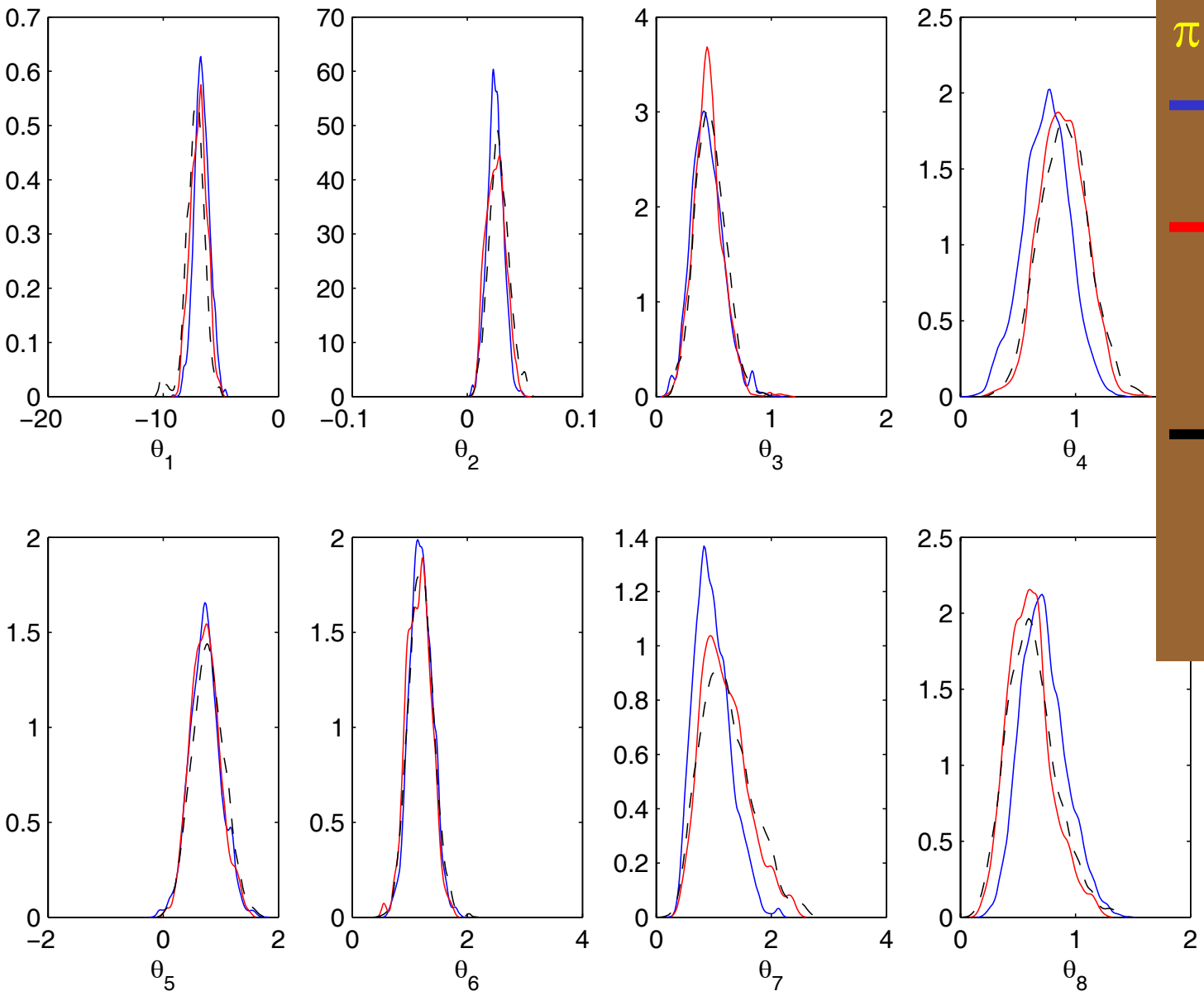




$\pi(\theta|u)$

- no missing (i.e. $u = x$)
- resp. 1 & 2 removed
- resp. ´1, 2, 3 & 4 removed

Lazega Lawyers – compare "naïve" (RO) approach



$\pi(\theta|u)$

- no missing (i.e. $u = x$)
- resp. ´1, 2, 3 & 4 removed REDUCED
- resp. ´1, 2, 3 & 4 removed BAYES AUG

- GWESP "global" measure
- 4 missing not enough to induce "phase transition"

Summary

We have outline some issues for missing data in SNA

We have illustrated the importance of treating missing data on ontological, epistemological as well as technical grounds

Presented an approach for making structural inference for data with missing edge indicators

- can we refine missing data mechanism?
- how much information can be missing?

Structure – missing – structure for missing edge indicators

- does this carry over to other forms of missingess… ?

References, I

Butts, C.T. (2003). Network inference, error, and informant (in)accuracy: a Bayesian approach. Social Networks, 25:103-140.

Butts, C.T. (2007). Network Inference from Unstructured Sources. Presented at Sunbelt XXVII, International Sunbelt Social Network Conference, 8th European Social Network Conference, Chandris Hotel Complex at Dassia Bay, Corfu Island, Greece, May 1-6, 2007.

Costenbader, E. , Valente, T.W. (2003). The Stability of Centrality Measures when Networks are Sampled. Social Networks 25: 283-307.

Frank, O. and Snijders, T.A.B. (1994). Estimating the Size of Hidden Populations Using Snowball Sampling. Journal of Official Statistics, 10: 53-67.

Gile, K. and Handcock, M.S. (2006) Model-based Assessment of the Impact of Missing Data on Inference for Networks, Working Paper no. 66, Center for Statistics and the Social Sciences, University of Washington

Handcock, M., (2007). Statistical models for dynamic networks for infectious disease spread based on egocentrically sampled data, The 8th Asia-Pacific Complex Systems Conference, July 2-5, 2007, Surfers Paradise, Gold Coast, QLD, Australia.

Hoff, P.D., and Raftery, A. E., and Handcock, M.S, (2002). Latent space approaches to social network analysis, Journal of the American Statistical Association, 97:1090-1098.

Huisman, Mark (2007). Imputation of missing data in social networks. Presented at Sunbelt XXVII, Corfu Island, Greece, May 1-6, 2007.

Karlberg, M. (1998). Triad count estimation in digraphs. Journal of Mathematical Sociology, 23:99-126.

References, II

Koskinen, J. (2007). Fitting models to social networks with missing data. Sunbelt XXVII, Corfu, Greece.

Koskinen, J., Jansson, I., and Spreen, M. (2002). The Role of Perceptional Data in Sampling Large Networks from Hidden Populations, in Hagberg (ed.), Contributions to Social Network Analysis, Information Theory, and Other Topics in Statistics; A Festschrift in honour of Ove Frank. Stockholm: Dept. of Statistics, Stockholm University.

Kossinets, G. (2006). Effects of missing data in networks. Social Networks, 28, 247-268.

Krackhardt, D. (1987), Cognitive social structures, Social Networks 9:109-134.

Laumann, E. O., Marsden, P. V., Prensky, D., 1983. The boundary specification problem in network analysis. In:Burt, R. S., Minor, M. J. (Eds.), Applied Network Analysis. Sage Publications, London, pp. 18-34.

Pattison, P.E., Robind, G., & Wang, P. (2007). Snowball sampling and exponential random graph models. Sunbelt XXVII, Corfu, Greece.

Robins, G.L., Woolcock, J., & Pattison, P. (2004). Models for social networks with missing data. Social Networks, 26, 257-283.

Snijders, T.A.B., P.E. Pattison, G.L. Robins and M.S. Handcock (2006) New specifications for exponential random graph models, Sociological Methodology 36: 99-153.

Schweinberger, Michael, and Snijders, Tom A.B. (2003). Settings in Social Networks: A Measurement Model, Sociological Methodology, 33:307-341.

Stork, D., Richards, W. D. (1992) Nonrespondents in communication network studies: Problems and possibilities. Group and Organization Management 17:193-209.