Dependence structures

Pip Pattison
University of Melbourne

or

Network configurations on the right hand side of the equation

Outline

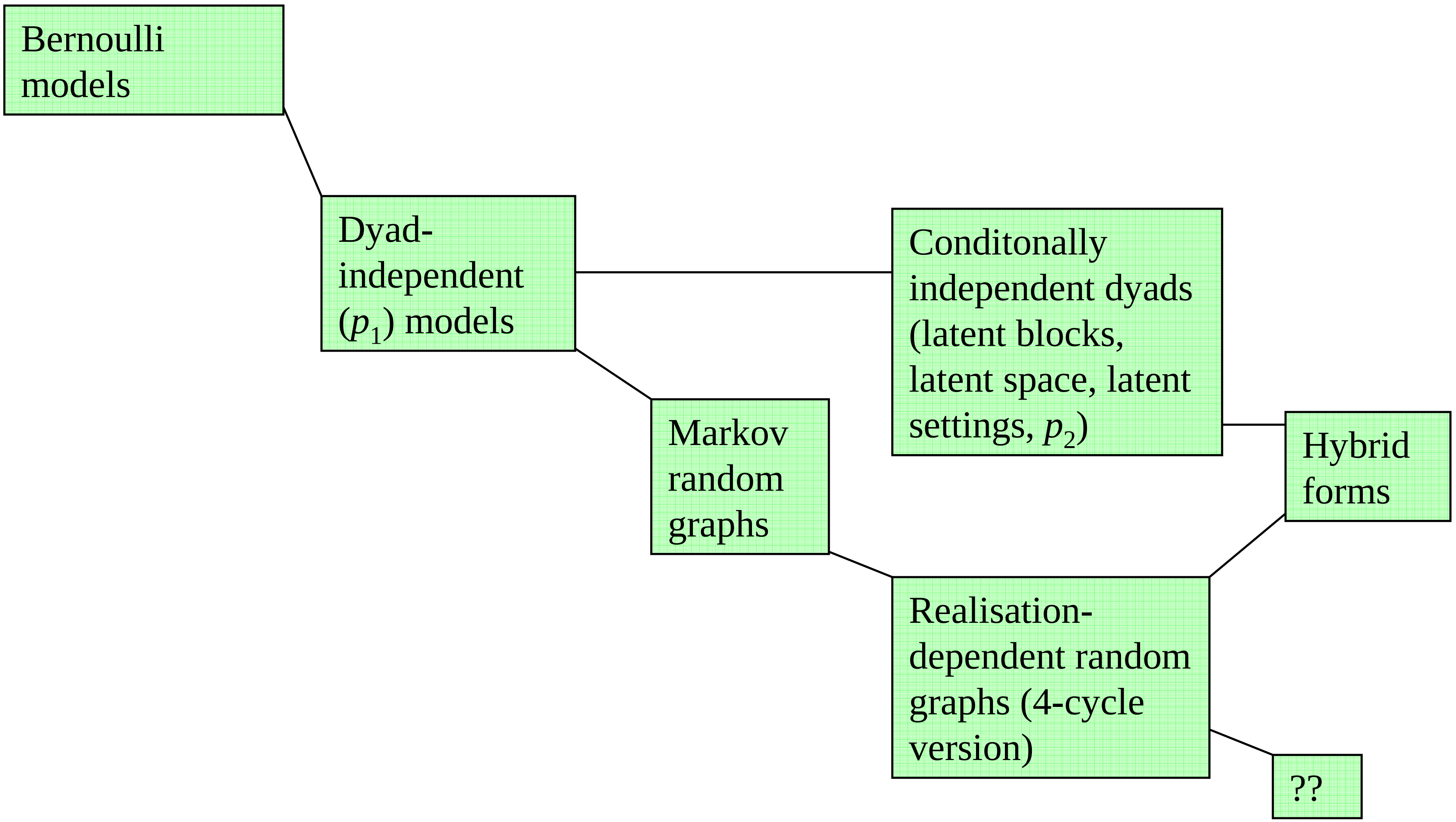
Four generations of network models

- Bernoulli (Erdös-Rényi) models
- p_1 models and latent variable models
- Markov random graphs
- Realisation-dependent models

Dependence structures, neighbourhoods and the Hammersley-Clifford theorem

Simplifying models by assuming relationships among related parameters

Statistical models for networks: the family tree



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

Exogenous effects:

Homophily, shared affiliations, spatial propinquity all matter (e.g., McPherson, Smith-Lovin & Cook, 2001; Koehly, Goodreau & Morris, 2005)

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 construct statistical models in which:

- network ties are variable and network nodes are fixed
- assumptions about the form of local interactions among tie variables 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 and dependence

We model ***tie variables***: $\mathbf{Y} = [Y_{ij}]$ $Y_{ij} = 1$ if i has a tie to j
0 otherwise

realisation of **Y** is denoted by $\mathbf{y} = [y_{ij}]$

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***

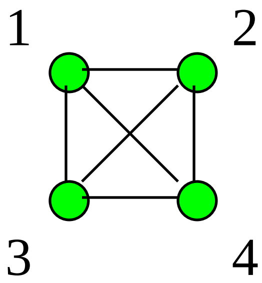
Reminder: Two variables A and B are ***conditionally independent*** given C if

$P(A,B|C)= P(A|C) \times P(B|C)$

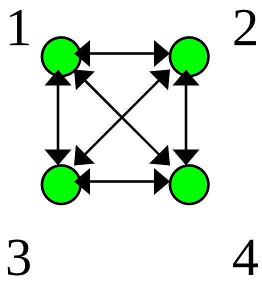
Dependence structure: hypothesis about ***neighbour*** assumptions

Random graph and random directed graphs on a node set N = {1,2,3,4}

Graph



Directed graph



Tie variables:

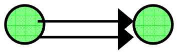
Y_{12}, Y_{13}, Y_{14}, Y_{23}, Y_{24}, Y_{34}

Tie variables:

Y_{12}, Y_{13}, Y_{14}, Y_{23}, Y_{24}, Y_{34},
Y_{21}, Y_{31}, Y_{41}, Y_{32}, Y_{42}, Y_{43}

The first generation

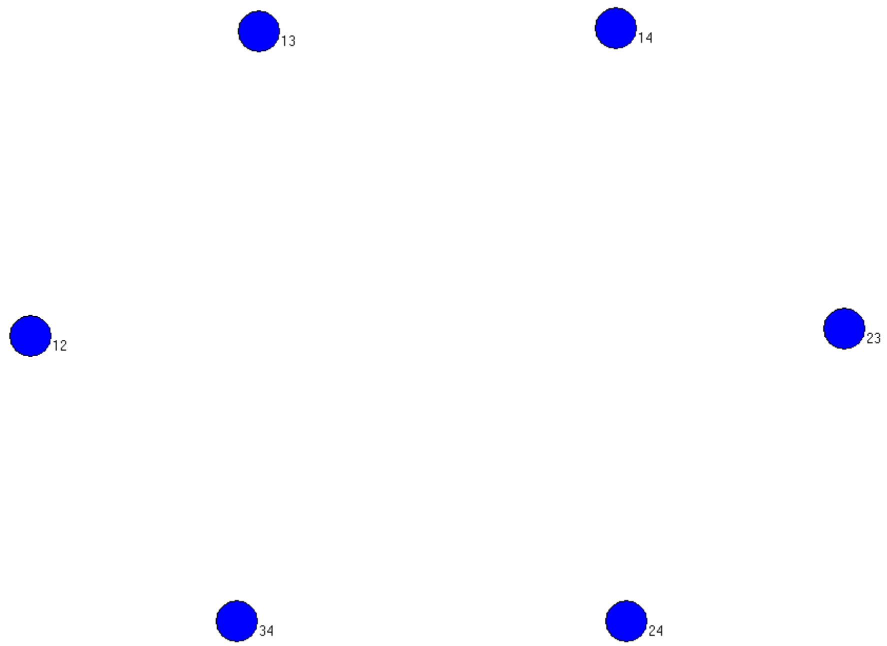
Tie variables Y_{ij} and Y_{kl} are conditionally independent unless:

$(i,j) = (k,l)$  **they involve the same ordered pair**

Bernoulli, or Erdös-Rényi random graphs

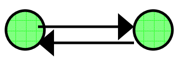
e.g., Erdös and Rényi (1959)

Dependence structure for a Bernoulli graph on node set $N = \{1,2,3,4\}$



The second generation

Tie variables Y_{ij} and Y_{kl} are conditionally independent unless:

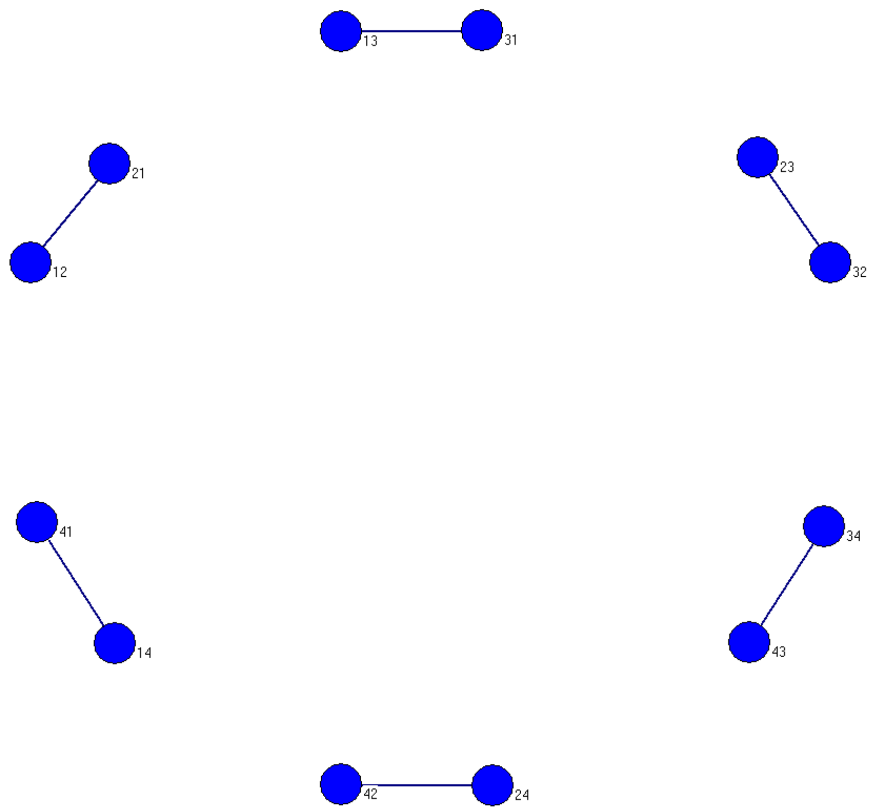
$\{i,j\} = \{k,l\}$  **they involve the same dyad**

p_1 model

Holland & Leinhardt (1981)

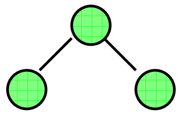
Wasserman & Galaskiewicz (1985)

Dependence structure for a p_1 model for a directed graph on 4 nodes

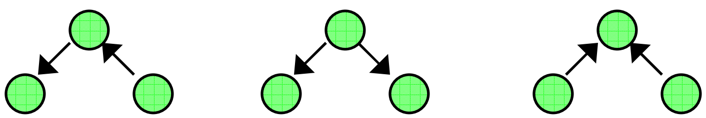


The third generation

Tie variables X_{ij} and X_{kl} are conditionally independent unless:

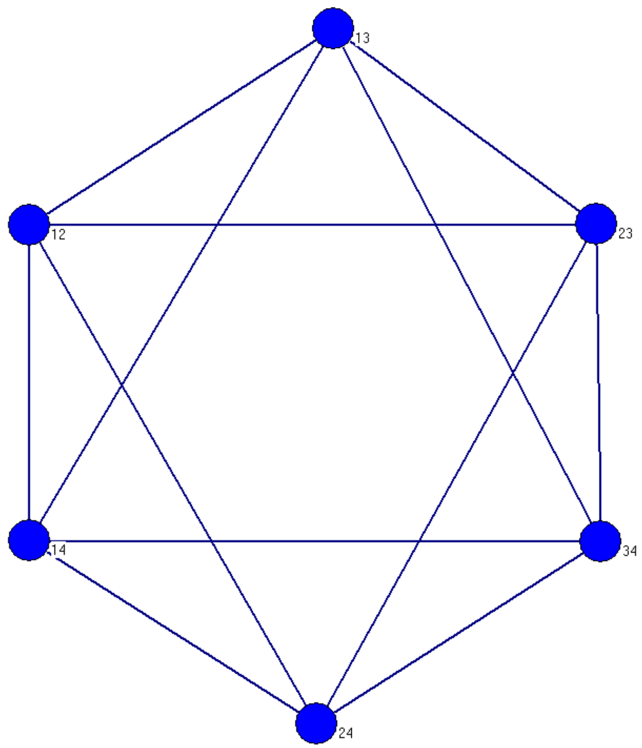
$\{i,j\} \cap \{k,l\} \neq \emptyset$  **they involve a shared actor**

Markov random graph

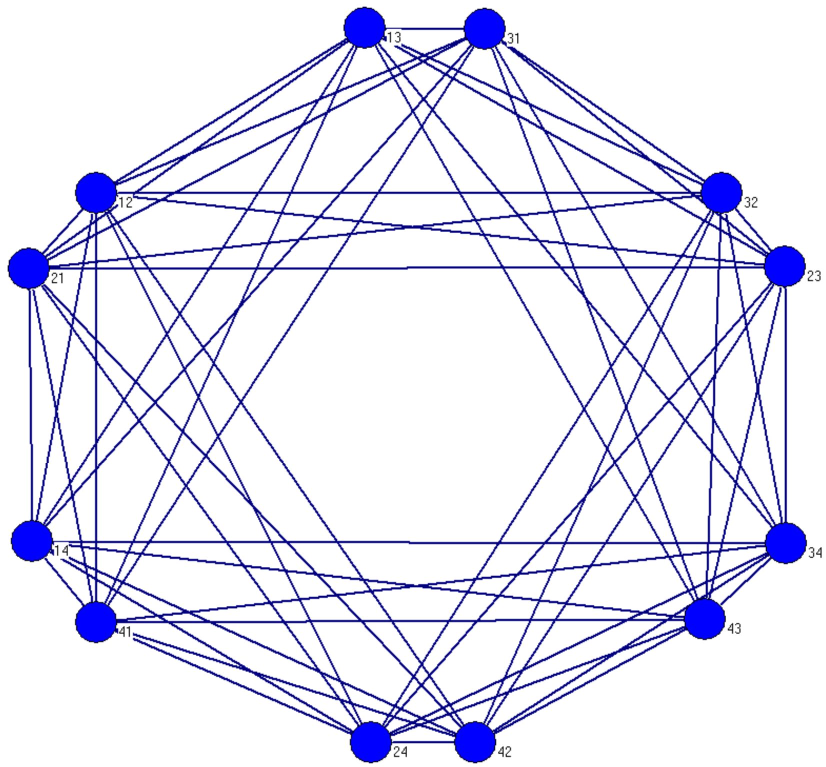


Markov random directed graph

Dependence structure for a Markov graph on $N = \{1,2,3,4\}$



Dependence structure for a Markov directed graph on 4 nodes



The fourth generation

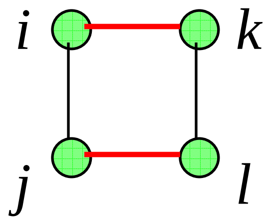
Tie variables X_{ij} and X_{kl} are conditionally independent for distinct *i*, *j*, *k*, *l* unless:

$y_{ik} = 1$ and $y_{jl} = 1$

or

$y_{il} = 1$ and $y_{jk} = 1$

(eg **red** ties are observed)



and so they complete a social circuit

realisation- dependent random graphs: embedded first-order zone model

Pattison & Robins, 2002; Hunter & Handcock, in press; Snijders, Pattison, Robins & Handcock, in press; forthcoming special edition of *Social Networks* (edited by Martina Morris and Garry Robins)

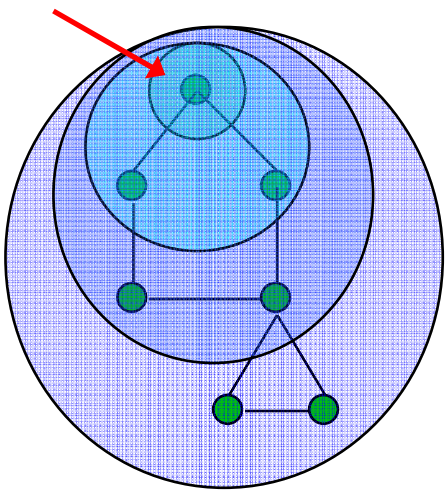
Zones of order *k*

Let **y** be an observed network on node set N *and c*onsider a subset $A \subseteq N$

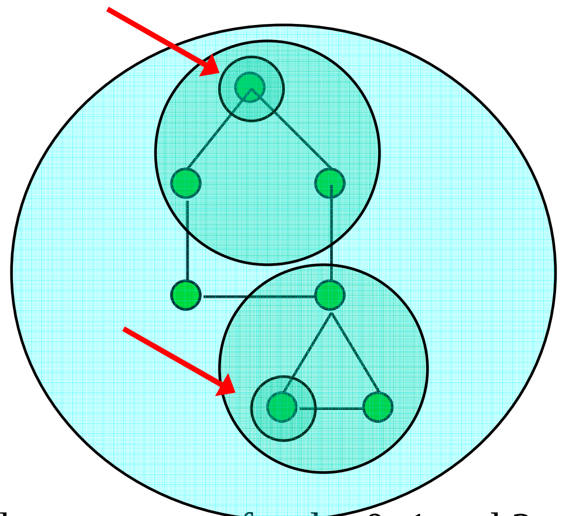
Define $Z_k(A)$ to be the **zone of order *k* of *A* in the network y** to be the set of nodes within k steps of a node in A:

$Z_k(A) = \{b \in N | d_{ab} \leq k$, for some $a \in A\}$ where d_{ab} is the geodesic distance from a to b

For sets comprising marked vertices:



Single node set: zones of order 0, 1, 2 and 3



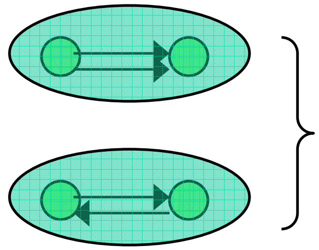
Two-node set: zones of order 0, 1 and 2

The four generations

Tie variables Y_{ij} and Y_{kl} are conditionally independent unless:

$(i,j) = (k,l)$

$Z_0(\{i,j\}) = Z_0(\{k,l\})$

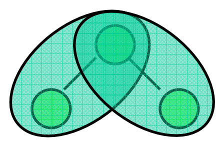


same zone of order 0 for {*i,j*} and {*k,l*}

Bernoulli

p_1

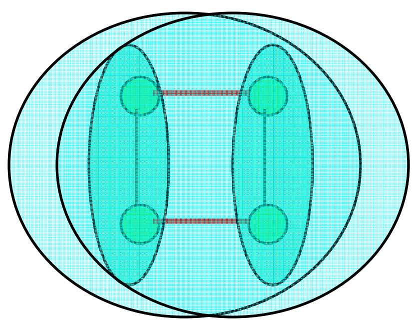
$Z_0(\{i,j\}) \cap Z_0(\{k,l\}) \neq \varnothing$



overlapping zones of order 0

Markov

$Z_1(\{i,j\}) \supseteq Z_0(\{k,l\})$
and
$Z_1(\{k,l\}) \supseteq Z_0(\{i,j\})$

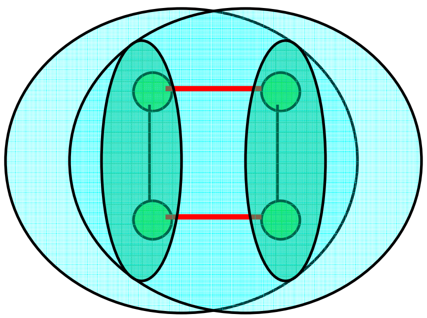


zero-order zones for {*i,j*) and {*k,l*} jointly embedded in first order zones for {*k,l*} and {*i,j*}

realisation-dependent

Why does embedding in first-order zones matter?

the potential ties are embedded in an emergent social setting



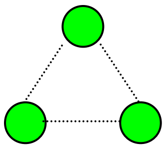
if realised, they would complete a social circuit

See Coleman (1988) on social capital and Bearman et al (2005) for arguments for and against closure in 4-cycles in different kinds of networks

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. $\{Y_{12}, Y_{13}, Y_{23}\}$ corresponds to



Hammersley-Clifford theorem (Besag, 1974):

A model for **Y** 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

$$P(Y = y) = (1/\kappa(\theta)) \exp\{\sum_Q \theta_Q g_Q(y)\}$$

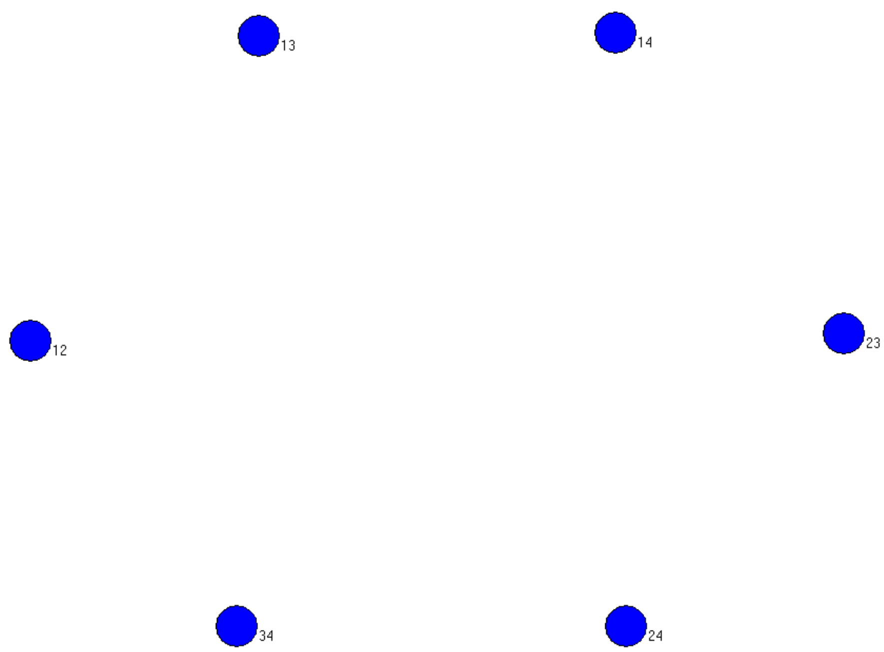
normalizing quantity ***parameter*** ***network statistic***

the summation is over all neighbourhoods *Q*

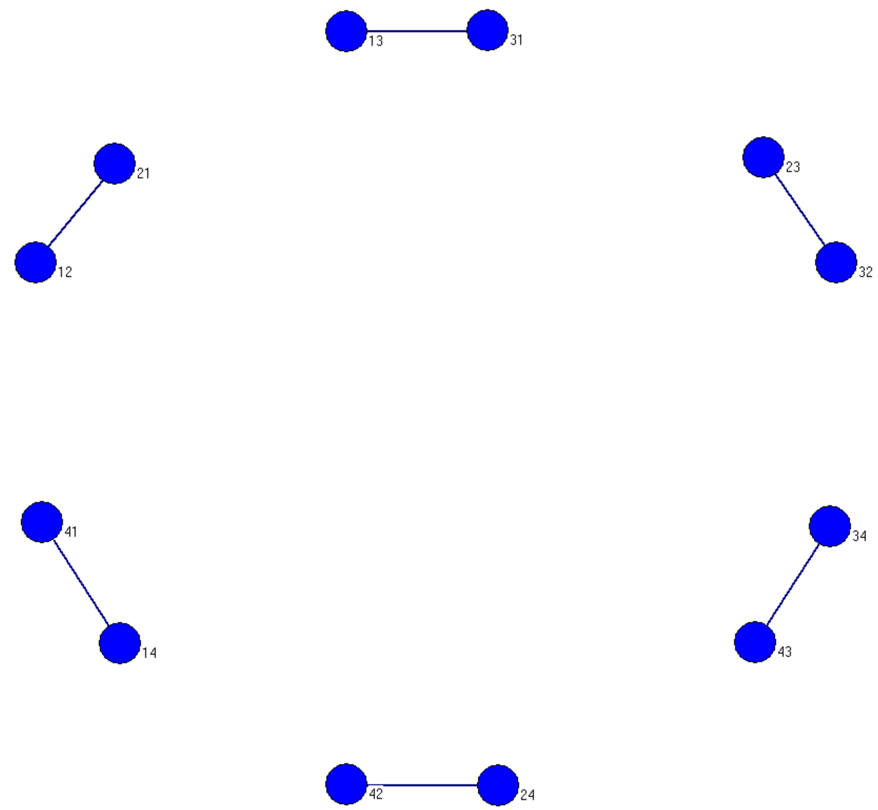
$g_Q(y) = \prod_{Y_{ij} \in Q} y_{ij}$ signifies whether all ties in *Q* are observed in **y**

$$\kappa(\theta) = \sum_y \exp\{\sum_Q \theta_Q g_Q(y)\}$$

Dependence structure for a Bernoulli graph on node set N = {1,2,3,4}



Dependence graph for a p_1 model for a directed graph on 4 nodes



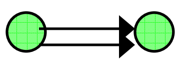
Neighbourhoods depend on proximity assumptions

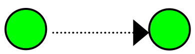
Assumptions: two ties are *neighbours*:

Configurations for neighbourhoods

if they share an ordered pair

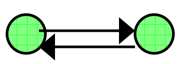
Bernoulli



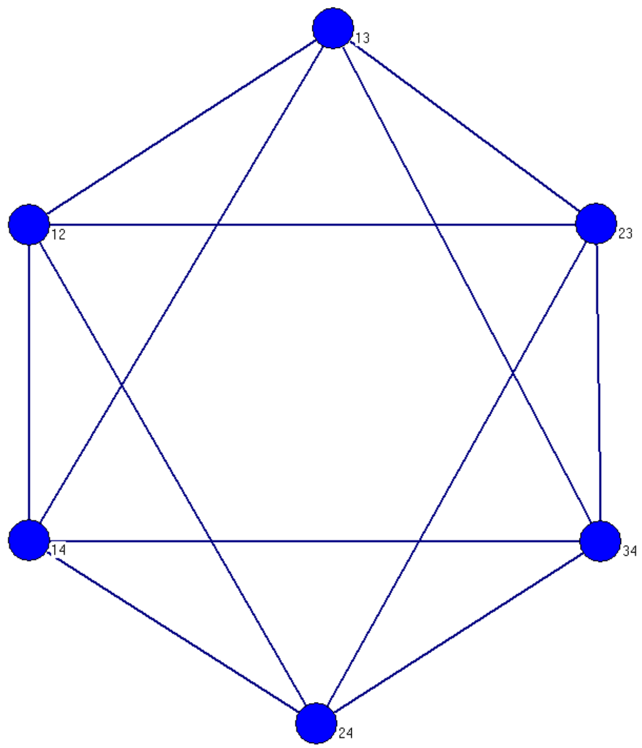


if they share a dyad

p_1



Dependence structure for a Markov graph on $N = \{1,2,3,4\}$

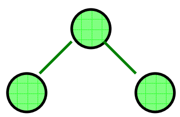


Neighbourhoods depend on proximity assumptions

Assumptions: two ties are *neighbours*:

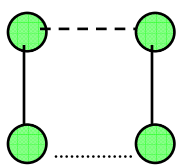
if they share an actor

Markov

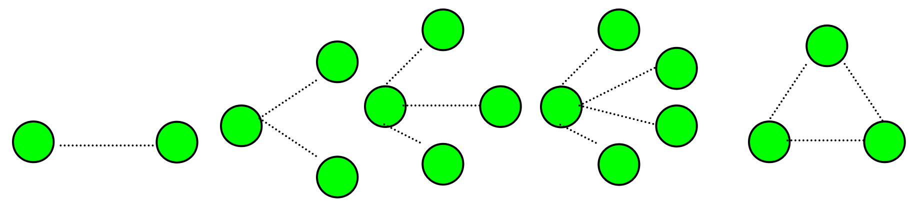


if they complete a 4-cycle

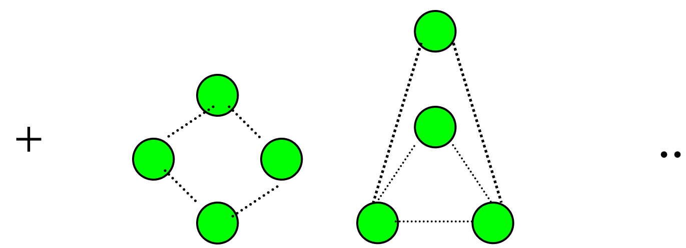
realisation-dependent



Configurations for neighbourhoods



edge 2-star 3-star 4-star …triangle



4-cycle 2-triangle and others

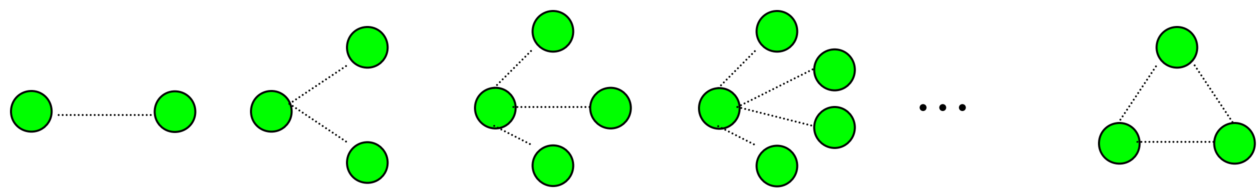
Homogenous models

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 **y**

E.g. for a Markov model:

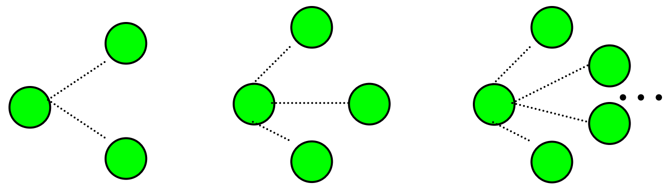
Configurations



Parameters	θ	σ_2	σ_3	σ_4	…	τ
Statistics	$S_1(\mathbf{y})$	$S_2(\mathbf{y})$	$S_3(\mathbf{y})$	$S_4(\mathbf{y})$	…	$T(\mathbf{y})$

Related model parameters

Star configurations



Parameters σ_2 σ_3 σ_4 …

If we assume that $\sigma_k = -\sigma_{k-1}/\lambda$, for $k > 2$ and $\lambda \geq 1$ a (fixed) constant

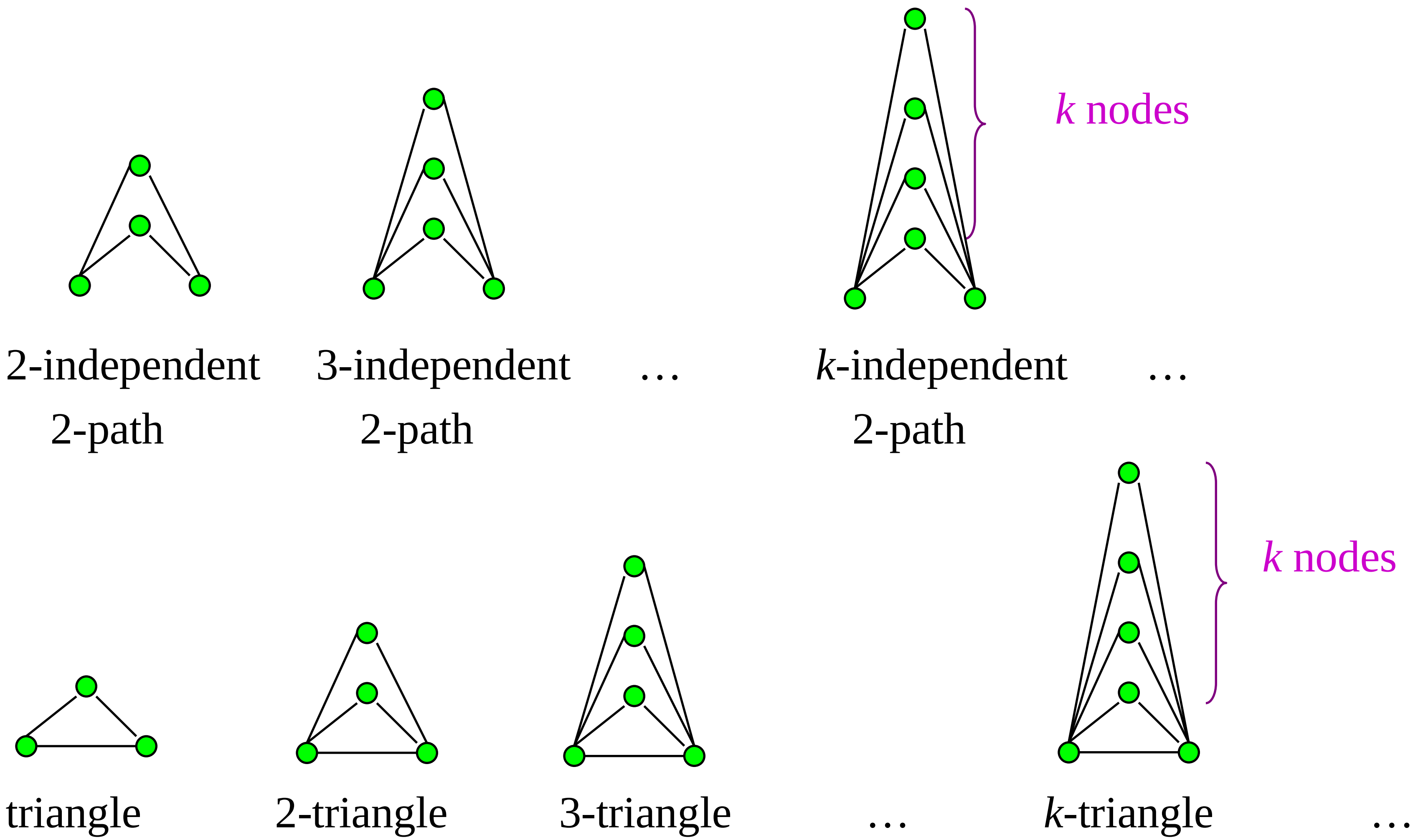
alternating k-star hypothesis

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

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

Hunter and Handcock (in press) show that, provided an edge parameter is included, this is equivalent to a geometrically weighted degree statistic (GWD)

Additional neighbourhoods for realisation-dependent model



Additional statistics

k-independent 2-paths

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

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

Statistic for ν_2:

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

alternating independent 2-path statistic

k-triangles

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

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

Statistic for τ_1:

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

alternating k-triangle statistic

Equivalently (Hunter & Handcock, in press):

Statistic: No of dyads with exactly k shared partners

Aggregate statistic: ***geometrically weighted dyadic shared partner statistic, GWDSP***

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

Aggregate statistic: ***Geometrically weighted edge-wise shared partner statistic, GWESP***

What have we learnt about the right-hand side of the equation?

1. Relevant exogenous variables 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”

$\rightarrow$ ***Mark Handcock***