New Specifications for Exponential Random Graph Models

Garry Robins
University of Melbourne, Australia

Workshop on exponential random graph models
Sunbelt XXVI
April 2006

http://www.sna.unimelb.edu.au/

- *What counts as a "good" model for a social network.*
- *Goodness of fit*
- *New specifications for exponential random graph (p*) models*
  - *Degree sequences; higher order triangulation; multiple connectivity*
  - *Nondirected and directed networks*
- *An empirical example*
- *Social selection models*
  - *example: exogenous attributes or endogenous network effects?*
- *Concluding remarks*

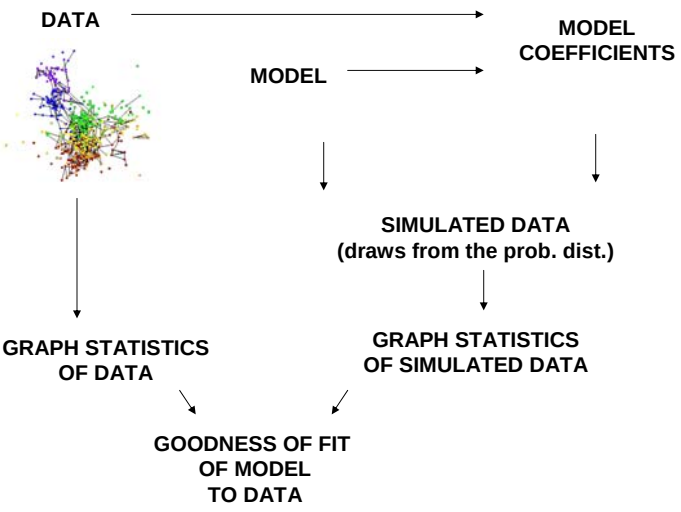
What counts as a "good" statistical model for an observed social network?

1. Models must be estimable from data.

2. Model parameters should imply model statistics consistent with those of the observed graph.

3. A good model will imply graphs with other features that are consistent with the observed graph.
   - *path lengths (geodesic distribution)*
   - *clustering (triangle formation)*
   - *degree distribution*
   - *denser regions (cohesive subsets of nodes)*

4. An excellent model will also successfully predict the presence or absence of network ties.



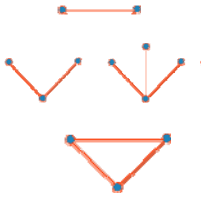
Markov random graph models

Notation:
Network variables: $Y_{ij} = 0, 1$
$\mathbf{Y} = [Y_{ij}]$, the set of all variables
$\mathbf{y} = [y_{ij}]$, the observed graph.

Sufficient statistics for a homogenous Markov graph model are the numbers of:

- *Edges*
- *Stars of different types*
- *Triangles*

in **y**

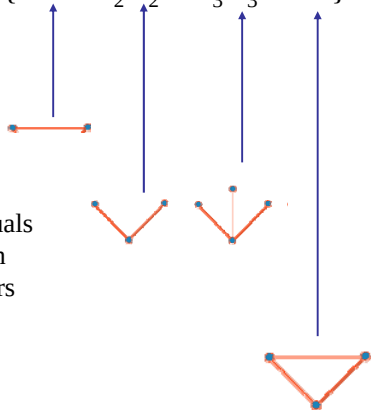


A Markov random graph model:

Nondirected networks

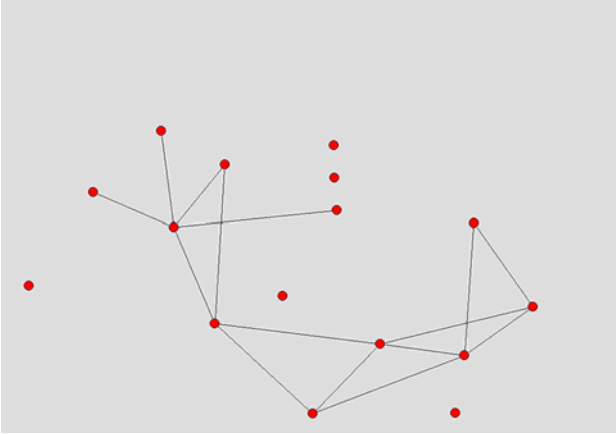
$$\Pr(\mathbf{Y} = \mathbf{y}) = (1/\kappa) \exp\{\theta L + \sigma_2 S_2 + \sigma_3 S_3 + \tau T\}$$

- *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 clustering



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

Example:
Florentine families business network
(Padgett & Ansell, 1993)



Markov model estimates:
Florentine families business network

Model containing edges, 2-stars, 3-stars, triangles

Monte Carlo Max. Likelihood estimates
(pnet)

Edge = - 4.27 (1.13)*

2-star = 1.09 (0.65)

3-star = -0.67 (0.41)

Triangle= 1.32 (0.65)*

Goodness of fit: Florentine Business network

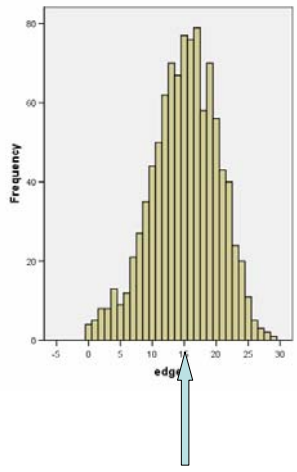
Comparing observed network to a sample from simulation of **Markov model**: *t*-statistics

Model statistics
Edges: t = -0.01
2-stars: t = -0.01
3-stars: t = 0.01
triangles: t = -0.03

Degree distribution
std dev degree dist: t = 0.51
skew deg dist: t = 0.17

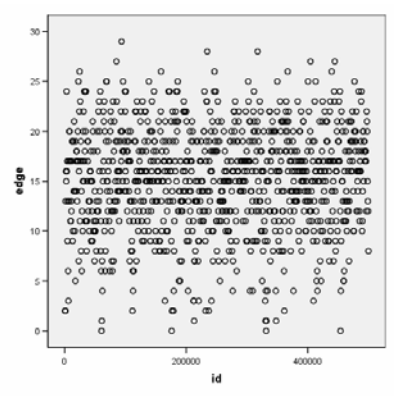
Clustering
global clustering: t = 0.12
local clustering: t = 0.68

Geodesic distribution:
None of the quartiles of the geodesic distribution for the observed graph are extreme in the distribution

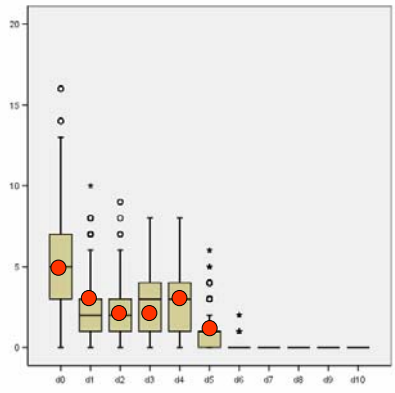


Observed graph

No of edges in simulated graphs

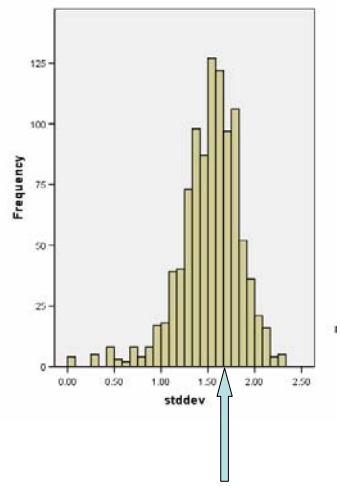


No of edges in simulated graphs against simulation number (*id*)



Degree distributions of simulated graphs

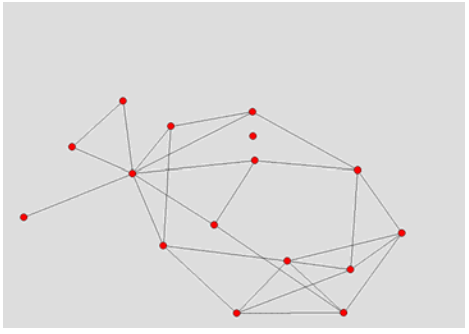
Observed degree distribution



Observed graph

Standard deviation of degree distribution of simulated graphs

But the same model is degenerate for the combined business and marriage networks: there are no coherent maximum likelihood parameter estimates



simulating from pseudo-likelihood estimates produces complete (or almost complete) graphs

Papers on new specifications

For advance copies, go to:
http://www.sna.unimelb.edu.au/

Original paper:

Snijders, Pattison, Robins, & Handock (2006). New specifications for exponential random graph models. *Sociological Methodology*. In press.

Forthcoming in *Social Networks*:

Goodreau (2006). Advances in Exponential Random Graph (*p**) Models Applied to a Large Social Network.

Hunter (2006). Curved exponential family models for social networks.

Robins, Pattison, Kalish, & Lusher (2006). An introduction to exponential random graph (*p**) models for social networks.

Robins, Snijders, Wang, Handcock, & Pattison (2006). Recent developments in exponential random graph (*p**) models for social networks.

Also see Hunter & Handcock (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*. In press.

New specifications for exponential random graph models

Estimation, simulation and goodness of fit software:

Statnet: http://csde.washington.edu/statnet

More tomorrow morning

SIENA: http://stat.gamma.rug.nl/siena.html

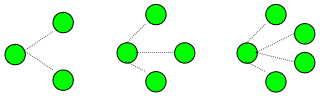
pnet: http://www.sna.unimelb.edu.au/pnet/pnet.html

New specifications for exponential random graph models

1. Parameters for degree sequences:
   a. Alternating *k*-stars
   b. Geometrically weighted degree distributions
2. Parameters for higher order triangulation
   a. Alternating *k*-triangles
   b. Edge-wise shared partner distributions
3. Parameters for higher order connectivity
   a. Alternating 2paths
   b. Dyad-wise shared partner distributions

Related model parameters

Star configurations



…

Parameters

σ_2 σ_3 σ_4 …

Markov models with star effects only (ignore triangles)

$\Pr(\mathbf{Y} = \mathbf{y}) = (1/\kappa) \exp\{\theta L + \sigma_2 S_2 + \sigma_3 S_3 + \sigma_4 S_3 + \ldots\}$

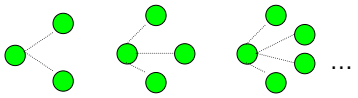
We can combine all the star effects into the one parameter by setting constraints among the star parameters:

$\sigma_2, \sigma_3, \sigma_4 \ldots$.

We have made an hypothesis about constraints based on experience with model parameters (and also on mathematical convenience.)

Related model parameters

Star configurations



Parameters

σ_2 σ_3 σ_4 …

Assume that $\sigma_k = -\sigma_{k-1}/\lambda$, for $k > 1$ and $\lambda \geq 1$ a (fixed) constant

alternating k-star hypothesis

For this presentation I assume that $\lambda = 2$

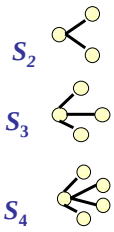
Then we obtain a single ***star*** parameter (σ_2) with statistic:

$S^{[\lambda]}(\mathbf{y}) = \sum_k (-1)^k S_k(\mathbf{y})/\lambda^{k-2}$ ***alternating k-star statistic***

Note that if $\lambda = 1$ and the edge parameter is included, the no. of isolated nodes is modeled separately

1. Parameters for degree sequences: a. alternating *k*-star parameters

$$z(\mathbf{y}) = S_2 - \frac{S_3}{\lambda} + \frac{S_4}{\lambda^2} - \ldots + (-1)^{n-2}\frac{S_{n-1}}{\lambda^{n-3}}$$



Interpretation:

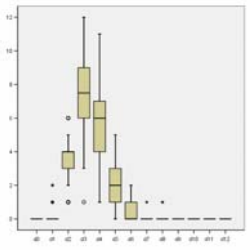
Positive parameter indicates centralization through a small number of high degree nodes

core-periphery based on popularity

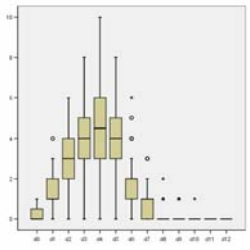
More dispersed degree distribution

Negative parameter: “truncated” (less dispersed) degree distribution; nodes tend not to have particularly high degrees.

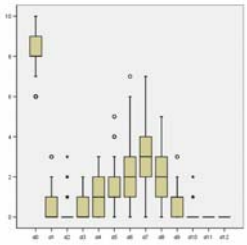
Simulated degree distributions:
20 nodes: fixed density = 0.2



Alt.kstar=-3



Alt.kstar=0



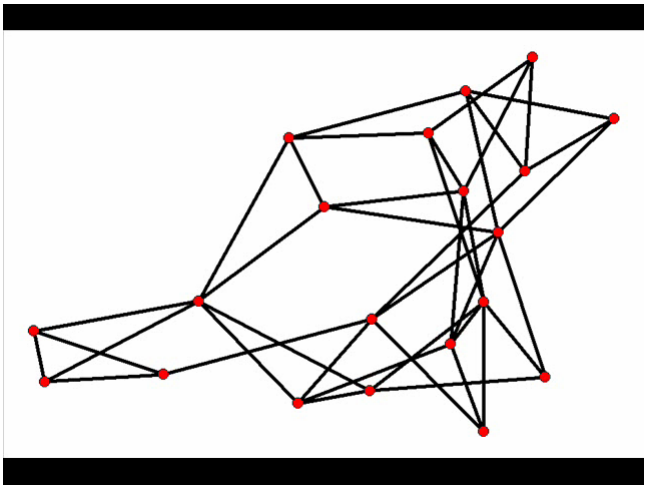
Alt.kstar=+3

Higher order models: **increasingly dispersed degree distribution**

20 nodes: Fix the density at 0.20

all parameters = 0
EXCEPT
Vary *alternating k-star* parameter from – 3.0 to +3.0
in steps of +0.5

The movie shows one representative graph from each simulated distribution



1. Parameters for degree sequences:
b. Geometrically weighted degree distributions

An equivalent characterisation:

Consider statistics $d_k(\mathbf{y})$, where $d_k(\mathbf{y})$ is the number of nodes in $\mathbf{y}$ of degree k (with corresponding parameters θ_k)

Assuming that $\theta_k = e^{-\alpha k}\gamma$ for $k = 1,2,\ldots, n-1$

yields the statistic:

$$D^{[\alpha]}(\mathbf{y}) = \sum_k e^{-\alpha k} d_k(\mathbf{y})$$ ***weighted degree distribution***

Relationship with alternating k-star statistic:

$S^{[\lambda]}(\mathbf{y}) = \lambda^2 D^{[\alpha]}(\mathbf{y}) + 2\lambda L(\mathbf{y}) - n\,\lambda^2$ See Hunter (2006) $\lambda = e^{\alpha}/(e^{\alpha}-1)$

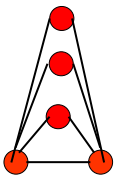
2. Parameters for higher order triangulation

Cohesive subsets of nodes

Require realization dependent neighbourhoods:

Alternating *k*-triangles:

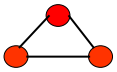
2-triangle ***(T_1)***
(T_3)

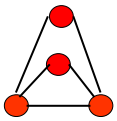


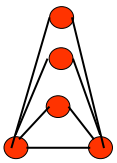
2. Parameters for higher order triangulation

Alternating *k*-triangles:

$$u(\mathbf{y}) = T_1 - \frac{T_2}{\lambda} + \frac{T_3}{\lambda^2} - \ldots + (-1)^{n-2}\frac{T_{n-2}}{\lambda^{n-3}}$$





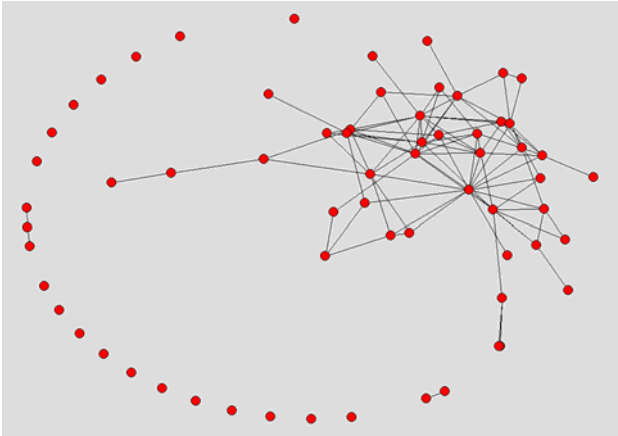


etc…

Interpretation:

a. Positive parameter suggests triangles tend to "clump" together in denser regions of the network (cohesive subsets).

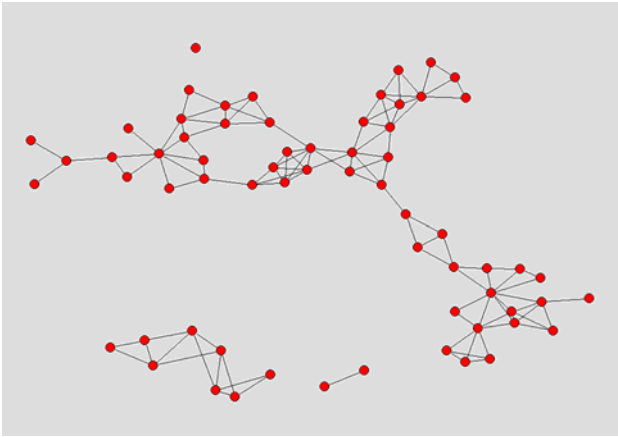
b. Models the ***edgewise shared partner distribution***:
For each pair of tied nodes, how many partners do they share? (Hunter, 2006)



Interpretation: With only positive *k*-triangle effect, there is a core-periphery type structure based on triangulation

Higher order model
Positive *k*-triangle parameter:

Parameters:
Edge = -4.5
Alt. k-triangle=1.3
65 nodes



Interpretation: With positive *k*-triangle effect, and negative *k*-star, various regions of greater density distributed across the network.

Higher order model
Positive *k*-triangle parameter & negative *k*-star parameter:

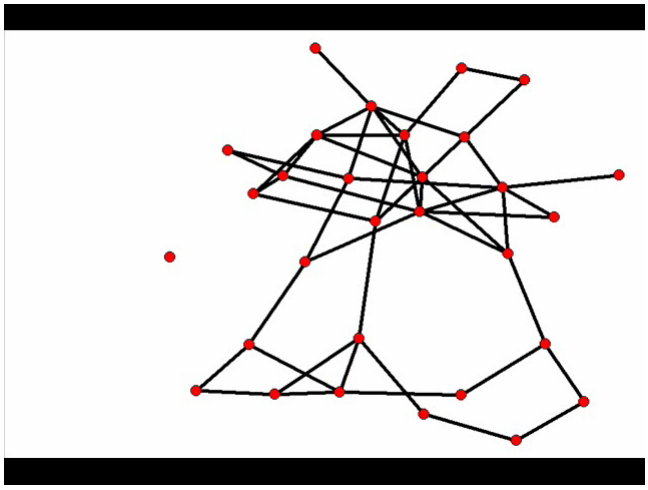
Parameters:
Edge = -0.5, alt. k-star = -1.5, alt. k-triangle = 2.0
65 nodes

Higher order models: **increasing triangulation**

30 nodes: Fix the density at 0.10

all parameters = 0
EXCEPT
Vary *alternating k-triangle* parameter from 0.0 to +3.0
in steps of +0.25

The movie shows one representative graph from each simulated distribution



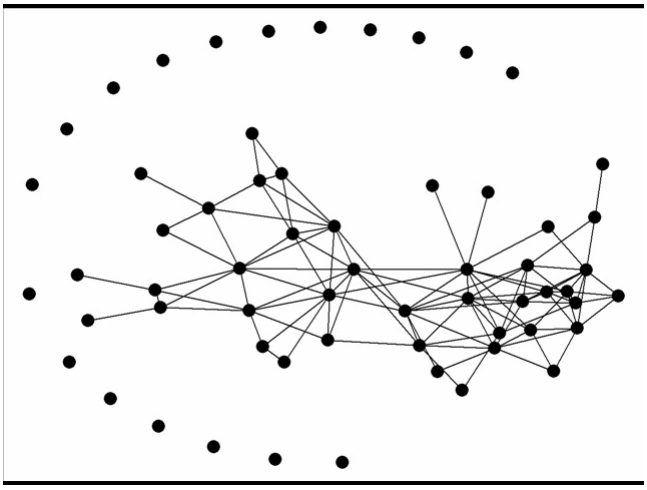
Higher order models:
From centralization to segmentation

60 nodes: Fix the density at 0.05 (i.e. 88 or 89 edges)

alternating k-triangle parameter = +2.0

Vary *alternating k-star* parameter from 0.0 to – 1.0
in steps of – 0.1

The movie shows one representative graph from each simulated distribution

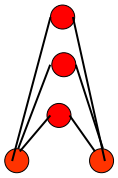


3. Parameters for multiple connectivity

Alternating independent 2-paths:

2-path (2-indpt. U_2) paths (U_3)

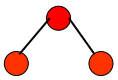
$$z(\mathbf{x}) = U_1 - \frac{2U_2}{\lambda} + \sum_{k=3}^{n-2}\left(\frac{-1}{\lambda}\right)^{k-1} U_k$$

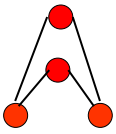


Models the ***dyadwise shared partner* distribution**:

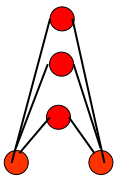
For each pair of nodes, how many partners do they share? (Hunter, 2006)

Why might we want Alternating independent 2-paths?

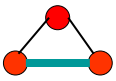


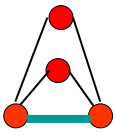


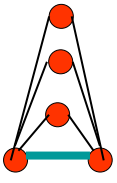
Indpt. 2-paths are lower order to *k*-triangles.



Helps assess whether a clustering (*k*-triangle) effect relates to the formation of the base of the k-triangle





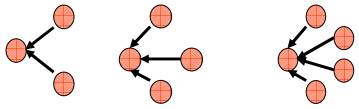


Directed networks: Alternating *k*-star parameters

Snijders, Pattison, Robins & Handcock (2006)

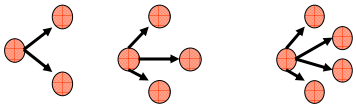
Alternating k-in-stars:

Statistic: Weighted summation of *k*-in-stars analogous to the non-directed model



Alternating k-out-stars:

Statistic: Weighted summation of *k*-out-stars analogous to the non-directed model

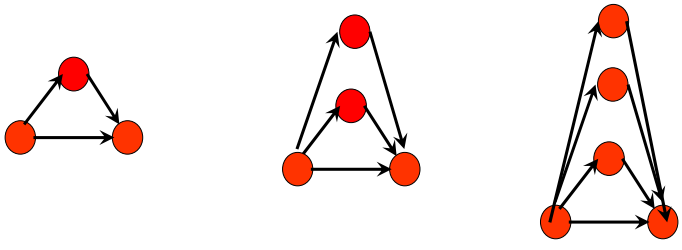


Directed network parameters

Snijders, Pattison, Robins & Handcock (2006)

Alternating directed k-triangles:

Statistic: weighted sum of directed *k*-triangles with 2-paths as sides of the *k*-triangle

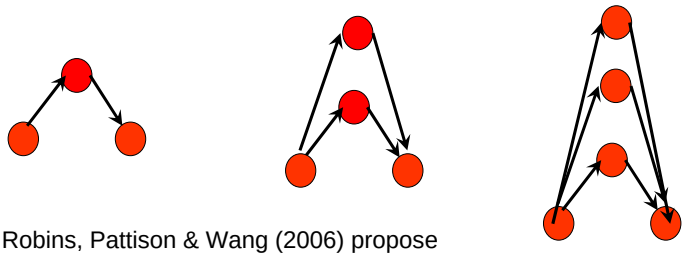


Directed network parameters

Snijders, Pattison, Robins & Handcock (2006)

Alternating directed k-2paths:

Statistic: weighted sum of *k-2paths*



NB: Robins, Pattison & Wang (2006) propose additional variations for directed network models

Classes of exponential random graph models

(Single binary networks without node attributes)

	Non-directed networks	Directed networks
Dyadic independence	Edges *(Bernoulli graphs; simple random graphs)*	Edges; Mutuality *(Also p1 models)*
Markov random graphs	Edges, stars, triangles	Edges, Mutuality, in-,out-,mixed-stars,transitive and cyclic triads
Higher order dependence	*The above plus:* Alternating k-stars, k-triangles, alt. 2-paths	*The above plus:* Alternating k-stars, k-triangles, alt. 2-paths

Fitting Models:
20 network data sets from UCINET5

(Borgatti, Everett & Freeman, 1999)

Non directed networks:
Kapferer mine: kapfmm, kapfmu (16 nodes)
Kapferer tailor shop: kapfts1, kapfts2 (39 nodes)
Padgett Florentine families: padgb, padgm (16 nodes)
Read Highland tribes: gamapos (16 nodes)
Zachary karate club: Zache (34 nodes)
Bank wiring room: rdpos, rdgam (14 nodes)
Taro exchange: Taro (22 nodes)
Thurman office: Thurm (15 nodes)

Fitting Models:
20 network data sets from UCINET5

(Borgatti, Everett & Freeman, 1999)

Directed networks:
Kapferer tailor shop: kapfti1, kapfti2 (39 nodes)
Wolf primates: wolfk (20 nodes)
Krackhardt hi-tech managers: friend, advice (21 nodes)
Bank wiring room: rdhlp (14 nodes)
Knoke bureaucracies: knokm, knoki (10 nodes)

Non directed networks

Data set	Markov
Kapfmm	OK
Kapfmu	OK
Kapfts1	Does not converge
Kapfts2	Does not converge
Padgm	OK
Padgb	OK
Gamapos	Does not converge
Zache	Does not converge
Rdpos	OK
Rdgam	OK
Taro	Does not converge
Thurm	OK
TOTAL	7/12

Directed networks

Data set	Markov
Kapfti1	Does not converge
Kapfti2	Does not converge
Wolfk	Does not converge
Krackhardt friend	OK
Krackhardt advice	OK
Rdhlp	OK
Knokm	OK
Knoki	OK
TOTAL (directed)	5/8
TOTAL (nondir.)	7/12
TOTAL	12/20

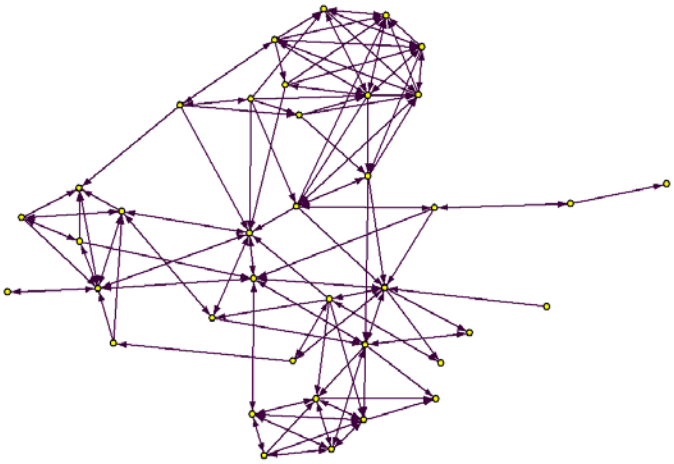
Non directed networks

Data set	Markov	Higher order
Kapfmm	OK	OK
Kapfmu	OK	OK
Kapfts1	Does not converge	OK
Kapfts2	Does not converge	OK
Padgm	OK	OK
Padgb	OK	OK
Gamapos	Does not converge	OK
Zache	Does not converge	OK
Rdpos	OK	OK
Rdgam	OK	OK
Taro	Does not converge	OK
Thurm	OK	OK
TOTAL	7/12	12/12

Directed networks

Data set	Markov	Higher order
Kapfti1	Does not converge	OK
Kapfti2	Does not converge	OK
Wolfk	Does not converge	OK
Krackhardt friend	OK	OK
Krackhardt advice	OK	OK
Rdhlp	OK	OK
Knokm	OK	OK
Knoki	OK	OK
TOTAL (directed)	5/8	
TOTAL (nondir.)	7/12	
TOTAL	12/20	20/20

Example: After hours socialising network



After hours network

Parameter	Estimate	Standard error	Convergence statistic
Arc	– 1.27	0.63	0.06 *
Reciprocity	2.42	0.34	0.06 *
k instar (2)	– 0.86	0.32	0.06 *
k outstar (2)	– 0.96	0.33	0.06 *
Alt. *k*-triangles (2)	1.09	0.14	0.06 *

Goodness of fit: After hours network

Model statistics
Arcs: t = 0.03
Reciprocity: t = 0.03
k instar (2) : t = 0.03
k outstar (2) : t = 0.02
Alt. *k*-triangles (2): t=0.03

Other Markov statistics
2-in-stars: t = - 0.14
3-in-stars : t = - 0.37
2-out-stars: t = - 0.29
3-out-stars : t = -0.53
2-paths: t= - 0.34
Transitive triads: t= - 0.07
Cyclic triads: t= - 0.37

Other higher order statistics
Alt. *2*-paths (2): t= - 0.37

Goodness of fit: After hours network

Degree distributions
standard deviations
indegrees: t = - 0.10
outdegrees : t = - 0.56
skew
indegrees: t = - 1.05
outdegrees : t = - 1.96

Clustering coefficients
Proportion of 2-stars in transitive triads:
2-in-stars: t= 0.40
2-out-stars: t= 1.18
2-paths: t= 1.57
Proportion of 2-paths
in cyclic triads: t=-0.13

Geodesic distribution:
None of the quartiles of the geodesic distribution for the observed graph are extreme in the distribution

Goodness of fit:
After hours network

Triad census

300: t = -0.5036

210: t = -0.1364

120C: t = -0.3707

120D: t = 1.4142

120U: t = 1.0437

201: t = -0.4520

111D: t = -0.1218

111U: t = -0.5027

030T: t = 2.2584

030C: t = -0.3287

102: t = 0.4598

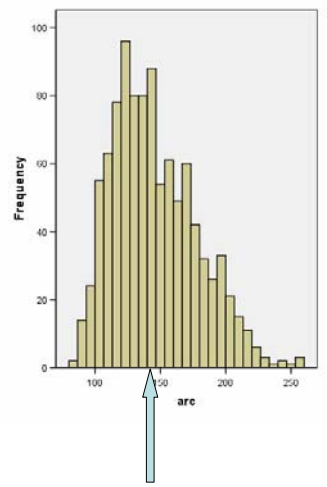
021D: t = -0.4971

021C: t = -0.9884

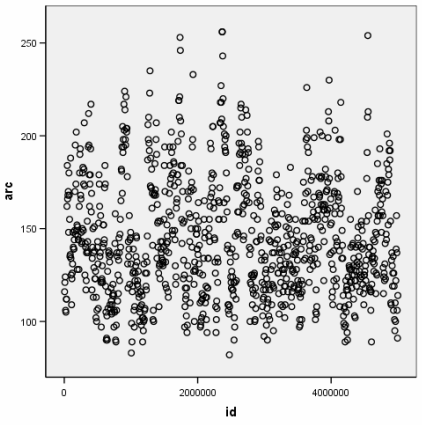
021U: t = 0.1030

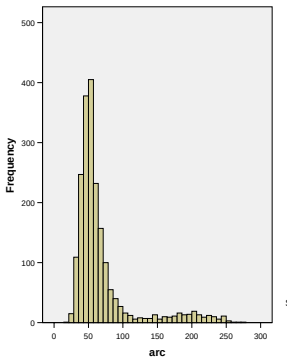
012: t = 0.3200

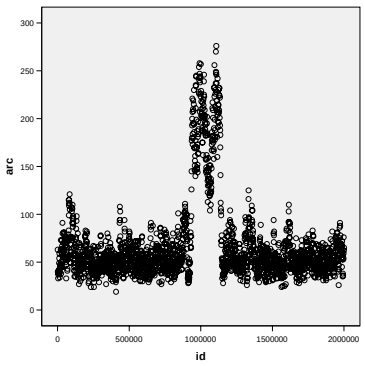
003: t = -0.1781



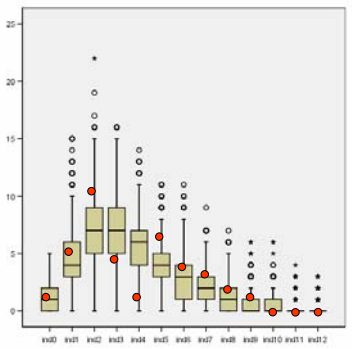
Observed graph



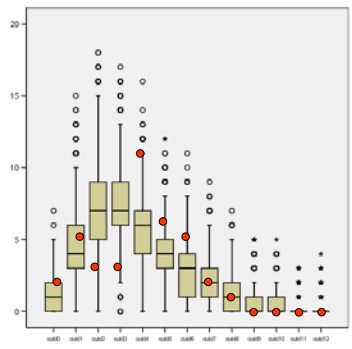




What we DON'T want to see:
the model exhibits two regions!
This would indicate a bad model.



indegree distribution

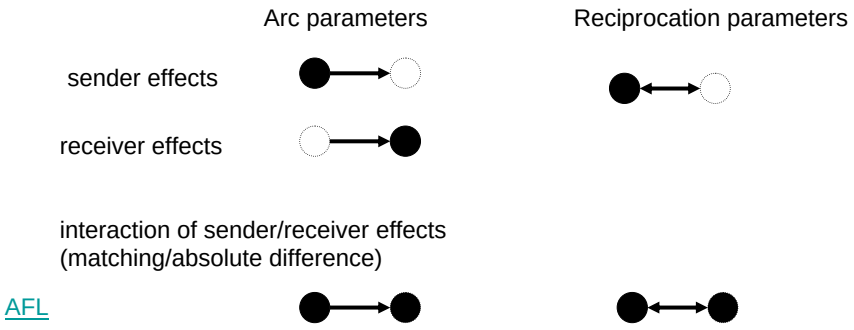


outdegree distribution

Models with nodal attributes: Social selection models

There are various ways to introduce actor attributes (binary or continuous) Robins, Elliott & Pattison (2001)

e.g. dyad level effects



Concluding remarks

- The new specifications are a dramatic improvement over the previous models. For many data sets we now have models that are quite convincing in reproducing major features of the network.
  - Larger networks are more difficult to fit; Directed networks are more difficult to fit.
- This is NOT to say all problems are solved
  - Degeneracy may still be a problem for these models applied to certain data sets (indicates the model specification is not right for that data.)
- So ongoing work will be required on model specification, BUT
  - Partial conditional dependence models give us a substantial way forward.
  - Combining related parameters into the one function through weighted constraints is parsimonious and helps with model convergence
- MCMCMLE methods of parameter estimation, and model simulation techniques are a crucial part of these recent developments