# Practical statistical modeling of networks: Recent developments in exponential random graph ($p^*$) models


***Co-chairs***

**Garry Robins, University of Melbourne**

**Martina Morris, University of Washington**

***Presenters***

**Pip Pattison, University of Melbourne**

**Mark S. Handcock, University of Washington**

**Garry Robins, University of Melbourne**

**Dave Hunter, Pennsylvania State University**

**Tom Snijders, University of Groningen**

**Steve Goodreau, University of Washington**



***Sunbelt XXV, Redondo Beach, California, Feb 16-20, 2005***

# *1. Conceptual foundations and some new model specifications*

**Pip Pattison, University of Melbourne**

***or***

**The left hand side of the equation learns to live with the right hand side**

# Three main points

We have learnt that:

1. We need slightly more subtle characterisations of dependencies among network ties than we initially thought

   ***realisation-dependent not simply Markov***

2. It is very helpful to simplify models by making (plausible) assumptions about relationships among related parameters

   ***alternating k-stars/weighted degree distribution***

   ***alternating k-independent 2-paths/weighted (dyad-wise) shared partner distribution***

   ***alternating k-triangles/weighted (edge-wise) shared partner distribution***

3. With these two insights, maximum likelihood estimation approaches work well

   **Siena** ***(Snijders, 2002)***

   **statnet** ***(Handcock, Hunter, Butts, Goodreau & Morris, 2004)***

# What do we think we know about processes underlying network tie formation?

***Exogenous effects***:

Homophily, shared affiliations, spatial propinquity all matter (e.g., Marsden, 1988; Morris 1993; 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)

# Modelling 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***: $\mathbf{X} = [X_{ij}]$ $X_{ij} = 1$ if $i$ has a tie to $j$

0 otherwise

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

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

**Two modelling steps**:

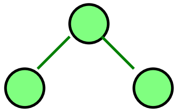
***Local interactions***: define two network tie variables to be ***neighbours*** if they are conditionally dependent, given the values of all other tie variables

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

# Network topologies:
# which tie variables are neighbours?

Two tie variables are ***neighbours*** if:

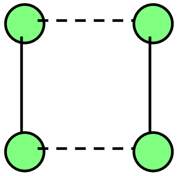
they share an actor



***Markov model***

(Frank & Strauss, 1986)

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



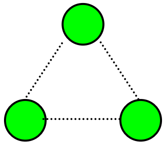
***realisation-dependent model***
(Pattison & Robins, 2002;
Snijders, Pattison, Robins & Handcock, 2004)

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

# Models for interactive systems of variables (Besag, 1974)

A ***neighbourhood*** is a set of mutually neighbouring 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 neighbourhoods

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

# Exponential random graph models

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

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

the summation is over all neighbourhoods *Q*

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

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

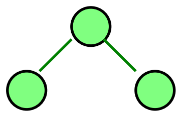
# Neighbourhoods depend on proximity assumptions

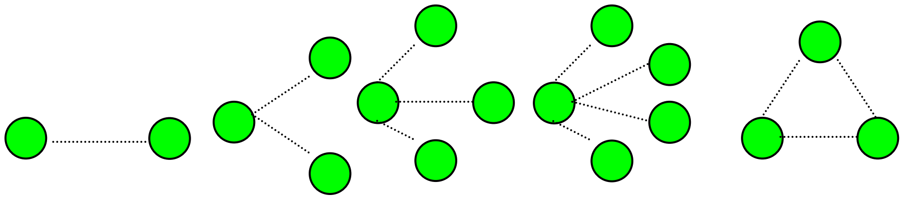
*Assumptions:* two ties are *neighbours*:

*Configurations for neighbourhoods*

if they share an actor

***Markov***

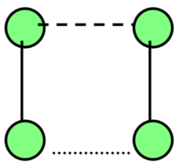




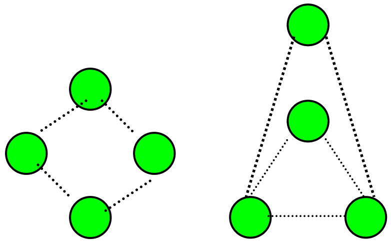
*edge 2-star 3-star 4-star …triangle*

if they complete a 4-cycle

***realisation-dependent***



\+



…

*4-cycle 2-triangle and others*

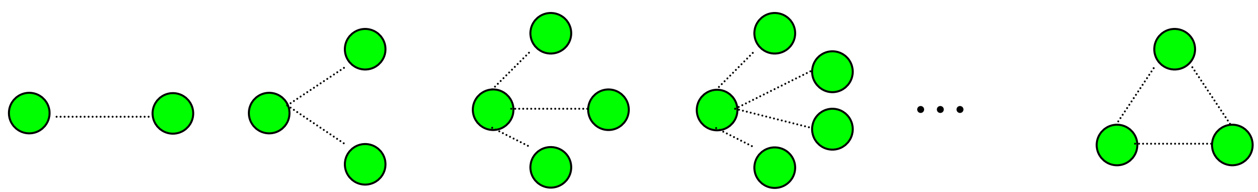
# Homogenous models

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:

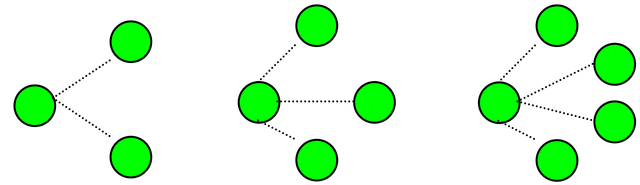
***Configurations***



| | | | | | | |
|---|---|---|---|---|---|---|
| ***Parameters*** | $\theta$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ | … | $\tau$ |
| ***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*** $\sigma_2$ $\sigma_3$ $\sigma_4$ …

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 ($\sigma_2$) with statistic:

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

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

# Geometrically decreasing degree distribution

***An equivalent characterisation***:

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

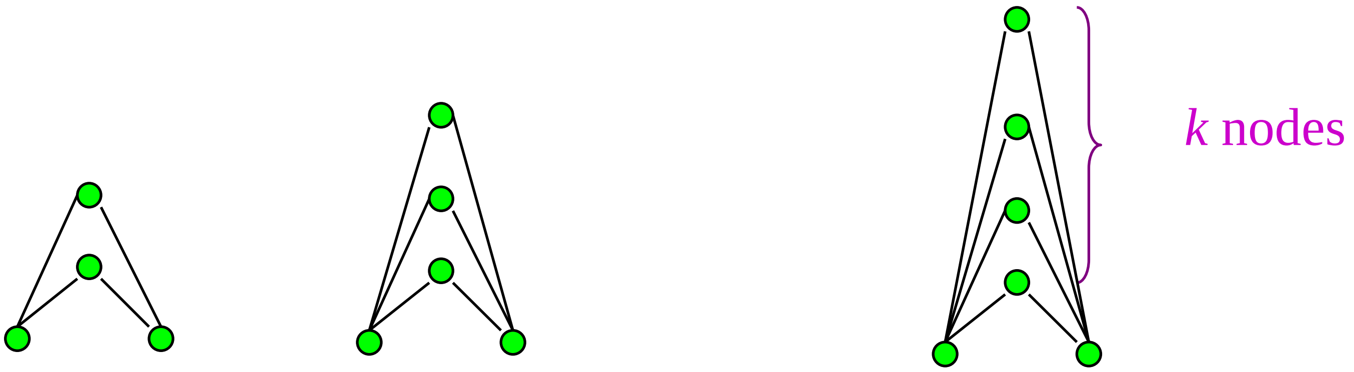
Assuming that

$\theta_k = e^{-\alpha k}\gamma$ for $k = 1,2,\ldots, n-1$ **$\lambda = e^{\alpha}/(e^{\alpha}-1)$**

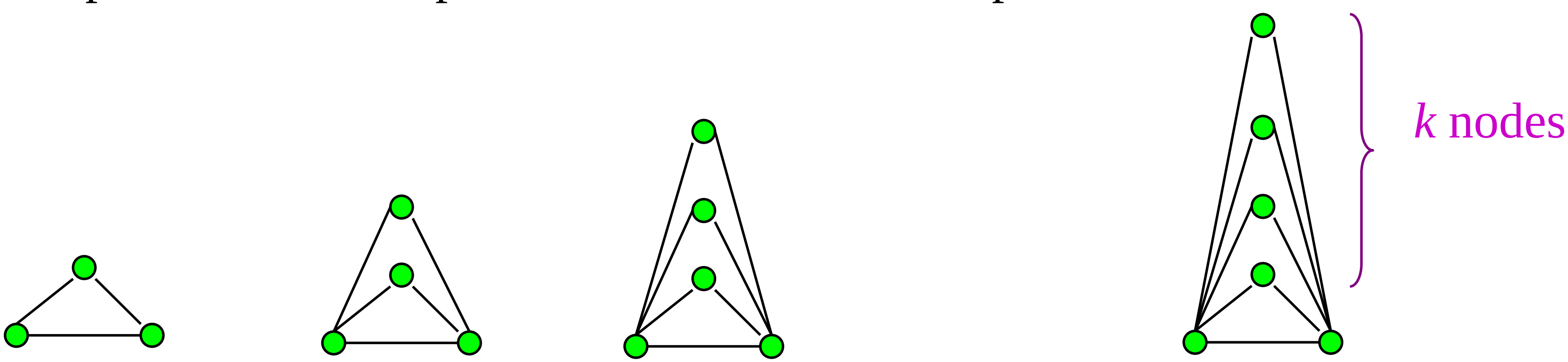
yields the statistic:

$\sum e^{-\alpha k} d_k(\mathbf{x})$ ***weighted degree distribution***

# Additional neighbourhoods for realisation-dependent model




2-independent 2-path | 3-independent 2-path | … | *k*-independent 2-path | …




triangle | 2-triangle | 3-triangle | … | *k*-triangle | …

# Additional statistics

***k-independent 2-paths***

$U_k(\mathbf{x})$ = no of $k$-independent 2-paths in $\mathbf{x}$, with parameter $\nu_k$

Let $\nu_{k+1} = -\nu_k/\lambda$

Statistic for $\nu_2$:

$U^{[\lambda]}(\mathbf{x}) = \sum_k (-1)^k \, U_k(\mathbf{x})/\lambda^{k-2}$

***alternating independent 2-path statistic***

***Equivalently (Hunter & Handcock, 2004):***

Statistic: No of dyads with exactly $k$ shared partners

Aggregate statistic: ***geometrically weighted dyad-wise shared partner distribution DSP***

***k-triangles***

$T_k(\mathbf{x})$ = no of $k$-triangles in $\mathbf{x}$, with parameter $\tau_k$

Let $\tau_{k+1} = -\tau_k/\lambda$

Statistic for $\tau_1$:

$T^{[\lambda]}(\mathbf{x}) = \sum_k (-1)^k \, T_k(\mathbf{x})/\lambda^{k-2}$

***alternating k-triangle statistic***

Statistic: No of dyads linked by an edge and having exactly $k$ shared partners

Aggregate statistic: ***Geometrically weighted edge-wise shared partner distribution ESP***

# Model specification: what have we learnt?

1. Relevant exogenous variables at node, tie and setting levels should be used

2. Realisation-dependent neighbourhoods appear to reflect social processes underlying network formation better than simple Markovian neighbourhoods

3. Hypotheses about relationships among the values of related parameters can provide practical and effective means of incorporating important higher-order effects without "death by parameter"

4. So we fit models with relevant ***exogenous variables*** as well as ***edge, alternating 2-star, alternating k-independent 2-path and alternating k-triangle parameters*** (**Siena**) or, equivalently, ***edge, weighted degree distribution, DSP and ESP parameters*** (**statnet**)

→ ***Mark Handcock***