Statistical Models for Social Networks with Application to HIV Epidemiology

Mark S. Handcock

Department of Statistics
University of Washington

Joint work with

Pavel Krivitsky
Martina Morris

and the

U. Washington Network Modeling Group

Supported by NIH NIDA Grant DA012831 and NICHD Grant HD041877

NIPS 2007, December 4 2007

Network modeling from a statistical perspective

- Networks are widely used to represent data on relations between interacting actors or nodes.

Network modeling from a statistical perspective

- Networks are widely used to represent data on relations between interacting actors or nodes.
- The study of social networks is multi-disciplinary
  - plethora of terminologies
  - varied objectives, multitude of frameworks

Network modeling from a statistical perspective

- Networks are widely used to represent data on relations between interacting actors or nodes.
- The study of social networks is multi-disciplinary
  - plethora of terminologies
  - varied objectives, multitude of frameworks
- Understanding the structure of social relations has been the focus of the social sciences
  - *social structure:* a system of social relations tying distinct social entities to one another
  - Interest in understanding how social structure form and evolve

Network modeling from a statistical perspective

- Networks are widely used to represent data on relations between interacting actors or nodes.
- The study of social networks is multi-disciplinary
  - plethora of terminologies
  - varied objectives, multitude of frameworks
- Understanding the structure of social relations has been the focus of the social sciences
  - *social structure:* a system of social relations tying distinct social entities to one another
  - Interest in understanding how social structure form and evolve
- Attempt to represent the structure in social relations via networks
  - the data is conceptualized as a realization of a network model

Network modeling from a statistical perspective

- Networks are widely used to represent data on relations between interacting actors or nodes.
- The study of social networks is multi-disciplinary
  - plethora of terminologies
  - varied objectives, multitude of frameworks
- Understanding the structure of social relations has been the focus of the social sciences
  - *social structure:* a system of social relations tying distinct social entities to one another
  - Interest in understanding how social structure form and evolve
- Attempt to represent the structure in social relations via networks
  - the data is conceptualized as a realization of a network model
- The data are of at least three forms:
  - individual-level information on the social entities
  - relational data on pairs of entities
  - population-level data

Network modeling from a statistical perspective

- Networks are widely used to represent data on relations between interacting actors or nodes.
- The study of social networks is multi-disciplinary
  - plethora of terminologies
  - varied objectives, multitude of frameworks
- Understanding the structure of social relations has been the focus of the social sciences
  - *social structure:* a system of social relations tying distinct social entities to one another
  - Interest in understanding how social structure form and evolve
- Attempt to represent the structure in social relations via networks
  - the data is conceptualized as a realization of a network model
- The data are of at least three forms:
  - individual-level information on the social entities
  - relational data on pairs of entities
  - population-level data

Deep literatures available

- Social networks community (Heider 1946; Frank 1972; Holland and Leinhardt 1981)

Deep literatures available

- Social networks community (Heider 1946; Frank 1972; Holland and Leinhardt 1981)
- Statistical Networks Community (Frank and Strauss 1986; Snijders 1997)

Deep literatures available

- Social networks community (Heider 1946; Frank 1972; Holland and Leinhardt 1981)
- Statistical Networks Community (Frank and Strauss 1986; Snijders 1997)
- Spatial Statistics Community (Besag 1974)

Deep literatures available

- Social networks community (Heider 1946; Frank 1972; Holland and Leinhardt 1981)
- Statistical Networks Community (Frank and Strauss 1986; Snijders 1997)
- Spatial Statistics Community (Besag 1974)
- Statistical Exponential Family Theory (Barndorff-Nielsen 1978)

Deep literatures available

- Social networks community (Heider 1946; Frank 1972; Holland and Leinhardt 1981)
- Statistical Networks Community (Frank and Strauss 1986; Snijders 1997)
- Spatial Statistics Community (Besag 1974)
- Statistical Exponential Family Theory (Barndorff-Nielsen 1978)
- Graphical Modeling Community (Lauritzen and Spiegelhalter 1988, . . .)

Deep literatures available

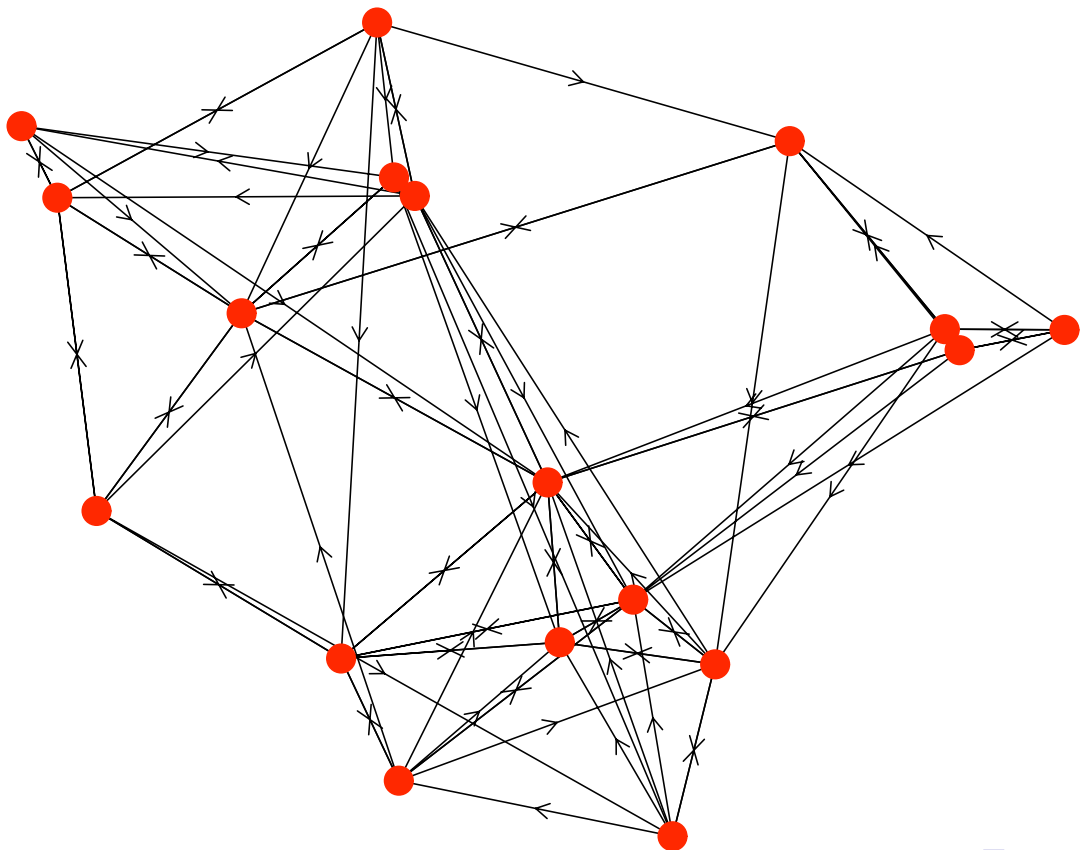
- Social networks community (Heider 1946; Frank 1972; Holland and Leinhardt 1981)
- Statistical Networks Community (Frank and Strauss 1986; Snijders 1997)
- Spatial Statistics Community (Besag 1974)
- Statistical Exponential Family Theory (Barndorff-Nielsen 1978)
- Graphical Modeling Community (Lauritzen and Spiegelhalter 1988, . . .)
- Machine Learning Community (Jordan, Jensen, Xing, ...)

Deep literatures available

- Social networks community (Heider 1946; Frank 1972; Holland and Leinhardt 1981)
- Statistical Networks Community (Frank and Strauss 1986; Snijders 1997)
- Spatial Statistics Community (Besag 1974)
- Statistical Exponential Family Theory (Barndorff-Nielsen 1978)
- Graphical Modeling Community (Lauritzen and Spiegelhalter 1988, . . .)
- Machine Learning Community (Jordan, Jensen, Xing, ...)
- Physics and Applied Math (Newman, Watts, . . .)

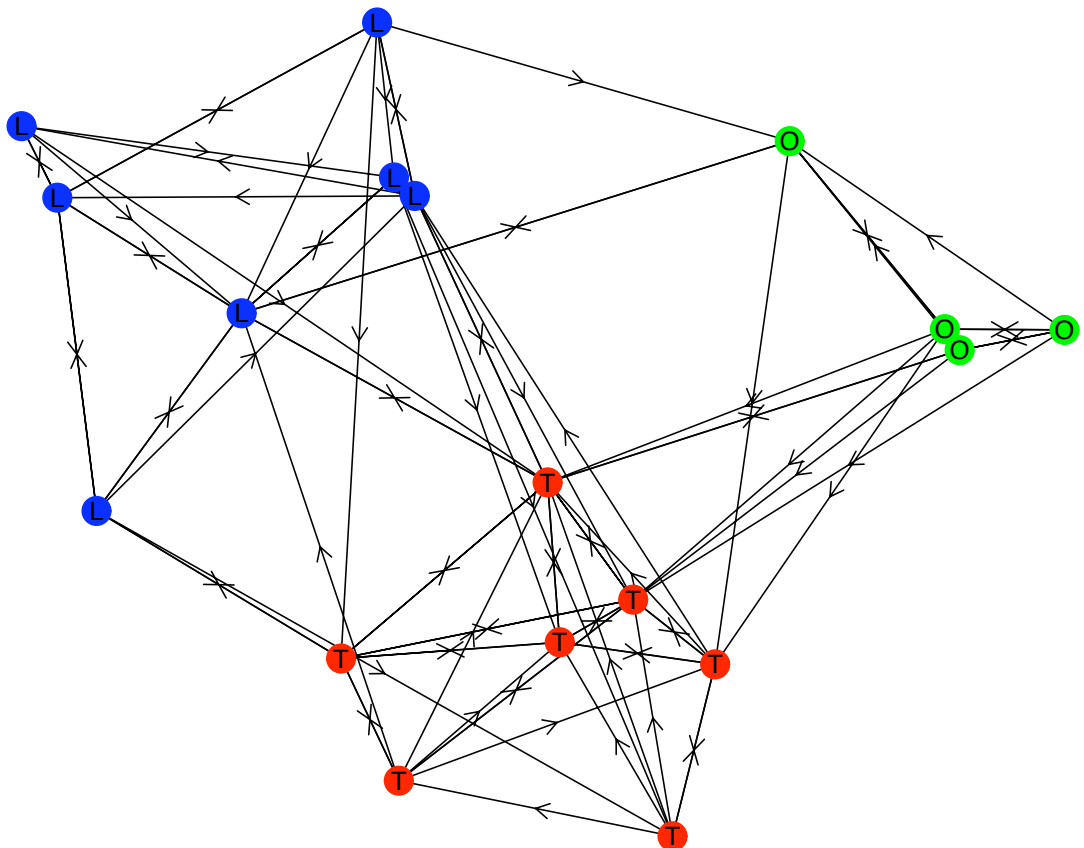
Example of Social Relationships between Monks

- Expressed "liking" between 18 monks within an isolated monastery
  $\Rightarrow$ Sampson (1969)
  - A directed relationship aggregated over a 12 month period before the breakup of the cloister.



Example of Social Relationships between Monks

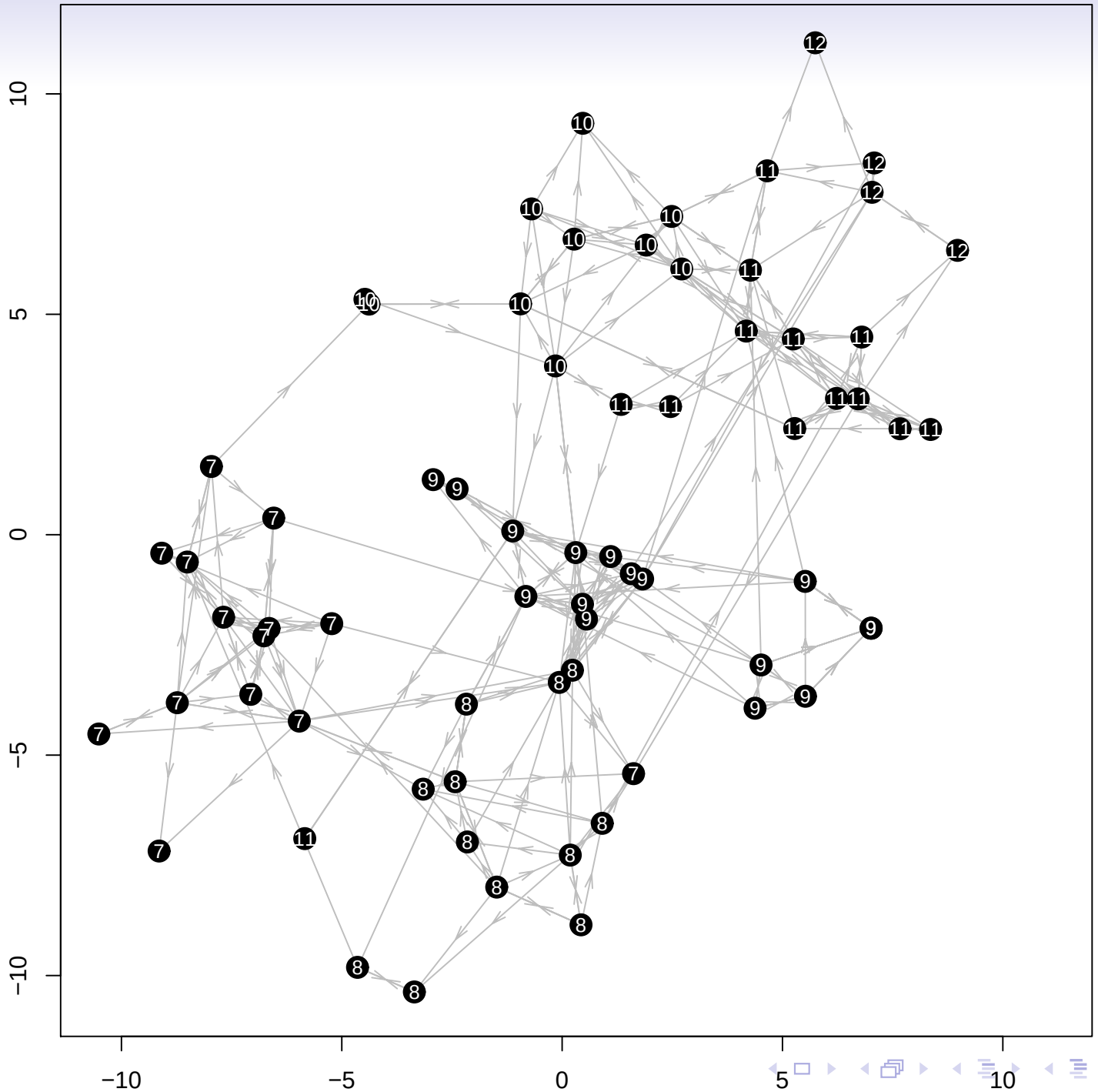
- Expressed "liking" between 18 monks within an isolated monastery
  $\Rightarrow$ Sampson (1969)
  - A directed relationship aggregated over a 12 month period before the breakup of the cloister.
- Sampson identified three groups plus:
  (T)urks, (L)oyal Opposition, (O)utcasts and (W)averers



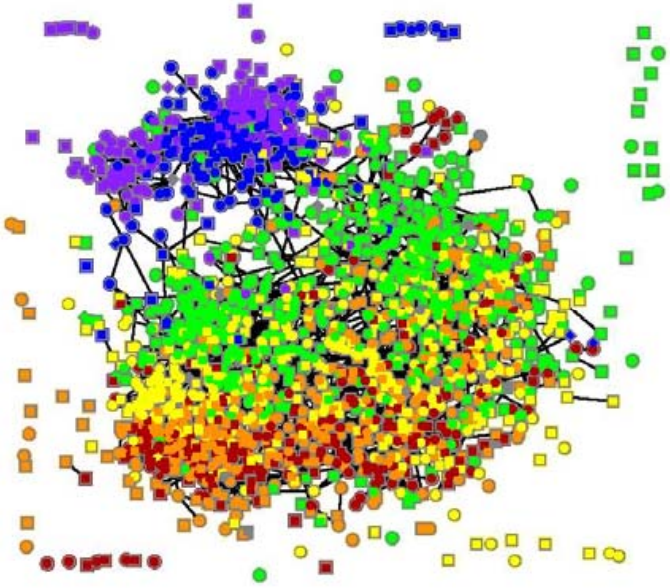
Examples of Friendship Relationships

Examples of Friendship Relationships

- The National Longitudinal Study of Adolescent Health
  $\Rightarrow$ `www.cpc.unc.edu/projects/addhealth`
  - "Add Health" is a school-based study of the health-related behaviors of adolescents in grades 7 to 12.
- Each nominated up to 5 boys and 5 girls as their friends
- 160 schools: Smallest has 69 adolescents in grades 7–12


10
5
0
−5
−10
−10
−5
0
5
10

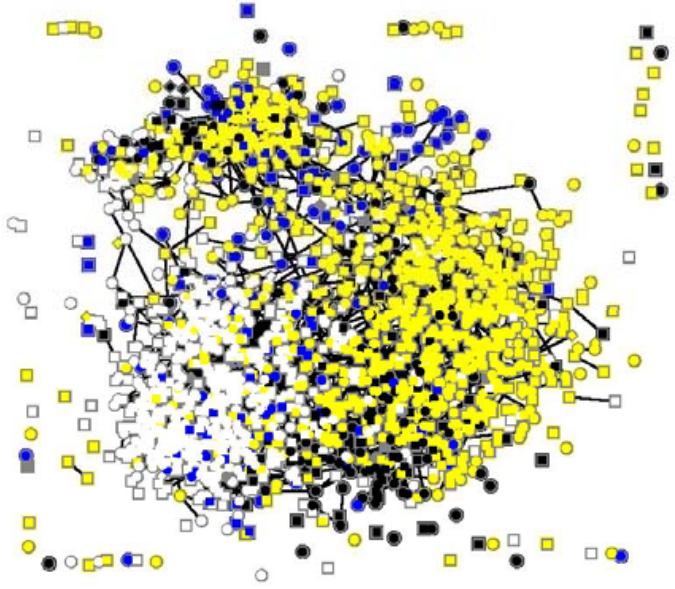
School Community Stratum 44
mutual friendships by Grade



2209 Students

- **Grade 7**
- **Grade 8**
- **Grade 9**
- **Grade 10**
- **Grade 11**

School Community Stratum 44
mutual friendships by Race



2209 Students

- **White** (non-Hispanic)
- **Black** (non-Hispanic)
- **Hispanic** (of any race)
- **Asian / Native Am / Other** (non-Hispanic)
- **Race NA**

Features of Many Social Networks

Features of Many Social Networks

- *Mutuality* of ties

Features of Many Social Networks

- *Mutuality* of ties
- *Individual heterogeneity* in the propensity to form ties

Features of Many Social Networks

- *Mutuality* of ties
- *Individual heterogeneity* in the propensity to form ties
- *Homophily* by actor attributes

  $\Rightarrow$ Lazarsfeld and Merton, 1954; Freeman, 1996; McPherson et al., 2001

  - higher propensity to form ties between actors with similar attributes
    e.g., age, gender, geography, major, social-economic status
  - attributes may be observed or unobserved

Features of Many Social Networks

- *Mutuality* of ties
- *Individual heterogeneity* in the propensity to form ties
- *Homophily* by actor attributes

  $\Rightarrow$ Lazarsfeld and Merton, 1954; Freeman, 1996; McPherson et al., 2001

  - higher propensity to form ties between actors with similar attributes
    e.g., age, gender, geography, major, social-economic status
  - attributes may be observed or unobserved
- *Transitivity* of relationships
  - friends of friends have a higher propensity to be friends

Features of Many Social Networks

- *Mutuality* of ties
- *Individual heterogeneity* in the propensity to form ties
- *Homophily* by actor attributes

  $\Rightarrow$ Lazarsfeld and Merton, 1954; Freeman, 1996; McPherson et al., 2001

  - higher propensity to form ties between actors with similar attributes
    e.g., age, gender, geography, major, social-economic status
  - attributes may be observed or unobserved
- *Transitivity* of relationships
  - friends of friends have a higher propensity to be friends
- *Balance* of relationships $\Rightarrow$ Heider (1946)
  - people feel comfortable if they agree with others whom they like

Features of Many Social Networks

- *Mutuality* of ties
- *Individual heterogeneity* in the propensity to form ties
- *Homophily* by actor attributes

  $\Rightarrow$ Lazarsfeld and Merton, 1954; Freeman, 1996; McPherson et al., 2001

  - higher propensity to form ties between actors with similar attributes
    e.g., age, gender, geography, major, social-economic status
  - attributes may be observed or unobserved
- *Transitivity* of relationships
  - friends of friends have a higher propensity to be friends
- *Balance* of relationships $\Rightarrow$ Heider (1946)
  - people feel comfortable if they agree with others whom they like
- *Context* is important $\Rightarrow$ Simmel (1908)
  - triad, not the dyad, is the fundamental social unit

The Choice of Models depends on the objectives

- Primary interest in the nature of relationships:
  - How the behavior of individuals depends on their location in the social network
  - How the qualities of the individuals influence the social structure

The Choice of Models depends on the objectives

- Primary interest in the nature of relationships:
  - How the behavior of individuals depends on their location in the social network
  - How the qualities of the individuals influence the social structure
- Secondary interest is in how network structure influences processes that develop over a network
  - spread of HIV and other STDs
  - diffusion of technical innovations
  - spread of computer viruses

The Choice of Models depends on the objectives

- Primary interest in the nature of relationships:
  - How the behavior of individuals depends on their location in the social network
  - How the qualities of the individuals influence the social structure
- Secondary interest is in how network structure influences processes that develop over a network
  - spread of HIV and other STDs
  - diffusion of technical innovations
  - spread of computer viruses
- Tertiary interest in the effect of *interventions* on network structure and processes that develop over a network

Perspectives to keep in mind

- Network-specific versus Population-process
  - *Network-specific*: interest focuses only on the actual network under study
  - *Population-process*: the network is part of a population of networks and the latter is the focus of interest
    - the network is conceptualized as a realization of a social process

Statistical Models for Social Networks

Notation

A *social network* is defined as a set of n social "actors" and a social relationship between each pair of actors.

Statistical Models for Social Networks

Notation

A *social network* is defined as a set of n social "actors" and a social relationship between each pair of actors.

$$Y_{ij} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \\ 0 & \text{otherwise} \end{cases}$$

Statistical Models for Social Networks

Notation

A *social network* is defined as a set of n social "actors" and a social relationship between each pair of actors.

$$Y_{ij} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \\ 0 & \text{otherwise} \end{cases}$$

- call $Y \equiv [Y_{ij}]_{n \times n}$ a *sociomatrix*
  - a $N = n(n-1)$ binary array

Statistical Models for Social Networks

Notation

A *social network* is defined as a set of n social "actors" and a social relationship between each pair of actors.

$$Y_{ij} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \\ 0 & \text{otherwise} \end{cases}$$

- call $Y \equiv [Y_{ij}]_{n \times n}$ a *sociomatrix*
  - a $N = n(n-1)$ binary array
- The basic problem of stochastic modeling is to specify a distribution for Y i.e., $P(Y = y)$

A Framework for Network Modeling

Let $\mathcal{Y}$ be the sample space of Y e.g. $\{0,1\}^N$
Any model-class for the multivariate distribution of Y
can be *parametrized* in the form:

$$P_\eta(Y = y) = \frac{\exp\{\eta \cdot g(y)\}}{\kappa(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

Besag (1974), Frank and Strauss (1986)

- $\eta \in \Lambda \subset R^q$ q-vector of parameters
- $g(y)$ q-vector of *network statistics*.
  $\Rightarrow$ $g(Y)$ are jointly sufficient for the model
- For a "saturated" model-class $q = 2^{|\mathcal{Y}|} - 1$
- $\kappa(\eta, \mathcal{Y})$ distribution normalizing constant

$$\kappa(\eta, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp\{\eta \cdot g(y)\}$$

Simple model-classes for social networks

Homogeneous Bernoulli graph (Rényi-Erdős model)

- Y_{ij} are independent and equally likely
  with log-odds $\eta = \text{logit}[P_\eta(Y_{ij} = 1)]$

$$P_\eta(Y = y) = \frac{e^{\eta \sum_{i,j} y_{ij}}}{\kappa(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

  where $q = 1$, $g(y) = \sum_{i,j} y_{ij}$, $\kappa(\eta, \mathcal{Y}) = [1 + \exp(\eta)]^N$
- homogeneity means it is unlikely to be proposed as a model for real phenomena

Dyad-independence models with attributes

- Y_{ij} are independent but depend on dyadic covariates $x_{k,ij}$

$$P_\eta(Y = y) = \frac{e^{\sum_{k=1}^{q} \eta_k g_k(y)}}{\kappa(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

Dyad-independence models with attributes

- Y_{ij} are independent but depend on dyadic covariates $x_{k,ij}$

$$P_\eta(Y = y) = \frac{e^{\sum_{k=1}^{q} \eta_k g_k(y)}}{\kappa(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

$$g_k(y) = \sum_{i,j} x_{k,ij} y_{ij}, \qquad k = 1, \ldots, q$$

Dyad-independence models with attributes

- Y_{ij} are independent but depend on dyadic covariates $x_{k,ij}$

$$P_\eta(Y = y) = \frac{e^{\sum_{k=1}^{q} \eta_k g_k(y)}}{\kappa(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

$$g_k(y) = \sum_{i,j} x_{k,ij} y_{ij}, \qquad k = 1, \ldots, q$$

$$\kappa(\eta, \mathcal{Y}) = \prod_{i,j} [1 + \exp(\sum_{k=1}^{q} \eta_k x_{k,ij})]$$

Of course,

$$logit[P_\eta(Y_{ij} = 1)] = \sum_k \eta_k x_{k,ij}$$

Some history of exponential family models for social networks

Holland and Leinhardt (1981) proposed a general dyad independence model

– Also an homogeneous version they refer to as the "p^1" model

$$P_\eta(Y = y) = \frac{\exp\{\rho \sum_{i<j} y_{ij} y_{ji} + \phi y_{++} + \sum_i \alpha_i y_{i+} + \sum_j \beta_j y_{+j}\}}{\kappa(\rho, \alpha, \beta, \phi)}$$

where $\eta = (\rho, \alpha, \beta, \phi)$.

– ϕ controls the expected number of edges
– ρ represent the expected tendency toward *reciprocation*
– α_i *productivity* of node i; β_j *attractiveness* of node j

Some history of exponential family models for social networks

Holland and Leinhardt (1981) proposed a general dyad independence model

– Also an homogeneous version they refer to as the "p^1" model

$$P_\eta(Y = y) = \frac{\exp\{\rho \sum_{i<j} y_{ij} y_{ji} + \phi y_{++} + \sum_i \alpha_i y_{i+} + \sum_j \beta_j y_{+j}\}}{\kappa(\rho, \alpha, \beta, \phi)}$$

where $\eta = (\rho, \alpha, \beta, \phi)$.

– ϕ controls the expected number of edges
– ρ represent the expected tendency toward *reciprocation*
– α_i *productivity* of node i; β_j *attractiveness* of node j

- Much related work and generalizations

Classes of statistics used for modeling

Actor Markov statistics

$\Rightarrow$ Frank and Strauss (1986)

– motivated by notions of "symmetry" and "homogeneity"

Classes of statistics used for modeling

Actor Markov statistics

$\Rightarrow$ Frank and Strauss (1986)

- motivated by notions of "symmetry" and "homogeneity"
- Y_{ij} in Y that do not share an actor are conditionally independent given the rest of the network

Classes of statistics used for modeling

Actor Markov statistics

$\Rightarrow$ Frank and Strauss (1986)

- motivated by notions of "symmetry" and "homogeneity"
- Y_{ij} in Y that do not share an actor are conditionally independent given the rest of the network

$\Rightarrow$ analogous to nearest neighbor ideas in spatial statistics

Classes of statistics used for modeling

Actor Markov statistics

$\Rightarrow$ Frank and Strauss (1986)

- motivated by notions of "symmetry" and "homogeneity"
- Y_{ij} in Y that do not share an actor are conditionally independent given the rest of the network

$\Rightarrow$ analogous to nearest neighbor ideas in spatial statistics

- Degree distribution: $\mathrm{d}_k(y) =$ proportion of actors of degree k in y.

Classes of statistics used for modeling

Actor Markov statistics

$\Rightarrow$ Frank and Strauss (1986)

- motivated by notions of "symmetry" and "homogeneity"
- Y_{ij} in Y that do not share an actor are conditionally independent given the rest of the network

$\Rightarrow$ analogous to nearest neighbor ideas in spatial statistics

- Degree distribution: $\mathrm{d}_k(y) =$ proportion of actors of degree k in y.
- k-star distribution: $\mathrm{s}_k(y) =$ proportion of k-stars in the graph y.
  (In particular,
  $\mathrm{s}_1 =$ proportion of edges that exist between pairs of actors.)

Classes of statistics used for modeling

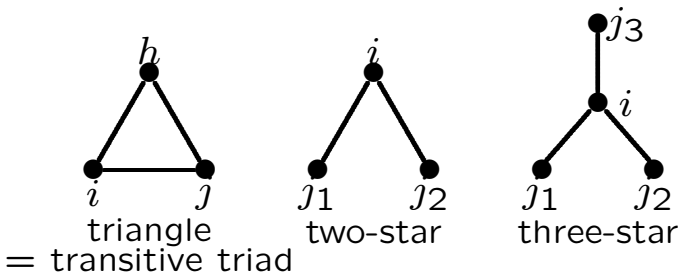
Actor Markov statistics

$\Rightarrow$ Frank and Strauss (1986)

- motivated by notions of "symmetry" and "homogeneity"
- Y_{ij} in Y that do not share an actor are conditionally independent given the rest of the network

$\Rightarrow$ analogous to nearest neighbor ideas in spatial statistics

- Degree distribution: $\mathrm{d}_k(y)$ = proportion of actors of degree k in y.
- k-star distribution: $\mathrm{s}_k(y)$ = proportion of k-stars in the graph y.
  (In particular,
  s_1 = proportion of edges that exist between pairs of actors.)
- triangles:
  $\mathrm{t}_1(y)$ = proportion of triads that from a complete sub-graph in y.



Other statistics motivated by conditional independence

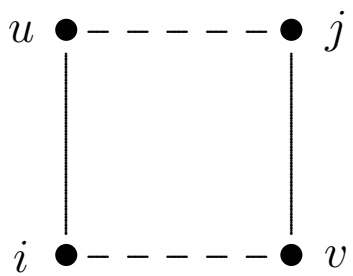
$\Rightarrow$ Pattison and Robins (2002), Butts (2005)
$\Rightarrow$ Snijders, Pattison, Robins and Handcock (2004)

- Y_{uj} and Y_{iv} in Y are conditionally independent given the rest of the network if they could not produce a cycle in the network

Other statistics motivated by conditional independence

$\Rightarrow$ Pattison and Robins (2002), Butts (2005)
$\Rightarrow$ Snijders, Pattison, Robins and Handcock (2004)

– Y_{uj} and Y_{iv} in Y are conditionally independent given the rest of the network if they could not produce a cycle in the network



Partial conditional dependence when four-cycle is created

This produces statistics of the form:

- edgewise shared partner distribution: $\mathrm{esp}_k(y) =$ proportion of edges between actors with exactly k shared partners $k = 0, 1, \ldots$

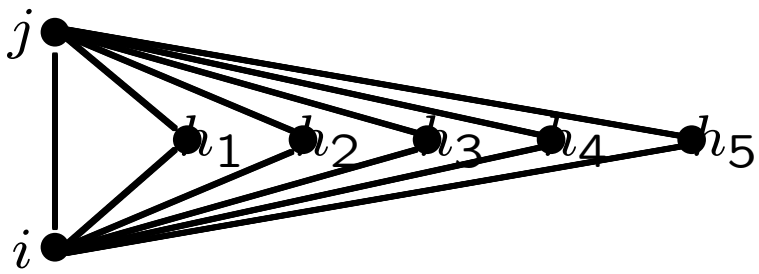


Figure: The actors in the non-directed (i, j) edge have 5 shared partners

- dyadwise shared partner distribution: $\mathrm{dsp}_k(y) =$ proportion of dyads with exactly k shared partners $k = 0, 1, \ldots$

Structural Signatures

- identify social constructs or features
- based on intuitive notions or partial appeal to substantive theory

Structural Signatures

– identify social constructs or features
– based on intuitive notions or partial appeal to substantive theory

- Clusters of edges are often *transitive*:
  Recall $t_1(y)$ is the proportion of triangles amongst triads

$$t_1(y) = \frac{1}{\binom{g}{3}} \sum_{\{i,j,k\} \in \binom{g}{3}} y_{ij} y_{ik} y_{jk}$$

Structural Signatures

– identify social constructs or features
– based on intuitive notions or partial appeal to substantive theory

- Clusters of edges are often *transitive*:
  Recall $t_1(y)$ is the proportion of triangles amongst triads

$$t_1(y) = \frac{1}{\binom{g}{3}} \sum_{\{i,j,k\} \in \binom{g}{3}} y_{ij} y_{ik} y_{jk}$$

A closely related quantity is the
proportion of triangles amongst 2-stars

$$C(y) = \frac{3 \times t_1(y)}{s_2(y)}$$

Structural Signatures

– identify social constructs or features
– based on intuitive notions or partial appeal to substantive theory

- Clusters of edges are often *transitive*:
  Recall $t_1(y)$ is the proportion of triangles amongst triads

$$t_1(y) = \frac{1}{\binom{g}{3}} \sum_{\{i,j,k\} \in \binom{g}{3}} y_{ij} y_{ik} y_{jk}$$

A closely related quantity is the
proportion of triangles amongst 2-stars

$$C(y) = \frac{3 \times t_1(y)}{s_2(y)}$$

Also called *mean clustering coefficient*

Example: A simple model-class with transitivity

$n = 50$ actors $\qquad$ $N = 1225$ pairs $\qquad$ 10^{369} graphs

$$P(Y = y) = \frac{\exp\{\eta_1 E(y) + \eta_2 C(y)\}}{\kappa(\eta_1, \eta_2)} \qquad y \in \mathcal{Y}$$

where

$E(x)$ is the density of edges $(0 - 1)$
$C(x)$ is the triangle percent $(0 - 100)$

- If we set the density of the graph to have about 50 edges then the expected triangle percent is 3.8%
- Suppose we set the triangle percent large to reflect transitivity in the graph: 38%

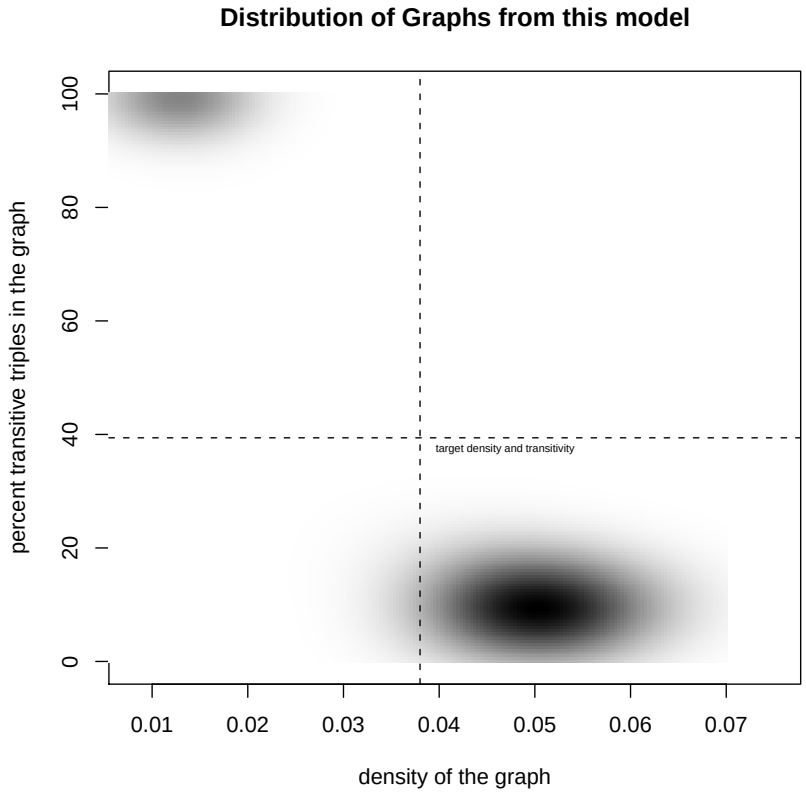
How can we tell if the model is useful?

- Does this model capture transitivity and density in a flexible way?

How can we tell if the model is useful?

- Does this model capture transitivity and density in a flexible way?

- By construction, on average, graphs from this model have average density 4% and average triangle percent 38%
- If the model is a good representation of transitivity and density we expect the graphs drawn from the model to be close to these values.
- What do graphs produced by this model look like?


Distribution of Graphs from this model
percent transitive triples in the graph
0
20
40
60
80
100
target density and transitivity
0.01
0.02
0.03
0.04
0.05
0.06
0.07
density of the graph

Curved Exponential Family Models

Suppose that η is modeled as a function of a lower dimensional parameter: $\theta \in R^p$

$$P(Y = y) = \frac{\exp\{\eta(\theta) \cdot g(y)\}}{\kappa(\theta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

Hunter and Handcock (2004)

Curved Exponential Family Models

Suppose that η is modeled as a function of a lower dimensional parameter: $\theta \in R^p$

$$P(Y = y) = \frac{\exp\{\eta(\theta) \cdot g(y)\}}{\kappa(\theta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

Hunter and Handcock (2004)

Suppose we focus on a model for network degree distribution and clustering

$$\log\left[P_\theta(Y = y)\right] = \eta(\phi) \cdot d(y) + \nu C(y) - \log c(\phi, \nu, \mathcal{Y}), \qquad (1)$$

where $d(x) = \{d_1(x), \ldots, d_{n-1}(x)\}$ are the network degree distribution counts.

Curved Exponential Family Models

Suppose that η is modeled as a function of a lower dimensional parameter: $\theta \in R^p$

$$P(Y = y) = \frac{\exp\{\eta(\theta) \cdot g(y)\}}{\kappa(\theta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

Hunter and Handcock (2004)

Suppose we focus on a model for network degree distribution and clustering

$$\log\left[P_\theta(Y = y)\right] = \eta(\phi) \cdot d(y) + \nu C(y) - \log c(\phi, \nu, \mathcal{Y}), \qquad (1)$$

where $d(x) = \{d_1(x), \ldots, d_{n-1}(x)\}$ are the network degree distribution counts.
Any degree distribution can be specified by $n - 1$ or less independent parameters.

Statistical Inference for η

Base inference on the loglikelihood function,

$$\ell(\eta) = \eta \cdot g(y_{\text{obs}}) - \log \kappa(\eta)$$

$$\kappa(\eta) = \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\eta \cdot g(z)\}$$

Mean-value representation of the model

Let $P_\nu(K = k)$ be the PMF of K, the number of ties that a randomly chosen node in the network has.
An alternative parameterization: (ϕ, ρ) where the mapping is:

$$\rho = \mathbf{E}_{\phi,\rho}\left[C(X)\right] = \sum_{y \in \mathcal{Y}} C(y) \exp\left[\eta(\phi) \cdot d(y) + \nu C(y)\right] \geq 0 \qquad (2)$$

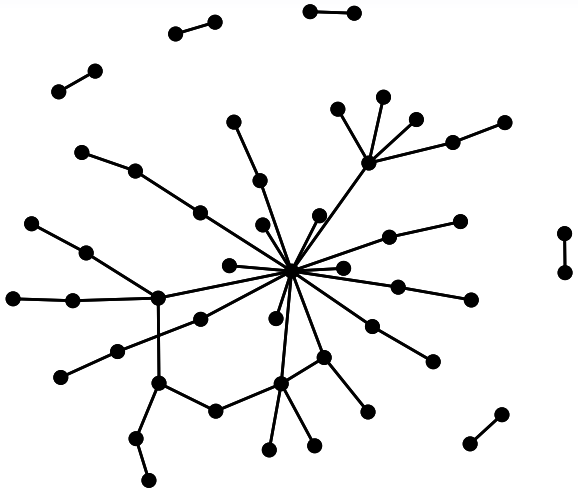
$$P_\nu(K = k) = \mathbf{E}_{\phi,\rho}\left[d_k(Y)\right] \qquad k = 0, \ldots, n-1 \qquad (3)$$

– ρ is the mean clustering coefficient over networks in $\mathcal{Y}$.
– ν controls the parametrization of the degree distribution

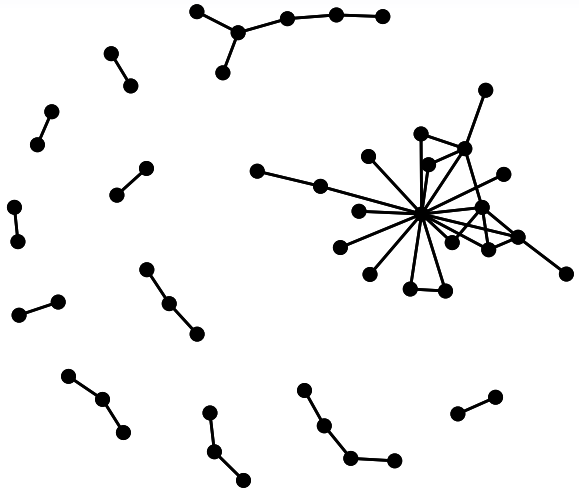
Illustrations of good models within this model-class

- village-level structure
  - $n = 50$
  - mean clustering coefficient $= 15\%$ – degree distribution: Yule with scaling exponent 3.
- larger-level structure
  - $n = 1000$
  - mean clustering coefficient $= 15\%$ – degree distribution: Yule with scaling exponent 3.
- Attribute mixing
  - Two-sex populations
  - mean clustering coefficient $= 15\%$ – degree distribution: Yule with scaling exponent 3.

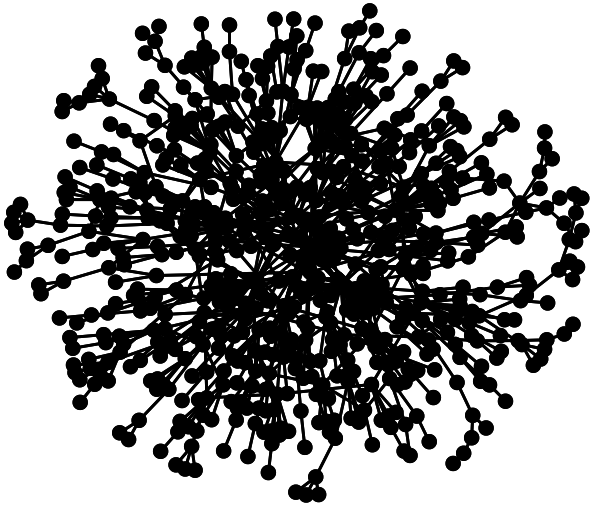
Yule with zero clustering coefficient conditional on degree



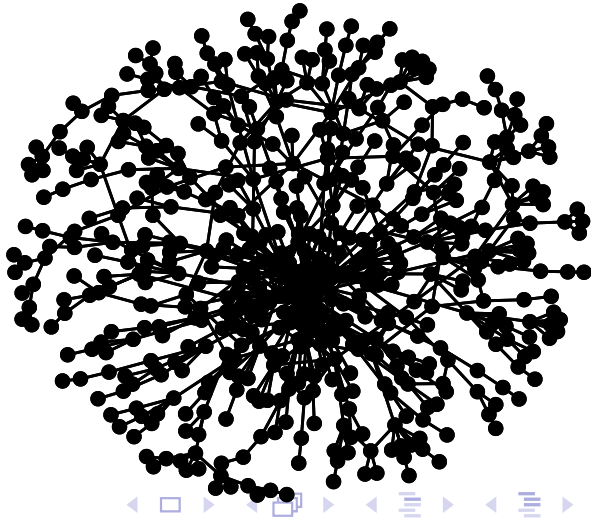
Yule with clustering coefficient 15%



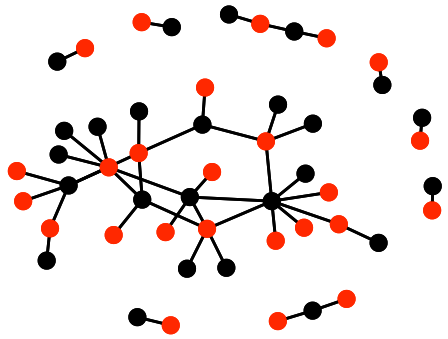
Yule with zero clustering coefficient conditional on degree



Yule with clustering coefficient 15%

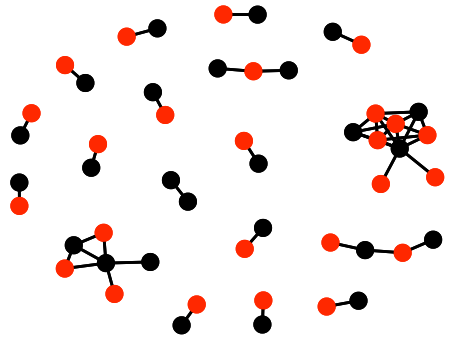


Heterosexual Yule with no correlation



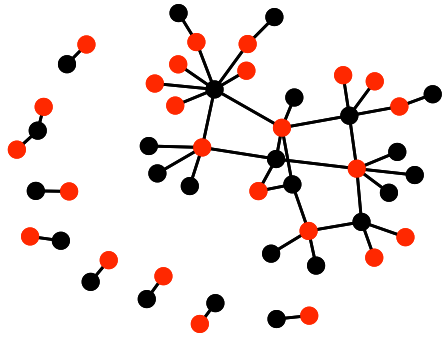
tripercent = 3

Heterosexual Yule with strong correlation

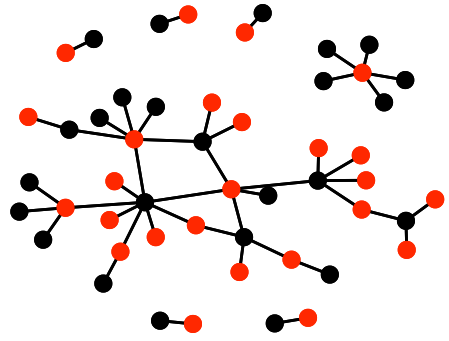


tripercent = 60.6

Heterosexual Yule with modest correlation



Heterosexual Yule with negative correlation



Application to a Protein-Protein Interaction Network

- By interact is meant that two amino acid chains were experimentally identified to bind to each other.
- The network is for *E. Coli* and is drawn from the "Database of Interacting Proteins (DIP)" `http://dip.doe-mbi.ucla.edu`
- For simplicity we focus on proteins that interact with themselves and have at least one other interaction
  – 108 proteins and 94 interactions.

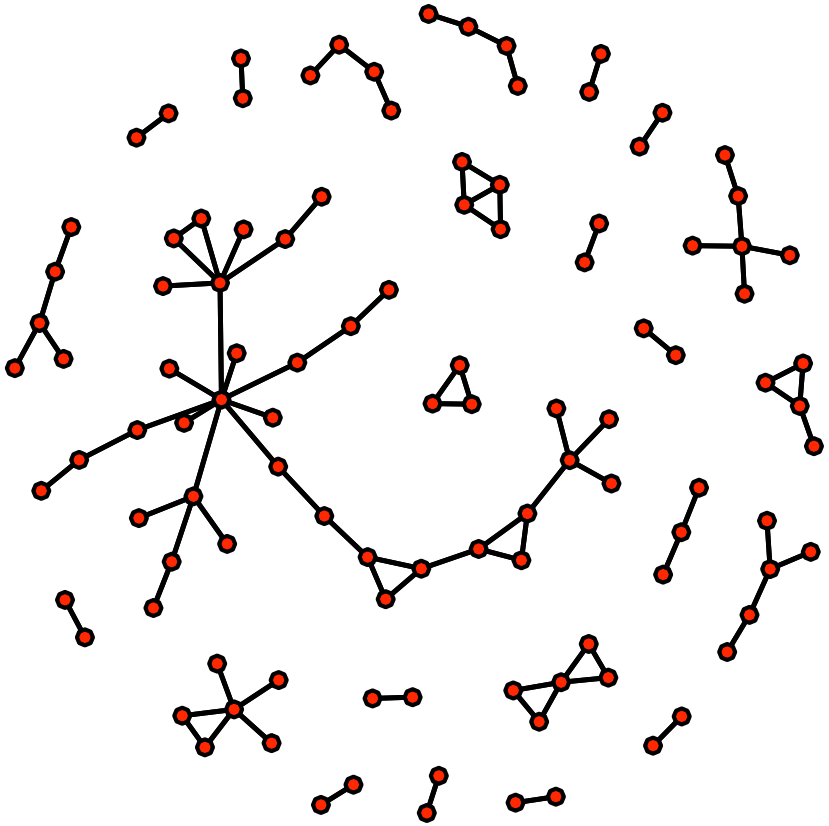


Figure: A protein - protein interaction network for *E. Coli*. The nodes represent proteins and the ties indicate that the two proteins are known to interact with each other.

Statistical Inference and Simulation

- Simulate using a Metropolis-Hastings algorithm (Handcock 2002).
- Here base inference on the likelihood function
- For computational reasons, approximate the likelihood via Markov Chain Monte Carlo (MCMC)
- Use maximum likelihood estimates (Geyer and Thompson 1992)

Parameter	est.	s.e.
Scaling decay rate (ϕ)	3.034	0.3108
Correlation Coefficient (ν)	1.176	0.1457

Table: MCMC maximum likelihood parameter estimates for the protein-protein interaction network.

Approximating the loglikelihood

Approximating the loglikelihood

- Suppose $Y_1, Y_2, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} P_{\eta_0}(Y = y)$ for some η_0.
- Using the LOLN, the difference in log-likelihoods is

$$\ell(\eta) - \ell(\eta_0) \quad = \quad \log \frac{\kappa(\eta_0)}{\kappa(\eta)}$$

Approximating the loglikelihood

- Suppose $Y_1, Y_2, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} P_{\eta_0}(Y = y)$ for some η_0.
- Using the LOLN, the difference in log-likelihoods is

$$\begin{aligned} \ell(\eta) - \ell(\eta_0) &= \log \frac{\kappa(\eta_0)}{\kappa(\eta)} \\ &= \log \mathbf{E}_{\eta_0} \left(\exp \left\{ (\eta_0 - \eta) \cdot g(Y) \right\} \right) \end{aligned}$$

Approximating the loglikelihood

- Suppose $Y_1, Y_2, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} P_{\eta_0}(Y = y)$ for some η_0.
- Using the LOLN, the difference in log-likelihoods is

$$\begin{aligned}
\ell(\eta) - \ell(\eta_0) &= \log \frac{\kappa(\eta_0)}{\kappa(\eta)} \\
&= \log \mathbf{E}_{\eta_0} \left(\exp \left\{ (\eta_0 - \eta) \cdot g(Y) \right\} \right) \\
&\approx \log \frac{1}{M} \sum_{i=1}^{M} \exp \left\{ (\eta_0 - \eta) \cdot (g(Y_i) - g(y_{\text{obs}})) \right\} \\
&\equiv \tilde{\ell}(\eta) - \tilde{\ell}(\eta_0).
\end{aligned}$$

Approximating the loglikelihood

- Suppose $Y_1, Y_2, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} P_{\eta_0}(Y = y)$ for some η_0.
- Using the LOLN, the difference in log-likelihoods is

$$\begin{aligned}
\ell(\eta) - \ell(\eta_0) &= \log \frac{\kappa(\eta_0)}{\kappa(\eta)} \\
&= \log \mathbf{E}_{\eta_0} \left(\exp \left\{ (\eta_0 - \eta) \cdot g(Y) \right\} \right) \\
&\approx \log \frac{1}{M} \sum_{i=1}^{M} \exp \left\{ (\eta_0 - \eta) \cdot (g(Y_i) - g(y_{\text{obs}})) \right\} \\
&\equiv \tilde{\ell}(\eta) - \tilde{\ell}(\eta_0).
\end{aligned}$$

- Simulate $Y_1, Y_2, \ldots, Y_m$ using a MCMC (Metropolis-Hastings) algorithm $\Rightarrow$ Handcock (2002).

Approximating the loglikelihood

- Suppose $Y_1, Y_2, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} P_{\eta_0}(Y = y)$ for some η_0.
- Using the LOLN, the difference in log-likelihoods is

$$\begin{aligned}
\ell(\eta) - \ell(\eta_0) &= \log \frac{\kappa(\eta_0)}{\kappa(\eta)} \\
&= \log \mathbf{E}_{\eta_0} \left(\exp \left\{ (\eta_0 - \eta) \cdot g(Y) \right\} \right) \\
&\approx \log \frac{1}{M} \sum_{i=1}^{M} \exp \left\{ (\eta_0 - \eta) \cdot (g(Y_i) - g(y_{\text{obs}})) \right\} \\
&\equiv \tilde{\ell}(\eta) - \tilde{\ell}(\eta_0).
\end{aligned}$$

- Simulate $Y_1, Y_2, \ldots, Y_m$ using a MCMC (Metropolis-Hastings) algorithm $\Rightarrow$ Handcock (2002).
- Approximate the MLE $\hat{\eta} = \text{argmax}_\eta \{\tilde{\ell}(\eta) - \tilde{\ell}(\eta_0)\}$ (MC-MLE) $\Rightarrow$ Geyer and Thompson (1992)

Modeling Network Dynamics

- Suppose we wish to represent the dynamics at $t = 0, 1, \ldots, T$ time points

$$Y_{ijt} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

Modeling Network Dynamics

- Suppose we wish to represent the dynamics at $t = 0, 1, \ldots, T$ time points

$$Y_{ijt} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

- Need a model that
  - has the correct cross-sectional statistics
  - has the correct durations for relationships
  - realistic dissolution and formation of relationships

A Naive Model for Longitudinal Network Data

Consider a dynamic variant of the cross-sectional ERGM:

$$P_\eta(Y_{t+1} = y_{t+1} | Y_t = y_t) = \frac{\exp\left(\eta_{t+1} \cdot g\left(y_{t+1)}; y_t\right)\right)}{\sum_{s \in \mathcal{Y}} \exp\left(\eta_{t+1} \cdot g\left(x; y_t\right)\right)} \quad t = 2, \ldots, T$$

where $g_k(y_{t+1}; y_t)$ are statistics formed from y_{t+1} given y_t

– Robins and Pattison (2000) Discrete temporal ERGM
– Morris and Handcock (2001) Discrete temporal ERGM
– Hanneke and Xing (2006) Discrete temporal ERGM
– Guo, Hanneke, Fu and Xing (2007)] Hidden temporal ERGM

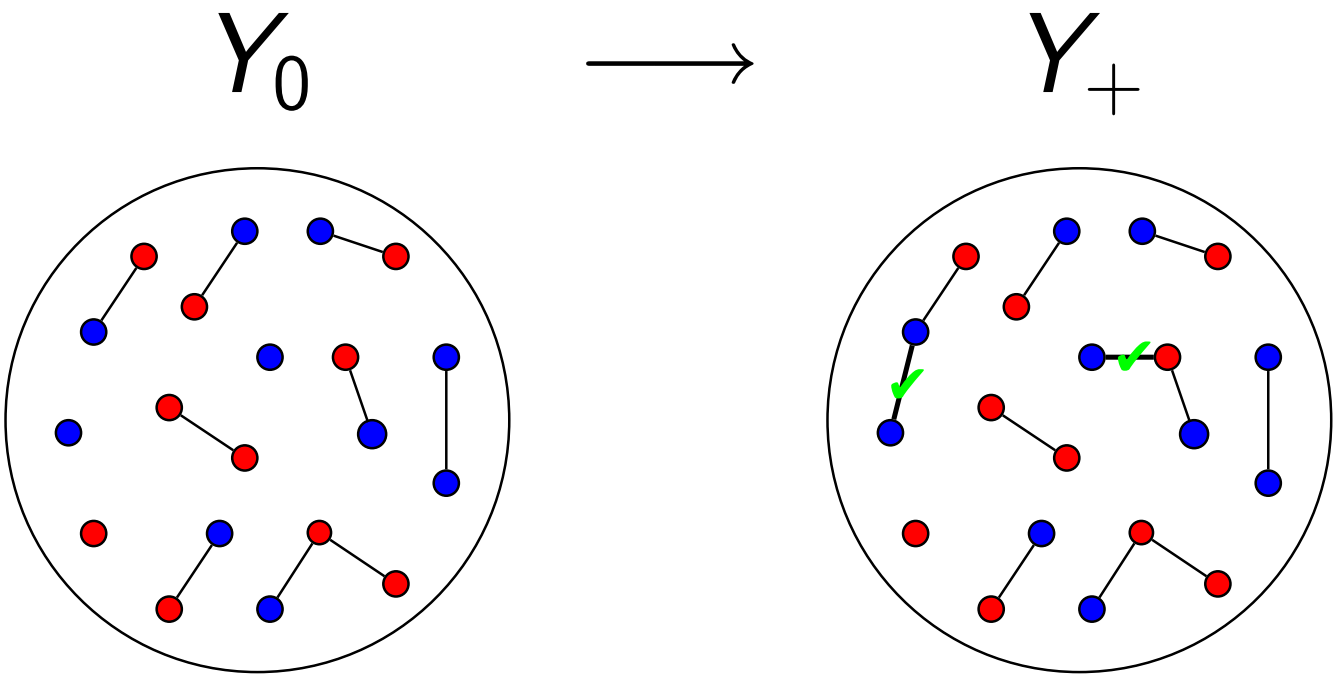
Two-Phase Dynamic Model

Consider a Markovian model with transition probabilities from Y_t to Y_{t+1}

governed by simultaneous
dissolution and *formation*
phases

Formation Phase

$$\Pr(Y_+ = y_+ | Y_0 = y_0; \beta) = \frac{e^{\beta \cdot g_+(y_+, y_0)} 1_{y_+ \supseteq y_0}}{c_+(\beta, y_0)}, \; y_+ \in \mathcal{Y}$$

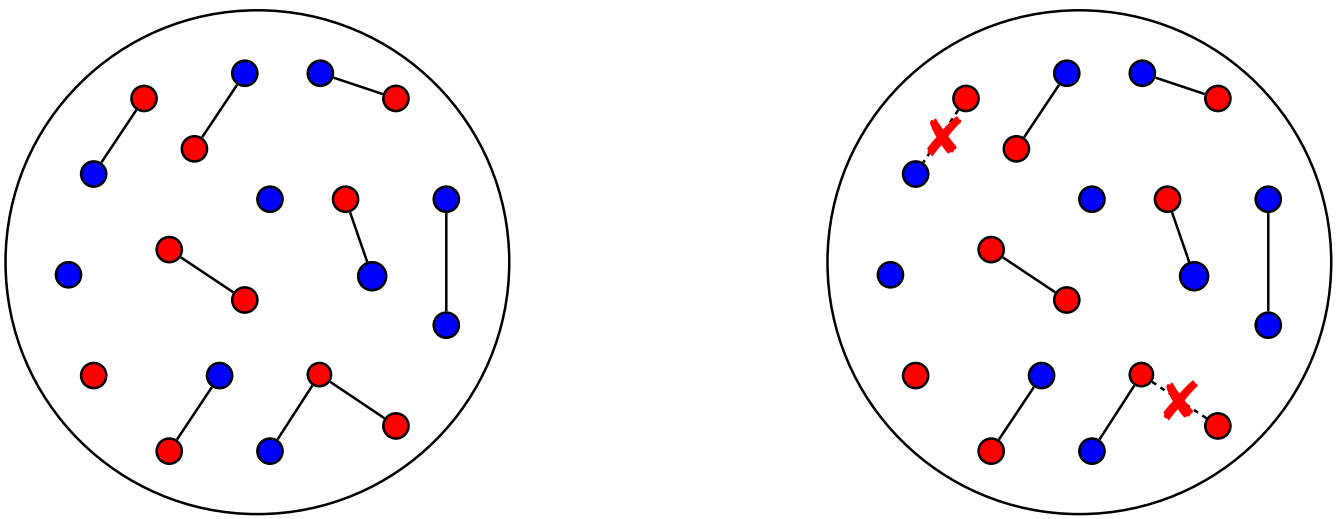


β completely controls the *incidence*

Dissolution Phase

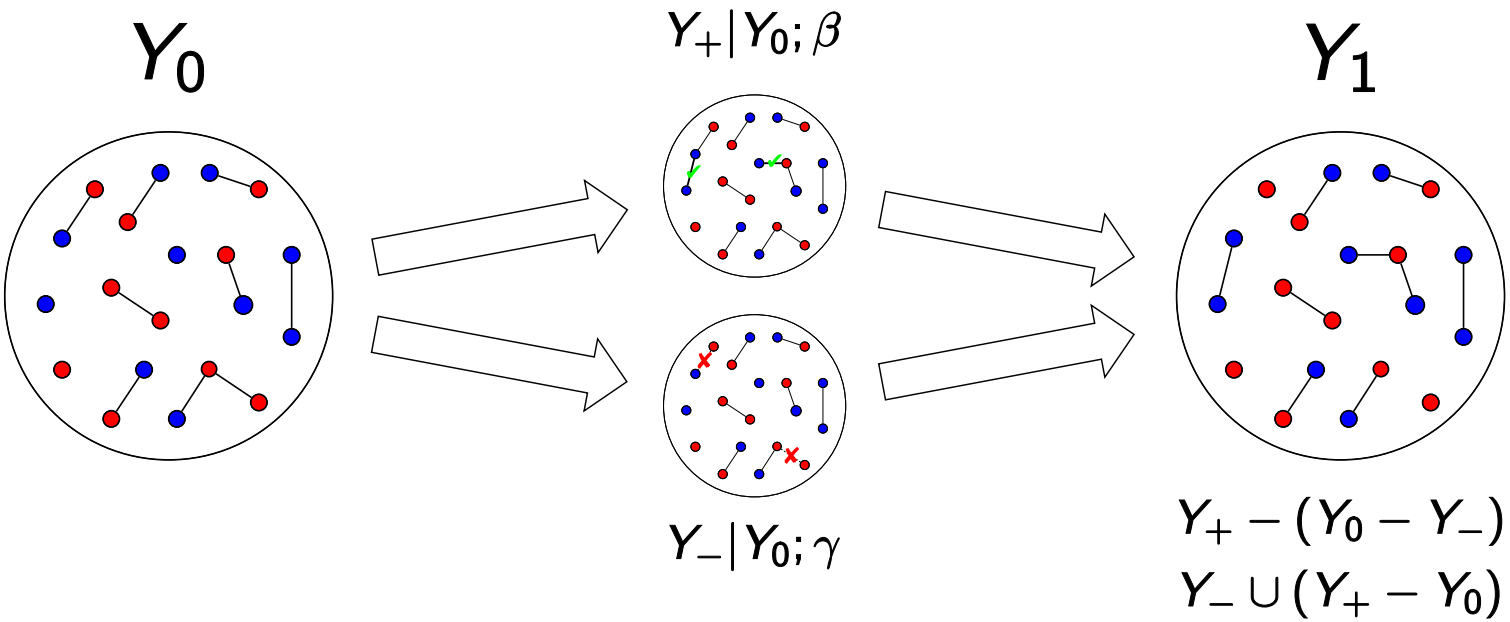
$$\Pr(Y_- = y_- | Y_0 = y_0; \gamma) = \frac{e^{\gamma \cdot g_-(y_-, y_0)} 1_{y_- \subseteq y_0}}{c_-(\gamma, y_0)}, \; y_- \in \mathcal{Y}$$

$Y_0 \longrightarrow Y_-$



γ complete controls the durations of partnerships

Simultaneous Formation and Dissolution



Simultaneous Formation and Dissolution

Transition Probability

$$\Pr(Y_1 = y_1 | Y_0 = y_0; \beta, \gamma) = p_-(y_1 \cap y_0 | y_0; \gamma) \times p_+(y_1 \cup y_0 | y_0; \beta)$$

Markov Process

$$Y_0 \longrightarrow Y_1 \longrightarrow Y_2 \longrightarrow Y_3 \longrightarrow \ldots$$

Equilibrium Distribution

$$Y_t \xrightarrow{\mathsf{D}} Y \sim \Pr(Y = y; \beta, \gamma)$$

Back to Prevalence

$$\text{Prevalence} = \text{Incidence} \times \text{Duration}$$

Back to Prevalence

$$\begin{array}{ccccc} \text{Prevalence} & = & \text{Incidence} & \times & \text{Duration} \\ & & \| & & \\ & & \text{Formation} & & \end{array}$$

Back to Prevalence

$$\text{Prevalence} = \text{Incidence} \times \begin{array}{c}\text{Duration} \\ || \\ \text{Dissolution}\end{array}$$

Back to Prevalence

$$\begin{array}{ccccc} \text{Prevalence} & = & \text{Incidence} & \times & \text{Duration} \\ \| & & & & \\ \text{Equilibrium} & & & & \end{array}$$

Back to Prevalence

$$\begin{array}{ccccc} \text{Prevalence} & = & \text{Incidence} & \times & \text{Duration} \\ \| & & \| & & \| \\ \text{Equilibrium} & & \text{Formation} & & \text{Dissolution} \end{array}$$

Application to the Dynamics of HIV Spread

- *Data*: The National Longitudinal Study of Adolescent Health
  - Wave III (with retrospective duration information)
  - Take into account age, sex, race (white/non-white) and age-sex "mixing" patterns

Application to the Dynamics of HIV Spread

- *Data*: The National Longitudinal Study of Adolescent Health
  - Wave III (with retrospective duration information)
  - Take into account age, sex, race (white/non-white) and age-sex "mixing" patterns
- Estimate the parameters of the model based on the likelihood.

Application to the Dynamics of HIV Spread

- *Data*: The National Longitudinal Study of Adolescent Health
  - Wave III (with retrospective duration information)
  - Take into account age, sex, race (white/non-white) and age-sex "mixing" patterns
- Estimate the parameters of the model based on the likelihood.
- Consider a (quasi)population of 10000 people (about half men and women)

Application to the Dynamics of HIV Spread

- *Data*: The National Longitudinal Study of Adolescent Health
  - Wave III (with retrospective duration information)
  - Take into account age, sex, race (white/non-white) and age-sex "mixing" patterns
- Estimate the parameters of the model based on the likelihood.
- Consider a (quasi)population of 10000 people (about half men and women)
- Simulate dynamics of sexual networks over 10 years
  - the time step is daily (3650 steps)

Application to the Dynamics of HIV Spread

- *Data*: The National Longitudinal Study of Adolescent Health
  - Wave III (with retrospective duration information)
  - Take into account age, sex, race (white/non-white) and age-sex "mixing" patterns
- Estimate the parameters of the model based on the likelihood.
- Consider a (quasi)population of 10000 people (about half men and women)
- Simulate dynamics of sexual networks over 10 years
  - the time step is daily (3650 steps)
- Simulate disease spread based on 10 "seeds" (2 non-white)
  - as daily have good control over micro-structure of transmission

Application to the Dynamics of HIV Spread

- *Data*: The National Longitudinal Study of Adolescent Health
  - Wave III (with retrospective duration information)
  - Take into account age, sex, race (white/non-white) and age-sex "mixing" patterns
- Estimate the parameters of the model based on the likelihood.
- Consider a (quasi)population of 10000 people (about half men and women)
- Simulate dynamics of sexual networks over 10 years
  - the time step is daily (3650 steps)
- Simulate disease spread based on 10 "seeds" (2 non-white)
  - as daily have good control over micro-structure of transmission
- Visualize only those that become infected

Conclusions and Challenges

- Network models are a very constructive way to represent (social) theory
- Some seemingly simple models are not so.
- Large and deep literatures exist are often ignored
- Simple models are being used to capture structural properties
- The inclusion of attributes is very important
  - actor attributes
  - dyad attributes e.g. homophily, race, location
  - structural terms e.g. transitive homophily
- Software: A suite of `R` packages to implement this is available: `statnetproject.org`
- See the papers at:`statnetproject.org/users_guide.shtml`
  To appear as a special Issue of the *Journal of Statistical Software*, Volume 24.