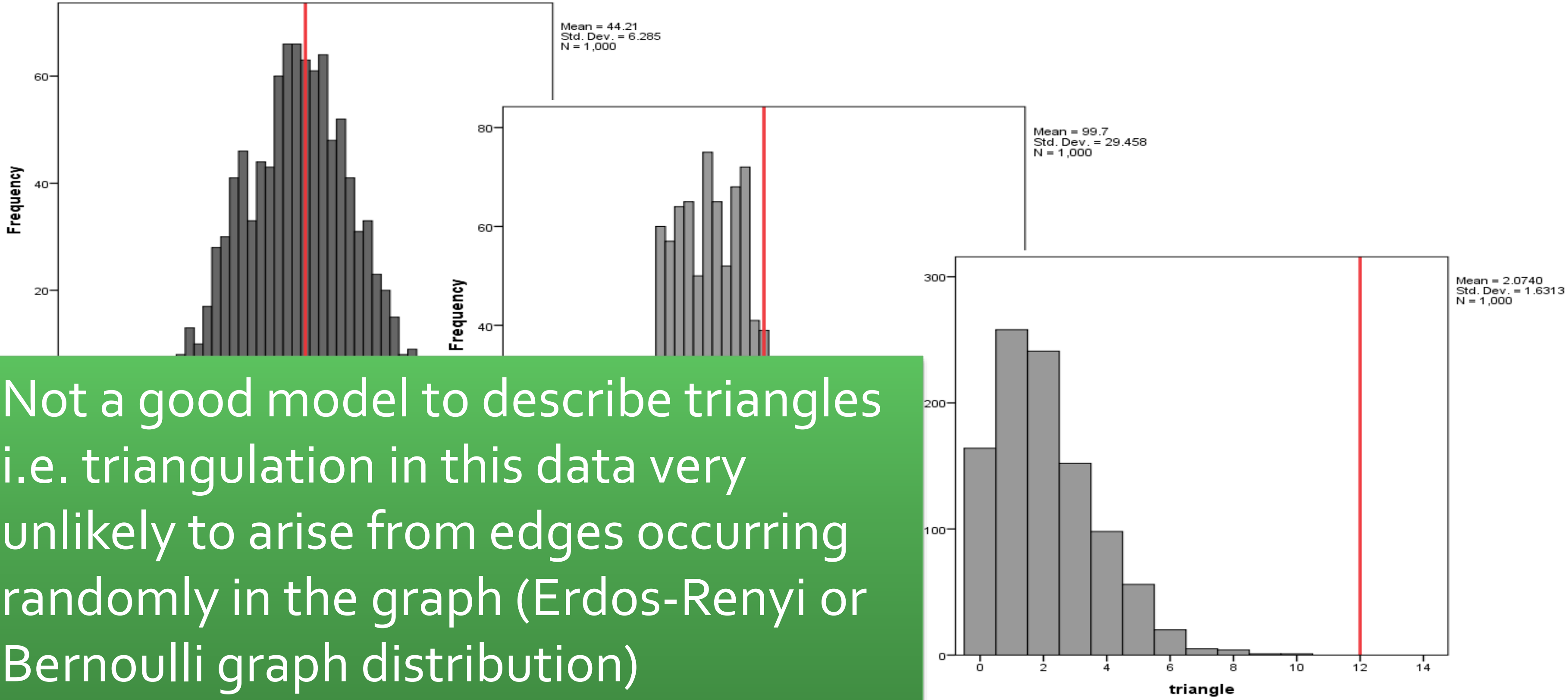
ERGM Specifications

Peng Wang

Simulating a Bernoulli graph distribution



Not a good model to describe triangles i.e. triangulation in this data very unlikely to arise from edges occurring randomly in the graph (Erdos-Renyi or Bernoulli graph distribution)

Simulating a Bernoulli graph distribution

Why is this a bad model?

Consider the implied tie-formation process:

For each pair

p=0.06259 is the probability coin comes up heads

Dependencies

Once we move beyond simple random graph models, we introduce dependencies among network tie variables

- These express various types of network self organization.
- A dependence assumption picks out certain types of network patterns – network configurations – that are possible in the model.
- In other words, we assume that the network is built up of these configurations.

Four generations of dependence assumptions

Bernoulli graphs: Network variables are independent of each other

Dyadic dependence: for directed graphs – dependence within dyads

Markov dependence: Network variables are (conditionally) independent unless they share at least one node.

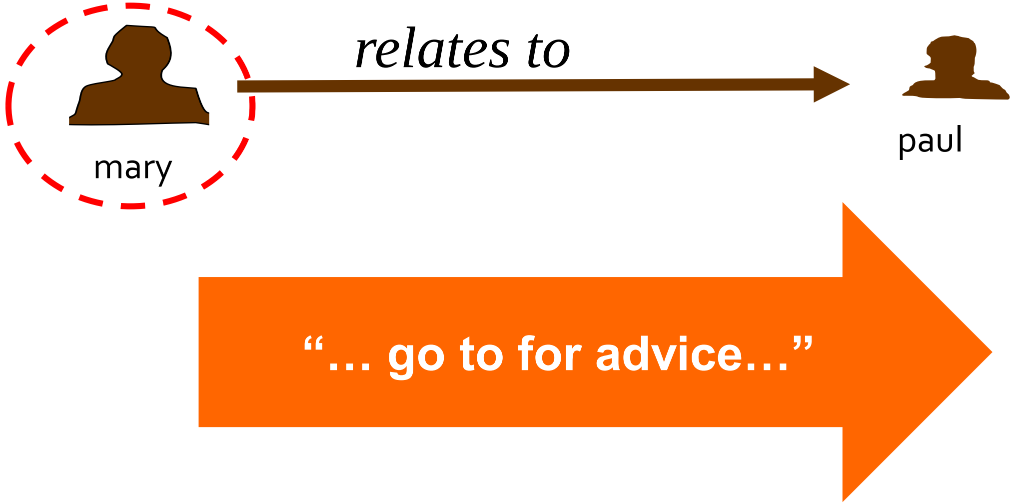
Social circuit dependence: Network variables are (conditionally) dependent if they create 4-cycles.

ERGMs and Dependence

- The general form of an ERGM
- Bernoulli models
- Markov models

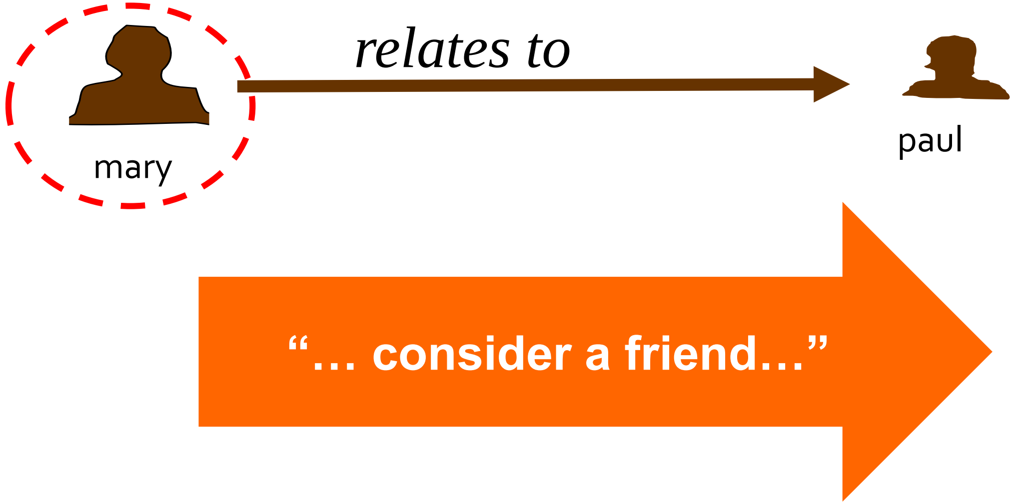
Some notation

We conceive of a network as a **Relation** defined on a collection of **individuals**



Some notation

We conceive of a network as a **Relation** defined on a collection of **individuals**



Some notation

We conceive of the **Graph** as a collection of

$$i \xrightarrow{(i\,,\,j)} j$$

Tie variables: $\{X_{ij} : i,j \in V\}$

$$x_{ij} = \begin{cases} 1 & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$$

Some notation

We conceive of the **Graph** as a collection of

$$i \xrightarrow{(i\,,\,j)} j$$

Tie variables: $\{X_{ij}: i,j \in V\}$

$$x_{ij} = \begin{cases} 1 & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$$

Generally binary

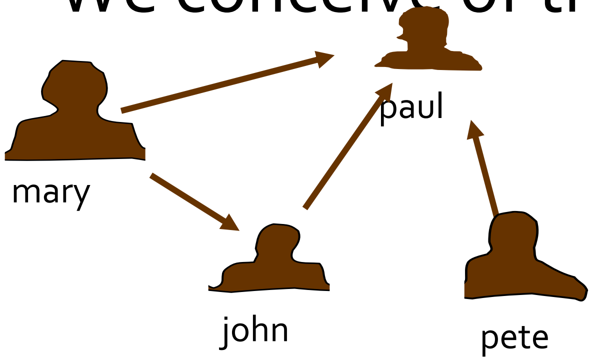
on $\longrightarrow$ $x_{ij} = 1$

off $\not\longrightarrow$ $x_{ij} = 0$

Some notation

We conceive of the **Graph** as a collection of

Tie variables: $\{X_{ij} : i,j \in V\}$

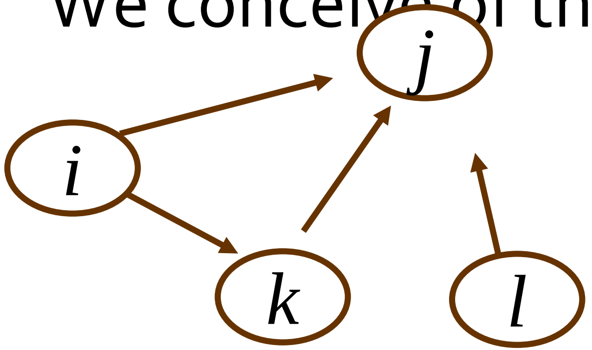


$$x_{ij} = \begin{cases} 1 & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$$

$$x = \begin{array}{c} i \\ j \\ k \\ l \end{array} \begin{pmatrix} - & x_{ij} & x_{ik} & x_{il} \\ x_{ji} & - & x_{jk} & x_{jl} \\ x_{ki} & x_{kj} & - & x_{kl} \\ x_{li} & x_{lj} & x_{lk} & - \end{pmatrix} = \begin{array}{c} i \\ j \\ k \\ l \end{array} \begin{pmatrix} - & 1 & 1 & 0 \\ 0 & - & 0 & 0 \\ 0 & 1 & - & 0 \\ 0 & 1 & 0 & - \end{pmatrix}$$

Some notation

We conceive of the **Graph** as a collection of



Tie variables: $\{X_{ij} : i,j \in V\}$

$$x_{ij} = \begin{cases} 1 & \text{if } i \to j \\ 0 & \text{otherwise} \end{cases}$$

$$x = \begin{array}{c} i \\ j \\ k \\ l \end{array} \begin{pmatrix} - & x_{ij} & x_{ik} & x_{il} \\ x_{ji} & - & x_{jk} & x_{jl} \\ x_{ki} & x_{kj} & - & x_{kl} \\ x_{li} & x_{lj} & x_{lk} & - \end{pmatrix} = \begin{array}{c} i \\ j \\ k \\ l \end{array} \begin{pmatrix} - & 1 & 1 & 0 \\ 0 & - & 0 & 0 \\ 0 & 1 & - & 0 \\ 0 & 1 & 0 & - \end{pmatrix}$$

Some notation

Regard each network tie as a *random variable* (often binary)

__Notation__

X_{ij} = 1 if there is a network tie from person i to person j
= 0 if there is no tie
for i, j members of some set of *actors N.*

A *directed network:* X_{ij} and X_{ji} are distinct.
A *non-directed network:* $X_{ij} = X_{ji}$

X … matrix of all variables
x … matrix of observed ties (the network)

Exponential random graph models

$$\Pr(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa} \exp\left\{ \sum_{Q} \lambda_Q z_Q(\mathbf{x}) \right\}$$

The summation is over all "configurations" *Q*

Local subgraphs that are hypothesized as the 'building blocks' of the network

$$z_Q(\mathbf{x}) = \prod_{X_{ij} \in Q} x_{ij} = 1 \text{ if } Q \text{ is observed in graph}$$

λ_Q parameter for the presence of Q

κ is a normalizing quantity.

Exponential random graph models

Parameter for Q
(how important is Q in the model)

$$\Pr(\mathbf{X}=\mathbf{x}) = \frac{1}{\kappa}\exp\left\{\sum_{Q}\lambda_{Q} z_{Q}(\mathbf{x})\right\}$$

Statistic for Q
(Is Q in the observed graph **x**?)

A probability distribution of graphs

Sum over all configurations Q
(dependence assumption determines which configurations)

A normalising constant
(probability of all graphs must add to 1)

What are we trying to do?

Estimate model parameters

- Positive parameter estimates indicate more configurations observed in the network than expected by chance.
- Negative parameter estimates indicate fewer configurations than expected by chance.

We want to know how the global network structure might have been built up out of small local substructures. The parameter estimates permit us to make inferences about this.

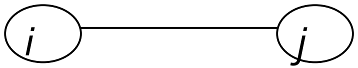
ERGMs and Dependence

- The general form of an ERGM
- Bernoulli models
- Markov models

Bernoulli dependence (independent edges)

Possible edges are independent of one another.

Configurations in this model relate to single possible edges (x_{ij}).



$$\Pr(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa} \exp\left(\sum_{ij} \lambda_{ij} x_{ij} \right)$$

Bernoulli graph model

$$\Pr(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa} \exp\left\{ \sum_{ij} \lambda_{ij} x_{ij} \right\}$$

Parameter for Q (edge ij)

Statistic for configuration
(Is the ij edge observed in **x**?)

A probability distribution of graphs

Sum over all configurations
(possible single edges ij)

A normalising constant
(probability of all graphs must add to 1)

Bernoulli graph model

But there is one parameter (λ_{ij}) for every possible edge – simply too many.

Homogeneity assumption: $\lambda_{ij} = \theta$ for all i,j

Assumes that the edge effect is the same across the entire network.

$$\Pr(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa}\exp\left(\sum \lambda_{ij} x_{ij}\right) = \frac{1}{\kappa}\exp\left(\theta \sum x_{ij}\right) = \frac{1}{\kappa}\exp(\theta L)$$

where *L* is the number of edges in the observed network

θ is an *edge* or *density* parameter

How to estimate the parameter θ

$$\Pr(\mathbf{X}=\mathbf{x}) = \frac{1}{\kappa}\exp(\theta L)$$

$$\hat{\theta} = \log\left(\frac{p}{1-p}\right)$$

where p is the density of the observed network

Network of 38 nodes, 44 edges (communication network)
$p = 0.06259$ so $\theta = -2.71$
This calculation only works because of independent edges

ERGMs and Dependence

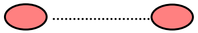
- The general form of an ERGM
- Bernoulli models
- Markov models

Markov random graphs
(Frank & Strauss, 1986, *JASA*)

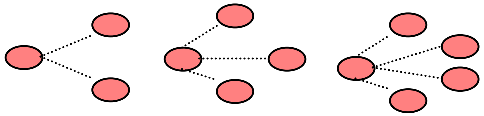
- Frank and Strauss drew on the work of Besag (1974) in spatial statistics
- They proposed a network dependence assumption (Markov dependence):
  - Two tie variables are conditionally independent unless they share a node.

Markov random graphs

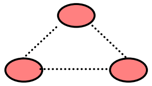
- Suppose that edges are conditionally dependent if and only if they share a node. (Frank & Strauss, 1986)
- Frank and Strauss showed that configurations in this model comprised edges, stars and triangles.



edge



2-star *3-star* *4-star* …

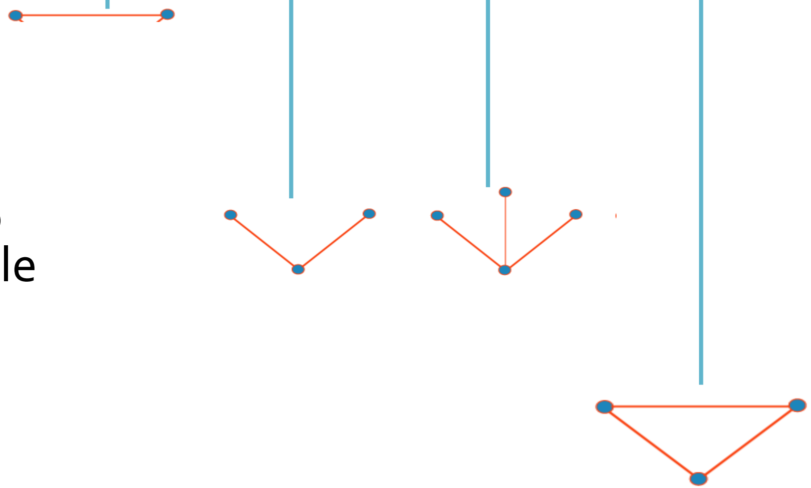


triangle

A Markov random graph model:

Undirected networks

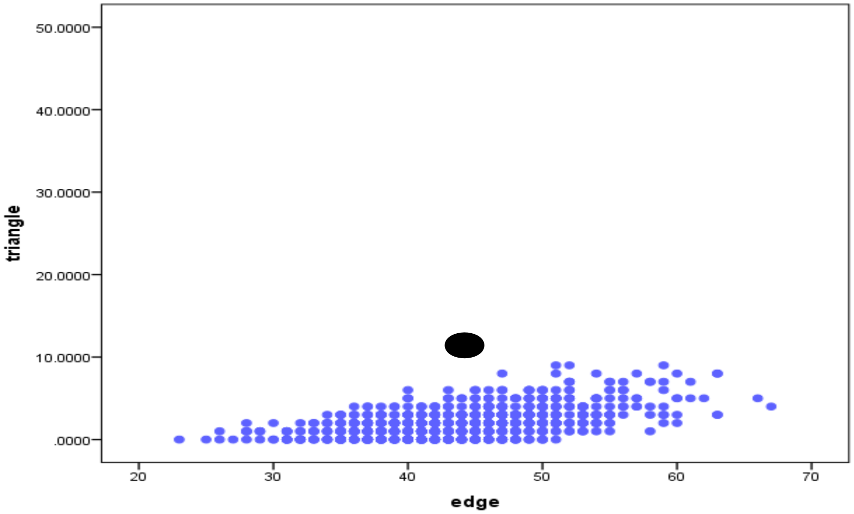
$$\Pr(\boldsymbol{X} = \boldsymbol{x}) = (1/\kappa)\exp\{\theta L + \sigma_2 S_2 + \sigma_3 S_3 + \tau T\}$$



- *Edge parameter (θ)*
  L ... number of edges
- *Star parameters (σ_k)*
  Propensities for individuals to have connections with multiple network partners
- *Triangle parameter (τ)*
  represents network closure

If θ is the only nonzero parameter, this is a Bernoulli random graph model.

Simulated results from Bernoulli graph model

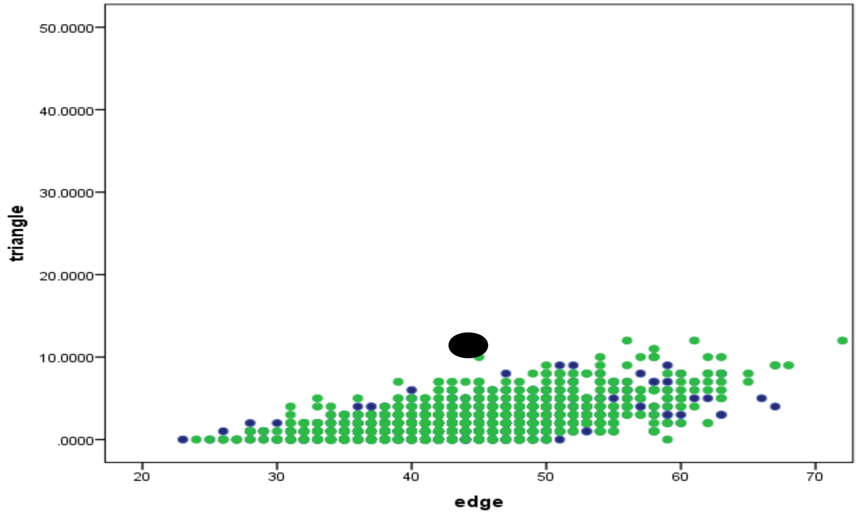


Statistics from simulated samples

Blue = Bernoulli

Observed statistics

Simulated results from Markov graph model Edge, 2star parameters



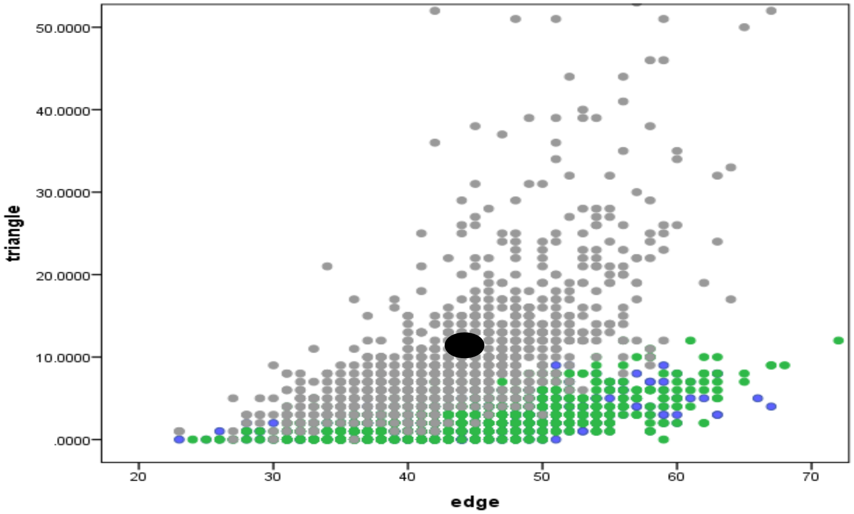
Statistics from simulated samples

Blue = Bernoulli

Green = L,2star

Observed statistics

Simulated results from Markov graph model Edge, 2star, 3star, triangle parameters



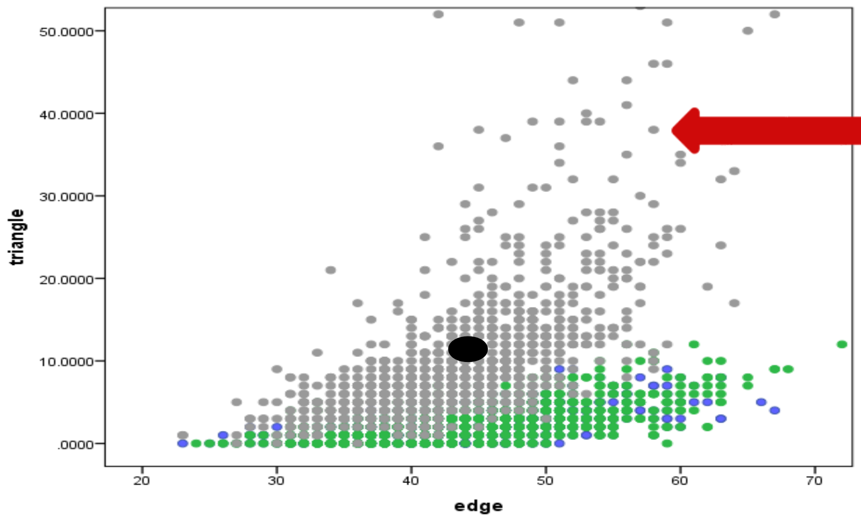
<u>Statistics from simulated samples</u>

Blue = Bernoulli

Green = L,2star

Observed statistics

Simulated results from Markov graph model
Edge, 2star, 3star, triangle parameters



This 'leakage' shows a common problem with Markov models – they are not always stable; and may be *degenerate.*

Observed statistics

Social Selection Models

Social selection

- Actors select network partners based on actor attributes.
- Possible mechanisms
  - Homophily: actors of similar attributes tend to form ties (McPherson et al, 2001).
  - homophily in itself cannot explain the emergence of hierarchy in relations (so difference may also be important)
  - more generalized selection: individuals select social positions for themselves.

Social selection

- Also actor *main effects*
  - Nondirected - activity: Actors with certain attributes might be more active (involved in more ties)

Terminology and notation: Network variables

For node set N

Let $X_{ij} = 1$ if there is a tie from node i to node j

$= 0$ if there is no tie from i to j.

A binary network

Let $X_{ii} = 0$ for all nodes i.

Define $\mathbf{X}$ as the matrix of variables $[X_{ij}]$

Define $\mathbf{x}$ as the *adjacency matrix*, the matrix of observed network ties

Terminology and notation: Attribute variables

For node set N

Let Y_i = 1 if node i has attribute Y

= 0 otherwise.

A binary attribute (e.g. gender)

Alternatively Y_i can represent categories (e.g. political party)

Or can be continuous (e.g. age)

Define $\mathbf{Y}$ as the vector of variables $[Y_i]$

Define $\mathbf{y}$ as the *attribute vector*, the vector of observed attributes.

Three types of attribute variables

1. **Binary**
   - eg male/female

2. **Categorical**
   - eg Workteams within a company

3. **Continuous**
   - eg Age, attitudes

Structural effects in the models

$$\Pr(\mathbf{X}=\mathbf{x}) = \frac{1}{\kappa}\exp\left\{\sum_Q \lambda_Q z_Q(\mathbf{x})\right\}$$

The summation is over all configuration types Q

$z_Q(\mathbf{x})$ is the network statistic for Q

λ_Q parameter for Q

κ is a normalizing quantity.

How to incorporate **y** into the model?

Social selection models

We want a model for: $\Pr(\mathbf{X}=\mathbf{x} \mid \mathbf{Y}=\mathbf{y})$

Probability of observing graph **x** GIVEN observed attribute vector **y**

$$\Pr(\mathbf{X}=\mathbf{x} \mid \mathbf{Y}=\mathbf{y}) = \frac{1}{\kappa}\exp\left\{ \underbrace{\sum_Q \lambda_Q z_Q(\mathbf{x})}_{\text{structural part – just as before}} + \underbrace{\sum_R \lambda_R z_R(\mathbf{x},\mathbf{y})}_{\text{selection part – interaction of ties and attributes}} \right\}$$

Second summation is over all selection configurations *R*

Possible binary attribute configurations (non-directed graphs)

Activity

Positive parameter indicates node with attribute has many ties

Statistic: For each tie, count the number of attributed nodes

$$X_{ij}(Y_i + Y_j)$$

Then sum across all ties:

Interaction

Positive parameter indicates nodes with attribute tend to share ties

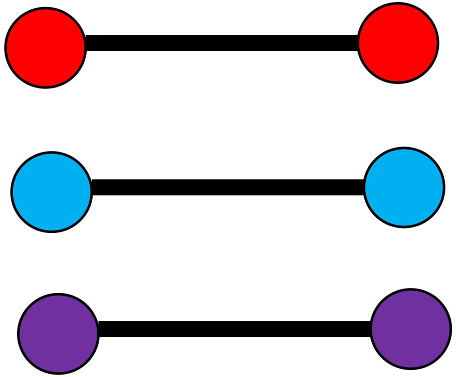
Statistic: For each tie, count those where both nodes are attributed

$$X_{ij}Y_i\,Y_j$$

Then sum across all ties:

Possible categorical attribute configurations (non-directed graphs)

Match category



Positive parameter indicates ties within categories are more likely

Possible continuous attribute configurations (non-directed graphs)

Difference

For each pair of tied nodes the statistic is the absolute difference between the attribute values – (then summed over all pairs)

Negative parameter indicates that a smaller absolute difference is associated with the presence of a tie:

$$X_{ij} \mid Y_i - Y_j \mid$$

Ties are more likely when nodes have similar attribute values - HOMOPHILY

Possible continuous attribute configurations (non-directed graphs)

Activity (sum)

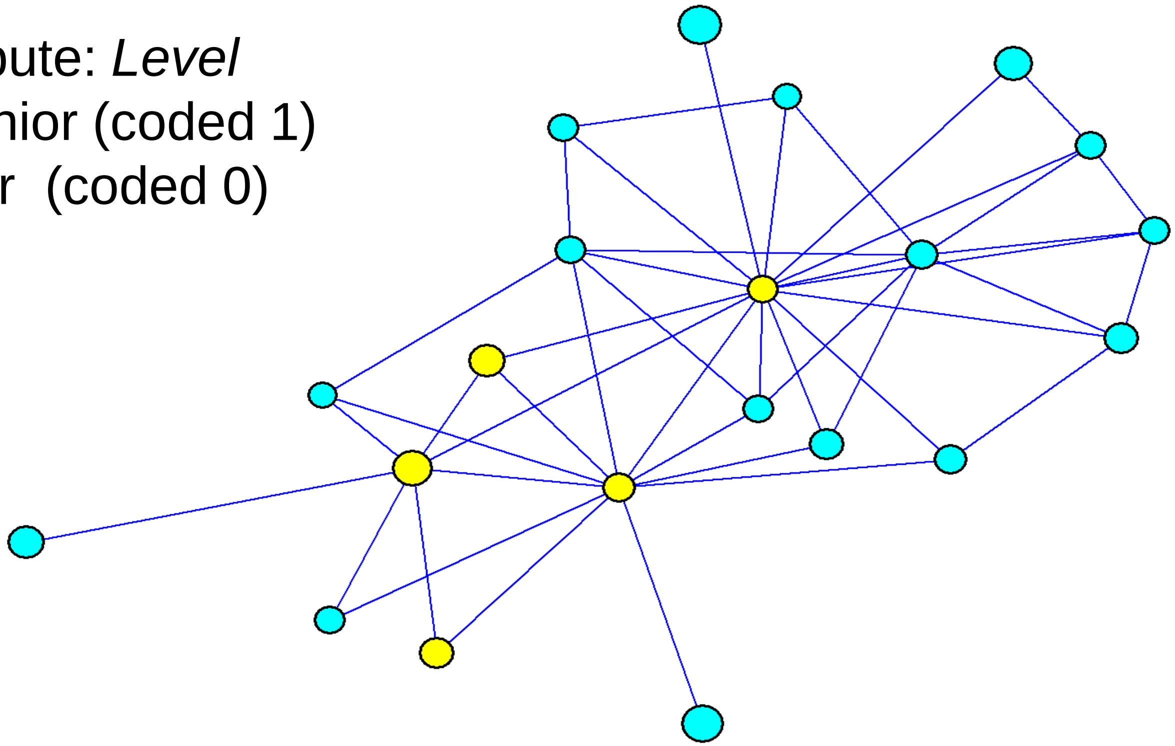
For each pair of tied nodes the statistic is the sum of the attribute values – (then summed over all pairs)

$$X_{ij}(Y_i + Y_j)$$

Positive parameter indicates that pairs of nodes with large (average) attribute values tend to be tied.

Example: Krackhardt hi-tech managers: Mutual advice network

Binary attribute: *Level*
yellow = senior (coded 1)
blue = junior (coded 0)



Example: Krackhardt hi-tech managers: Mutual advice network

Bernoulli model for *level*

Parameter	Estimate	Standard error	Convergence
Edge	-1.96*	0.29	0.07
Interaction (Homophily)	1.40	1.00	0.002
Activity	0.97*	0.40	0.02

Example: Krackhardt hi-tech managers: Mutual advice network

Edge-based model for *level*
Parameter estimates from pnet:

Parameter	Estimate	Standard error	Convergence
Edge:	-1.96*	0.29	0.07
Rb (Homophily)	1.40	1.00	0.002
R (Activity)	0.97*	0.40	0.02

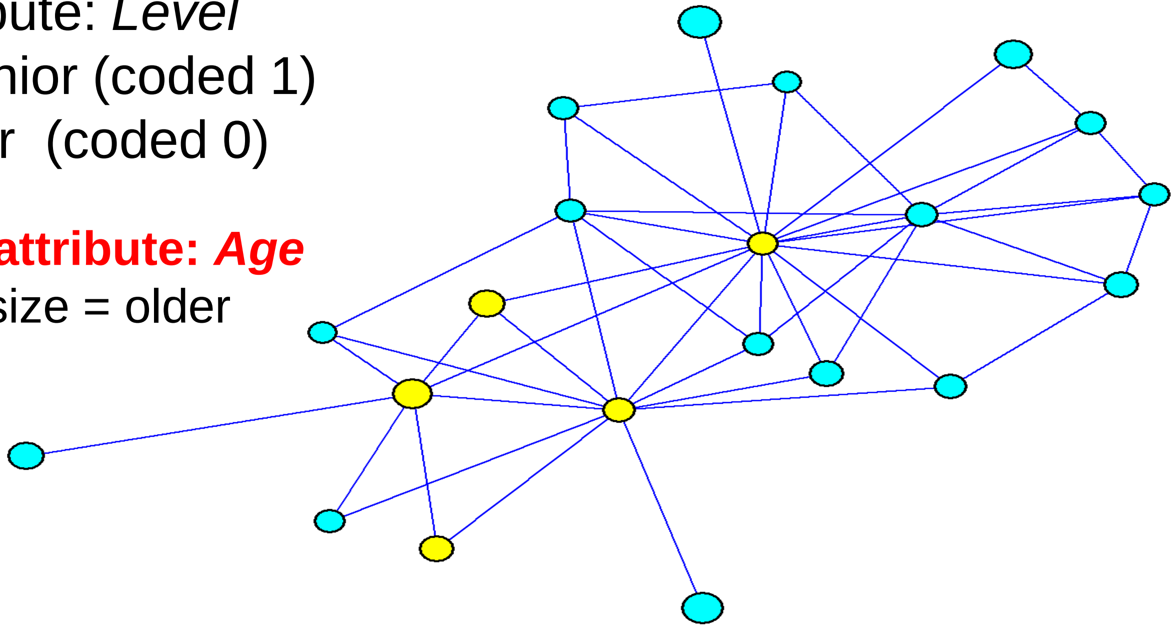
Conditional log-odds of mutual advice between juniors = – 1.96
between junior/senior = -1.96 + 0.97 = – 0.99
between seniors = -1.96 + 0.97 + 0.97 + 1.40 =1.38

Advice links between juniors are the most unlikely; between junior and senior are more common; between seniors are the most common

Example: Krackhardt hi-tech managers: Mutual advice network

Binary attribute: *Level*
yellow = senior (coded 1)
blue = junior (coded 0)

Continuous attribute: *Age*
Larger node size = older



REMEMBER: Continuous attributes

Difference

For each pair of tied nodes the statistic is the absolute difference between the attribute values – (then summed over all pairs)

$$X_{ij} \, | Y_i - Y_j |$$

REMEMBER: Continuous attributes

Activity (sum)

For each pair of tied nodes the statistic is the sum of the attribute values – (then summed over all pairs)

$$X_{ij}(Y_i + Y_j)$$

Example: Krackhardt hi-tech managers: Mutual advice network

Bernoulli model for *age*

Parameter estimates from pnet:

Parameter	Estimate	Standard error	Convergence
Edge:	0.85	1.31	-0.01
sum *age*	-0.022	0.018	-0.003
difference *age*	-0.045	0.029	0.01

Effects of age are not significant, but the negative effect for difference suggests there may be some age homophily.

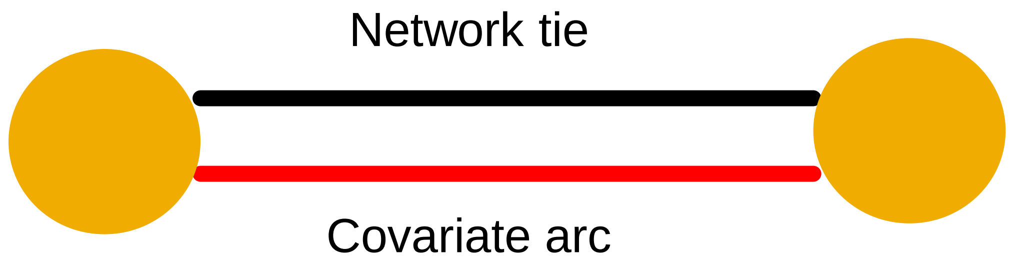
Dyadic covariate
(or dyadic attribute)

- Some other relationship among nodes that could influence the network structure:

  Examples:
  - Formal organisation structure
  - Geography
  - Another network

Configurations in MPnet

- There is one configuration, *covariate-arc*, in MPnet – a parameter for when the network tie and the covariate co-occur.



For instance, do people who have advice ties also have communication ties?