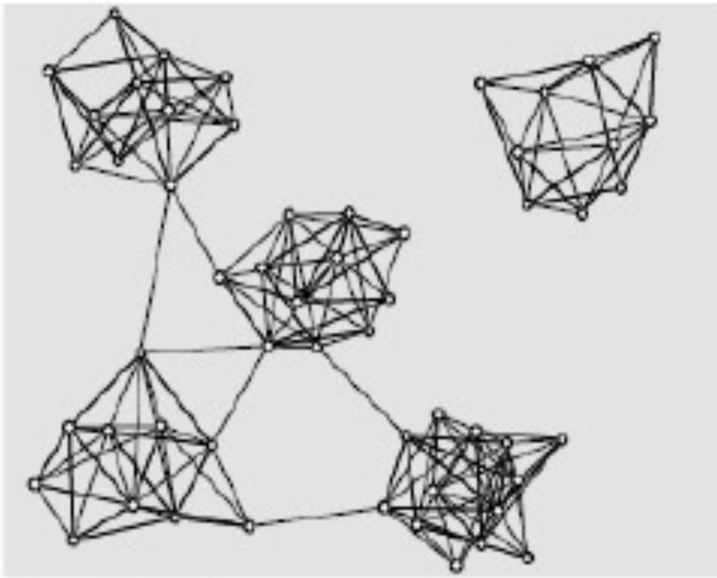
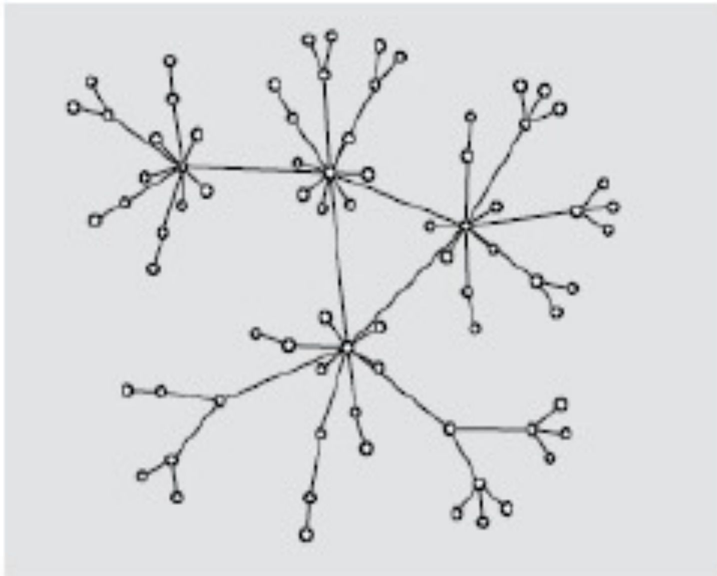
Statistical Analysis of Networks

Lecture ERGM

Models accounting for specific tie-formation mechanisms



"Small World"



"Scale-free"

Small World configuration (SW) (Watts and Strogatz, 1998)

- high node connectivity with low average distance among regions of the network: $\ell(G) \leq \ell(G_{Rand})$, G_{Rand} = random networks of equal size
- high tendency towards clustering, $\Gamma(G) \gg \Gamma(G_{Rand})$

[Random graph (Erdos & Renyi, 1959)]

Scale free networks model (or Barabasi-Albert model, 1999) based on:

- Preferential Attachment (PA) tie formation mechanism ⇒ the probability that a new node will be connected to i is proportional to the degree d_i
- degree distribution $P(x)$ with fat tails approximating a power law with parameter α $(2 < \alpha < 3)$
  1. Standard (de Solla Price, 1963): $P(x) = Cx^{-\alpha}$
  2. Truncated (Clauset *et al.*, 2009): $P(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha}$

COMMON FEATURES OF INTEREST

Beyond nodal and dyadic attributes, **many real networks** exhibit the following features:

- Reciprocity of ties
- Degree heterogeneity among actors
  Activity, Popularity
- Homophily by actor attributes
  Higher propensity to form ties between actors with similar attributes
- Transitivity of relationships
  Friends of friends have a higher propensity to be friends
- Equivalence of nodes
  Some nodes may have identical/similar patterns of relationships

NETWORK MODELS

*"A good [statistical network graph] model needs to be both estimable **from data** and a **reasonable representation of that data**, to be theoretically plausible about the **type of effects that might have produced the network**, and to be amenable to examining which **competing effects** might be the best explanation of the data."* (Robins and Morris, 2007)

Small-word and Preferential attachment models:

→ not really intended to meet such criteria

Statistical modeling: evaluation and fitting of network models

- Testing: evaluation of competing theories of network formation
- Estimation: evaluation of parameters in a presumed network model
- Description: summaries of main network patterns
- Prediction: prediction of missing or future network relations

RANDOM GRAPH AND RANDOM MATRIX

- Let $G = (V, E)$ be a graph. If E (and perhaps V) is a random set, then G is a random graph
  - Can consider G to be a random variable on some set $\mathbf{G}$ of possible graphs ("multinomial" representation)
- Let *Y* be the adjancency matrix of random graph G, then *Y* is a random matrix

$$Y \equiv \left[Y_{ij}\right]_{nxn} \qquad Y_{ij} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \\ 0 & \text{otherwise} \end{cases}$$

valued and signed ties can be considered

– a $N = n(n-1)$ array of binary random variables

– Y represents a random network with nodes the actors and edges the relationship

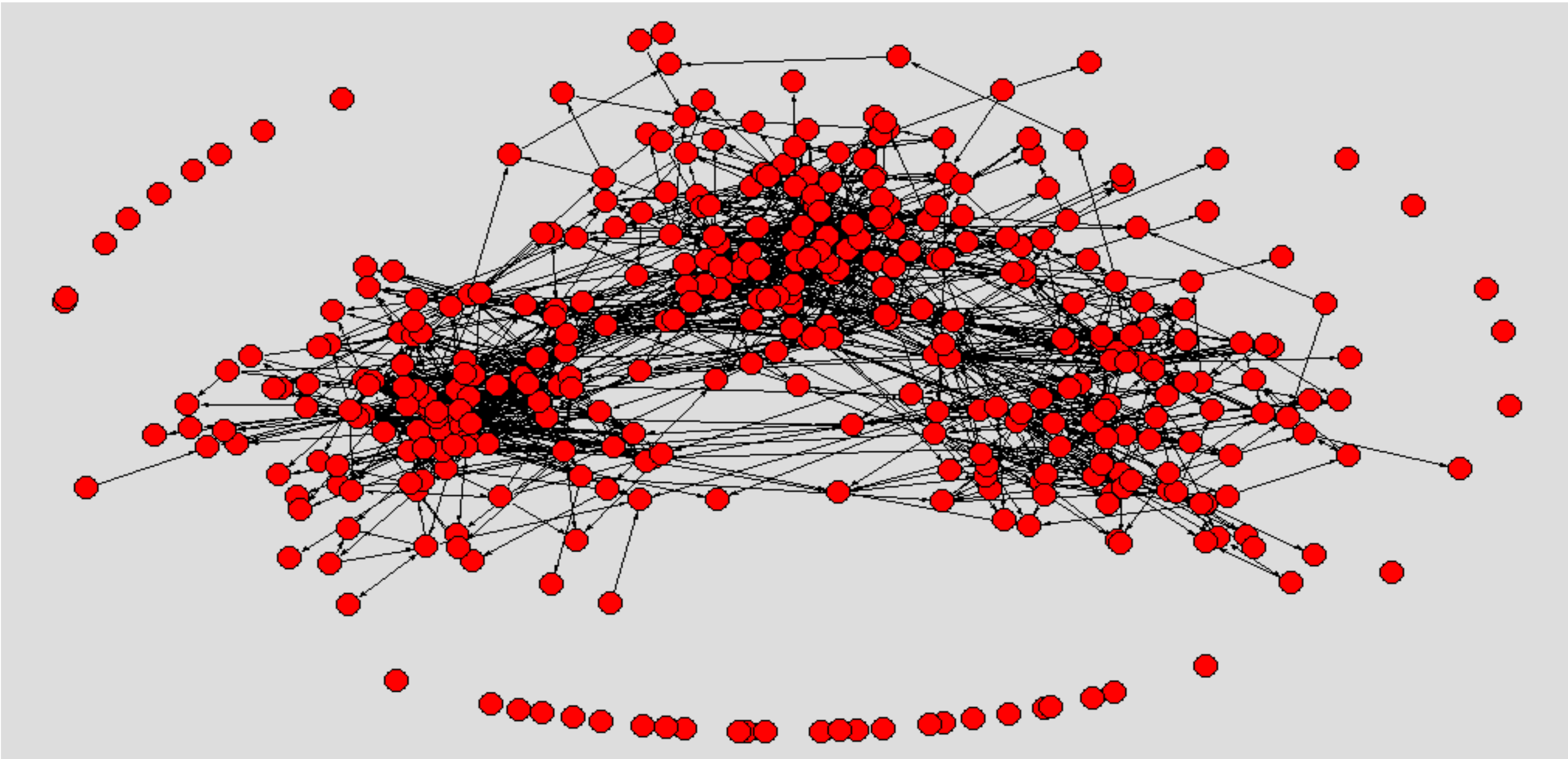
- The basic problem of stochastic modeling is to specify a distribution for Y i.e., $P(Y = y)$

STATISTICAL MODELS FOR NETWORK DATA

- Statistical model for the ties in a network

But

- The overall structure – the network – is evident



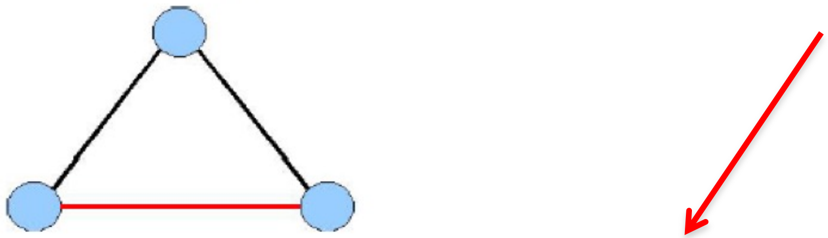
- What kind of **structural** elements can be included in a model for the tie variable ?

(MAIN) NETWORK DEPENDENCIES

1. *Reciprocation*: dependencies between Y_{ij} and Y_{ji}
2. *Homophily: tendency of similar actor to relate to each other (assortative mixing by attribuite)*
3. *Transitivity:* $Y_{ij} = Y_{jh} = 1$ will lead to increase $P(Y_{ih} = 1)$ (triad/triangle closure: *"friends of my friends are my friends"*)
4. *Degree differentials:* some actors are higly connected and others have only few connections (*sociality*)

Problem: Type 2, 3, 4, and partially 1: all can lead to similar macro signatures (network configurations, e.g. "clustering")

So, for three actors of the same type:

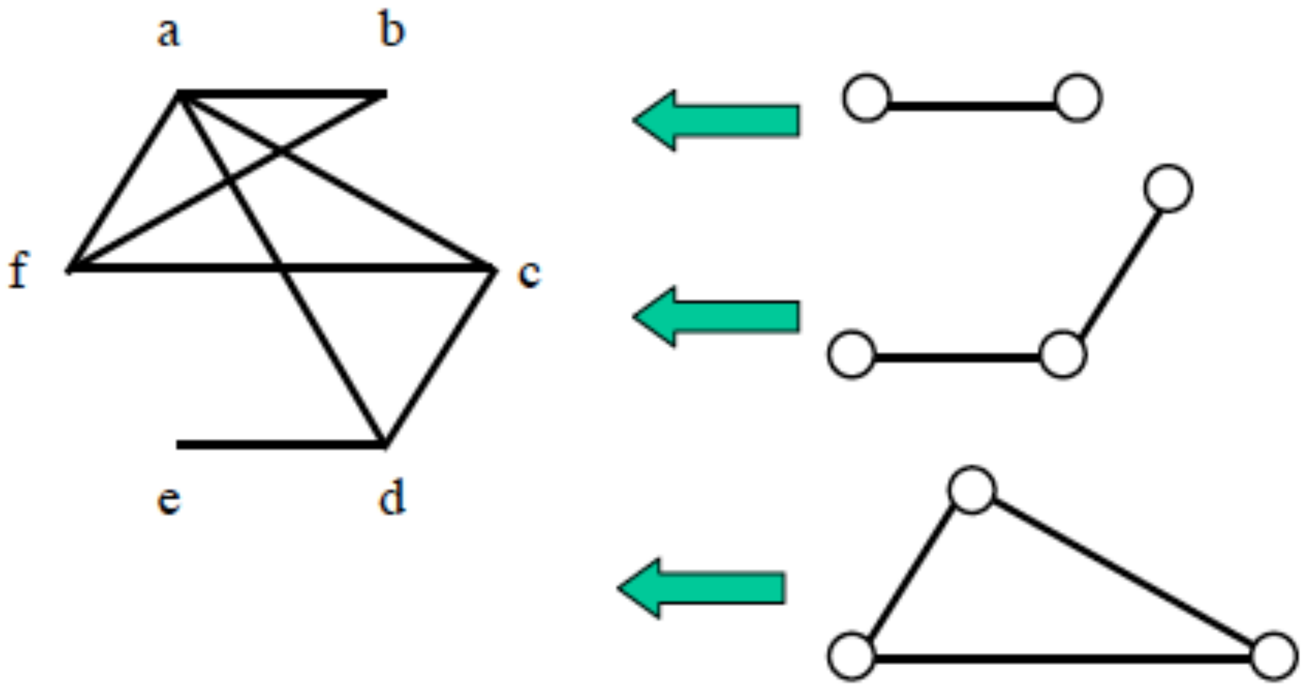


Cycle-closing tie may form because of *transitivity* but also *homophily*

Aim: to be able to fit these terms simultaneously and identify the effects of each mechanism on the overall outcome.

MODEL CONSTRUCTION - GENERAL IDEA

The probability of observing a specific graph (*Y_{ob}*) is **dependent** on **local** characteristis of the graph (*$f(Y_{ob})$*)



EXPONENTIAL RANDOM GRAPH (ERGM)

Probability distribution of the set of possible graphs

$$P(Y = y) = \frac{exp\left\{\sum_{k=1}^{K} \theta_k g_k(y)\right\}}{c(\theta)}$$

network statistics
(network feutures)

where $\theta_{1,2\ldots k}$ are parameters $g_{1,2\ldots k}(y)$ are statistics, and $c(\theta)$ is a normalizing constant:

$$c(\theta) = \sum_{y \in \mathcal{Y}} exp\left\{\sum_{k=1}^{K} \theta_k g_k(y)\right\}$$

In other words,

$$P(Y = y) \propto \theta_1 g_1(y) + \theta_2 g_2(y) + \theta_3 g_3(y) + \ldots + \theta_k g_k(y)$$

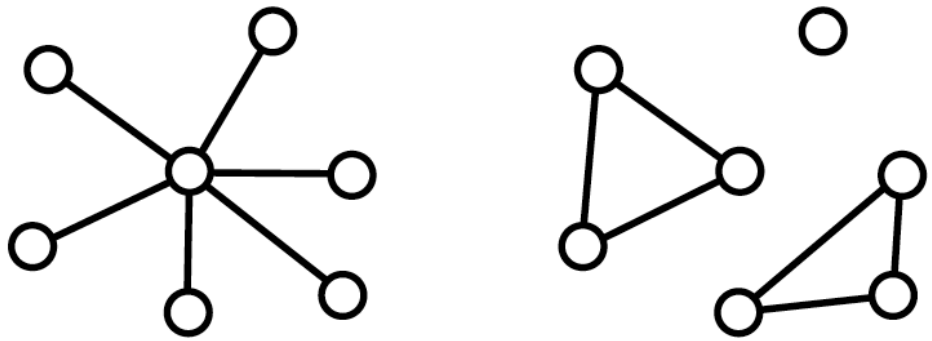
Intuition: the ERGM places more/less weight on graphs with certain features, as determined by θ, g

network statistics = number of local congurations of a specific type

ERGM PROBABILITY

$$P(Y = y) = \frac{exp\left\{\sum_{k=1}^{K} \theta_k g_k(y)\right\}}{c(\theta)}$$

- The probability of a graph y is an exponential family model.
- Parameter vector θ (weights), statistics vector (counts of ties, reciprocal ties, transitive triplets, degree distribution, homophilic ties, . . .)
- The probability of a graph thus depends on the structures that it includes, given the parameters.
- The following two graphs have a different probability depending on the terms and parameters of the model:



ERGM specifies the probability of the entire network (the left hand side), as a function of terms that represent network features we hypothesize may occur more or less likely than expected by chance (the right hand side)

MODEL CONSTRUCTION - GENERAL FRAMEWORK

1. *Step 1: each network tie is a random variable.*
2. *Step 2: a dependence hypothesis is proposed, defining contingencies among the tie variables.*
3. *Step 3: the dependence hypothesis implies a specific form to the model.*
4. *Step 4: simplification of parameters through homogeneity or other constraints.*
5. *Step5: estimate and interpret parameters*

E.g. friendship: are there more reciprocated ties than would be expected by chance ?

Model will include a density parameter (randomness occurance of ties) and a reciprocation parameter

MLE will be the parameter value such that the most probable degree of reciprocation is that which occurs in the observed network

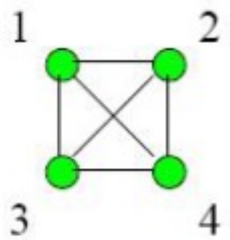
ERGMs:
superficially resembling linear regression or GLMs

TIES AS RANDOM VARIABLES

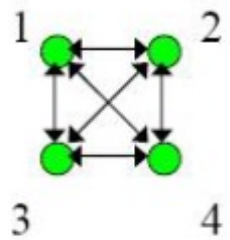
We model tie variables: $\mathbf{Y} = [Y_{ij}]$
$Y_{ij} = 1$ if i has a tie to j, 0 otherwise
The realization of $\mathbf{Y}$ is denoted by $y = [y_{ij}]$

Random graph and random directed graphs on a node set $N = \{1,2,3,4\}$

Graph



Directed graph



Tie variables:

$Y_{12}, Y_{13}, Y_{14}, Y_{23}, Y_{24}, Y_{34}$

Tie variables:

$Y_{12}, Y_{13}, Y_{14}, Y_{23}, Y_{24}, Y_{34},$
$Y_{21}, Y_{31}, Y_{41}, Y_{32}, Y_{42}, Y_{43}$

Model: statistical analysis

1. Estimate parameters of the process
   - Joint estimation of multiple, possibly correlated, effects
2. Inference
   - Is a certain parameter significantly different from zero?
   - Uncertainty in parameter estimates
3. Goodness of fit
   - Traditional diagnostics
     - Model fit (BIC, AIC)
     - Estimation diagnostics (MCMC performance)
   - Network-specific goodness-of-fit
     - Network statistics already in the model as covariates
     - Network properties not in the model

ERGM: SOME NOTATIONS

Probability distribution of the set of possible graphs

$$P(Y = y) = \frac{exp\left\{\sum_{k=1}^{K} \theta_k g_k(y)\right\}}{c(\theta)}$$

probability
of a single graph

Since each network tie is a random variable, the goal is to re-express the probability of the graph in terms of the probabilities of an individual tie:

- this gives a "local" view of the model
- and some insight into what the coefficients mean

In order to re-express the probability of the graph in terms of the probabilities of a tie, we need to introduce some notation:

- $Y_{ij}^{+} = \{Y \text{ with } Y_{ij} = 1\}$ the graph w/ the (i, j)th dyad set to 1
- $Y_{ij}^{-} = \{Y \text{ with } Y_{ij} = 0\}$ the graph w/ the (i, j)th dyad set to 0
- $Y_{ij}^{c} = \{Y_{kl} \text{ with } (k, l) \neq (i, j)\}$ all dyads except (i,j)

ERGM: THE CONDITIONAL PROBABILITY OF A LINK

A simple logical re-expression of $P(Y = y) = \frac{exp\left\{\sum_{k=1}^{K} \theta_k g_k(y)\right\}}{c(\theta)}$

$$\Pr(Y_{ij} = 1 | Y_{ij}^c) = \frac{\Pr(Y = y_{ij}^+)}{\Pr(Y = y_{ij}^+) + \Pr(Y = y_{ij}^-)}$$

$$= \frac{\exp\{\theta^T g(y_{ij}^+)\}}{\exp\{\theta^T g(y_{ij}^+)\} + \exp\{\theta^T g(y_{ij}^-)\}}$$

Note:
the costant term $c(\theta)$ has canceled out, but . . . an even simpler expression, in terms of the odds, can be used

ERGM: THE CONDITIONAL LOG-ODDS PROBABILITY OF A LINK

Reminder: $\operatorname{logit}(p) = \log\left(\frac{p}{(1-p)}\right)$

- Given, $\Pr(Y_{ij} = 1 | Y_{ij}^c) = \frac{\exp\{\theta^T g(y_{ij}^+)\}}{\exp\{\theta^T g(y_{ij}^+)\} + \exp\{\theta^T g(y_{ij}^-)\}}$
- Then

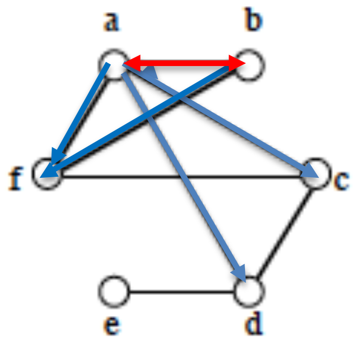
$$\log\left\{\frac{\Pr(Y_{ij} = 1 | Y_{ij}^c)}{\Pr(Y_{ij} = 0 | Y_{ij}^c)}\right\} = \theta^T[g(y_{ij}^+) - g(y_{ij}^-)]$$
$$= \theta^T \delta(y_{ij})$$

Note: $\delta(y_{ij})$ is known as the change statistic

- Useful implication: each unit change in g_k for (i, j) tie present (versus absent) increases the conditional log-odds of (i,j) by θ_k
- θ is the impact of the covariate on the log-odds of a tie

[Prob= odds/(1+odds)]

An undirected network and graph:



	a	b	c	d	e	f
a	0	1	1	1	0	1
b	1	0	0	0	0	1
c	1	0	0	1	0	1
d	1	0	1	0	1	0
e	0	0	0	1	0	0
f	1	1	1	0	0	0

What if ab = 0 instead of 1?

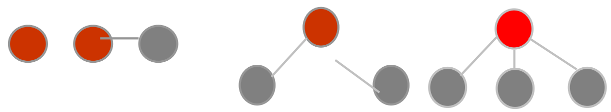
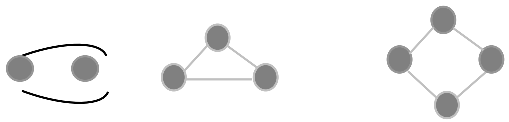
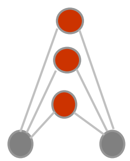
2-stars {b,a,c}, {b,a,d} , {b,a,f} and {a,b,f} would disappear, and triangle {a,b,f} also!

= Change scores

Dyad	value	density	#2-star	#triangle
ab	1	1	3+1	1
ac	1	1	3+2	2
ad	1	1	3+2	1
ae	0	1	4+1	0
af	1	1	3+2	2
bc	0	1	2+3	2
bd	0	1	2+3	1
be	0	1	2+1	0
bf	1	1	1+2	1
cd	1	1	2+2	1
ce	0	1	3+1	1
cf	1	1	2+2	1
de	1	1	2+0	0
df	0	1	3+3	2
ef	0	1	1+3	0

TYPES OF COVARIATES - G(Y) TERMS IN THE MODEL

What creates heterogeneity in the probability of a tie being formed?

attributes of nodes	Heterogeneity by group – Average activity – Mixing by group Individual heterogeneity	Dyad **Independent** Terms
attributes of links	eterogeneity in – Duration – Types (sex, drug…)	
configurations	Degree distributions (or stars)  Cycle distributions (2, 3, 4, etc.)  Shared partner distributions 	Dyad **(In)dependent** Terms

ERG TYPES OF STATISTICAL MODELS (BASED ON DIFFERENT DEPENDENCE HYPOTHESES) /1

Exponential Random Graph Model – ERGM class (also known as p model, especially in SNA literature)*

dependency generated by specific structural configurations, including:

- ***Erdős-Rény** (Bernoulli) model (edge independence)*
- ***p₁** model (dyad independence with attributes and reciprocity, Holland and Leinhardt, 1981)*

$$P(Y = y) = \frac{\exp\{\rho \sum_{i<j} y_{ij} y_{ji} + \theta \sum_{i,j} y_{ij} + \sum_i \alpha_i \sum_j y_{ij} + \sum_j \beta_j \sum_i y_{ij}\}}{c(\rho, \alpha, \beta, \theta)}$$

where

- θ controls the expected number of edges
- ρ represent the expected tendency toward *reciprocation*
- α_i *productivity* of node i; β_j *attractiveness* of node j

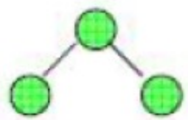
ERG TYPES OF STATISTICAL MODELS (BASED ON DIFFERENT DEPENDENCE HYPOTHESES) /2

Exponential Random Graph Model – ERGM class (also known as p model, especially in SNA literature)*

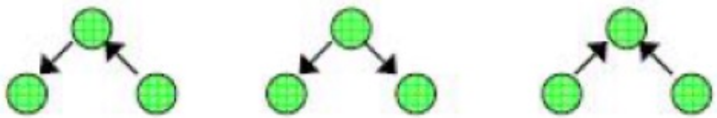
dependency generated by specific structural configurations, including:

– ***Markov Random Graph** model (Markov dependence: edges share a vertex, Frank and Strauss,1986)*

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



they involve a shared actor (undirected)



(directed)

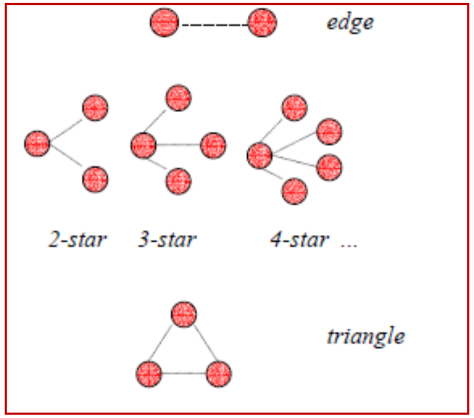
$$P(\mathbf{Y}=\mathbf{y};\boldsymbol{\theta}) = \frac{1}{\kappa(\boldsymbol{\theta})} \exp\left\{ \sum_{k=1}^{N_v-1} \theta_k S_k(\mathbf{y}) + \theta_\tau T(\mathbf{y}) \right\}$$ (with triangles)

MARKOV RANDOM GRAPH MODEL (BASIC CONFIGURATIONS FOR UN/DIRECTED GRAPHS)

$$P(\mathbf{Y}=\mathbf{y};\boldsymbol{\theta}) = \frac{1}{\kappa(\boldsymbol{\theta})}\exp\left\{\sum_{k=1}^{N_v-1}\theta_k S_k(\mathbf{y}) + \theta_\tau T(\mathbf{y})\right\}$$

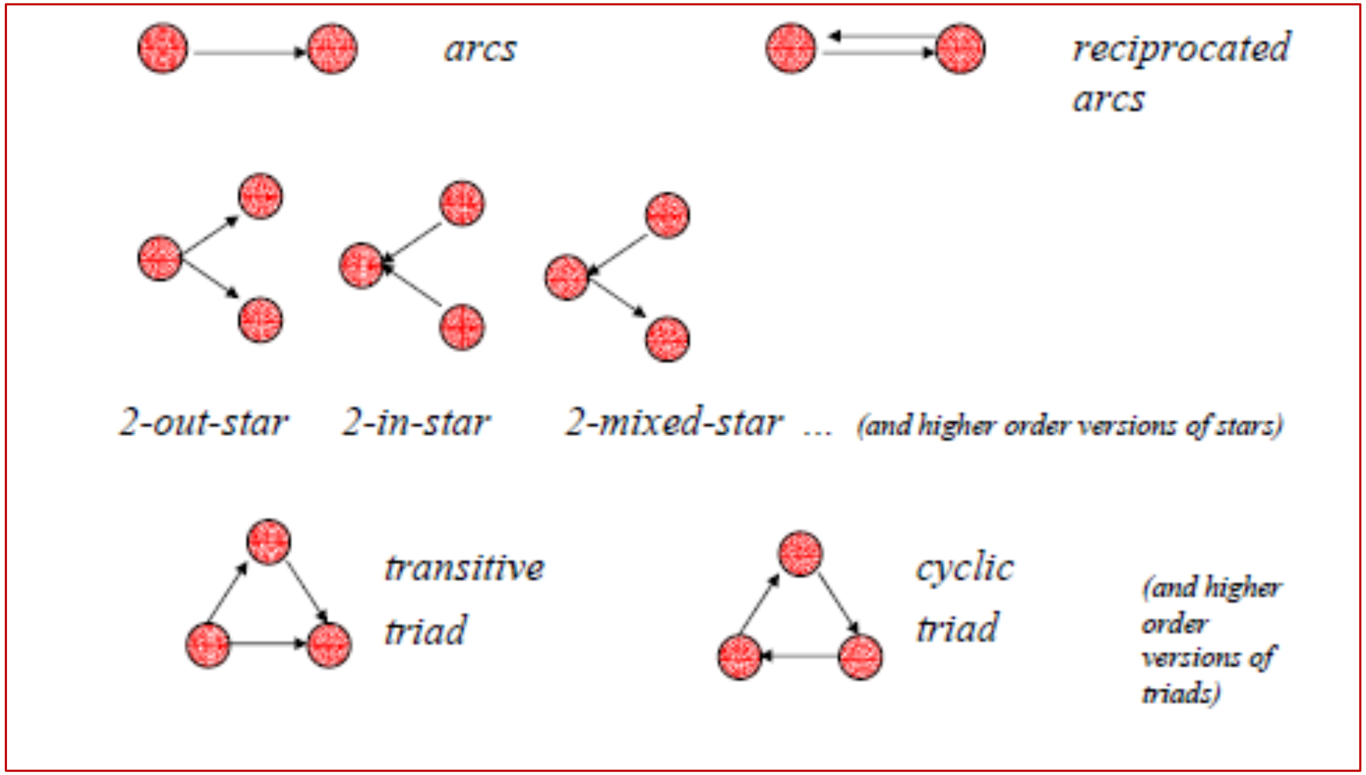
$$(1/\kappa)\exp\{\theta L + \sigma_2 S_2 + \sigma_3 S_3 + \tau T\}$$

Typical spefication



(undirected)

(directed)



ERG TYPES OF STATISTICAL MODELS (BASED ON DIFFERENT DEPENDENCE HYPOTHESES) /3

Exponential Random Graph Model – ERGM class (also known as p model, especially in SNA literature)*

dependency generated by specific structural configurations, including:

- ***Social circuit** model (dependence might arise from the presence of other edges - partial conditional dependence)*

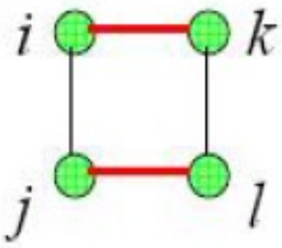
Tie variables $\mathbf{Y}_{ij}$ and $\mathbf{Y}_{kl}$ are conditionally independent for distinct i, j, k, l unless:

$y_{ik} = 1$ and $y_{jl} = 1$

or

$y_{il} = 1$ and $y_{jk} = 1$

(eg **red** ties are observed)



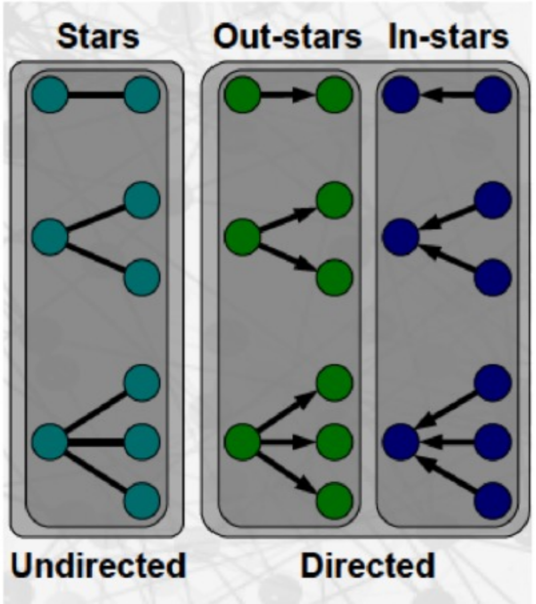
and so they complete a social circuit

INTERPRETATION OF STAR AND TRIAD EFFECTS

- $k-$stars: number of subgraphs with one/two/three... endpoints (in/out in directed ties) with respects to node i

  interpretation:

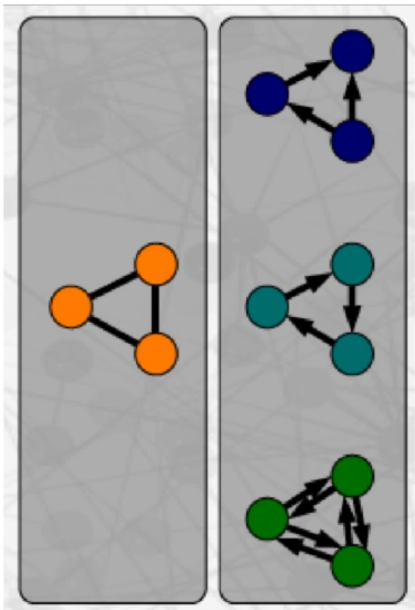
- tendency of edges " to stick together" on endpoints ("edge clustering")



- Most basic terms for endogeneous clustering
- each term counts triads of a given type (triangles, cycles, ...)

  interpretation:

- tendency towards transitive closure

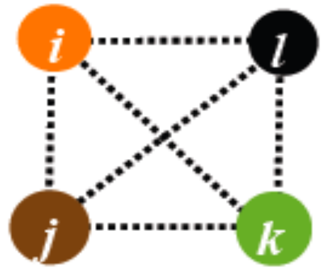


ERGM PARAMETRIZATION (DIRECTED NETWORK)

- **ERG form is just a way of writing models – to use it, we must choose a set of terms ($\mathbf{t}$)**
- **Some basics (*dyad independence* terms):**
  - Edge term: $\sum_i \sum_j \mathbf{y}_{ij}$
    - Captures overall tendency of ties to form/not (density effect)
  - Row-sum term: $\sum_i \mathbf{y}_{ij}$
    - Captures net tendency to send ties (sender/expansiveness effect)
  - Col-sum term: $\sum_j \mathbf{y}_{ij}$
    - Captures net tendency to receive ties (receiver/popularity effect)
  - Mutuality term: $\sum_i \sum_{j>i} \mathbf{y}_{ij} \mathbf{y}_{ji}$
    - Captures tendency of ties to reciprocate one another (reciprocity effect)
  - Linear covariates: $\sum_i \sum_j \mathbf{y}_{ij} X_{ij}$
    - Captures tendency of $\mathbf{y}_{ij}$ edges to covary with X_{ij} (covariate effect)

ERGM PARAMETRIZATION AND NETWORK STATISTICS /1

If we believe that the frequency of interaction/**density** is an important aspect of the network

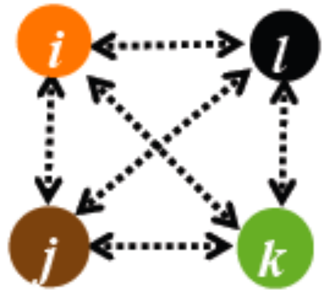


We should include

Counts of  the number of ties in our model

ERGM PARAMETRIZATION AND NETWORK STATISTICS /2

If we believe that the **reciprocity** is an important aspect of the (directed) network

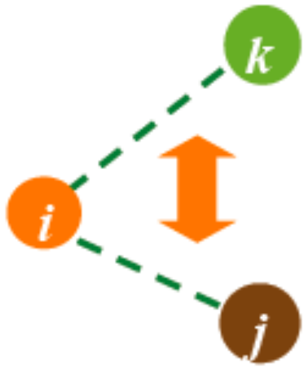


We should include

Counts of  the number of mutual ties in our model

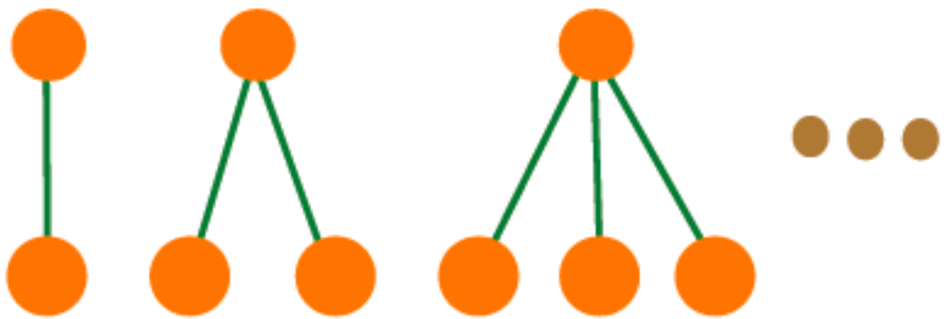
ERGM PARAMETRIZATION AND NETWORK STATISTICS /3

If we believe that an important aspect of the network is that

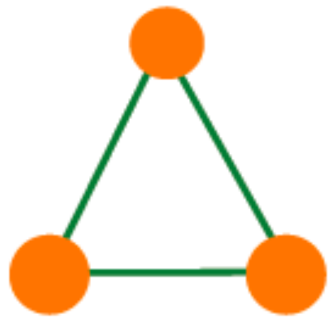


two edge indicators $\{i,j\}$ and $\{i',k\}$ are conditionally **dependent** if $\{i,j\} \cap \{i',k\} \neq \emptyset$

We should include counts of



degree distribution; preferential attachment, etc

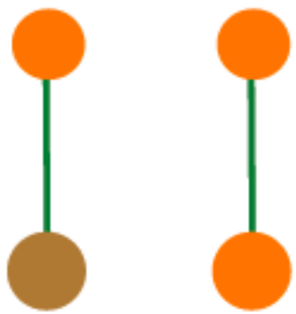


friends meet through friends; clustering; etc

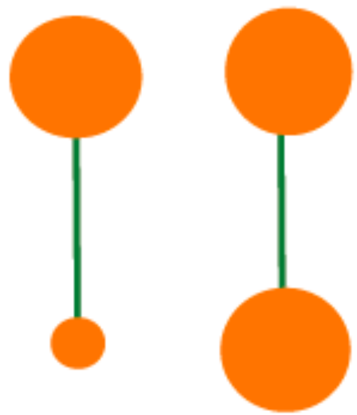
ERGM PARAMETRIZATION AND NETWORK STATISTICS /4

If we believe that the **attributes** of the actors are important (selection effects, homophily, etc)

We should include counts of



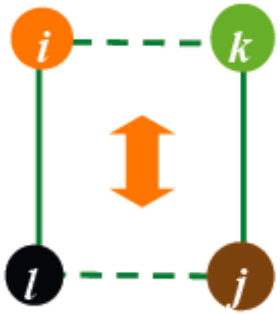
Heterophily/homophily



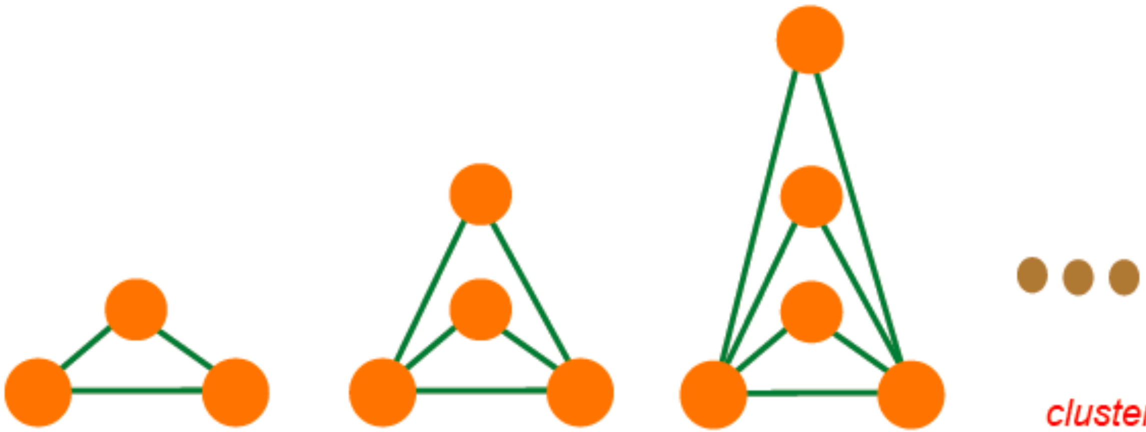
Distance/similarity

ERGM PARAMETRIZATION AND NETWORK STATISTICS /5

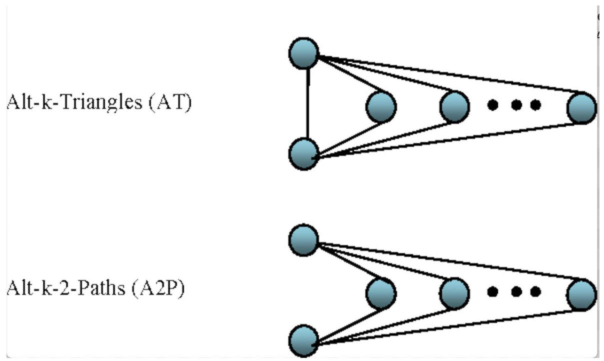
If we believe that (Snijders, et al., 2006)



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



clustered regions

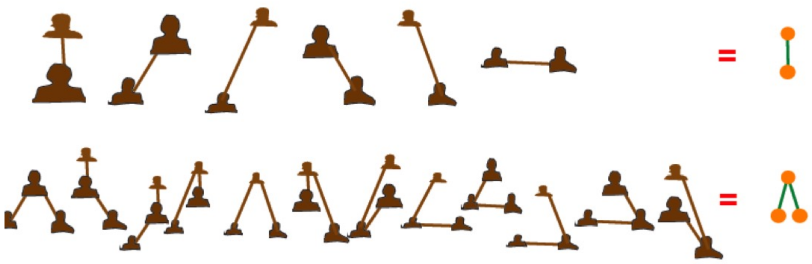


ERGM SPECIFICATION

Model specification involves:

1. choosing the set of network statistics *g(y)*
   - *a. from minimal: # of edges*
   - *b. to satured: one term for every dyad in the network*
2. choosing homogeneity constraints on the parameter θ

   i.e, for edges:
   - *a. all homogeneus*
   - *b. group specific (by sex, age,…)*
   - *c. dyad specific*



- There is one parameter for each class of network configurations
- The corresponding statistic is the number of configurations in y

ERGM ESTIMATION

- MLE: really hard to compute the constant $\kappa(\eta)$
- Simulated ML: relatively "simple" to simulate a sample of *m* random networks (via MCMC, also in Bayesian framework) from an ERGM with a fixed parameter η_0 (P_{η_0}) and thus approximate and then maximize loglikehood
- Pseudo MLE (PMLE): the same of logit model estimation with DV *Y* and covariate matrix given by:

$$\Delta = \left\{ g(1, \mathbf{y}_{(-ij)}) - g(0, \mathbf{y}_{(-ij)}) \right\}_{i,j}$$

- PMLE: usually works well in the choice of η_0

Amati V., Lomi A., Mira A. (2018) Social Network Modelling, *Annual Review of Statistics and Its Application*, 5, 343-369.

FITTING ERGM TO DATA IN R

- Dedicated statnet package for fitting, simulating models in ERG form
- Basic call structure:

```
ergm(y ~ term1(arg) + term2(arg))
```

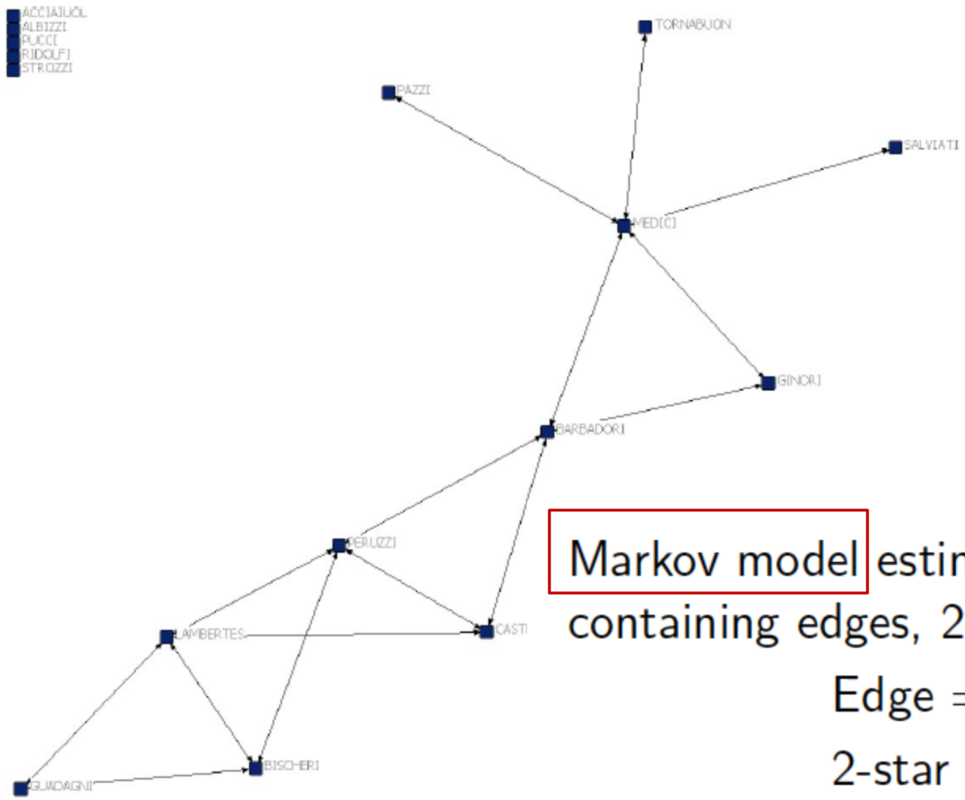
- All available terms can be found in:

```
help("ergm-terms")
```

FITTING ERGM TO DATA IN R

- `statnet` employs MCMC methods and MCMC-MLE methods to perform likelihood-based inference
- What happens when you run `ergm`?
  - First guess at θ done using the MPLE
  - Simulation of $y_1, \ldots, y_n$ based on the initial guess
  - The simulated sample is used to find θ using MLE
  - Previous two steps are iterated for good measure (since initial estimate is likely off)

FLORENTINE FAMILIES: BUSINESS NETWORK



Markov model estimates and standard errors, for a model containing edges, 2-stars, 3-stars, triangles

Edge = - 4.27 (1.13)*

2-star = 1.09 (0.65)

3-star = -0.67 (0.41)

Triangle= 1.32 (0.65)*

Interpretation:

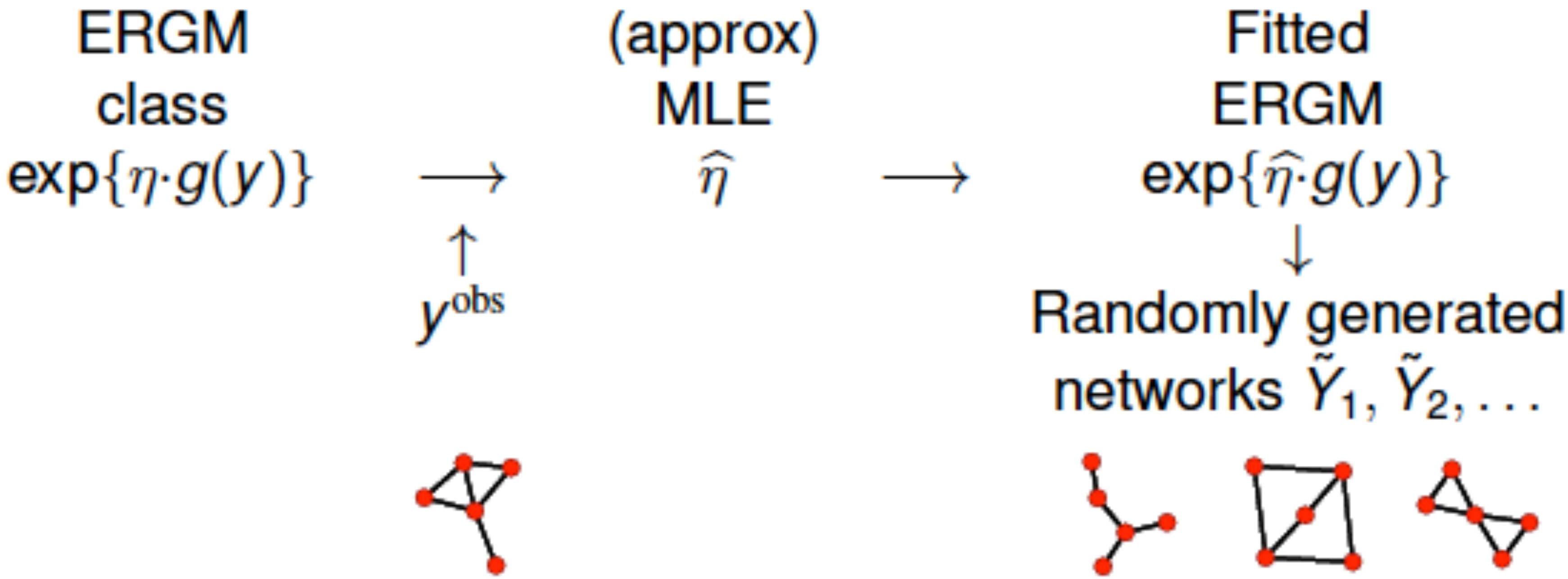
- Edges occur relatively rarely (negative edge parameter)
- Business ties tend to occur in triangular clusters
- Although not significant, star effects suggest that there is a tendency for a limited number of business partners

MODEL EVALUATION

- Is the model-class itself able to represent a range of realistic networks?
  – *model degeneracy:* small range of graphs covered as
    the parameters vary (Handcock 2003)
- What are the properties of different methods of estimation?
  – e.g, MLE, psuedolikelihood, Bayesian framework
  – *computational failure*: estimates do not exist for
    certain observable graphs
- *Can we assess the goodness-of-fit of models?*
  – appropriate measures and tests
  (Besag 2000; Hunter, Goodreau, Handcock 2007)

ERGM: GOODNESS OF FIT

General reasoning:



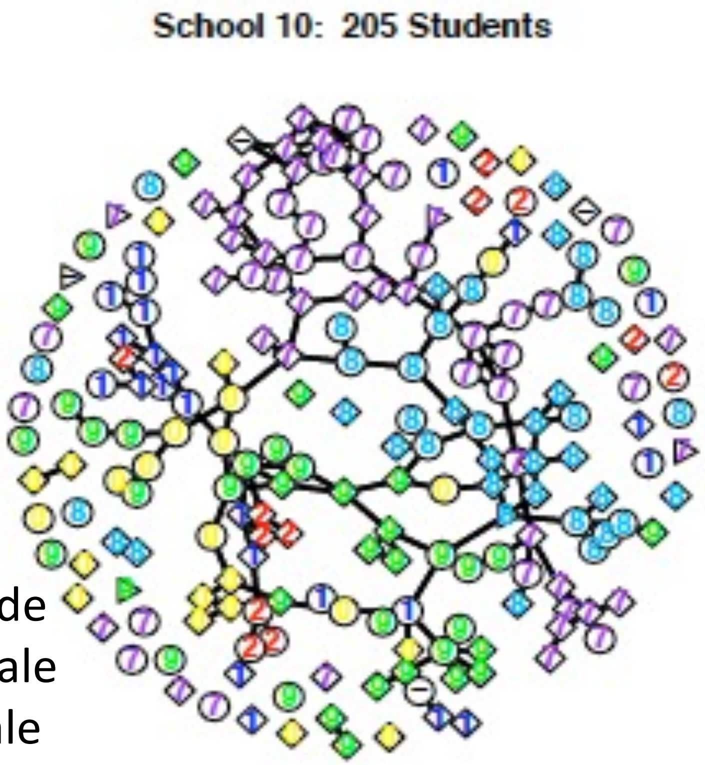
How does the observed network is “representative of the sample Y_1, Y_2, …?

ADD HEALTH DATA SET

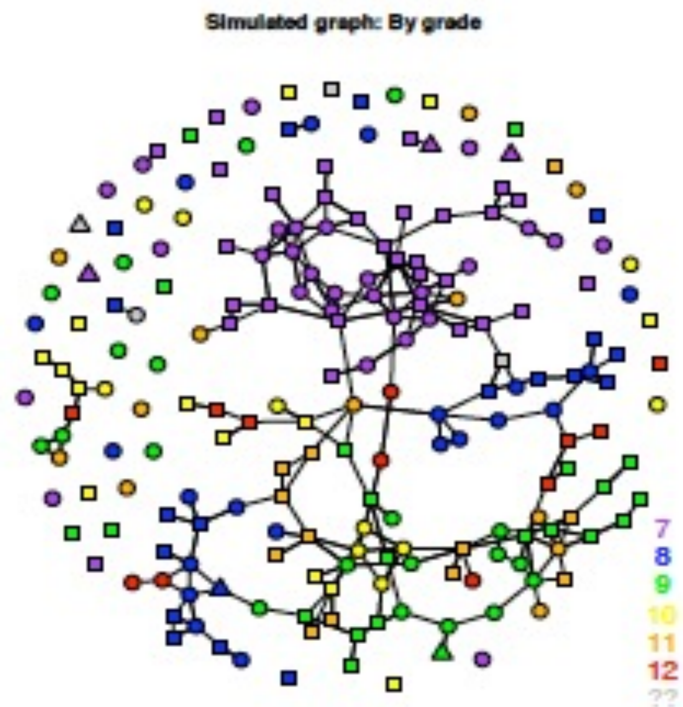
– "Add Health" is a school-based study of the health-related behaviors of adolescents in grades 7 to 12.

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

The data:



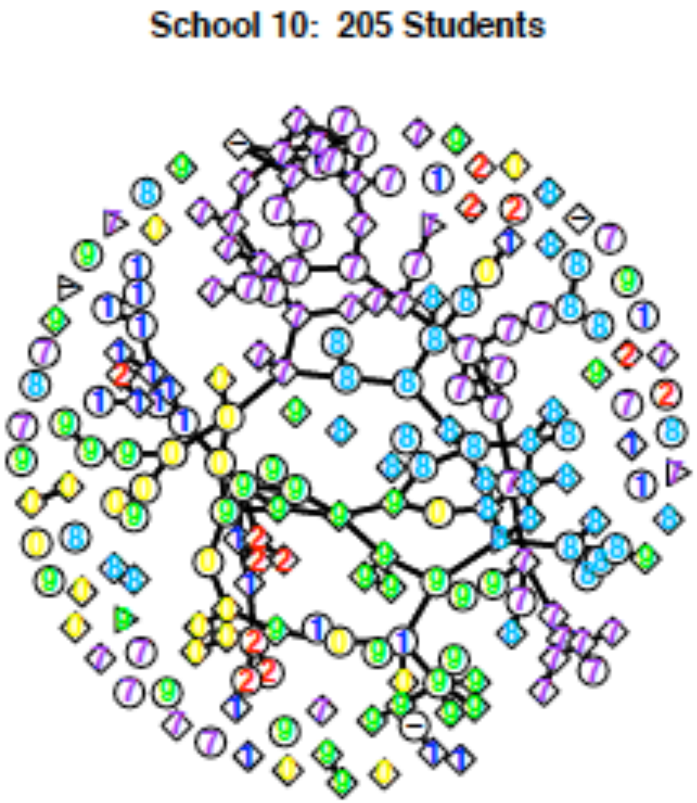
Simulated network, model A:



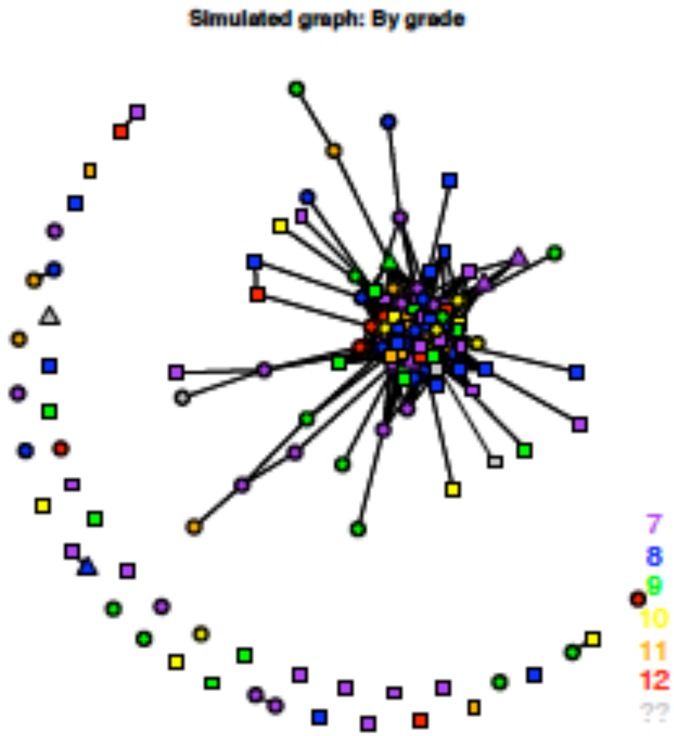
Colour: grade
Circle: Female
Square: Male
Triangle: missing

ADD HEALTH DATA SET

The data:



Simulated network, model B:

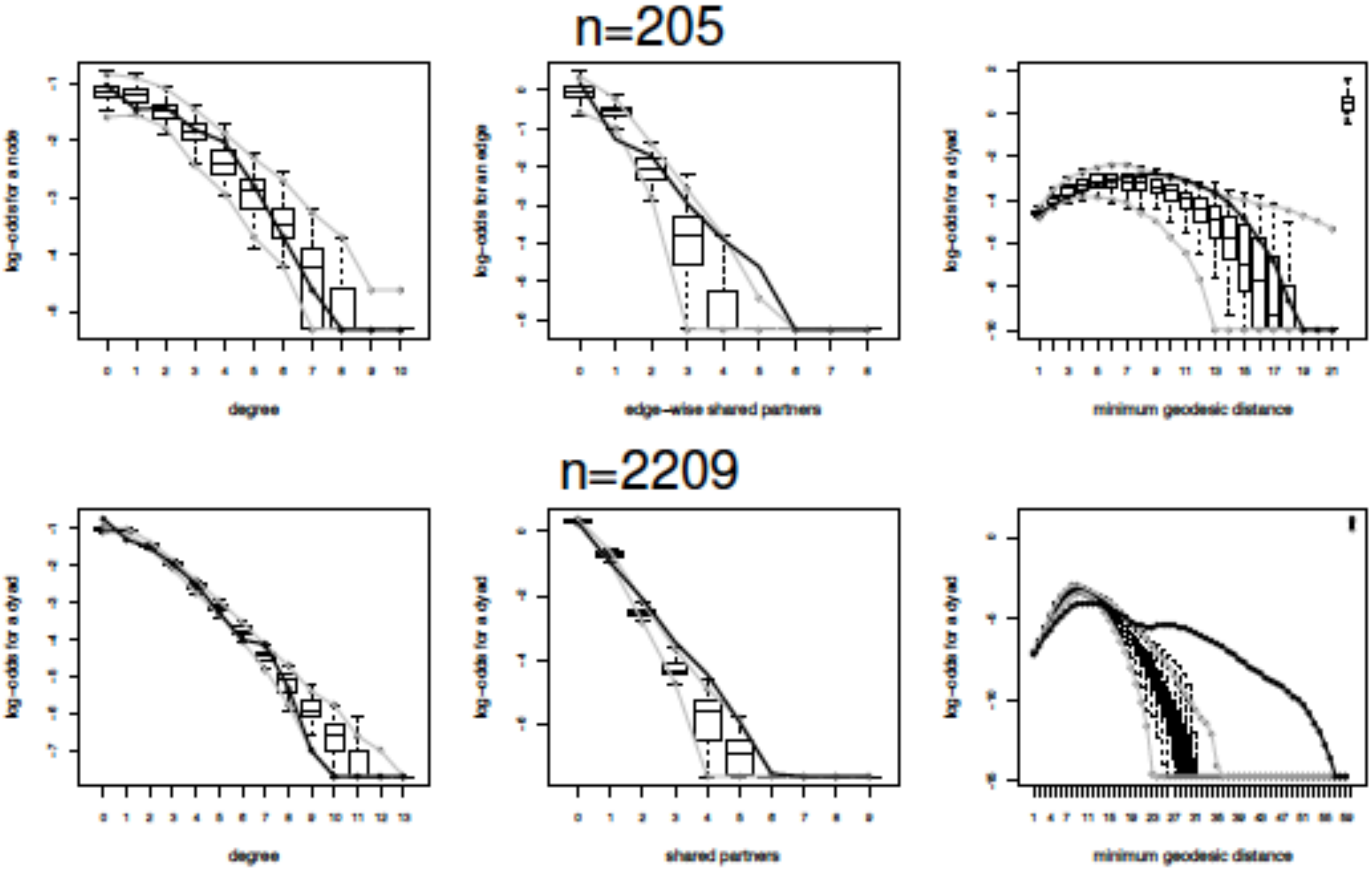


ADD Health data set

- Model A: $g(y)$ contains terms for
  - # of edges
  - Homophily effects of grade, sex, and race factors
  - Main effects of grade, sex, and race factors
  - $\sum_i (.632)^i \mathrm{EP}_i$, where $\mathrm{EP}_i =$# edges with i shared partners

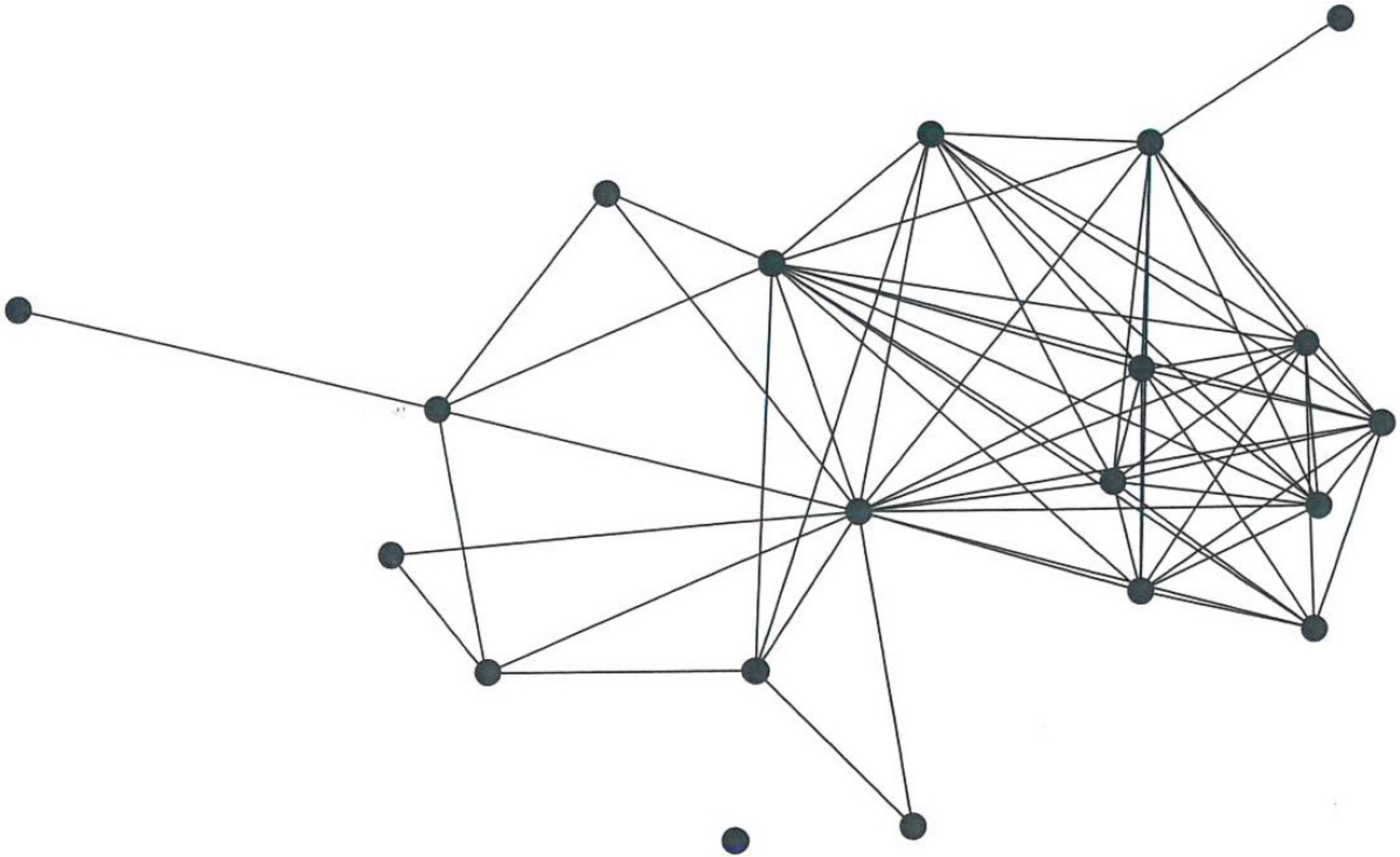
- Model B: $g(y)$ contains terms for
  - # of edges
  - # of neighbors of the same sex (homophily effect)
  - # of 2-stars
  - # of triangles

GRAPHICAL GOF: ADD HEALTH DATA



See also Kolaczyk (2009) on Lazega's network of collaborative working ties (case study 6.5.4, pp. 188-193) for parameter and GOF interpretation)

ERGM EXAMPLE: ACQUAINTANCESHIP UNDIRECTED NETWORK (N= 20)

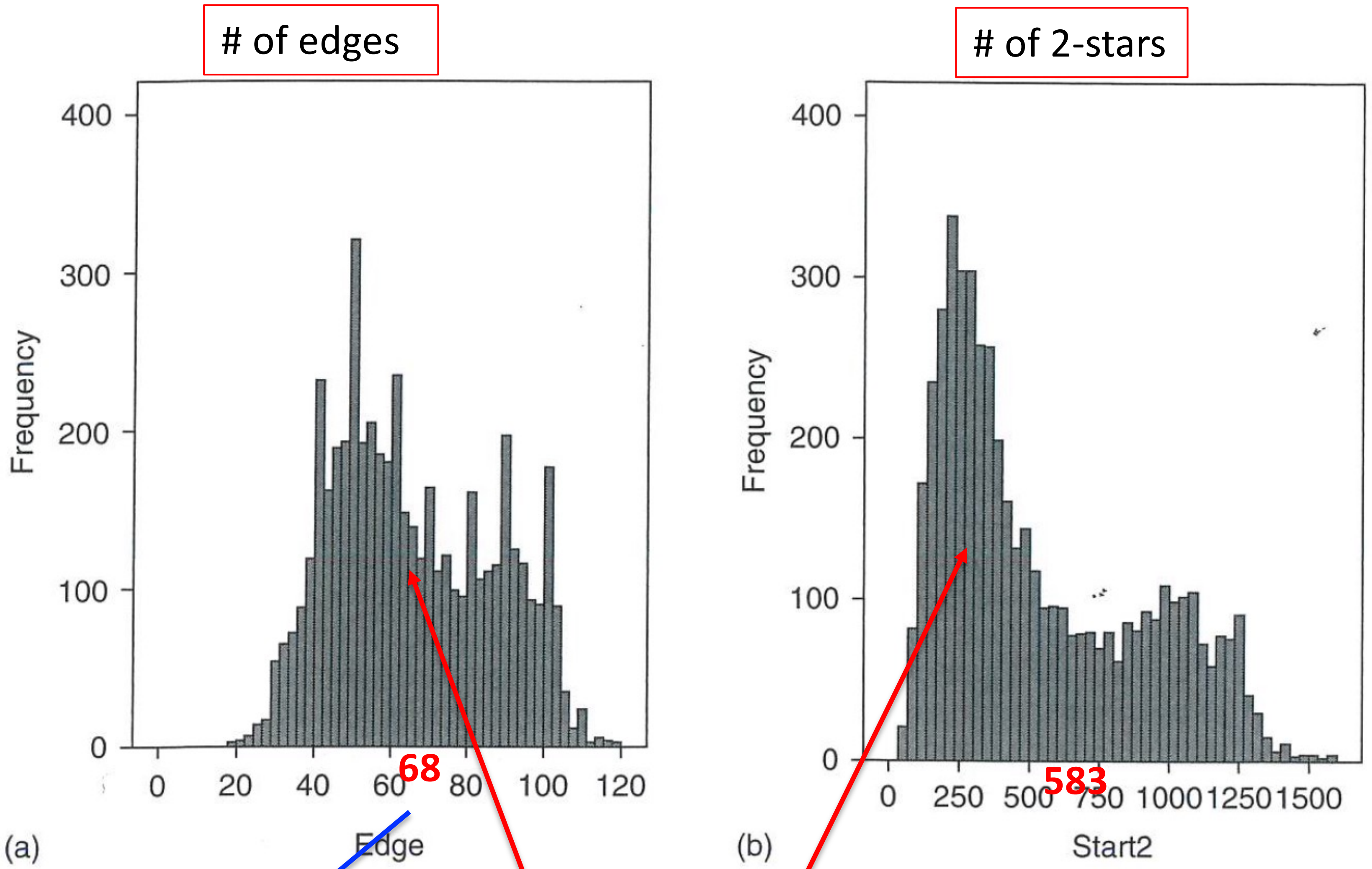


Parameter estimates: 1. Markov model, 2. Social circuit model

Parameter	*Estimate*	*Standard error*	*Convergence*
Markov model			
Edge	−2.165*	0.716	−0.046
2-star	0.108	0.100	−0.038
3-star	−0.044*	0.012	−0.032
Triangle	0.733*	0.104	−0.022
Social circuit model			
Edge	−2.232	1.270	−0.015
Popularity/Activity (*k*-stars)	−0.165	0.476	−0.015
Multiple triangulation (*k*-triangles)	1.355*	0.535	−0.016
Multiple connectivity (*k*-2paths)	−0.252*	0.088	−0.025

- Good convergence for both models (statistics not reported here)
- Edges are uncommon in both models (negative edge parameter, although a large *se* in SC model) unless they are part of higher order configuration (as in SC model)
- No high-degree actors (unless involved in triangulation or multiple connectivity effects)
- Triangulation occurs through the formation of *k*-triangle bases rather than edges: sharing several partners tends towards a direct tie

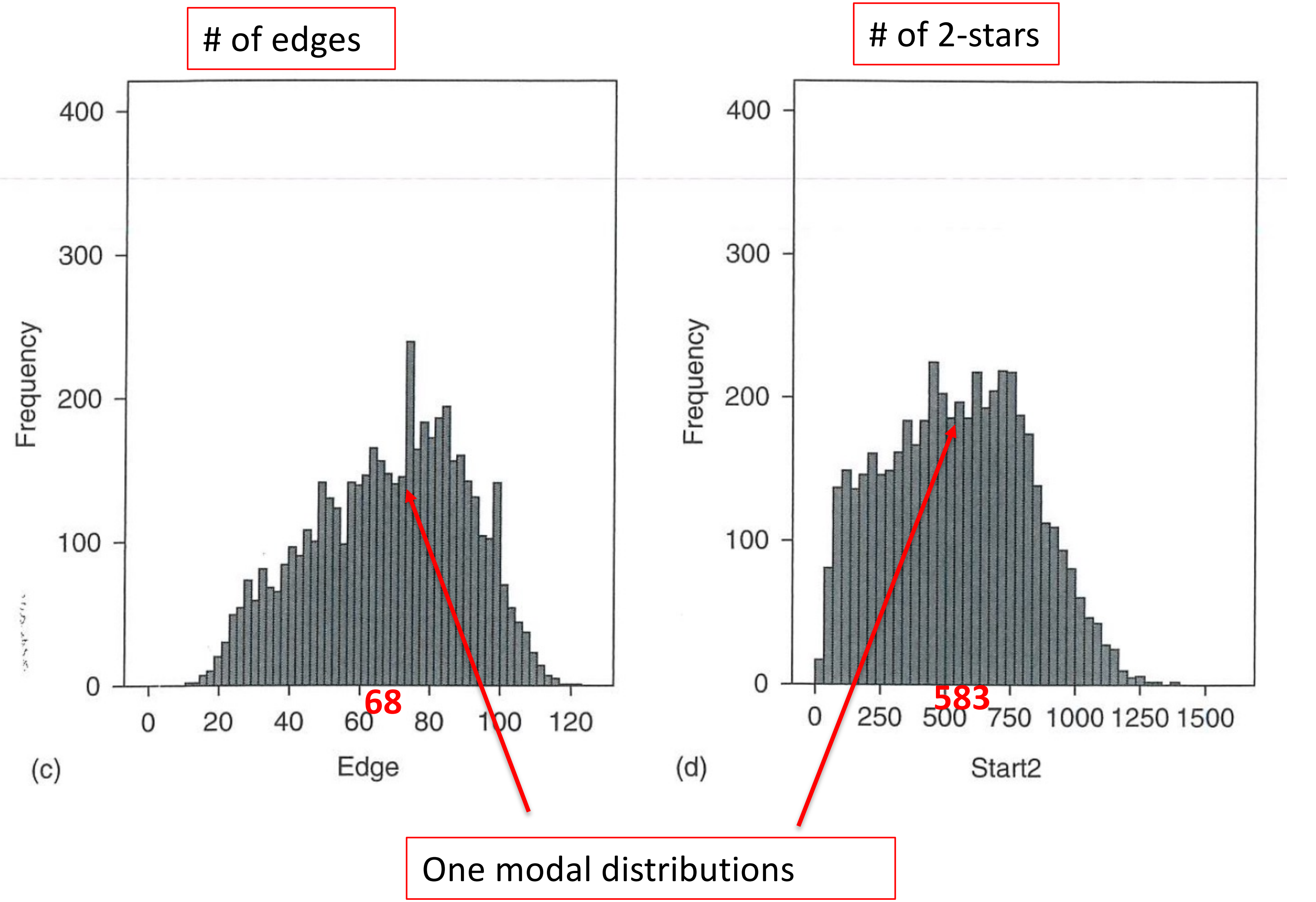
SIMULATION RESULTS: MARKOV MODEL (N=20, EDGES=68, 2-STARS=583)



Bimodal distributions

Observed network is not «typical» of the networks by the estimated Markov model

SIMULATION RESULTS: SOCIAL CIR. MODEL (N=20, EDGES=68, 2-STARS=583)



Social circuit model is to be preferred for the observed network

More on ERGM and network modeling

Other types of relational data (network):

- Valued / weighted (*Generalized ERGM*)

- Bipartite data

- Multiplex data

Longitudinal data (*Temporal Exponential Random Graph Model, TERGM*)

Egocentric network (ergm.ego in R, also simulation of complete networks from these egodata that are consistent with the observed model statistics)

Other modeling approach:

- Latent network models: assuming the existence of latent (i.e., unobserved) variables, such that the observed variables have a simple probability distribution given the latent variables.
- An important class: stochastic blocks models