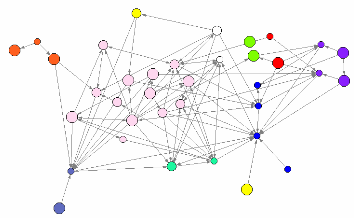
Exponential random graph (p^*) models for social networks

NetSci06



Garry Robins,
University of Melbourne, Australia
http://www.sna.unimelb.edu.au/

With acknowledgements to
Pip Pattison, Tom Snijders, Stanley Wasserman, Mark Handcock

Overview

A. Introduction
- Why statistical models of social networks?
- The conceptual basis for exponential random graph models

B. Bernoulli, Dyadic independence and Markov models
- Simulation and estimation

C. Goodness of fit

D. Degeneracy

E. New specifications and Goodness of fit
- Degree sequences
- Higher order triangulation

F. Social selection models: incorporating attributes

Overview

A. Introduction

- Why statistical models of social networks?
- The conceptual basis for exponential random graph models

B. Bernoulli, Dyadic independence and Markov models

- Simulation and estimation

C. Goodness of fit

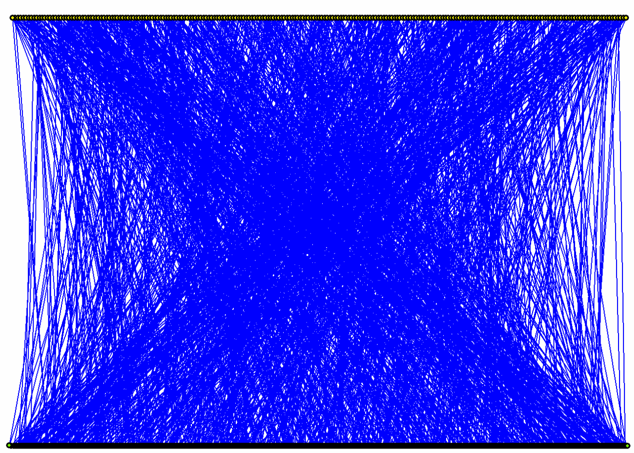
D. Degeneracy

E. New specifications

- Degree sequences
- Higher order triangulation

F. Social selection models: incorporating attributes

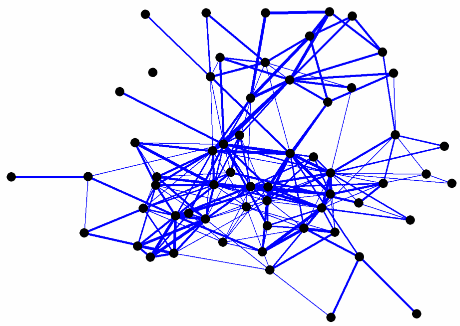
Why statistical models for social networks?



US companies/directors 1996

Why statistical models for social networks?

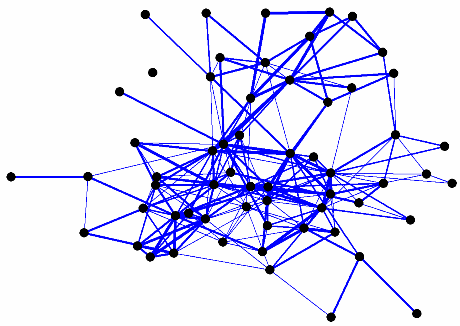
There are many descriptive techniques that may provide insight into an observed social network:
Density, Centrality, Cohesive subsets, etc



Frequency of work interactions

Why statistical models for social networks?

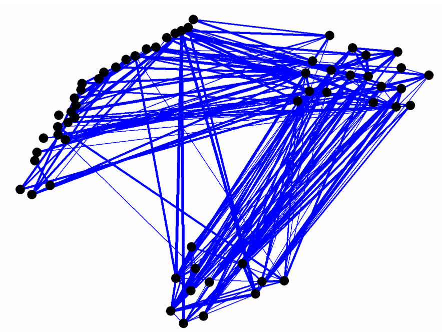
But models may provide some clarity of major effects



Frequency of work interactions

Why statistical models for social networks?

But models may provide some clarity of major effects



Frequency of work interactions

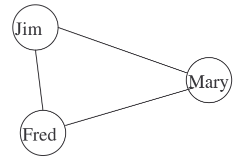
Why statistical models for social networks?

Competing explanations for structural effects.
For instance, the appearance of triangles can occur through different social processes:

1. a *balance-type* effect (triadic closure)
2. an *homophily* effect (people with similar attributes attract).

Only a model including both effects can help decide whether one or both are important.

Friendly Fred introduces Mary to Jim – a balance- or transitivity-type effect.



The outcome is a triangle.

Why statistical models for social networks?

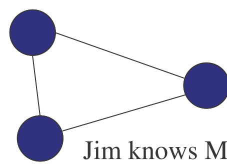
Competing explanations for structural effects.
For instance, the appearance of triangles can occur through different social processes:

1. a *balance-type* effect (triadic closure)
2. an *homophily* effect (people with similar attributes attract).

Only a model including both effects can help decide whether one or both are important.

The outcome is a triangle.

Fred knows Mary because they support the same football team.



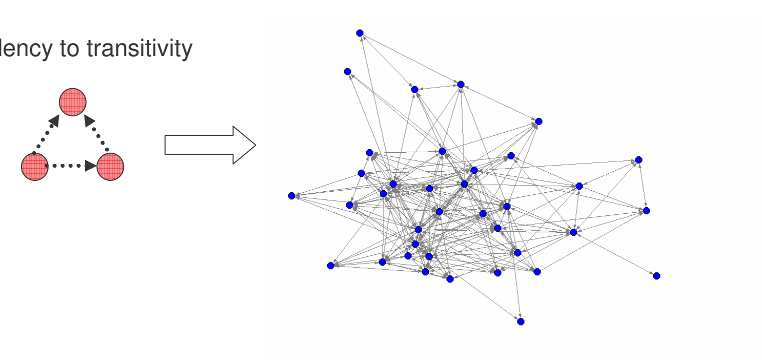
Fred knows Jim because they support the same football team.

Jim knows Mary because they support the same football team.

Why statistical models for social networks?

Models may help traverse the micro-macro divide:
What is the range of global outcomes likely from given local network processes?

Tendency to transitivity

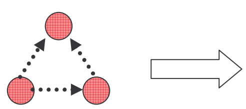


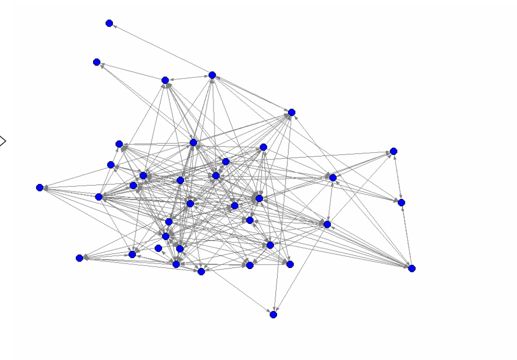
Why statistical models for social networks?

Models may help traverse the micro-macro divide:

What is the range of global outcomes likely from given local network processes?

Tendency to transitivity



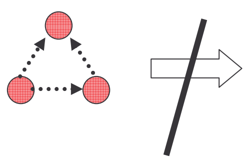


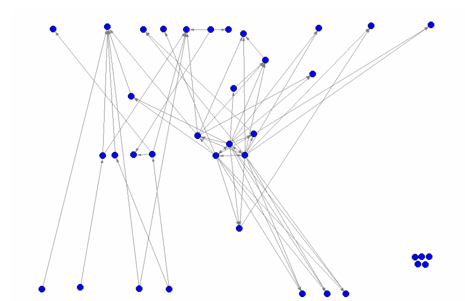
Why statistical models for social networks?

Models may help traverse the micro-macro divide:

What is the range of global outcomes likely from given local network processes?

Tendency to transitivity





For an observed network, which local network processes may be implicated in its emergence?

Why statistical models for social networks?

Models may help traverse the micro-macro divide:
What is the range of global outcomes likely from given local network processes?

And so models (especially statistical models) may help us understand the range of possible outcomes for *processes on networks*

CS

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)*

Overview

A. Introduction

- Why statistical models of social networks?
- The conceptual basis for exponential random graph models

B. Bernoulli, Dyadic independence and Markov models

- Simulation and estimation

C. Goodness of fit

D. Degeneracy

E. New specifications

- Degree sequences
- Higher order triangulation
- F. Social selection models: incorporating attributes

What do we know about processes underlying network tie formation?

Exogenous effects:

Homophily, shared affiliations, spatial propinquity all matter (e.g., McPherson, Smith-Lovin & Cook, 2001)

More generally: network tie formation is often set within ***foci*** (Feld, 1981) or ***settings*** (White, 1995; Pattison & Robins, 2002)

Endogenous network effects:

Clustering: tie formation is often more likely when actors have network partners in common (e.g., Cartwright & Harary, 1956; Granovetter, 1973; and many others)

More generally: "a social tie … is subject to, and known to be subject to, the hegemonic pressures of others engaged in the social construction of that network." (White, 1998)

Interactions between exogenous and endogenous effects:

General social selection (Robins, Elliott & Pattison, 2001)

Modeling endogenous network processes

Guiding principles:

Network ties are the outcome of (unobserved) social processes that tend to be local and interactive

There are both regularities and irregularities in these local interactive processes

We hence construct statistical models in which:

local interactivity is permitted and assumptions about form of "local interactions" are explicit

regularities are represented by model parameters and estimated from data

consequences of local regularities for global network properties can be understood *and can also provide an exacting approach to model evaluation*

Local interactivity

We model ***tie variables***: $\mathbf{X} = [X_{ij}]$ $\quad X_{ij} = 1$ if i has a tie to j

0 otherwise

realization of $\mathbf{X}$ is denoted by $\mathbf{x} = [x_{ij}]$

We also incorporate node-level exogenous ***attribute variables***: $\mathbf{Y} = [Y_i]$

Two modeling steps:

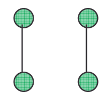
Local interactions: define two network tie variables to be ***neighbors*** if they are conditionally dependent, given the values of all other tie variables

But: what are appropriate ***neighbor*** assumptions?

Dependency assumptions

Network dependencies:
which tie variables are neighbors?
Simple dyadic dependence hypotheses

(1) There are no neighborhood relations ***Bernoulli random graphs***
All edges are independent

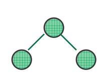


Two tie variables are ***neighbors***:
(2) if they are in the same dyad ***dyad independence***
Directed graphs

Network dependencies:
which tie variables are neighbors?
More realistic dependence hypotheses

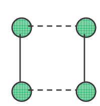
Two tie variables are ***neighbors*** if:

(3) they share an actor



Markov model
(Frank & Strauss, 1986)

(4) they share connections with two existing ties
(completing a social circuit)

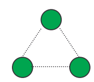


realization-dependent model
(Pattison & Robins, 2002; Snijders, Pattison, Robins & Handcock, 2006)

There are other possibilities, but these two get us a long way

Models for interactive systems of variables (Besag, 1974)

A ***neighborhood*** is a set of mutually neighboring variables and corresponds to a potential ***network configuration***:

e.g. $\{X_{12}, X_{13}, X_{23}\}$ corresponds to 

Hammersley-Clifford theorem (Besag, 1974):

A model for **X** has a form determined by its neighborhoods

This general approach leads to *exponential random graph* or ***p* models***
(Frank & Strauss 1986; extended by Wasserman, Pattison & Robins)

Exponential random graph models

$$\mathbf{P(X = x) = (1/}c\mathbf{)\, \exp\{\textstyle\sum_Q \gamma_Q} z_Q\mathbf{(x)\}}$$

normalizing quantity ***parameter*** ***network statistic***

the summation is over all neighbourhoods Q

$z_Q(\mathbf{x}) = \Pi_{X_{ij} \in Q} x_{ij}$ signifies whether all ties in Q are observed in **x**

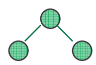
$c = \sum_{\mathbf{x}} \exp\{\sum_Q \gamma_Q z_Q(\mathbf{x})\}$

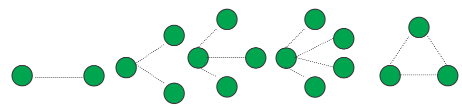
Neighborhoods depend on dependence assumptions

Assumptions: two ties are *neighbors*:

Configurations for neighborhoods

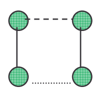
if they share an actor
Markov



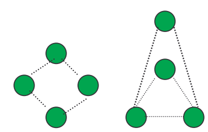


edge 2-star 3-star 4-star …triangle

if they complete a 4-cycle
realization-dependent



\+



…

4-cycle 2-triangle and others

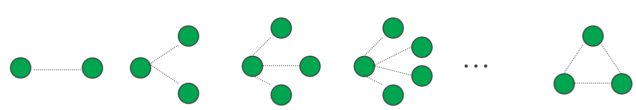
Homogenous models

If we assume that ***isomorphic neighbourhoods have equal parameters***, then:

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

E.g. for a Markov model:

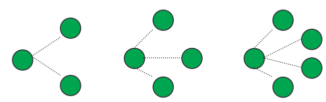
Configurations



Parameters	θ	σ_2	σ_3	σ_4	…	τ
Statistics	$L(\mathbf{x})$	$S_2(\mathbf{x})$	$S_3(\mathbf{x})$	$S_4(\mathbf{x})$	…	$T(\mathbf{x})$

Related model parameters

Star configurations



…

Parameters σ_2 σ_3 σ_4 …

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

alternating k-star hypothesis

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

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

Overview

A. Introduction
- Why statistical models of social networks?
- The conceptual basis for exponential random graph models

B. Bernoulli, Dyadic independence and Markov models
- Simulation and estimation

C. Goodness of fit

D. Degeneracy

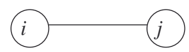
E. New specifications
- Degree sequences
- Higher order triangulation

F. Social selection models: incorporating attributes

Dependence assumptions:
1. Bernoulli (independent edges)

Possible edges are independent of one another.

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



Hammersley-Clifford theorem:

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

Dependence assumptions:
1. Bernoulli (independent edges)

One parameter (λ_{ij}) for each possible edge – simply too many.

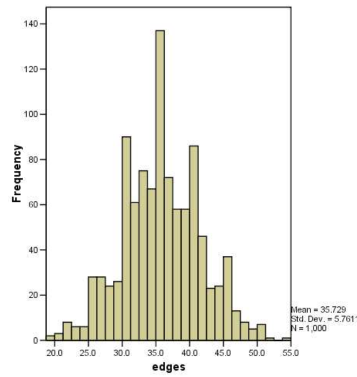
Homogeneity assumption: $\lambda_{ij} = \theta$ for all i,j

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

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

where L is the number of edges in the observed network

Bernoulli graph distribution: distribution of no. of edges



Dependence assumptions: 2. Dyadic independence (directed graphs)

Dyads are independent of one another.

Configurations in this model relate to arcs (x_{ij}) and reciprocated arcs ($x_{ij}\, x_{ji}$)

$$\Pr(\mathbf{X}=\mathbf{x}) = \frac{1}{\kappa}\exp\left(\theta L + \rho M\right)$$

where L is the number of arcs and M the number of reciprocated arcs in the observed network.

More relaxed homogeneity assumptions:Two-blocks

a. Homogeneous Bernoulli:

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

b. Bernoulli blockmodel:

$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp\{\theta_{11} L_{11}(\mathbf{x})+\theta_{12} L_{12}(\mathbf{x})+\theta_{21} L_{21}(\mathbf{x})+\theta_{22} L_{22}(\mathbf{x})\}$

c. Homogeneous Dyad-independent:

$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp\{\theta L(\mathbf{x}) + \rho M(\mathbf{x})\}$

d. Dyad-independent block

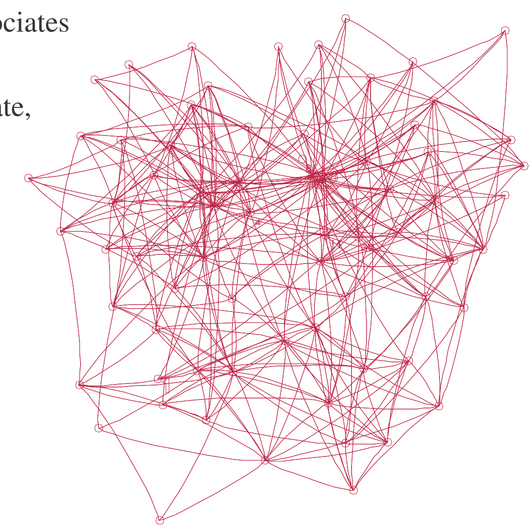
$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp\{\sum_{r,s=1,2} \theta_{rs} L_{rs}(\mathbf{x})+ \rho M(\mathbf{x})\}$

e. Dyad-independent block with block reciprocity parameters:

$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp\{\sum_{r,s=1,2} (\theta_{rs} L_{rs}(\mathbf{x})+ \rho_{rs} M_{rs}(\mathbf{x}))\}$

Collaboration network in a US law firm (Lazega)

36 partners, 35 associates
3 offices
2 practices (corporate, litigation)



Parameter estimates for Dyad-independent block model: Lazega lawyer data

Block 1 = partners; block 2 = associates

d. Dyad-independent block

$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp\{\sum_{r,s=1,2} \theta_{rs} L_{rs}(\mathbf{x}) + \rho M(\mathbf{x}) \}$

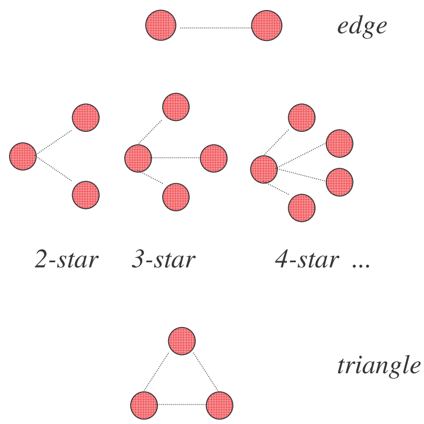
Parameter	*Estimate*
θ_{11}	*-1.34*
θ_{12}	*-3.51*
θ_{21}	*-1.53*
θ_{22}	*-1.98*
ρ	*1.52*

Dependence assumptions: 3. Markov random graphs

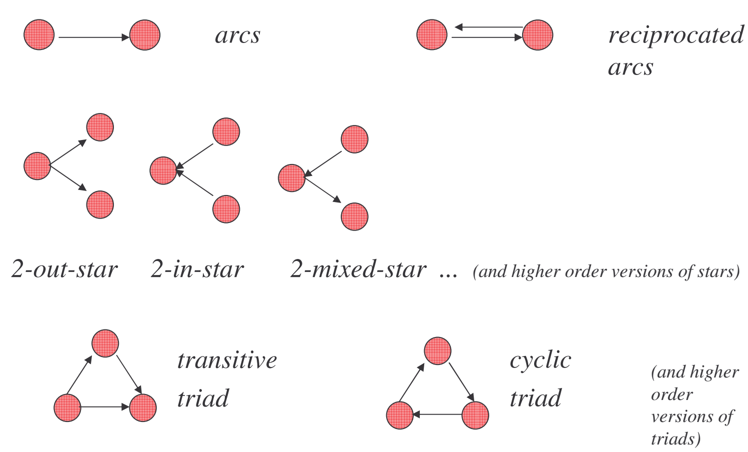
Suppose that edges are conditionally independent if and only if they share a node. (Frank & Strauss, 1986)

Frank and Strauss showed that configurations in this model comprised edges, stars and triangles.

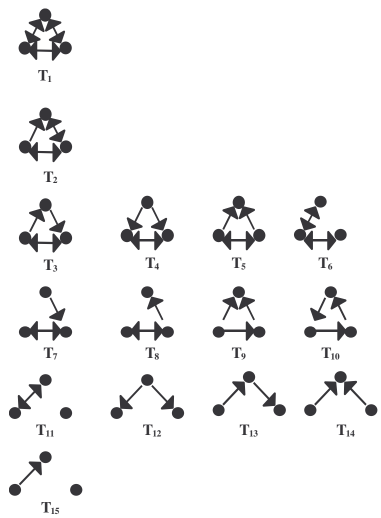
Network configurations for non-directed Markov random graph models



Some network configurations for directed Markov random graph models



All triadic network configurations for directed Markov random graph models



A Markov random graph model:
Nondirected networks

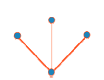
$$\Pr(\mathbf{X} = \mathbf{x}) = (1/\kappa) \exp\{\theta L + \sigma_2 S_2 + \sigma_3 S_3 + \tau T\}$$

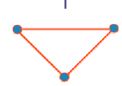
Edge parameter (θ)

– L … number of edges

Star parameters (σ_k)

– Propensities for individuals to have connections with multiple network partners





Triangle parameter (τ)

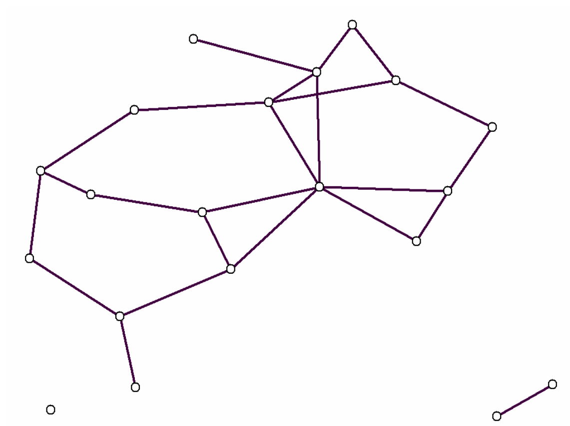
- represents clustering

Examples of graphs from Markov graph distributions

Bernoulli

$\Pr(\boldsymbol{X} = \boldsymbol{x}) = (1/\kappa) \exp\{-2L\}$

Graph statistics:
Edges = 24
2-stars = 51
3-stars = 34
Triangles = 3

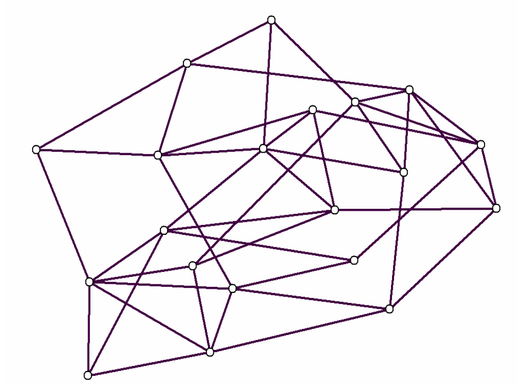


Examples of graphs from Markov graph distributions

Markov: edge/star model

$\Pr(\boldsymbol{X} = \boldsymbol{x}) = (1/\kappa) \exp\{-2L + 0.5 S_2 - 0.3 S_3\}$

Graph statistics:
Edges = 44
2-stars = 158
3-stars = 148
Triangles = 12

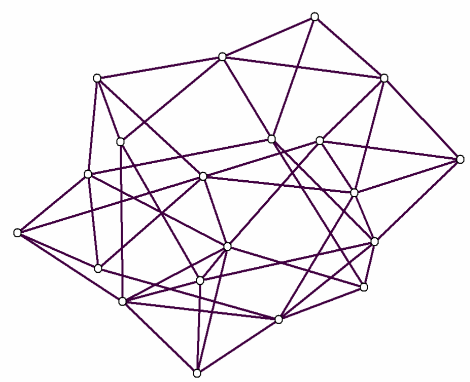


Examples of graphs from Markov graph distributions

Markov with triangles

$\Pr(\boldsymbol{X} = \boldsymbol{x}) = (1/\kappa) \exp\{-2L + 0.5 S_2 - 0.3 S_3 + T\}$

Graph statistics:
Edges = 49
2-stars = 201
3-stars = 224
Triangles = 21



Overview

A. Introduction
- Why statistical models of social networks?
- The conceptual basis for exponential random graph models

B. Bernoulli, Dyadic independence and Markov models
- Simulation and estimation

C. Goodness of fit

D. Degeneracy

E. New specifications
- Degree sequences
- Higher order triangulation

F. Social selection models: incorporating attributes

Simulating Markov random graph distributions

Using the *Metropolis algorithm*:

random starting graph on fixed number of nodes

many iterations (e.g. 1,000,000)

sample every 1000th graph

The algorithm sets up a Markov Chain on the space of all possible nondirected networks that has Pr(**X**=**x**) as its stationary distribution

at each step, we consider changing the value of the (*i*,*j*) tie (from 1 to 0 or 0 to 1) for a randomly selected pair (and relation): $\mathbf{x} \rightarrow \mathbf{x}'_{ij}$

the change is made with probability

$$\min[1,\exp(\{\sum_Q \lambda_Q (z_Q(\mathbf{x}'_{ij})-z_Q(\mathbf{x}))\})]$$

burnin

Estimation of Markov random graph models

1. Pseudo-likelihood estimation: An approximate technique based on logistic regression. Maybe indicates major effects but needs to be used with caution.

2. Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMCMLE): simulation of a distribution of random graphs from a starting set of parameter values, and subsequent refinement of the parameter values by comparing the distribution of graphs against the observed graph, with this process repeated until the parameter estimates stabilize.

estimation

Estimation software (MCMCMLE)

Statnet:

csde.washington.edu/statnet

SIENA:

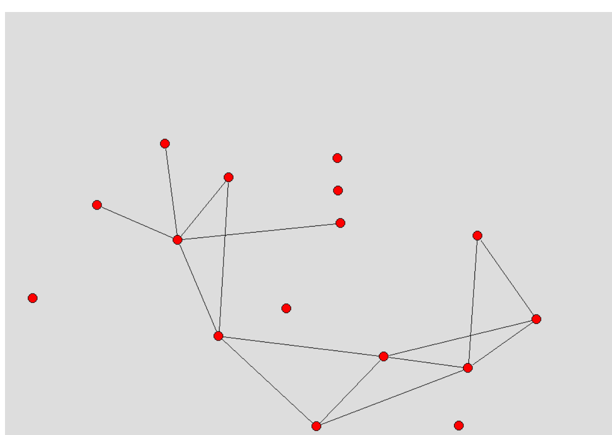
stat.gamma.rug.nl/siena.html

pnet:

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

Example:
Florentine families business network

(Padgett & Ansell, 1993)



Markov model estimates: Florentine families business network

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

Maximum Likelihood estimates

Edge = - 4.27 (1.13)*

2-star = 1.09 (0.65)

3-star = -0.67 (0.41)

Triangle= 1.32 (0.65)*

Overview

A. Introduction
- Why statistical models of social networks?
- The conceptual basis for exponential random graph models

B. Bernoulli, Dyadic independence and Markov models
- Simulation and estimation

C. Goodness of fit

D. Degeneracy

E. New specifications
- Degree sequences
- Higher order triangulation

F. Social selection models: incorporating attributes

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)*

Markov model estimates: Florentine families business network

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

Maximum Likelihood estimates

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

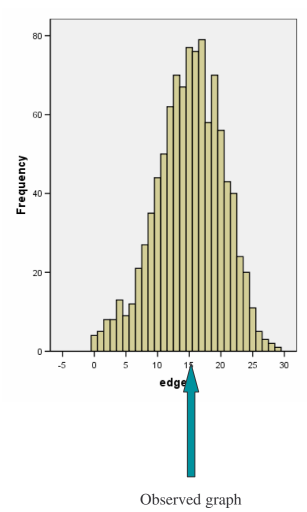
Comparing observed network to a sample from simulation of **Markov model**: *t*-statistics

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

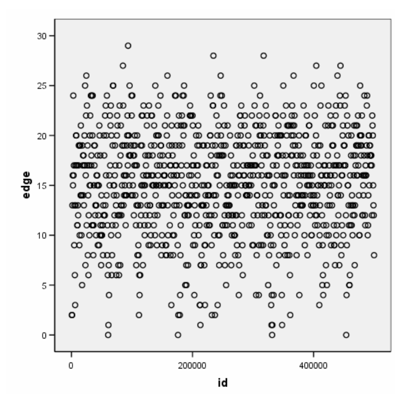
Degree distribution
std dev degree dist: t = 0.51
skew deg dist: t = 0.17

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

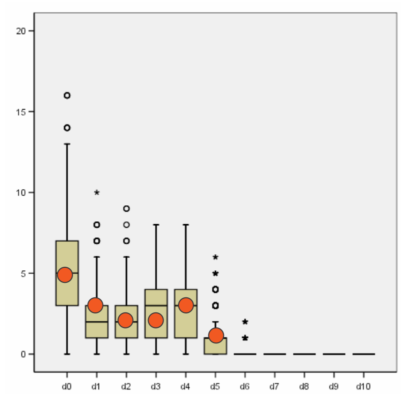
Geodesic distribution:
None of the quartiles of the geodesic distribution for the observed graph are extreme in the distribution



No of edges in simulated graphs

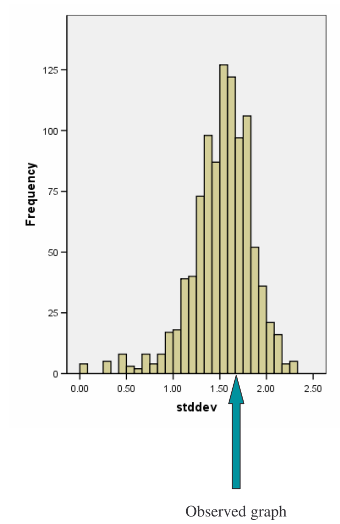


No of edges in simulated graphs against simulation number (*id*)



Degree distributions of simulated graphs

Observed degree distribution



Standard deviation of degree distribution of simulated graphs

Overview

A. Introduction
- Why statistical models of social networks?
- The conceptual basis for exponential random graph models

B. Bernoulli, Dyadic independence and Markov models
- Simulation and estimation

C. Goodness of fit

D. Degeneracy

E. New specifications
- Degree sequences
- Higher order triangulation

F. Social selection models: incorporating attributes

Edge/Triangle model:

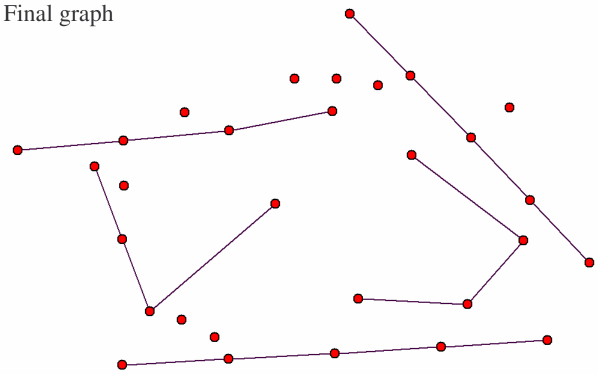
$\theta = -3; \tau = 0.25$

Start from empty graph

Burn-in 50,000

200,000 simulations

Final graph



Edge/Triangle model:

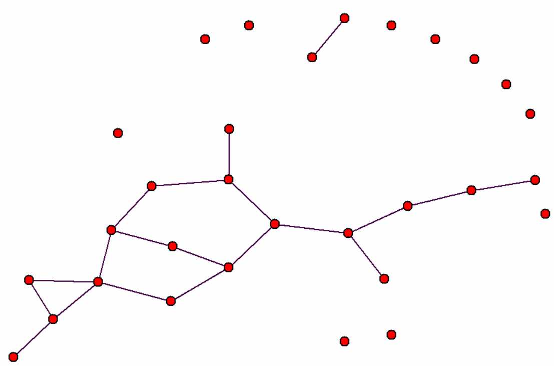
$\theta = -3; \tau = \mathbf{0.75}$

Start from empty graph

Burn-in 50,000

200,000 simulations

Final graph



Edge/Triangle model:

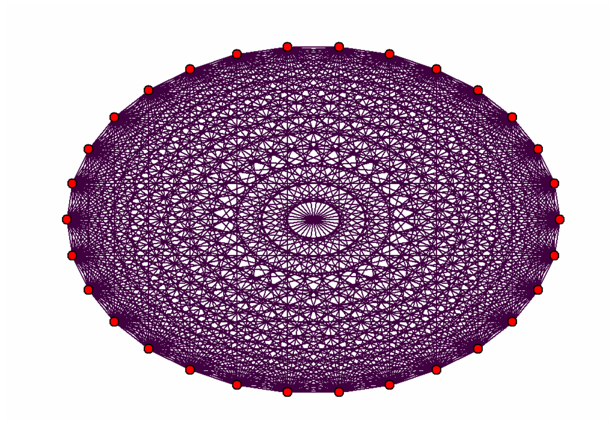
$\theta = -3; \tau = 0.75$

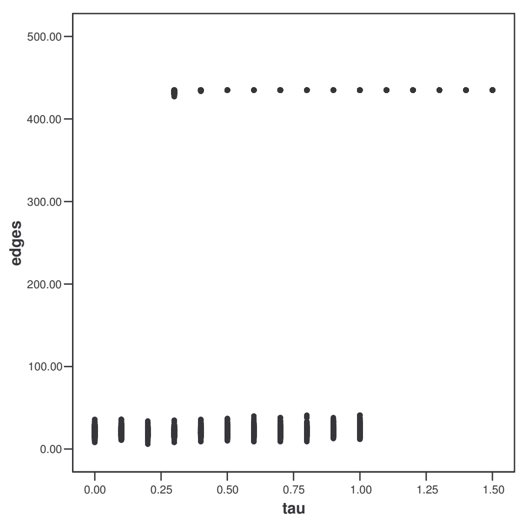
Start from **complete** graph

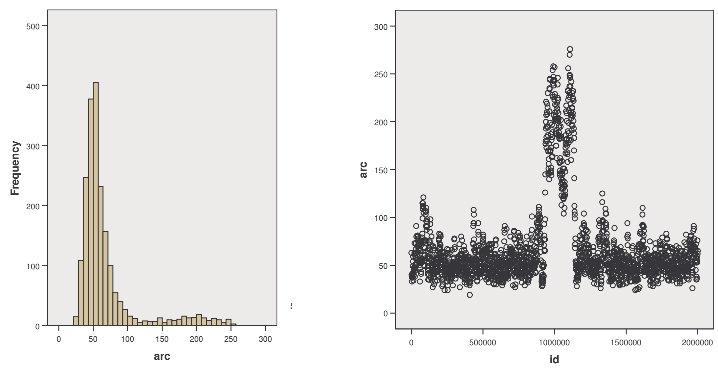
Burn-in 50,000

200,000 simulations

Final graph







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

Model degeneracy

For certain parameter values, a model may imply that only one or two graphs with non-zero probability.

Often such graphs are the empty or full graph

(or a graph of complete disconnected components).

These *near degenerate* models cannot be estimated.

Markov models are often degenerate when clustering is high

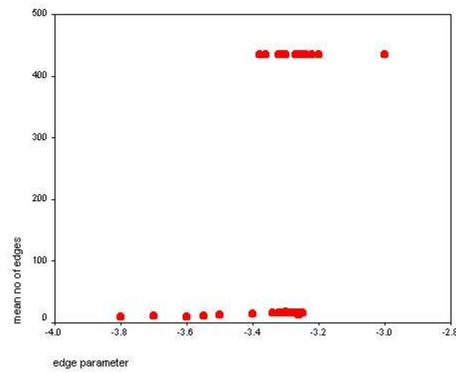
Model degeneracy

Model with only an edge and triangle parameter

$$\Pr(\boldsymbol{X} = \boldsymbol{x}) = (1/c)\exp\{\theta L + \tau T\}$$

Set $\tau = 1.0$

30 node graph



Model degeneracy

Model with edge, triangle parameter and star parameters

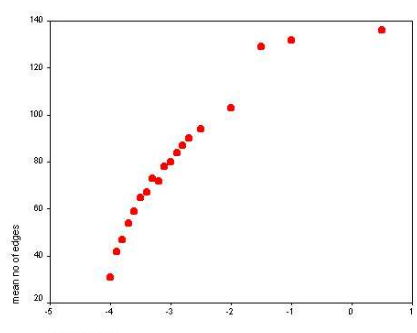
$$\Pr(\boldsymbol{X} = \boldsymbol{x}) = (1/c)\exp\{\theta L + \sigma_2 S_2 + \sigma_3 S_3 + \tau T\}$$

Set $\tau = 1.0$

$\sigma_2 = 0.5$

$\sigma_3 = -0.2$

30 node graph



Model degeneracy

For Markov random graphs, the bottom line is:

- Edge/triangle models are NOT going to fit observed data:
  - They imply graphs that are either close to empty or close to complete (*degenerate)*
- Models with edges, triangles, 2-star and 3-star parameters will do better:
  - They may fit data if there is a negative 3-star parameter estimate
  - But they are also likely to fail when clustering is high.

Model degeneracy

Why do Markov models fail when clustering is high in observed networks?

An attempt at an intuitive explanation:

- Markov models suppose that triangles are spread rather evenly throughout the graph (homogeneous models).
- They have trouble when triangles tend to clump into *cohesive subsets of nodes.* (The dense part of the graph suggests a very strong triangle effect – but applying this effect to the not-so-dense part tends to "fill the graph up".)

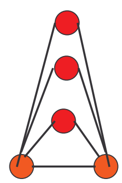
One solution

Go beyond Markov dependence assumptions:

Higher order dependence assumptions producing, most importantly, *alternating k-triangle* statistics

(plus some other higher order statistics.)

3-triangle (T_3)



Snijders, Pattison, Robins & Handcock, 2006;

Overview

A. Introduction
- Why statistical models of social networks?
- The conceptual basis for exponential random graph models

B. Bernoulli, Dyadic independence and Markov models
- Simulation and estimation

C. Goodness of fit

D. Degeneracy

E. New specifications
- Degree sequences
- Higher order triangulation

F. Social selection models: incorporating attributes

Papers on new specifications

Advance copies: 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.

Overview

A. Introduction
- Why statistical models of social networks?
- The conceptual basis for exponential random graph models

B. Bernoulli, Dyadic independence and Markov models
- Simulation and estimation

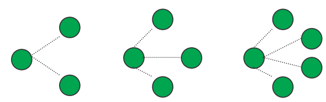
C. Goodness of fit

D. Degeneracy

E. New specifications
- Degree sequences
- Higher order triangulation

F. Social selection models: incorporating attributes

Related model parameters

Star configurations  …

Parameters σ_2 σ_3 σ_4 …

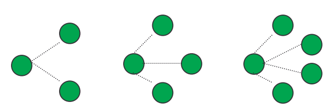
Markov models with star effects only (ignore triangles)

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

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

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

Related model parameters

Star configurations  …

Parameters σ_2 σ_3 σ_4 …

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

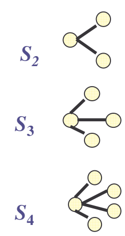
alternating k-star hypothesis

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

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

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

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



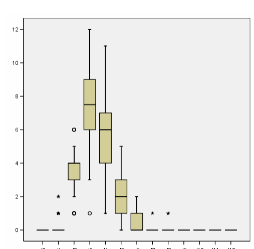
Interpretation:

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

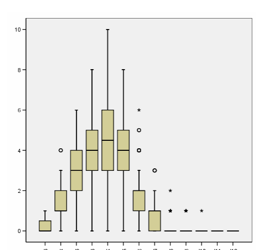
- core-periphery based on popularity
- More dispersed degree distribution

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

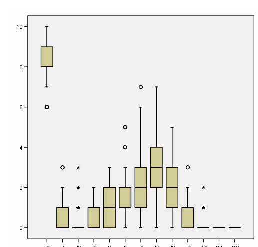
Simulated degree distributions:
20 nodes: fixed density = 0.2



Alt.kstar=-3



Alt.kstar=0



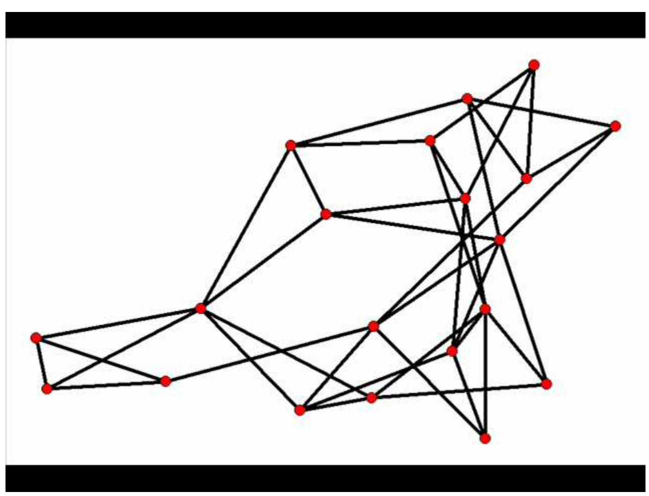
Alt.kstar=+3

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 **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) $\boldsymbol{\lambda = e^{\alpha}/(e^{\alpha}-1)}$

Overview

A. Introduction
- Why statistical models of social networks?
- The conceptual basis for exponential random graph models

B. Bernoulli, Dyadic independence and Markov models
- Simulation and estimation

C. Goodness of fit

D. Degeneracy

E. New specifications
- Degree sequences
- Higher order triangulation

F. Social selection models: incorporating attributes

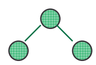
Neighborhoods depend on proximity assumptions

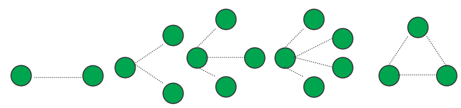
Assumptions: two ties are *neighbors*:

Configurations for neighborhoods

if they share an actor

Markov

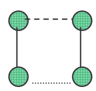


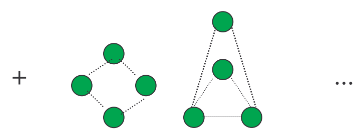


edge *2-star* *3-star* *4-star …triangle*

if they complete a 4-cycle

realization-dependent





4-cycle *2-triangle* *and others*

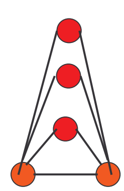
2. Parameters for higher order triangulation

Cohesive subsets of nodes

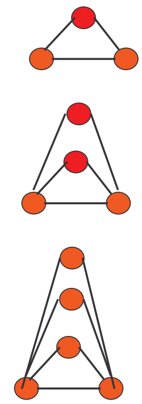
Require realization dependent neighbourhoods:

Alternating k-triangles:

3-triangle (T_3)



2. Parameters for higher order triangulation



Alternating k-triangles:

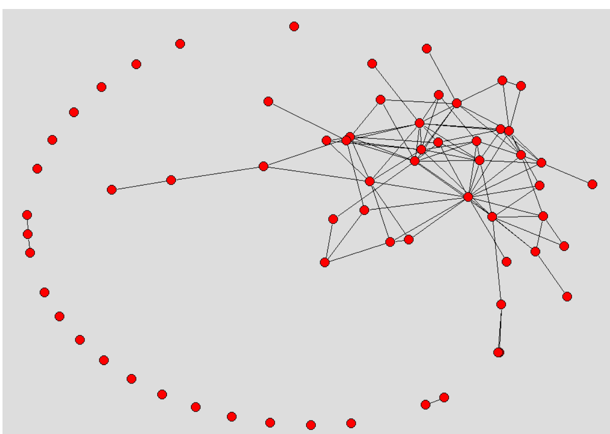
$$u(\mathbf{y}) = T_1 - \frac{T_2}{\lambda} + \frac{T_3}{\lambda^2} - \ldots + (-1)^{n-2}\frac{T_{n-2}}{\lambda^{n-3}}$$

etc…

Interpretation:

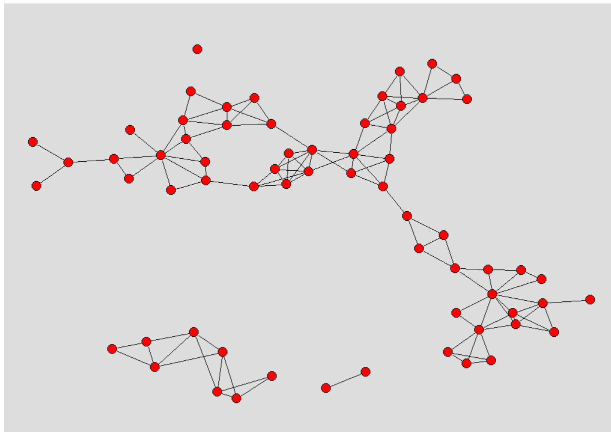
a. Positive parameter suggests triangles tend to "clump" together in denser regions of the network (cohesive subsets).

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:

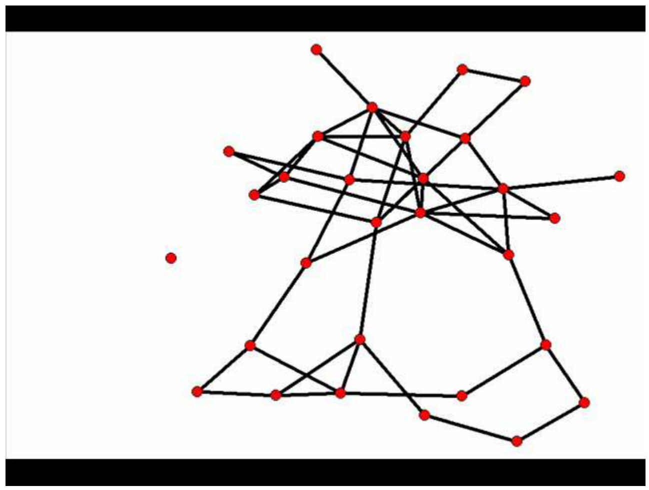
Parameters:
Edge = -0.5, alt. k-star = -1.5, alt. k-triangle = 2.0
65 nodes

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



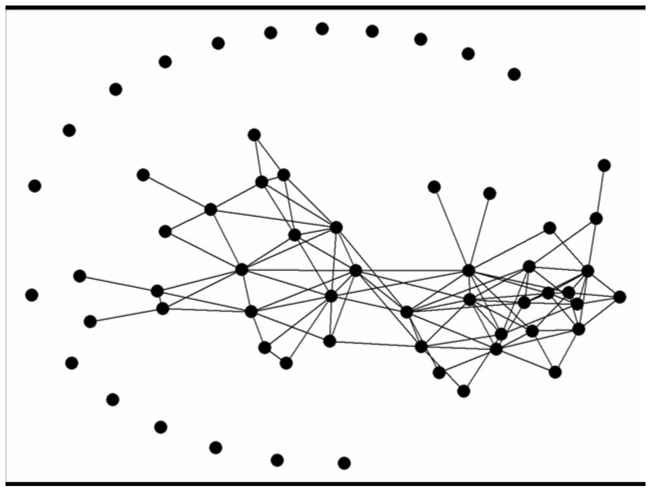
Higher order models:
From centralization to segmentation

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

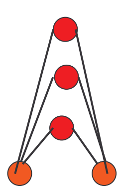


3. Parameters for multiple connectivity

Alternating independent 2-paths:

3-indept.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**:

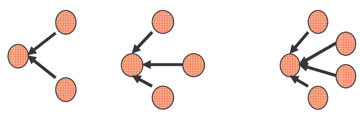
For each pair of nodes, how many partners do they share? (Hunter, 2006)

Directed networks:
Alternating *k*-star parameters

Snijders, Pattison, Robins & Handcock (2006)

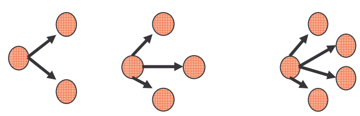
Alternating k-in-stars:

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

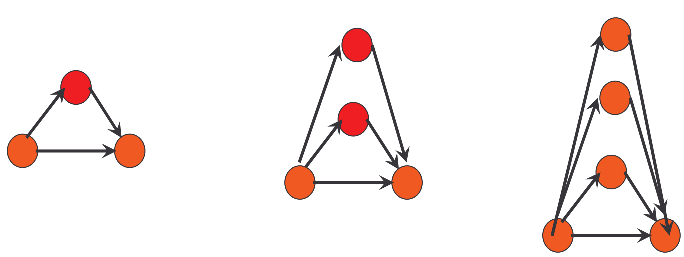


Directed networks:

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

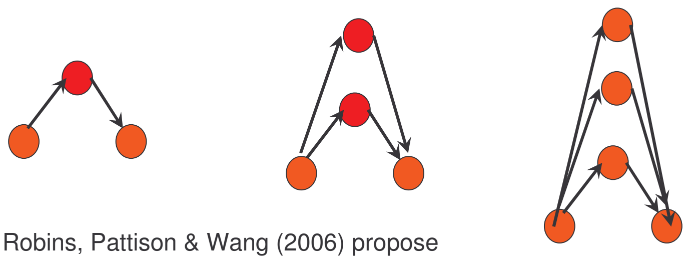


Directed network parameters

Snijders, Pattison, Robins & Handcock (2006)

Alternating directed <u>k-2paths</u>:

Statistic: weighted sum of *k-2paths*



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

Fitting Models:
20 network data sets from UCINET5

(Borgatti, Everett & Freeman, 1999)

Non directed networks:

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

Directed networks:

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

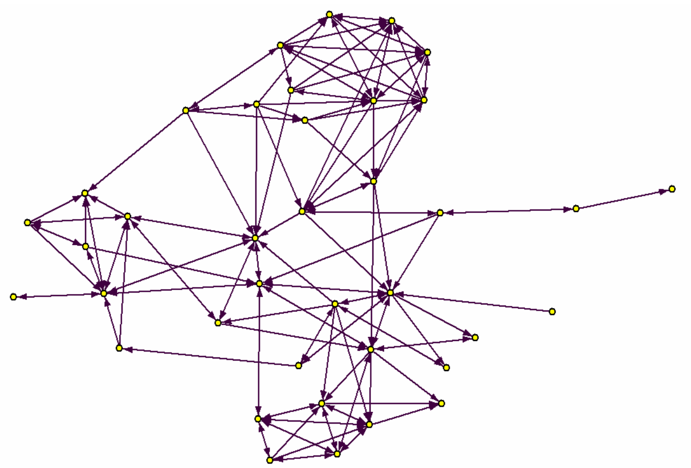
Non directed networks

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	– 0.86*	0.32	0.06
k outstar	– 0.96*	0.33	0.06
Alt. *k*-triangles	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

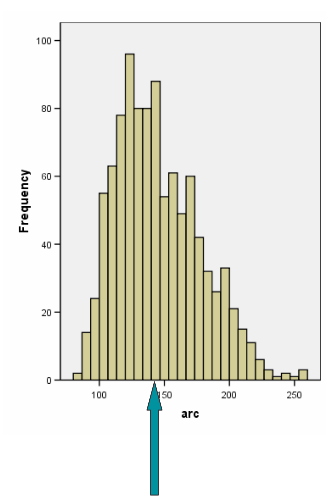
Clustering coefficients
Proportion of 2-stars in transitive triads:
2-in-stars: t= 0.40
2-out-stars: t= 1.18
2-paths: t= 1.57
Proportion of 2-paths in cyclic triads: t=-0.13

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

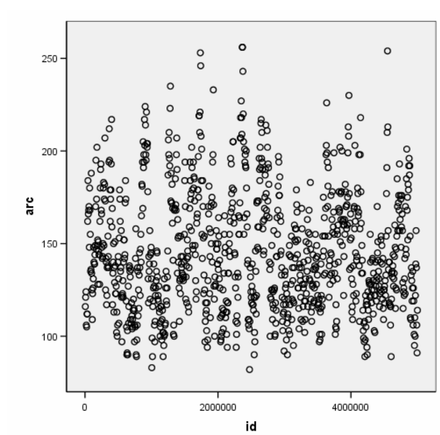
Goodness of fit: After hours network

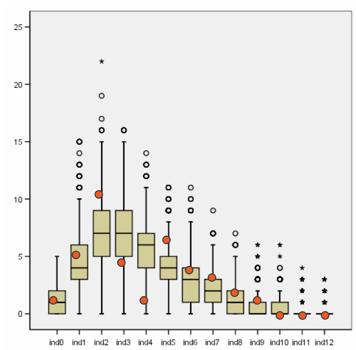
<u>Triad census</u>

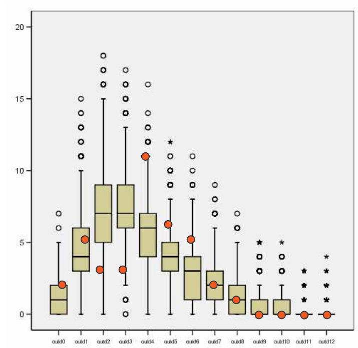
300: t = -0.5036
210: t = -0.1364
120C: t = -0.3707
120D: t = 1.4142
120U: t = 1.0437
201: t = -0.4520
111D: t = -0.1218
111U: t = -0.5027
030T: t = 2.2584
030C: t = -0.3287
102: t = 0.4598
021D: t = -0.4971
021C: t = -0.9884
021U: t = 0.1030
012: t = 0.3200
003: t = -0.1781



Observed graph







indegree distribution

outdegree distribution

Overview

A. Introduction

- Why statistical models of social networks?
- The conceptual basis for exponential random graph models

B. Bernoulli, Dyadic independence and Markov models

- Simulation and estimation

C. Goodness of fit

D. Degeneracy

E. New specifications

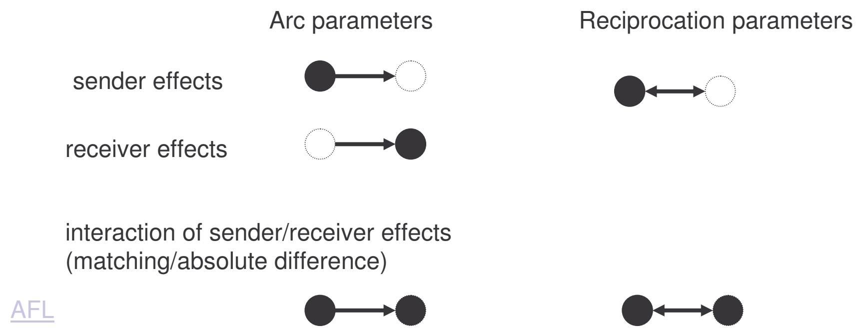
- Degree sequences
- Higher order triangulation

F. Social selection models: incorporating attributes

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 required on model specification, BUT
  - Realization 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

Work in progress

- Directed networks
- Bipartite graphs (affiliation networks)
  - Also tripartite graphs
- Multiple networks
- Longitudinal models

To come:

- Missing data
- Network sampling