Overview course module "Stochastic Modelling"

I. Introduction

II. Actor-based models for network evolution

III. Co-evolution models for networks and behaviour

IV. **Exponential Random Graph Models**

   A. Definition

   B. Example: gossip about the boss

   C. Estimation / goodness of fit issues

A: Exponential Random Graph Models

- history: Frank & Strauss “Markov random graph” model (1986), Frank (1991) and Wasserman & Pattison (1996) generalised to exponential family distribution:

$$\Pr(X = x) \propto \exp\left(\sum_{k=1}^{K} \beta_k s_k(x)\right)$$

‘proportional to’
proportionality constant unknown and practically uncalculable(!)

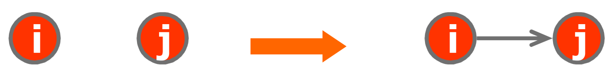
linear combination of network statistics
very flexible!

*Likelihood of macro structure **x** is explained by prevalence of micro structures **s(x)**, testable via parameters **β**.*

Some possible statistics

- tie count statistic

$$s_k(x) = \sum_{i,j=1}^{n} x_{ij}$$



- reciprocity statistic
  (for directed graphs)

$$s_k(x) = \sum_{i,j=1}^{n} x_{ij} x_{ji}$$

- transitive closure

$$s_k(x) = \sum_{i,j,k=1}^{n} x_{ij} x_{jk} x_{ik}$$

A positive parameter β attached to any of these statistics means: the right configuration is more likely than the left one.

Example

- Consider an ERG model for an undirected network with 3 parameters for these statistics:

(1) number of edges $s_{edges}(x) = \frac{1}{2} \sum_{i,j=1}^{n} x_{ij}$

$s_{2-stars}(x) = \frac{1}{2} \sum_{i,j,k=1}^{n} x_{ij} x_{ik}$ number of 2-stars (2)

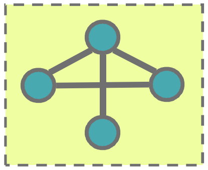
(3) number of triangles $s_{triangles}(x) = \frac{1}{6} \sum_{i,j,k=1}^{n} x_{ij} x_{jk} x_{ik}$

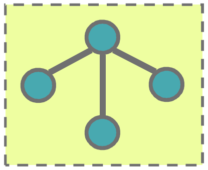
- Then the 3-parameter ERG distribution function is this one:

$$\Pr(X = x) \propto \exp(\beta_{edges} \times s_{edges}(x) + \beta_{2-stars} \times s_{2-stars}(x) + \beta_{triangles} \times s_{triangles}(x))$$

Example

› ...and consider the following two 4-node-networks & their statistics:

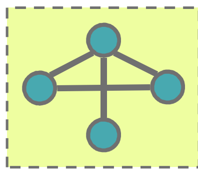




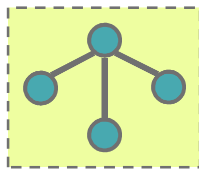
	x^a	x^b
s_{edges}	4	3
$s_{2\text{-}stars}$	5	3
$s_{triangles}$	1	0

Example

› Probabilities are (only) given up to a proportionality factor:



x^a

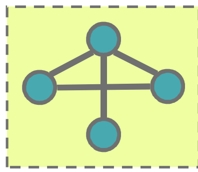


x^b

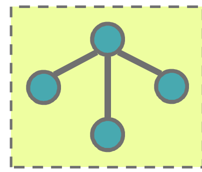
$$\Pr(x^a) \propto \exp(4\times\beta_{edges} + 5\times\beta_{2\text{-}stars} + 1\times\beta_{triangles})$$

$$\Pr(x^b) \propto \exp(3\times\beta_{edges} + 3\times\beta_{2\text{-}stars})$$

…but the ratio of probabilities can be calculated!



x^a



x^b

$$\frac{\Pr(x^a)}{\Pr(x^b)} = \frac{\exp(4\beta_{edges} + 5\beta_{2-stars} + \beta_{triangles})}{\exp(3\beta_{edges} + 3\beta_{2-stars})}$$

$$= \exp\big((4-3)\beta_{edges} + (5-3)\beta_{2-stars} + \beta_{triangles}\big)$$

$$= \exp(\beta_{edges} + 2\beta_{2-stars} + \beta_{triangles})$$

Local characterisation of ERGMs

- ERG distribution is considered as a collection of local conditional tie probabilities

$$\frac{\Pr(x^a)}{\Pr(x^b)} = \exp\left(\sum_{k=1}^{K} \beta_k \left(s_k(x^a) - s_k(x^b)\right)\right)$$

compared are a network x^a
and a "neighbouring network" x^b

model parameters

differences in model statistics between x^a and x^b

What this equation refers to is the (conditional) probability of the one tie by which they differ!

Recall similar formula for dynamic networks

- Local characterisation of ERG models:

$$\frac{\Pr(x^a)}{\Pr(x^b)} = \exp\left(\sum_{k=1}^{K} \beta_k \left(s_k(x^a) - s_k(x^b)\right)\right)$$

- Local characterisation for actor-based network evolution models:

$$\frac{\Pr(x^c \rightarrow_i x^a)}{\Pr(x^c \rightarrow_i x^b)} = \exp\left(\sum_{k=1}^{K} \beta_k \left(s_{ik}(x^a) - s_{ik}(x^b)\right)\right)$$

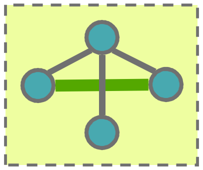
compared are two moves ('micro steps') made by actor ***i*** from a network $\boldsymbol{x^c}$ to two "neighbouring networks" $\boldsymbol{x^a}$ and $\boldsymbol{x^b}$

model parameters

difference in model statistics of actor ***i*** between the two compared moves

› …so how do these ‘conditional odds’ for the middle tie to exist (vs. not to exist) look like in applications?

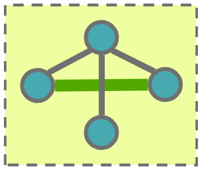
- Suppose, in a (larger) trade network, estimation gave:
  - redundant ties are avoided: $\beta_{triangles} = -0.4$
  - positive degree variance: $\beta_{2\text{-}stars} = 0.1$
  - low density: $\beta_{edges} = -1.5$



- Then the equation becomes:

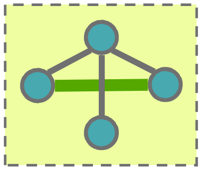
$$\frac{\Pr(\text{middle tie})}{\Pr(\text{no middle tie})} = \exp(\beta_{edges} + 2\beta_{2-stars} + \beta_{triangles})$$
$$= \exp(-1.5 + 2 \times 0.1 - 0.4)$$
$$= \exp(-1.7)$$
$$\approx 0.183$$
$$\approx 1/5.5$$

i.e., the middle tie is about
5.5 times as likely NOT to exist
as to exist [given the rest of the network].



› Same model in a different application:

- Suppose, in a (larger) friendship network, estimates were:
  - closure: $\beta_{triangles} = 0.4$
  - pos. degree variance: $\beta_{2\text{-}stars} = 0.1$
  - low density: $\beta_{edges} = -1.5$

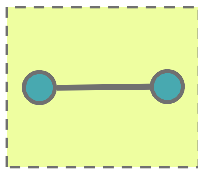


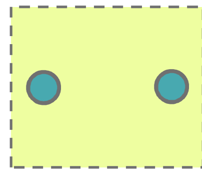
- Then the equation becomes:

$$\frac{\Pr(\text{middle tie})}{\Pr(\text{no middle tie})} = \exp(-0.9) \approx 0.407 \approx 1/2.5$$

- Still two½ times as likely NOT to exist – but always keep in mind that *all* ties are random…

...random ties are benchmark for evaluating such conditional odds! Hence, consider these “networks” as baseline comparison:





$$\frac{\Pr(\text{tie})}{\Pr(\text{no tie})} = \exp(\beta_{edges}) = \exp(-1.5) \approx 0.223 \approx 1/4.5$$

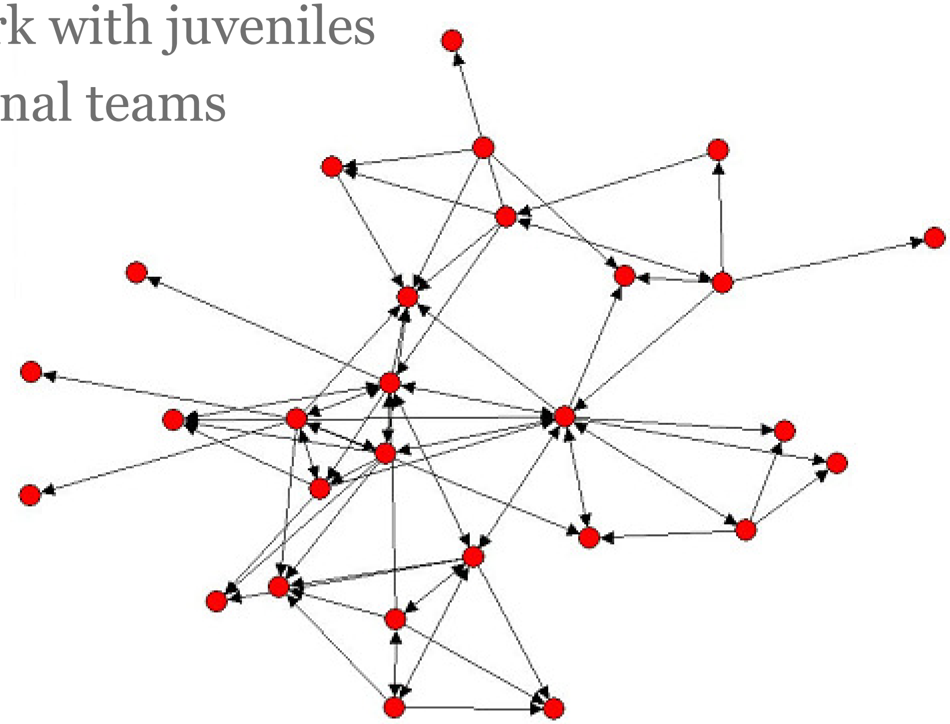
- In the *friendship context* (odds=2.5), the tie in question is *relatively likely* to be present,
- In the *trade context* (odds=5.5), it is *relatively unlikely* to be present, given the network’s density.

Exponential Random Graph Models

- high flexibility due to the many possibilities of choosing statistics & controlling effects for each other **+**
- can express small worlds (see Robins et al. 2005) but also other network types
- network probability model: not tied to any particular algorithm**+**
- problems: estimation, model specification, interpretation **–**

B: Example: gossip about the boss

› **data from Lea Ellwardt (2008)**
  - organisation for social work with juveniles
  - one department with internal teams
  - 28 employees



› *gossip about the boss in this department was largely* ***negative***

Hypotheses:

1. closeness / proximity is associated with gossip
    - friendship / liking predicts gossip
    - communication frequency predicts gossip
    - team structure predicts gossip
    - gossip occurs in local clusters
2. information asymmetry is associated with gossip
    - dissimilarity in / lack of contacts with the boss predicts g'sp.
3. "negative attitude" is associated with gossip
    - distrust in management predicts gossip
    - disliking the boss predicts gossip
4. attitude similarity is associated with gossip

Results from an exponential random graph analysis:

Parameter	*Estimate*	*SE*	
NETWORK EFFECTS			
1. Reciprocity	0.9356	0.6254	
2. Transitive triplets	-0.1851	0.2873	… (1)
3. 3-cycles	-0.1362	0.3619	… (1)
4. Alternating out-k-stars	**0.8823***	**0.3179**	← – – – Unpredicted: heterogeneity of outdegrees
5. Alternating in-k-stars	-0.2607	0.3701	
6. Alternating k-triangles	**0.9762***	**0.4834**	✓ (1)
7. Team membership	**0.7866***	**0.2835**	✓ (1)
8. Communication ego-alter (symmetric)	**0.9111***	**0.3254**	✓ (1)
9. Liking of alter by ego (out-degree)	**1.8782***	**0.3305**	✓ (1)

Results from an exponential random graph analysis:

Parameter	*Estimate*	*SE*	
SENDER EFFECTS			
10. Trust in management	-0.1261	0.1337	… (3)
11. Liking of manager	**-0.5071***	**0.2289**	✓ (3)
12. Communication with manager	**-0.3239***	**0.0915**	✓ (2)
RECEIVER EFFECTS			
13. Trust in management	0.2187	0.1996	… (3)
14. Liking of manager	0.2244	0.3588	… (3)
15. Communication with manager	-0.0445	0.1411	… (2)

Results from an exponential random graph analysis:

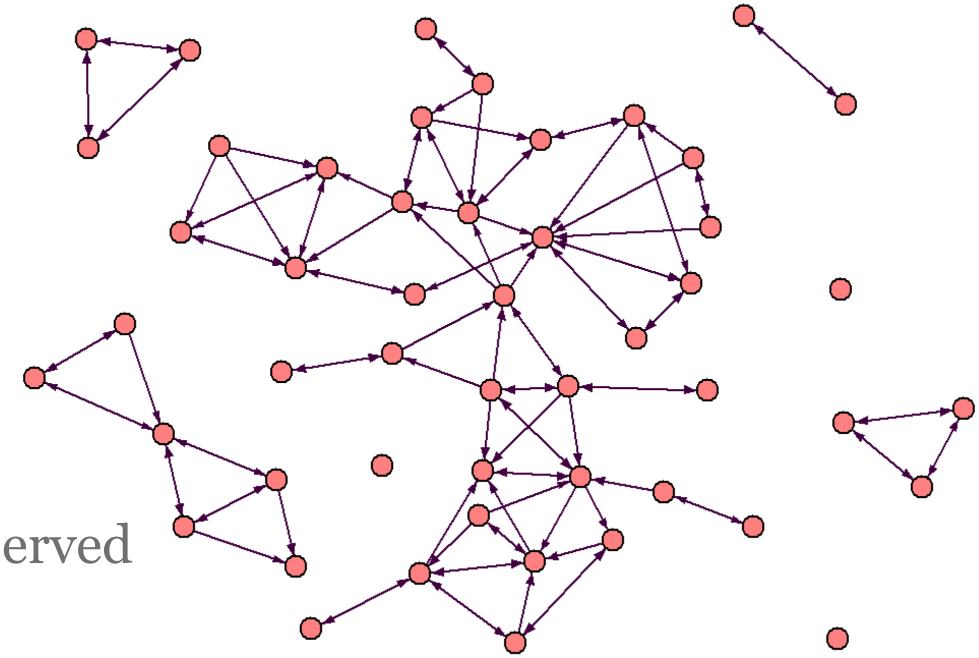
Parameter	*Estimate*	*SE*	
SIMILARITY EFFECTS			
16. Trust in management (similar)	**-1.3708***	**0.7921**	✓ (4)
17. Liking of manager (same)	0.3919	0.2878	… (4)
18. Communication with manager (similar)	-0.4359	0.6343	… (4)

* $p<0.05$ (one-sided test)

Conclusion: by and large, the analysis supports the hypotheses…

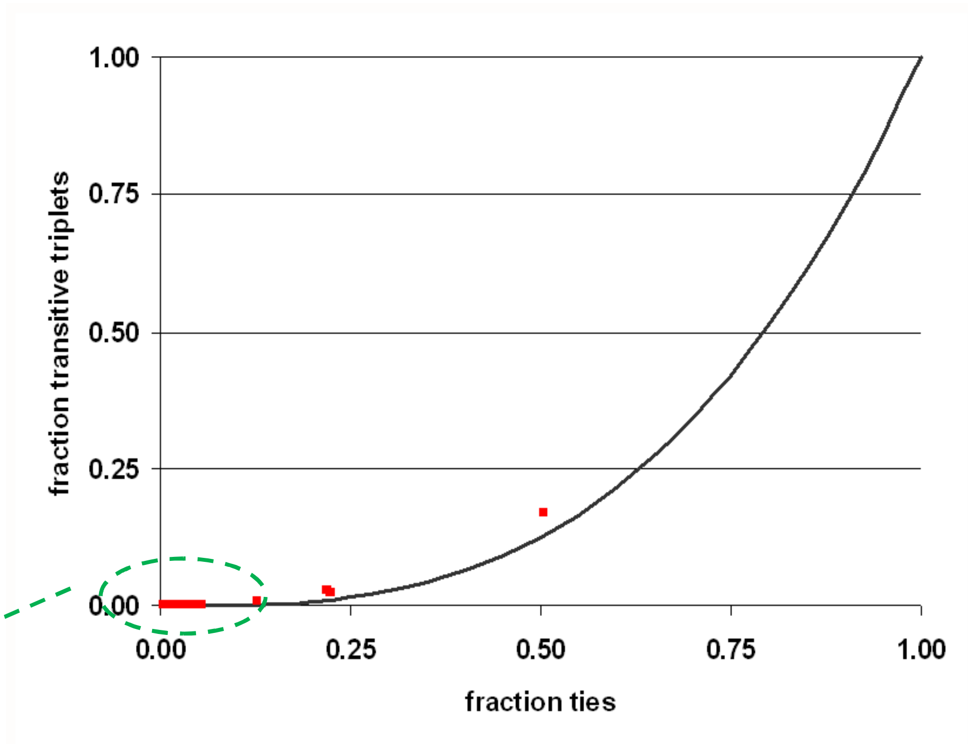
Example: 50 Scottish girls' nominated friendship

› **data from Michell & Amos (1997)**
  - 50 girls
  - Scotland 1995
  - 1st year secondary
  - 13 years old



› ***triangulation*** can be observed

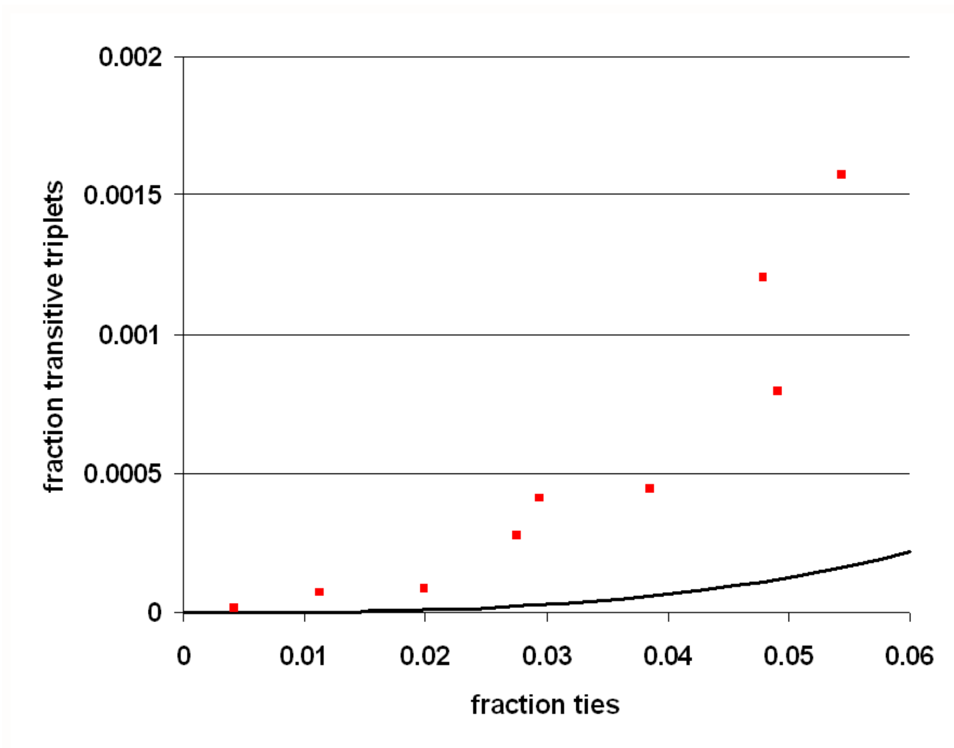
› triangulation general: typically more triangles in data than expected under random tie formation…



zooming in on the next slide …

Red markers stand for empirically observed data sets, black line for expectations under randomness.

› Red markers stand for observed values in several empirical data sets; the black line indicates expectations under randomness.

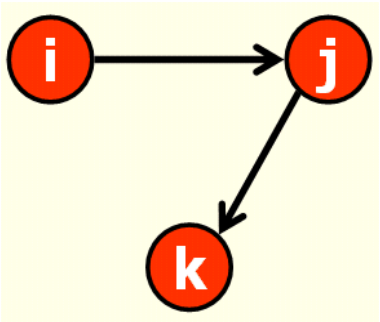


Red markers lie way above the random expectation.

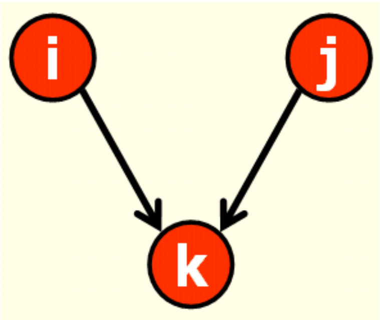
So: include triangulation in your model for this type of data!

First try: estimate a simple model with triangulation / transitive closure tendencies

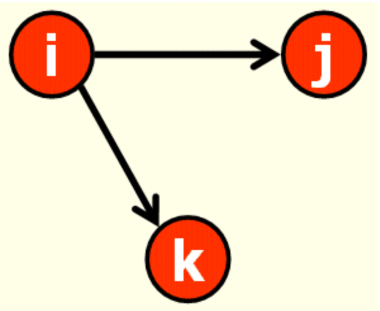
› include *precursor configurations* in attempt to arrive at a meaningful model (Snijders 2002)



2-path



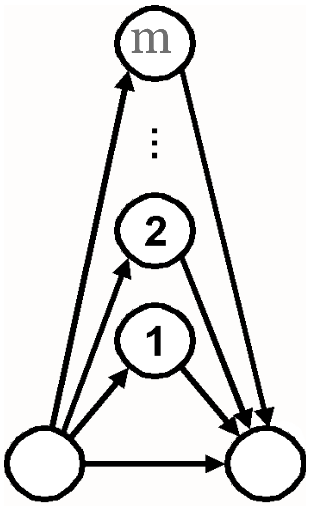
in-2-star



out-2-star

```
Good convergence is indicated by the t-statistics being close to zero.
One or more of the t-statistics are rather large.
Convergence of the algorithm is doubtful.

@2
Estimation results.
-------------------

Regular end of estimation algorithm. Total of 2092 iteration steps.

@3
Estimates and standard errors

 1. u: number of arcs                             4.9212  (   1.7544)
 2. u: reciprocity                                3.9062  (   0.3581)
 3. u: in-2-stars                                -1.2377  (   0.3367)
 4. u: 2-paths                                   -0.8630  (   0.1349)
 5. u: transitive triplets                      -12.7060  (  99.0000)
```

Second try: estimate another model with 'new specification' (Snijders et al. 2006)



$$s_k(x) = \sum_{m=1}^{n-2} (-1)^{m+1} \frac{\tau_m}{2^{m-1}}$$

"alternating m-triangles statistic"

τ_m = number of m-triangles (configurations on the left)

```
Good convergence is indicated by the t-statistics being close to zero.

@2
Estimation results.
-------------------

Regular end of estimation algorithm.
Total of 2015 iteration steps.

@3
Estimates and standard errors

 1. u: number of arcs                          -3.2498  (   0.3456)
 2. u: reciprocity                              4.4430  (   0.3867)
 3. u: alternating m-triangles, parameter 2     0.6734  (   0.0758)
```

Why does estimation seem so difficult for some models?
Why so complex statistics like alternating m-triangles?

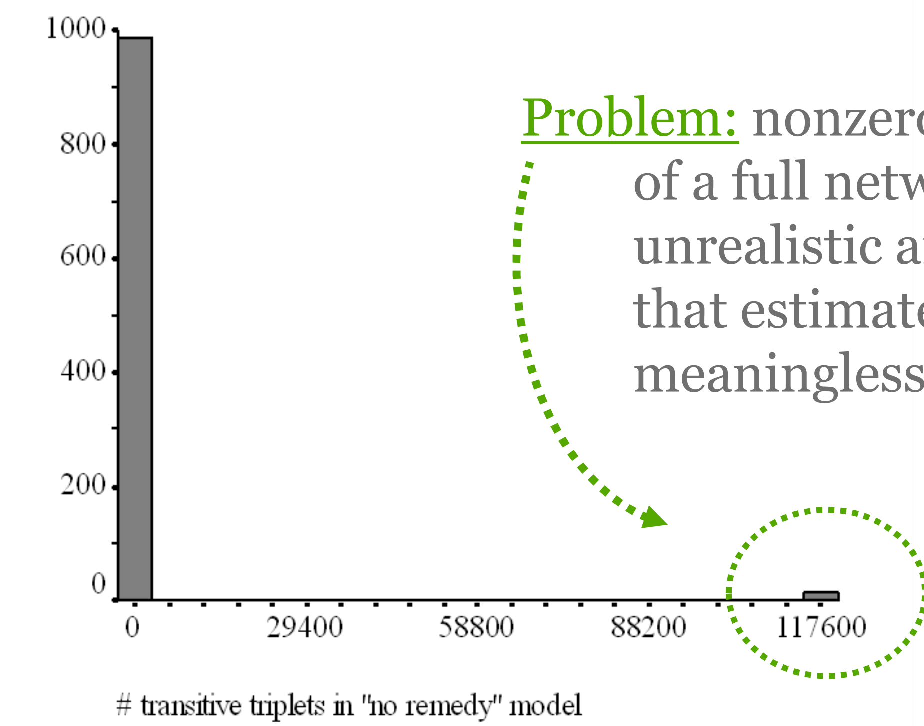
C: Estimation of ERGMs

› ***estimation is based on simulation:***

1. assume starting values for parameters
2. simulate networks that are drawn from this parametrisation's ERG distribution
3. compare them to actually observed (and to-be-modelled) network
4. adjust parameters in order to get **better fit** [on desired dimensions]
5. repeat from 2. until parameter values **converge**

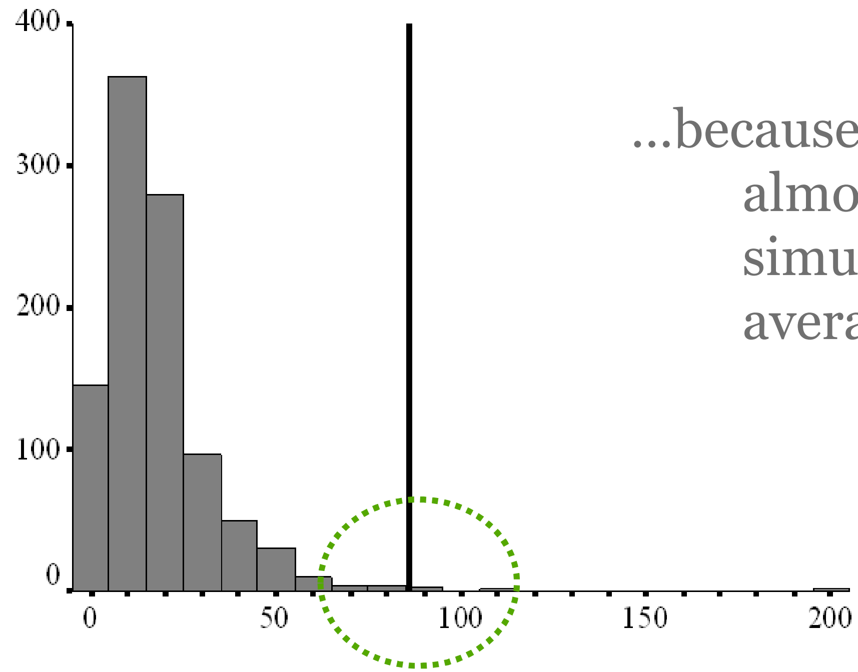
› this unfortunately does not always work,
and notably triangulation effects cause problems...

› see here the distribution of the number of transitive triplets over simulated networks from a **model like the first** we estimated...



Problem: nonzero probability of a full network! This is unrealistic and implies that estimates are meaningless...

› ...zooming in on the 'low end' (i.e., realistic networks; vertical line = observed value):



same graph with emphasis on the 'empty end'

...because observed value is almost never assumed by simulations – only "on average"!

To keep in mind about ERGMs:

- Visualise ERG models as probability distributions on a (huge) space of all possible network,
- one observed network is modelled as drawn from that distribution.
- Model parameters β are
  - attached to network statistics $\boldsymbol{s}$,
  - these statistics in general correspond to subgraph counts (local patterns, 'motifs'),
  - the parameters describe the relative prevalence of the corresponding subgraph in the total graph.
- Interpretation of parameters needs to take into account other parameters (notably edges / arcs).

References (in chronological order of appearance)

Small worlds and scale-free networks:

Barabási, A.L., and Albert, R. (1999). Emergence of scaling in random networks *Science*, 286:509-512.

Caldarelli, G.; Capocci, A.; De Los Rios, P.; Munoz, M. A. (2002). Scale-free Networks without Growth or Preferential Attachment: Good get Richer. eprint arXiv:cond-mat/0207366.

Guare, J. (1990). *Six Degrees of Separation: A Play*. New York: Vintage Books.

Kasturirangan, R. (1999). Multiple scales in small-world graphs. Massachusetts Institute of Technology AI Lab Memo 1663. Also cond-mat: 9904055.

Kleinberg, J.M. (2000). Navigation in a small world. *Nature* 406, 845.

Milgram, S. (1967). The Small World Problem. *Psychology Today*, 2, 60-67.

Newman, M.E.J. (2000). Models of the Small World. *Journal of Statistical Physics*, 101, 819-841.

Watts, D.J. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology*, 105, 493-527.

Watts, D.J., and Strogatz, S.H. (1998). Collective Dynamics of ‘small world’ networks. *Nature*, 393, 440-442.

Classic paper on testing the triad census against tie-independence:

Holland, P.W. & Leinhardt, S. (1978). An Omnibus Test for Social Structure Using Triads. *Sociological Methods & Research* 7: 227-256. *(see also Wasserman & Faust, chp.14)*

Null model assumes tie-independence, given the degree distribution:

Karlberg,M. (1999). Testing Transitivity in Digraphs. *Sociological Methodology* 29 (1): 225–251.

Papers on permutation tests:

Hubert, L.J., & Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology* 29: 190-241.

Krackhardt, D. (1988). Predicting with networks: nonparametric multiple regression analysis of dyadic data. *Social Networks* 10: 359-381.

Dekker, D., Krackhardt, D., & Snijders, T.A.B. (2003). Multicollinearity robust QAP for multiple regression. Forthcoming in *Psychometrika*.

Butts, C.T. (2005). Permutation models for relational data. Forthcoming in *Sociological Methodology*.

Exponential Random Graph Models:

Frank, Ove. 1991. Statistical Analysis of Change in Networks. *Statistica Neerlandica*, 45: 283-293.

Frank, Ove, and David Strauss. 1986. Markov Graphs. *Journal of the American Statistical Association*, 81: 832-842.

Robins, G.L., Woolcock, J., and Pattison, P. (2005) Small and other worlds: Global network structures from local processes. *American Journal of Sociology*, 110, 894-936.

Snijders, Tom A.B., Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock (2006). New specifications for exponential random graph models. To be published, *Sociological Methodology*.

Handcock, Mark S. (2003). Assessing degeneracy in statistical models of social networks. Working paper no. 39, Center for Statistics and the Social Sciences. Seattle: University of Washington.

Snijders, Tom A.B. (2002) Markov Chain Monte Carlo Estimation of Exponential Random Graph Models Journal of Social Structure. Volume 3, number 2.

Snijders, Tom A.B., Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock (2006). New specifications for exponential random graph models. *Sociological Methodology*.

Network dynamics:

Snijders, T.A.B., Ch. Steglich, M. Schweinberger and M.E. Huisman (2007). Manual for SIENA version 3.

Torlò, V., Steglich, Ch., Lomi, A. and Snijders, T.A.B. (2007). Network Influence, Social Selection and Individual Performance in Organizations. Working paper. http://www.gmw.rug.nl/~steglich/pdf/TSLS2007.pdf

Co-evolution:

Steglich, Christian ,Tom Snijders, and Patrick West, (2006). Applying SIENA: An illustrative analysis of the co-evolution of adolescents' friendship networks, taste in music, and alcohol consumption. Methodology 2(1), 48-56.

The s50 data set:

Michell, L., and Amos, A. (1997) Girls, pecking order and smoking. *Social Science and Medicine*, 44, 1861-1869.