Statistical Modeling of Complex Networks Within the ERGM family

Mark S. Handcock

Department of Statistics
University of Washington

U. Washington network modeling group
Research supported by NIDA Grant DA012831 and NICHD Grant HD041877

Sunbelt 2006

Available online as CSSS Working Papers #39, #42 and #43 from
`http://www.csss.washington.edu/Papers`

Statistical Models for Social Networks

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

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

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

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*;
  call the graphical representation of Y a *sociogram*
  – a $N = n(n-1)$ array of binary random variables
  – Y represents a random network with nodes the actors and edges the relationship

Statistical Models for Social Networks

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*;
  call the graphical representation of Y a *sociogram*
  – a $N = n(n-1)$ array of binary random variables
  – Y represents a random network with nodes the actors and edges the relationship
- 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 *parameterized* in the form:

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

Besag (1974), Bahadur (1961), Frank and Strauss (1986)

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

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

Simple model-classes for social networks

Homogeneous Bernoulli graph (Renyi-Erdos model)

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

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

  where $q = 1$, $g(y) = \sum_{i,j} y_{ij}$, $\kappa(\eta) = [1 + \exp(\eta)]^N$
- 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 covariates dyadic $x_{k,ij}$

$$P(Y = y) = \frac{\exp\{\eta_1 g_1(y) + \eta_2 g_2(y) \ldots + \eta_K g_K(y)\}}{\kappa(\eta)} \qquad y \in \mathcal{Y}$$

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

$$\text{logit}[P(Y_{ij} = 1)] = \eta_1 x_{1,ij}(y) + \eta_2 x_{2,ij} \ldots + \eta_K x_{K,ij}$$

$$\kappa(\eta) = \prod_{i,j} [1 + \exp(\sum_k \eta_k x_{k,ij})]$$

Bernoulli graphs

- Y_{ij} are independent but have arbitrary distributions

$$P(Y = y) = \frac{\exp\left\{\sum_{i,j} \eta_{ij} y_{ij}\right\}}{\kappa(\eta)} \qquad y \in \mathcal{Y}$$

$$t_{i,j}(y) = y_{ij}, \quad i,j = 1,\ldots,n \qquad q = N$$

$\eta_{ij} = \text{logit}[P(Y_{ij} = 1)]$

$$\kappa(\eta) = \prod_{i,j} [1 + \exp(\eta_{ij})]$$

- Y_{ij} can depend on dyadic covariates x_{ij}

$$\eta_{ij} = X_{ij}\beta$$

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
  – Wasserman (1980), Fienberg, Meyer, and Wasserman (1985)

Nodal Markov statistics

$\Rightarrow$ Frank and Strauss (1986)

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

Nodal Markov statistics

$\Rightarrow$ Frank and Strauss (1986)

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

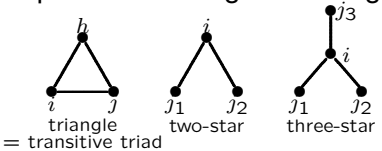
Nodal Markov statistics

⇒ Frank and Strauss (1986)

– motivated by notions of “symmetry" and “homogeneity"
– edges in Y that do not share an actor are conditionally independent given the rest of the network

⇒ analogous to nearest neighbor ideas in spatial statistics

- Degree distribution: $\mathrm{d}_k(\mathbf{y}) =$ proportion of nodes of degree k in $\mathbf{y}$.
- k-star distribution: $\mathrm{s}_k(\mathbf{y}) =$ proportion of k-stars in the graph $\mathbf{y}$.
- triangles: $\mathrm{t}_1(\mathbf{y}) =$ proportion of triangles in the graph $\mathbf{y}$.



Other conditional independence statistics

⇒ Pattison and Robins (2002), Butts (2005)
⇒ Snijders, Pattison, Robins and Handcock (2004)

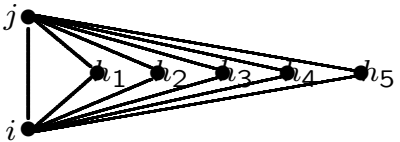
– edges in Y that are not tied are conditionally independent given the rest of the network

Other conditional independence statistics

⇒ Pattison and Robins (2002), Butts (2005)
⇒ Snijders, Pattison, Robins and Handcock (2004)

– edges in Y that are not tied are conditionally independent given the rest of the network

- k-triangle distribution: $\mathrm{t}_k(\mathbf{y}) =$ proportion of k-triangles in the graph $\mathbf{y}$.
- edgewise shared partner distribution:
  $\mathrm{p}_k(\mathbf{y}) =$ proportion of nodes with exactly k edgewise shared partners in $\mathbf{y}$.



k-triangle for $k = 5$, i.e., 5-triangle

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*:
  Include $t_1(y)$, the proportion of triangles amongst triads
  A closely related quantity is the *percent of complete triangles*
  or *mean clustering coefficient*

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

Structural Signatures

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

- Clusters of edges are often *transitive*:
  Include $t_1(y)$, the proportion of triangles amongst triads
  A closely related quantity is the *percent of complete triangles*
  or *mean clustering coefficient*

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

Example: A simple model-class with transitivity

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

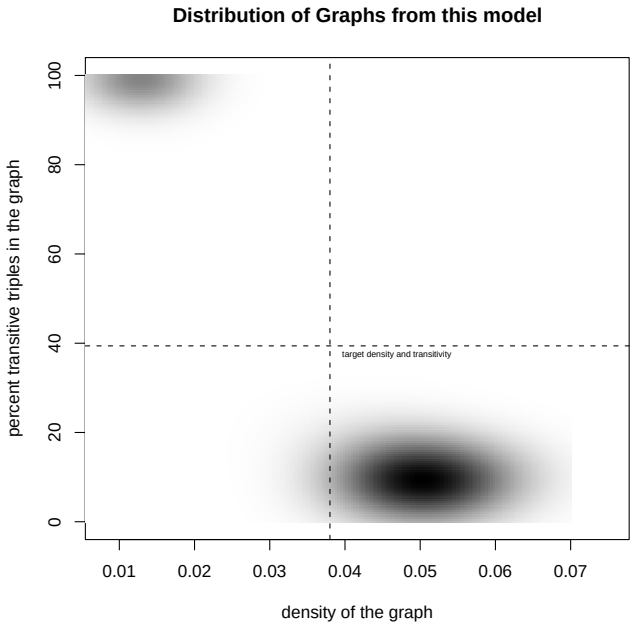
$E(x)$ is the density of edges $(0 - 1)$
$C(x)$ is the triangle percent $(0 - 100)$

- 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?

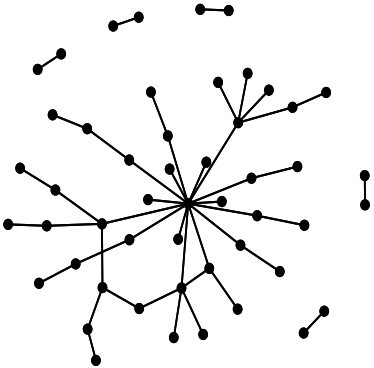
- By construction, on average, graphs from this model have average density 4% and average triangle percent 38%
- If the model is a good representation of transitivity and density we expect the graphs drawn from the model to be close to these values.
- What do graphs produced by this model look like?


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

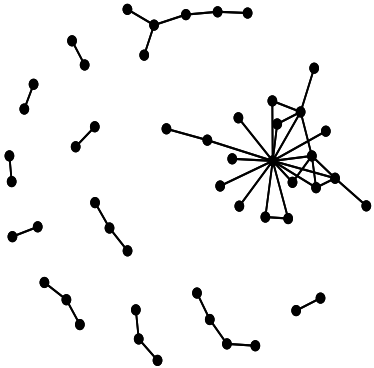
Illustrations of good models within this model-class

- village-level structure
  - $n = 50$
  - mean clustering coefficient = 15%
- larger-level structure
  - $n = 1000$
  - mean clustering coefficient = 15%
- Attribute mixing
  - Two-sex populations
  - mean clustering coefficient = 15%

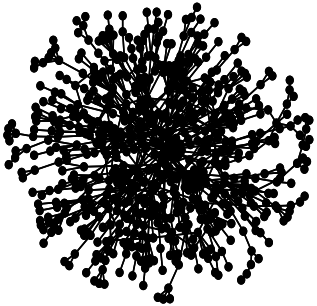
Yule with zero clustering coefficient conditional on degree



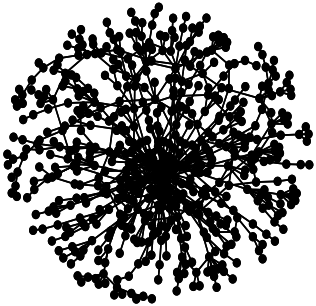
Yule with clustering coefficient 15%



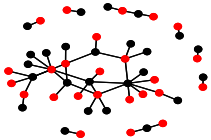
Yule with zero clustering coefficient conditional on degree



Yule with clustering coefficient 15%

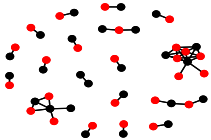


Heterosexual Yule with no correlation



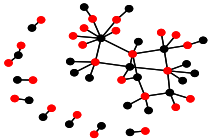
tripercent = 3

Heterosexual Yule with strong correlation



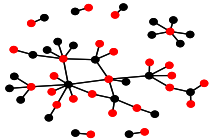
tripercent = 60.6

Heterosexual Yule with modest correlation



tripercent = 0

Heterosexual Yule with negative correlation



tripercent = 0

Conclusions and Challenges

- Models are a very constructive way to represent theory
- Homogeneity is a foundation to build models on
- Some seemingly simple models are not so.
- Useful models require additional development
- 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

- latent class and trait models are important
  - an underlying latent “social space" of actors
    $\Rightarrow$ Hoff, Raftery and Handcock (2002)
    $\Rightarrow$ Hoff (2003, 2004 ,...)
  - latent class models are very promising
    $\Rightarrow$ Nowicki and Snijders (2001)
  - latent class and trait models
    $\Rightarrow$ Tantrum, Handcock, Raftery (2004)