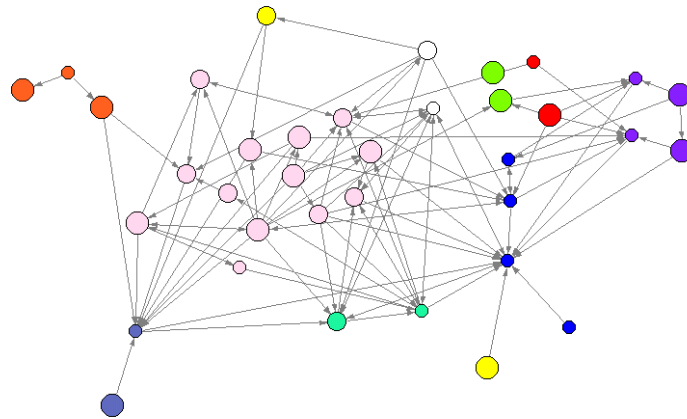


Neighbourhood-based models for social networks: model specification issues



Pip Pattison, University of Melbourne
[with Garry Robins, University of Melbourne
Tom Snijders, University of Gröningen]

IMA Workshop, November 17-21, 2003

Outline

- 1. Exponential random graph models**
- 2. Model specification and homogeneous Markov random graphs**
A critical analysis
- 3. Model specification: two suggestions***
Alternating k -star hypothesis
Independent 2-paths and k -triangles
- 4. Example: modelling mutual collaboration in a law firm***
- 5. Model specification: what have we learnt?**

* *based on Snijders, Pattison & Robins (in preparation)*

1. Random graph models

Why is it important to *model* networks?

Modelling allows

precise inferences about the nature of regularities in networks and network-based processes from empirical observations

Quantitative estimates of these regularities (and their uncertainty) are important

small changes in these regularities can have substantial effects on global system properties

Modelling allows

an understanding of the relationship between (local) interactive network processes and aggregate (eg group, community) outcomes (and can be assessed by how well it predicts global outcomes)

Approach to modelling networks

Guiding principles:

- 1. Network ties are the outcome of unobserved processes that tend to be local and interactive*
- 2. There are both regularities and irregularities in these local interactive processes*

Hence we aim for a stochastic model formulation in which:

local interactivity is permitted and assumptions about “locality” 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)

What do we model?

We model observations at the level of nodes, network ties, settings, ...

For example:

node attribute variables: $\mathbf{Y} = [Y_i]$ $Y_i =$ attribute of node i

tie variables: $\mathbf{X} = [X_{ij}]$ $X_{ij} = 1$ if i has a tie to j
0 otherwise

setting variables: $\mathbf{S} = [S_{ij,kl}]$ $S_{ij,kl} = 1$ if X_{ij} and X_{kl} share a setting
0 otherwise

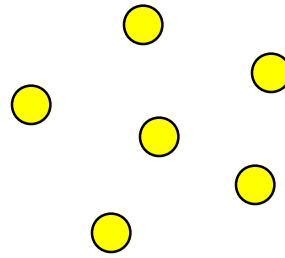
realisations of node-level variables \mathbf{Y} , tie-level variables \mathbf{X} and setting-level variables \mathbf{S} are denoted \mathbf{y} , \mathbf{x} and \mathbf{s} , respectively

A *simplified* multi-layered framework

Social units (y)

individuals

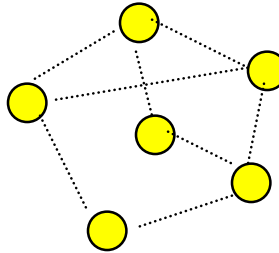
...



Ties among social units (x)

person-to-person

...

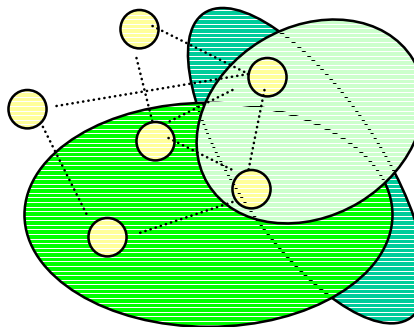


Settings (s)

geographical

sociocultural

...



For example:

Interactions between tie variables depend on node attributes

social selection effects

Interactions between ties depend on proximity through settings

context effects

Local interactivity

Two modelling steps:

methodological: choose a notion of “local” that it is convenient from a modelling point of view:

proximity \leftrightarrow interactivity

define two variable entities (eg network tie variables) to be **neighbours** if they are conditionally dependent, given the values of all other entities

substantive: what are appropriate assumptions about proximity in this sense?

Some assumptions about proximity

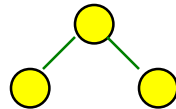
Two tie variables are neighbours if:

they share a dyad



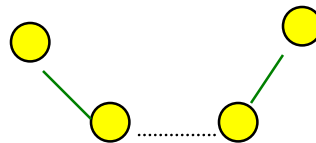
dyad-independent model

they share an actor



Markov model

they share a connection
with the same tie



realisation-dependent model
(catalysis, of a sort)

etc.

Models for interactive systems of variables (Besag, 1974)

Two variables are *neighbours* if they are conditionally dependent given the observed values of all other variables

A *neighbourhood* is a set of mutually neighbouring variables

A model for a system of variables has a form determined by its neighbourhoods

Hammersley-Clifford theorem

This general approach leads to:

$\Pr(\mathbf{X} = \mathbf{x})$ *exponential random graph models* Frank & Strauss 1986
(extended by Wasserman, Robins & Pattison)

Extension to directed dependence assumptions:

$\Pr(\mathbf{X} = \mathbf{x} \mid \mathbf{Y} = \mathbf{y})$ *social selection models* Robins et al 2001

$\Pr(\mathbf{X} = \mathbf{x} \mid \mathbf{S} = \mathbf{s})$ *setting-dependent models* Pattison & Robins 2002

Exponential random graph (p^*) models (Frank & Strauss, 1986)

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

normalizing quantity

parameter

network statistic

the summation is over all neighbourhoods Q

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

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

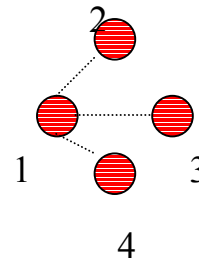
What is a neighbourhood?

A *neighbourhood* is a subset of tie variables, each pair of which are neighbours

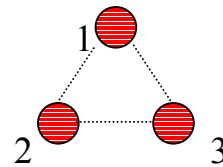
Each neighbourhood corresponds to a *network configuration*:

for example:

$\{X_{12}, X_{13}, X_{14}\}$ corresponds to the configuration of tie variables:



$\{X_{12}, X_{13}, X_{23}\}$ corresponds to:



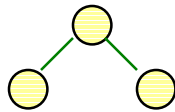
Neighbourhoods depend on proximity assumptions

Assumptions: two ties are neighbours:

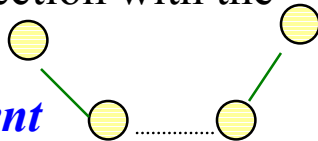
if they share a dyad
dyad-independence



if they share an actor
Markov



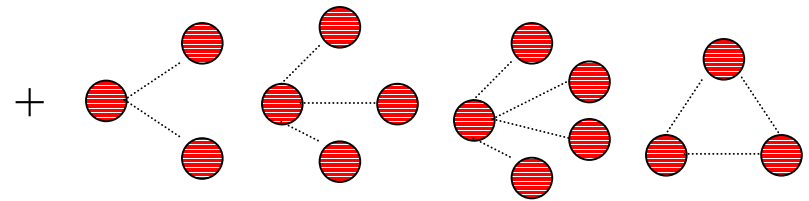
if they share a connection with the same tie
realisation-dependent



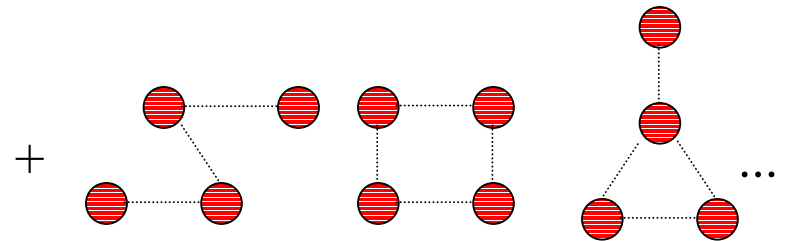
Configurations for neighbourhoods



edge



2-star 3-star 4-star ... triangle

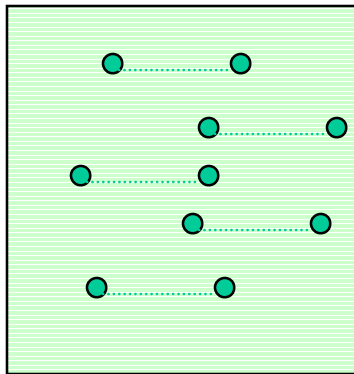


3-path 4-cycle "coathanger"

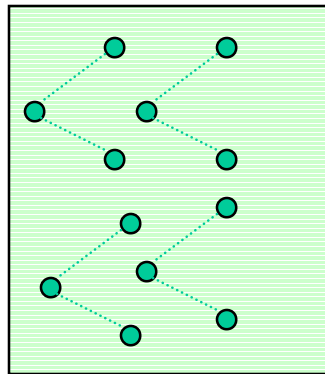
Homogeneous network models

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/c) \exp\{\sum_{Q^*} \gamma_{Q^*} z_{Q^*}(\mathbf{x})\}$$

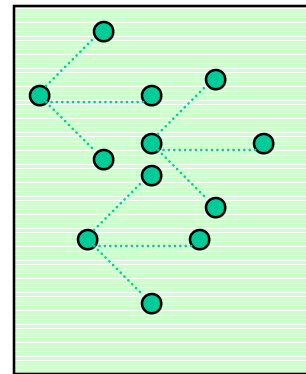
If we assume that parameters for *isomorphic* configurations are the same:



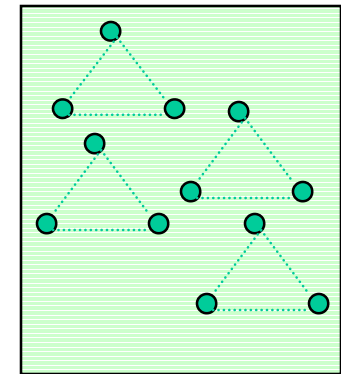
edges [θ]



2-stars [σ_2]



3-stars [σ_3] ...



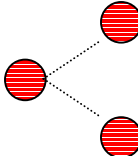
triangles [τ]

then there is one parameter γ_{Q^*} for each *class* Q^* of isomorphic configurations and the corresponding statistic $z_{Q^*}(\mathbf{x})$ is a *count* of such observed configurations in \mathbf{x}

Homogeneous Markov random graphs (Frank & Strauss, 1986)

$$\Pr(\mathbf{X} = \mathbf{x}) = (1/c) \exp\{\theta L(\mathbf{x}) + \sigma_2 S_2(\mathbf{x}) + \dots + \sigma_k S_k(\mathbf{x}) + \dots + \tau T(\mathbf{x})\}$$

where: $L(\mathbf{x})$ no of *edges* in \mathbf{x} 

$S_2(\mathbf{x})$ no of *2-stars* in \mathbf{x} 

...

$S_k(\mathbf{x})$ no. of *k-stars* in \mathbf{x} 

...

$T(\mathbf{x})$ no of *triangles* in \mathbf{x} 

3. Model specification: edge and dyad parameters

In the beginning: the edge

very large literature on random graphs

And then: the dyad

p_1 (Holland, Leinhardt, Wasserman, Fienberg)

p_2 (van Duijn, Snijders & Zijlstra, 2004)

latent blocks (Nowicki & Snijders, 2001)

latent space (Hoff, Raftery & Handcock, 2002)

latent ultrametric space (Schweinberger & Snijders, 2003)

Can dyadic models suffice?

They capture important actor-level and homophily processes (which should not be ignored)

But we hypothesise network-based extra-dyadic processes (on theoretical grounds, with suggestive supporting evidence)

Nonetheless this is still a somewhat open (and important) question

Homogeneous Markov models

with $\sigma_p = 0$, for $p > p_0$

Homogeneous Markov random graph models

Alluring potential for modelling theorised triadic effects, *but*

Handcock (2002) defines models to be *degenerate* if most of the probability mass is concentrated in small parts of the state space

Regions of parameter space that are not degenerate may be quite small (Handcock, 2002)

We appear not to have dealt satisfactorily with star parameters

Simulation-based parameter estimation methods often wander into degenerate regions of parameter space (unless steering is excellent)

Robins (2003) showed empirically that parameters estimated from data (using SIENA [Snijders, 2002]) can be quite close to degenerate regions

Even where parameters can be estimated successfully, the model may fail to reproduce important features of the data (eg degree distribution, connectivity)

There are theoretical reasons to doubt the adequacy of a homogeneous Markov assumption

3. New specifications

I: *the alternating k-star hypothesis*

Suppose that: $\sigma_k = -\sigma_{k-1}/\lambda$ where $\lambda \geq 1$ is a (fixed) constant
alternating k-star hypothesis

Then we may write

$$\sum_k S_k(\mathbf{x})\sigma_k = S^{[\lambda]}(\mathbf{x}) \sigma_2$$

where $S^{[\lambda]}(\mathbf{x}) = S_2(\mathbf{x}) - S_3(\mathbf{x})/\lambda + S_4(\mathbf{x})/(\lambda^2) - \dots + (-1)^{n-2}S_{n-1}(\mathbf{x})/(\lambda^{n-3})$

The statistic $S^{[\lambda]}(\mathbf{x})$ may be expressed in the form

$$S^{[\lambda]}(\mathbf{x}) = \lambda^2 \sum_i \{(1 - 1/\lambda)^{d(i)} + d(i)/\lambda - 1\}$$

where $d(i)$ denote the degree of node i *alternating k-star statistic*

Properties of alternating k-star models

The case of $\lambda = 1$: $S^{[\lambda]}(\mathbf{x}) = 2L(\mathbf{x}) - n + \#\{i \mid d(i) = 0\}$

so if a model also includes an edge parameter θ , then no. of isolated nodes is modelled separately

Change statistic (change in $S^{[\lambda]}(\mathbf{x})$ if x_{ij} is changed from 1 to 0):

$$\begin{aligned}\Delta(S^{[\lambda]}(\mathbf{x}))_{ij} &= \lambda \{2 - (1 - 1/\lambda)^{[d(i)-1]} - (1 - 1/\lambda)^{[d(j)-1]}\} & \lambda > 1 \\ &= I\{d(i) \geq 0\} + I\{d(j) \geq 0\} & \lambda = 1\end{aligned}$$

Note that:

$$0 \leq \Delta(S^{[\lambda]}(\mathbf{x}))_{ij} \leq 2\lambda$$

if $\lambda > 1$ and $\sigma^{[\lambda]} > 0$, the conditional log-odds of a tie is enhanced the higher the degree of i or j , but the marginal gain is nonlinear in degree

Robins presents some simulations based on models with positive $\sigma^{[\lambda]}$ to investigate heterogeneity in degrees compared to Bernoulli models and models with higher-order star parameters set to zero

Other functions of degree

Other hypotheses could of course also be entertained ...

For example, Tom Snijders has suggested:

$$S(\mathbf{x}) = \sum_i 1/(d(i)+c)_r$$

where $(q)_r = q(q+1)\dots(q+r-1)$ is *Pochhammer's symbol*

longer tail for degree distribution

Alternating k -star models

Are homogeneous Markov random graph models with alternating k -stars likely to suffice?

Probably not, because:

they overstate likely dependencies: possible ties such as X_{ij} and X_{jk} may be conditionally independent if they occupy distinct “social locales” (eg i may not know that k exists) *use exogenous setting information where possible*

they understate likely dependencies: X_{ij} and X_{kl} may be conditionally dependent in some cases, e.g., where they become linked either exogenously or by endogenous network processes *leads to a consideration of realisation-dependent models*

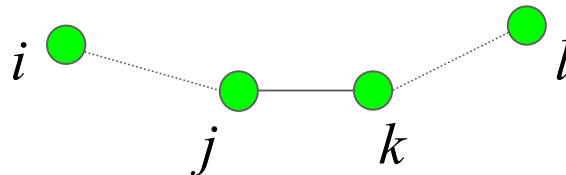
Realisation-dependent models

Consider the relation ties X_{ij} and X_{kl} and the following neighbourhood structure:

If $\{i,j\} \cap \{k,l\} = \emptyset$, assume, in general, that X_{ij} and X_{kl} are conditionally independent (occupy distinct neighbourhoods)

but suppose that if $x_{jk} = 1$ then X_{ij} and X_{kl} are conditionally dependent

3-path model



In the latter case the neighbourhood is *generated* by the observed relation x_{jk} (c.f. Baddeley & Möller, 1989)

a modified Hammersley-Clifford theorem

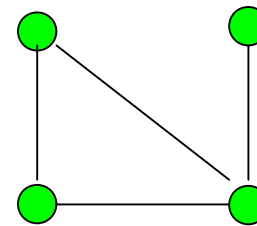
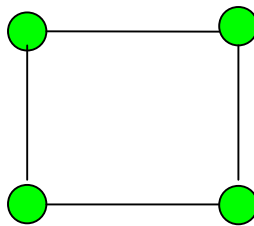
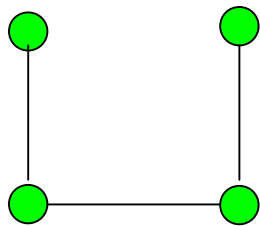
neighbourhoods are connected configurations with longer paths

Generalised exchange: 4-cycles in networks (Pattison & Robins, 2002)

The 3-path model leads to simple models for *generalised exchange*

Generalised exchange establishes a system of operations conducted ‘on credit’ (Lévi-Strauss, 1969)

Neighbourhoods now include configurations of the form:



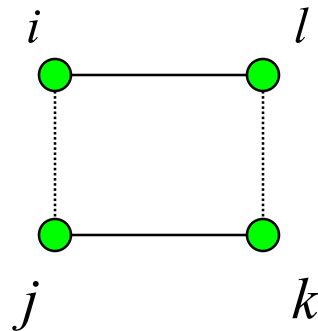
New specifications II. *Realisation-dependent models for higher-order clustering effects*

Consider the following neighbourhood structure:

If $\{i,j\} \cap \{k,l\} \neq \emptyset$, assume that X_{ij} and X_{kl} are conditionally dependent (Markov assumption)

If $\{i,j\} \cap \{k,l\} = \emptyset$, assume, in general, that X_{ij} and X_{kl} are conditionally independent

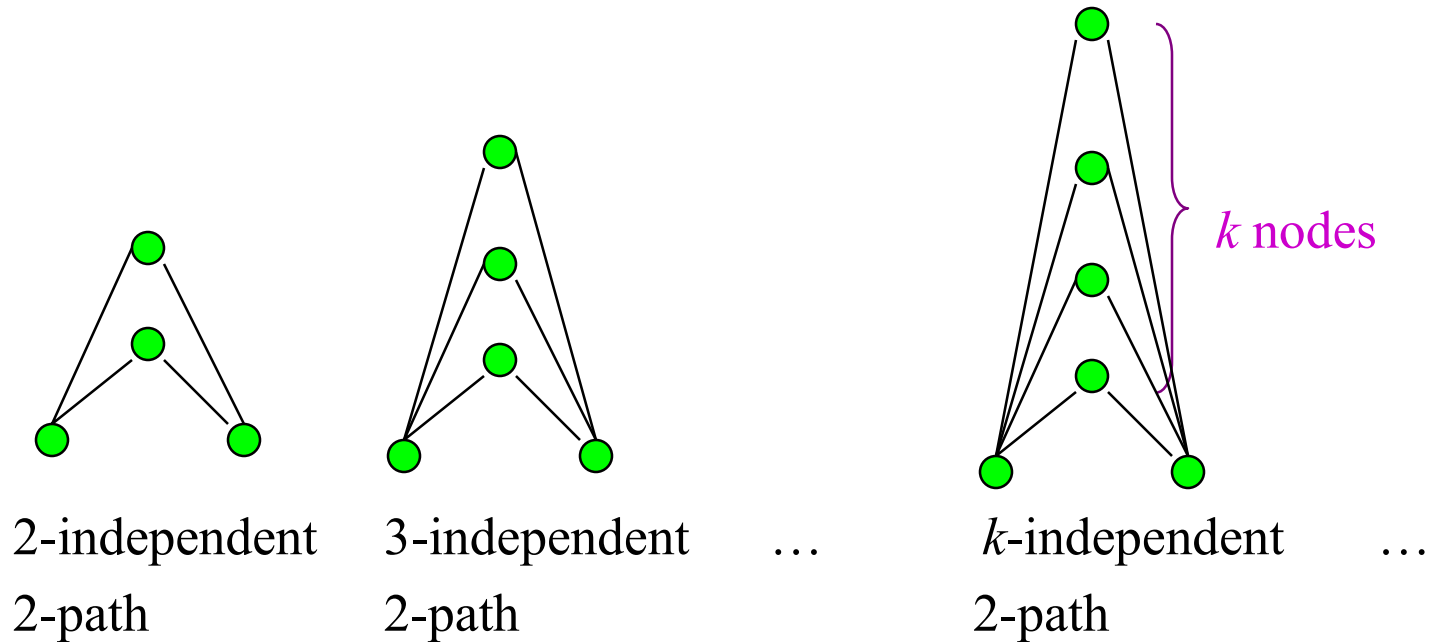
But assume that if $x_{jk} = 1 = x_{il}$ then X_{ij} and X_{kl} are conditionally dependent



the 4-cycle model

Some neighbourhoods for 4-cycle model: independent 2-paths

The 4-cycle model includes neighbourhoods of the form:



In general, a k -independent 2-path is a configuration comprising k independent 2-paths between two nodes
can model 2-path connectivity effects

Independent 2-path statistics

Let $\mathbf{p} = \mathbf{x}^2$ and let $\mathbf{x}^{[ij0]} = \mathbf{x}$ but with $\mathbf{x}^{[ij0]}_{ij} = \mathbf{x}^{[ij0]}_{ji} = 0$

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

$$U_2(\mathbf{x}) = \{\sum_{i<j} (p_{ij}-1)p_{ij}\}/4 \quad \textit{number of 2-independent 2-paths}$$

$$U_k(\mathbf{x}) = 1/2 \sum_{i<j} \binom{p_{ij}}{k} \quad \textit{number of k-independent 2-paths}$$

$$\text{Suppose that } v_{k+1} = -v_k/\lambda \quad \textit{alternating independent 2-path hypothesis}$$

Then the statistic corresponding to v_2 is:

$$U^{[\lambda]}(\mathbf{x}) = \lambda \sum_{i,j} \{1 - (1 - 1/\lambda)^{p_{ij}}\} \quad \textit{alternating independent 2-path statistic}$$

Change statistic for $U^{[\lambda]}(\mathbf{x})$

The case of $\lambda = 1$:

$$U^{[1]}(\mathbf{x}) = \sum_{i < j} I\{q_{ij} \geq 1\}$$

number of pairs at least indirectly linked by a 2-path

Change statistic:

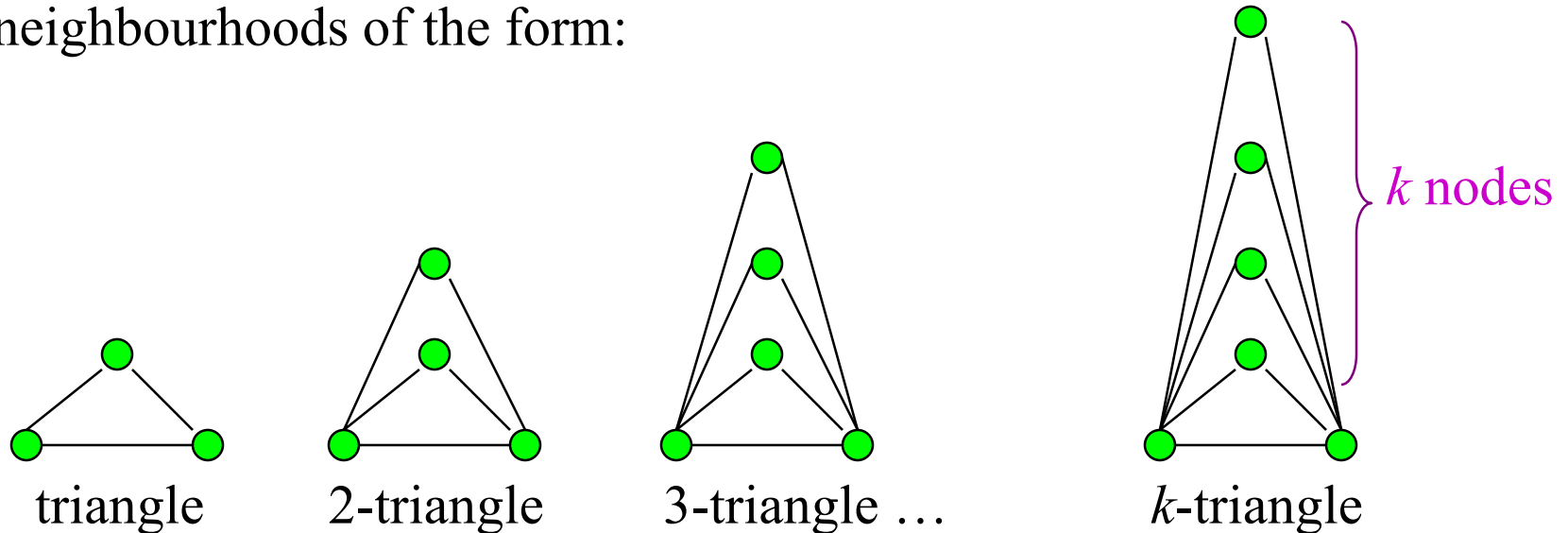
Let $\mathbf{q} = (\mathbf{x}^{[ij^0]})^2$

2-paths computed with x_{ij} and x_{ji} set to 0

$$\Delta(U^{[\lambda]}(\mathbf{x}))_{ij} = \sum_{h \neq i, j} \{x_{jh}(1 - 1/\lambda)^{q_{ih}} + x_{hi}(1 - 1/\lambda)^{q_{hj}}\}$$

More neighbourhoods for 4-cycle model

A model with Markovian and 4-cycle dependencies also has neighbourhoods of the form:



In general, a k -triangle comprises k triangles sharing a common base

can model 2-path induced cohesion effects

Let $T_k(\mathbf{x}) =$ no of k -triangles in \mathbf{x}

Triangle and k -triangle statistics

Let $\mathbf{p} = \mathbf{x}^2$ and let $\mathbf{x}^{[ij0]} = \mathbf{x}$ but with $\mathbf{x}^{[ij0]}_{ij} = \mathbf{x}^{[ij0]}_{ji} = 0$

$$T_1(\mathbf{x}) = \{\sum_{i,j} x_{ij} p_{ij}\} / 3 \quad \textit{number of 1-triangles}$$

$$T_k(\mathbf{x}) = \sum_{i < j} x_{ij} \binom{p_{ij}}{k} \quad \textit{number of } k\text{-triangles}$$

Suppose that $\tau_{k+1} = -\tau_k / \lambda$ *alternating k -triangle hypothesis*

Then the statistic corresponding to τ_1 is:

$$T^{[\lambda]}(\mathbf{x}) = \lambda \sum_{i,j} x_{ij} \{1 - (1 - 1/\lambda)^{p_{ij}}\} \quad \textit{alternating } k\text{-triangle statistic}$$

Change statistic for $T^{[\lambda]}(\mathbf{x})$

Let $\mathbf{q} = (\mathbf{x}^{[ij^0]})^2$

2-paths computed with x_{ij} and x_{ji} set to 0

Change statistic

$$\Delta(T^{[\lambda]}(\mathbf{x}))_{ij} = \lambda \{1 - (1 - 1/\lambda)^{q_{ij}}\} + \sum_h x_{ih} x_{jh} (1 - 1/\lambda)^{q_{ih}} + \sum_h x_{hi} x_{hj} (1 - 1/\lambda)^{q_{jh}}$$

The case of $\lambda=1$

$$T^{[\lambda]}(\mathbf{x}) = \sum_{i < j} x_{ij} I\{p_{ij} \geq 1\}$$

number of pairs that lie on at least one triangle

**MCMC parameter estimates for mutual
collaborations among partners of a law firm
(Lazega, 1999; SIENA, conditioning on total ties)**

	Model 1		Model 2	
<i>Parameter</i>	<i>est</i>	<i>s.e.</i>	<i>est</i>	<i>s.e.</i>
alternating k -stars ($\lambda=3$)	-0.083	0.316		
Alternating ind. 2-paths ($\lambda=3$)	-0.042	0.154		
Alternating k -triangles ($\lambda=3$)	0.572	0.190	0.608	0.089
No pairs connected by a 2-path	-0.025	0.188		
No pairs lying on a triangle	0.486	0.513		
Seniority main effect	0.023	0.006	0.024	0.006
Practice (corp. law) main effect	0.391	0.116	0.375	0.109
Same practice	0.390	0.100	0.385	0.101
Same gender	0.343	0.124	0.359	0.120
Same office	0.577	0.110	0.572	0.100

Features of model fit

Simulating from estimates for Model 2, we find that the model recovers well the number of 2-stars, 3-stars, 4-stars and triangles (even though these are not *directly* fitted)

Good results also for $\lambda = 2, 4$ and 5 (but not for 1 or 6)

SIENA obtained estimates for model with covariates, triangles, and 2-stars, but less satisfactory reproduction of low-level statistics (indeed, more careful scrutiny raises further questions)

5. Model specification: what have we learnt?

Relevant exogenous variables at node, tie, group and setting levels should be used!

Realisation-dependent neighbourhoods may reflect social processes of exchange and cohesion better than simple Markovian neighbourhoods

- Cycles and generalised exchange

- k -triangles and cohesion

- independent 2-paths and connectivity

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” setting in

- Alternating k -stars

- Other functions of degree

- Alternating independent-2-paths

- Alternating k -triangles

Next steps

Learning from interaction with data, especially from data with careful and well-designed measurements at node, tie and setting levels (Tom Snijders' **SIENA** and Mark Handcock's **ergm** make this possible)

Continue to be open to realisation-dependent neighbourhood forms

Explore other hypotheses on relationships among parameters