

Statistical Modeling of Networks using the statnet suite of packages in R

Mark S. Handcock

Department of Statistics
University of California - Los Angeles

Joint work with

Martina Morris Steve Goodreau
Krista Gile Pavel Krivitsky
David Hunter

Supported by NIH NIDA Grant DA012831, NICHD Grant HD041877, NSF award MMS-0851555 and the DoD ONR MURI award N00014-08-1-1015.

Statistical Modeling of Networks using the statnet suite of packages in R

Mark S. Handcock

Department of Statistics
University of California - Los Angeles

Joint work with

Martina Morris Steve Goodreau
Krista Gile Pavel Krivitsky
David Hunter

Supported by NIH NIDA Grant DA012831, NICHD Grant HD041877, NSF award MMS-0851555 and the DoD ONR MURI award N00014-08-1-1015.

Working Papers available at

<http://www.stat.ucla.edu/~handcock>
<http://statnet.org>

UseR! 2010, July 21 2010

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

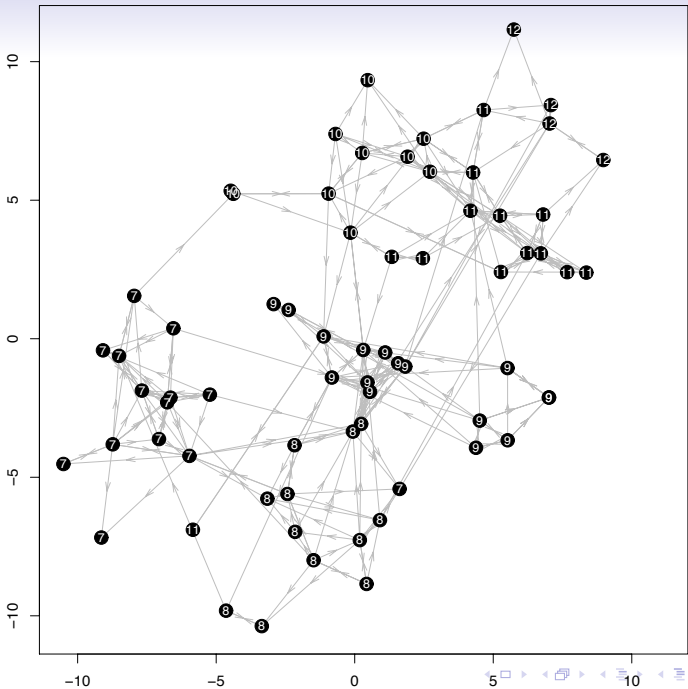
Deep literatures available

- 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, ...)

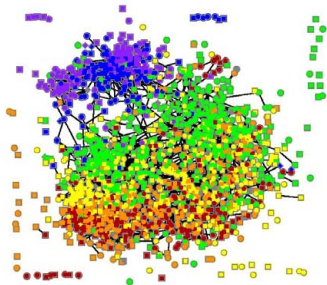
Examples of Friendship Relationships

Examples of Friendship Relationships

- The National Longitudinal Study of Adolescent Health
 - ⇒ 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



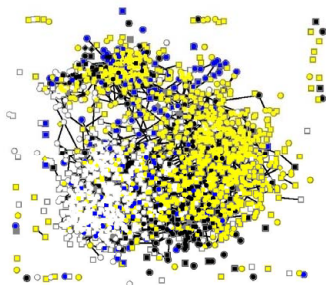
School Community Stratum 44
mutual friendships by Grade



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
 - ⇒ 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
 - ⇒ 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
 - ⇒ 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 ⇒ 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
 - ⇒ 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 ⇒ Heider (1946)
 - people feel comfortable if they agree with others whom they like
- *Context* is important ⇒ 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

The statnet project (2000-present)

- Mission: Develop new statistical methodology for the representation, visualization, analysis and simulation of (social) network data
 - develop computational methods for these statistical methods
 - implement these methods within a coherent suite of user-friendly R packages
 - make them open-source and foster a community outside the developers
- Primary sources of information
 - <http://statnet.org>:
website, software, manuals, documentation, community
 - <http://www.jstatsoft.org/v24>:
Special volume of the *Journal of Statistical Software* on statnet

Statnet Commons

Statnet Commons, a collaborative effort among individual statnet developers and their institutions.

The Statnet Commons aims to:

- coordinate development of the Statnet software by contributing organizations
- to manage the resulting work for the advancement of public benefit
- provide for an environment of continuous sharing and collaborative work among individual members
- provide a mechanism for releasing stable versions of the software under GPL at regular intervals.
- Community activities
 - Pedagogical efforts: tutorials, workshops, seminars
 - Make it easy to add new packages that can add functionality to statnet

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:

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

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 $q = |\mathcal{Y}| - 1$ e.g. $2^N - 1$
- $\kappa(\eta, \mathcal{Y})$ 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 (Erdős-Rényi model)

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

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

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 $x_{k,ij}$

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

Dyad-independence models with attributes

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

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

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

Dyad-independence models with attributes

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

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

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

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

Of course,

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

Generative Theory for Network Structure

Actor Markov statistics

⇒ Frank and Strauss (1986)

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

Generative Theory for Network Structure

Actor Markov statistics

⇒ 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

Generative Theory for Network Structure

Actor Markov statistics

⇒ 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

⇒ analogous to nearest neighbor ideas in spatial modeling

Generative Theory for Network Structure

Actor Markov statistics

⇒ 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

⇒ analogous to nearest neighbor ideas in spatial modeling

- Degree distribution: $d_k(y) =$ proportion of actors of degree k in y .

Generative Theory for Network Structure

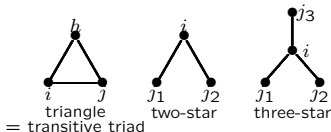
Actor Markov statistics

⇒ 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

⇒ analogous to nearest neighbor ideas in spatial modeling

- Degree distribution: $d_k(y)$ = proportion of actors of degree k in y .
- triangles: $\text{triangle}(y)$ = number of triads that form a complete sub-graph in y .

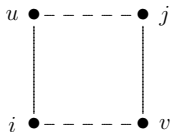


More General mechanisms motivated by conditional independence

- ⇒ Pattison and Robins (2002), Butts (2005)
- ⇒ Snijders, Pattison, Robins and Handcock (2006)
- 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

More General mechanisms motivated by conditional independence

- ⇒ Pattison and Robins (2002), Butts (2005)
- ⇒ Snijders, Pattison, Robins and Handcock (2006)
- 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 features on configurations of the form:

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

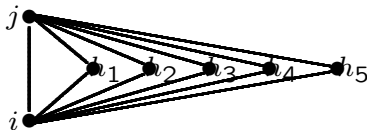


Figure: The actors in the non-directed (i, j) edge have 5 shared partners

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

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 $\text{triangle}(y)$ is the number of triangles amongst triads

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

Structural Signatures

- identify social constructs or features
- based on intuitive notions or partial appeal to substantive theory
- Clusters of edges are often *transitive*:
Recall $\text{triangle}(y)$ is the number of triangles amongst triads

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

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

$$C(y) = \frac{3 \times \text{triangle}(y)}{\text{two-star}(y)}$$

Structural Signatures

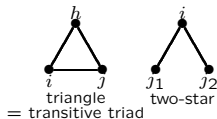
- identify social constructs or features
- based on intuitive notions or partial appeal to substantive theory
- Clusters of edges are often *transitive*:
Recall $\text{triangle}(y)$ is the number of triangles amongst triads

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

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

$$C(y) = \frac{3 \times \text{triangle}(y)}{\text{two-star}(y)}$$

mean clustering coefficient



Statistical Inference for ERGM parameter η

Base inference on the loglikelihood function,

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

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

Approximating the loglikelihood

Approximating the loglikelihood

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

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

Approximating the loglikelihood

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

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

Approximating the loglikelihood

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

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

Approximating the loglikelihood

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

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

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

Approximating the loglikelihood

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

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

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

How can we tell if a model class is useful?

Many aspects:

- Is the model-class itself able to represent a range of realistic networks?
 - *model degeneracy*: small range of graphs covered as the parameters vary (Handcock 2003)

How can we tell if a model class is useful?

Many aspects:

- Is the model-class itself able to represent a range of realistic networks?
 - *model degeneracy*: small range of graphs covered as the parameters vary (Handcock 2003)
- What are the properties of different methods of estimation?
 - e.g, MLE, psuedolikelihood, Bayesian framework
 - *computational failure*: estimates do not exist for certain observable graphs

How can we tell if a model class is useful?

Many aspects:

- Is the model-class itself able to represent a range of realistic networks?
 - *model degeneracy*: small range of graphs covered as the parameters vary (Handcock 2003)
- What are the properties of different methods of estimation?
 - e.g, MLE, psuedolikelihood, Bayesian framework
 - *computational failure*: estimates do not exist for certain observable graphs
- Can we assess the goodness-of-fit of models?
 - appropriate measures and tests
(Besag 2000; Hunter, Goodreau, Handcock 2007)

Model Degeneracy

idea: A random graph model is *near degenerate* if the model places almost all its probability mass on a small number of graph configurations in \mathcal{Y} .

e.g. empty graph, full graph, an individual graph, no 2-stars, mono-degree graphs

Model Degeneracy

idea: A random graph model is *near degenerate* if the model places almost all its probability mass on a small number of graph configurations in \mathcal{Y} .

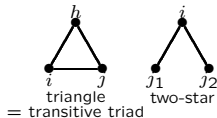
e.g. empty graph, full graph, an individual graph, no 2-stars, mono-degree graphs

- Example: The *two-star* model

$$P(Y = y) = \frac{\exp\{\eta_1 \text{edge}(y) + \eta_2 \text{two-star}(y)\}}{c(\eta_1, \eta_2)} \quad y \in \mathcal{Y}$$

is near-degenerate for most values of $\eta_2 > 0$

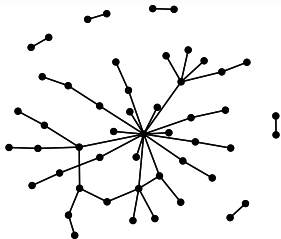
$$\text{edge}(y) = \sum_{i < j} y_{ij} \quad \text{two-star}(y) = \sum_{i < j < k} y_{ij} y_{ik}$$



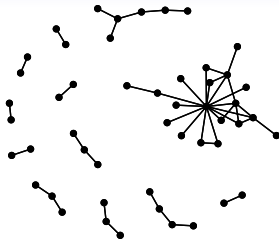
Illustrations of models within this model-class

- 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.

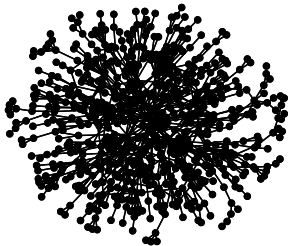
Yule with zero clustering coefficient conditional on degree



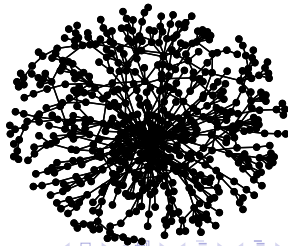
Yule with clustering coefficient 15%



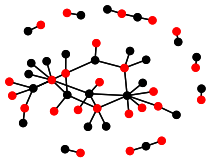
Yule with zero clustering coefficient conditional on degree



Yule with clustering coefficient 15%

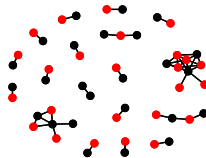


Heterosexual Yule with no correlation



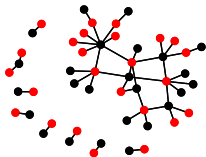
trippercent = 3

Heterosexual Yule with strong correlation

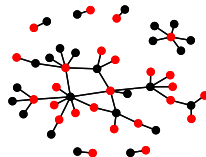


trippercent = 60.6

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.

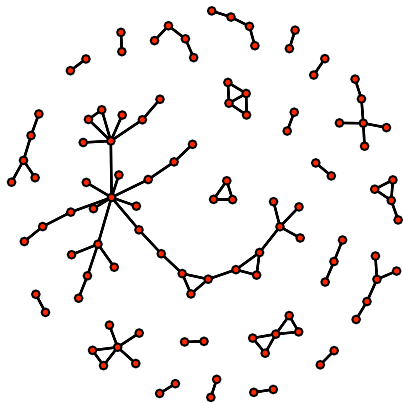


Figure: A protein - protein interaction network for *E. Coli*. The nodes represent proteins and the ties indicate that the two proteins are known to interact with each other.

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.

Clustering and Social Networks

Clustering and Social Networks

- Three types of clustering in social networks:
 - transitivity of relationships
 - homophily of actors with similar *observed* characteristics
 - further clustering that could be due to:
 - homophily on unobserved attributes, or
 - “self-organization” into groups

Clustering and Social Networks

- Three types of clustering in social networks:
 - transitivity of relationships
 - homophily of actors with similar *observed* characteristics
 - further clustering that could be due to:
 - homophily on unobserved attributes, or
 - “self-organization” into groups
- Drawing conclusions about clustering of social actors is often a focus of interest in social network analysis

Clustering and Social Networks

- Three types of clustering in social networks:
 - transitivity of relationships
 - homophily of actors with similar *observed* characteristics
 - further clustering that could be due to:
 - homophily on unobserved attributes, or
 - “self-organization” into groups
- Drawing conclusions about clustering of social actors is often a focus of interest in social network analysis
- But most methods don't address it directly

Clustering and Social Networks

- Three types of clustering in social networks:
 - transitivity of relationships
 - homophily of actors with similar *observed* characteristics
 - further clustering that could be due to:
 - homophily on unobserved attributes, or
 - “self-organization” into groups
- Drawing conclusions about clustering of social actors is often a focus of interest in social network analysis
- But most methods don't address it directly
- Instead conclusions about clustering are often drawn by informally eyeballing results from other methods

Clustering and Social Networks

- Three types of clustering in social networks:
 - transitivity of relationships
 - homophily of actors with similar *observed* characteristics
 - further clustering that could be due to:
 - homophily on unobserved attributes, or
 - “self-organization” into groups
- Drawing conclusions about clustering of social actors is often a focus of interest in social network analysis
- But most methods don't address it directly
- Instead conclusions about clustering are often drawn by informally eyeballing results from other methods
- We present a statistical model of social networks that incorporates clustering and allows formal inference about:
 - whether or not there is clustering (beyond transitivity)
 - if so, how many groups there are
 - who is in what group
 - uncertainty about group memberships

Positing Latent Social Structure via Random Effects

- model an underlying latent “social space” of actors
 - Latent space models: Hoff, Raftery and Handcock (2002)
Hoff (2003, 2004 ,...)
 - Latent class and space models: Tantrum, Handcock, Raftery (2004)
 - GLM and actor heterogeneity: Krivitsky et al (2009)

Positing Latent Social Structure via Random Effects

- model an underlying latent “social space” of actors
 - Latent space models: Hoff, Raftery and Handcock (2002)
Hoff (2003, 2004 ,...)
 - Latent class and space models: Tantrum, Handcock, Raftery (2004)
 - GLM and actor heterogeneity: Krivitsky et al (2009)
- Hierarchical model for the network:
 - Actors i and j are an unknown distance apart in social space
 - Conditional on the distances the ties are independent

Positing Latent Social Structure via Random Effects

- model an underlying latent “social space” of actors
 - Latent space models: Hoff, Raftery and Handcock (2002)
Hoff (2003, 2004 ,...)
 - Latent class and space models: Tantrum, Handcock, Raftery (2004)
 - GLM and actor heterogeneity: Krivitsky et al (2009)
- Hierarchical model for the network:
 - Actors i and j are an unknown distance apart in social space
 - Conditional on the distances the ties are independent

Let:

- $\{\delta_i\}$ individual propensity of the actors to form ties

Positing Latent Social Structure via Random Effects

- model an underlying latent “social space” of actors
 - Latent space models: Hoff, Raftery and Handcock (2002)
Hoff (2003, 2004 ,...)
 - Latent class and space models: Tantrum, Handcock, Raftery (2004)
 - GLM and actor heterogeneity: Krivitsky et al (2009)
- Hierarchical model for the network:
 - Actors i and j are an unknown distance apart in social space
 - Conditional on the distances the ties are independent

Let:

- $\{\delta_i\}$ individual propensity of the actors to form ties
- $\{\gamma_i\}$ individual propensity of the actors to receive ties

Positing Latent Social Structure via Random Effects

- model an underlying latent “social space” of actors
 - Latent space models: Hoff, Raftery and Handcock (2002)
Hoff (2003, 2004 ,...)
 - Latent class and space models: Tantrum, Handcock, Raftery (2004)
 - GLM and actor heterogeneity: Krivitsky et al (2009)
- Hierarchical model for the network:
 - Actors i and j are an unknown distance apart in social space
 - Conditional on the distances the ties are independent

Let:

- $\{\delta_i\}$ individual propensity of the actors to form ties
- $\{\gamma_i\}$ individual propensity of the actors to receive ties
- $\{z_i\}$ be the positions of the actors in the social space \mathbf{R}^k

Positing Latent Social Structure via Random Effects

- model an underlying latent “social space” of actors
 - Latent space models: Hoff, Raftery and Handcock (2002)
Hoff (2003, 2004 ,...)
 - Latent class and space models: Tantrum, Handcock, Raftery (2004)
 - GLM and actor heterogeneity: Krivitsky et al (2009)
- Hierarchical model for the network:
 - Actors i and j are an unknown distance apart in social space
 - Conditional on the distances the ties are independent

Let:

- $\{\delta_i\}$ individual propensity of the actors to form ties
- $\{\gamma_i\}$ individual propensity of the actors to receive ties
- $\{z_i\}$ be the positions of the actors in the social space \mathbf{R}^k
- $\{x_{i,j}\}$ denote observed characteristics that may be dyad-specific and vector-valued

Positing Latent Social Structure via Random Effects

- model an underlying latent “social space” of actors
 - Latent space models: Hoff, Raftery and Handcock (2002)
Hoff (2003, 2004 ,...)
 - Latent class and space models: Tantrum, Handcock, Raftery (2004)
 - GLM and actor heterogeneity: Krivitsky et al (2009)
- Hierarchical model for the network:
 - Actors i and j are an unknown distance apart in social space
 - Conditional on the distances the ties are independent

Let:

- $\{\delta_i\}$ individual propensity of the actors to form ties
- $\{\gamma_i\}$ individual propensity of the actors to receive ties
- $\{z_i\}$ be the positions of the actors in the social space \mathbf{R}^k
- $\{x_{i,j}\}$ denote observed characteristics that may be dyad-specific and vector-valued

Positing Latent Social Structure via Random Effects

- model an underlying latent “social space” of actors
 - Latent space models: Hoff, Raftery and Handcock (2002)
Hoff (2003, 2004 ,...)
 - Latent class and space models: Tantrum, Handcock, Raftery (2004)
 - GLM and actor heterogeneity: Krivitsky et al (2009)
- Hierarchical model for the network:
 - Actors i and j are an unknown distance apart in social space
 - Conditional on the distances the ties are independent

Let:

- $\{\delta_i\}$ individual propensity of the actors to form ties
- $\{\gamma_i\}$ individual propensity of the actors to receive ties
- $\{z_i\}$ be the positions of the actors in the social space \mathbf{R}^k
- $\{x_{i,j}\}$ denote observed characteristics that may be dyad-specific and vector-valued

Specifically:

$$\log \text{odds}(Y_{ij} = 1 | z_i, z_j, x_{ij}, \beta) = \beta^T x_{ij} - |z_i - z_j| + \delta_i + \gamma_j$$

where β denotes parameters to be estimated.

Model-based Clustering of Social Networks

Model-based Clustering of Social Networks

- Model the latent positions as clustered into G groups:

$$z_i \stackrel{i.i.d.}{\sim} \sum_{g=1}^G \lambda_g \text{MVN}_d(\mu_g, \sigma_g^2 I_d)$$

Model-based Clustering of Social Networks

- Model the latent positions as clustered into G groups:

$$z_i \stackrel{i.i.d.}{\sim} \sum_{g=1}^G \lambda_g \text{MVN}_d(\mu_g, \sigma_g^2 I_d)$$

- Spherical covariance motivated by invariance

Model-based Clustering of Social Networks

- Model the latent positions as clustered into G groups:

$$z_i \stackrel{i.i.d.}{\sim} \sum_{g=1}^G \lambda_g \text{MVN}_d(\mu_g, \sigma_g^2 I_d)$$

- Spherical covariance motivated by invariance
- captures position, transitivity, homophily on attributes, and clustering

Model-based Clustering of Social Networks

- Model the latent positions as clustered into G groups:

$$z_i \stackrel{i.i.d.}{\sim} \sum_{g=1}^G \lambda_g \text{MVN}_d(\mu_g, \sigma_g^2 I_d)$$

- Spherical covariance motivated by invariance
- captures position, transitivity, homophily on attributes, and clustering
- captures individual propensities to form and receive ties

$$\delta_i \stackrel{i.i.d.}{\sim} \text{N}(0, \sigma_\delta^2) \quad i = 1, \dots, n,$$

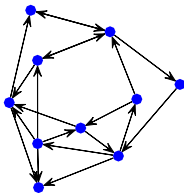
$$\gamma_i \stackrel{i.i.d.}{\sim} \text{N}(0, \sigma_\gamma^2) \quad i = 1, \dots, n,$$

Summary of latent cluster model

- Model-based clustering of latent positions for social networks provides a formal model of social networks that incorporates clustering
- It permits inference about:
 - whether there is clustering
 - how many groups there are
 - who is in what group
 - uncertainty about group memberships
 - the actors' latent social positions
- It gave reasonable results for two examples
- **Software: The R package latentnet, available on CRAN**

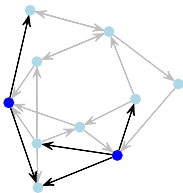
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
“Ask 10% of employees about their collaborations”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole company, but someone is out sick”
- **Boundary Specification Problem:**
“Should a contractor be considered a part of the company?”



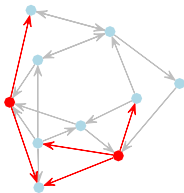
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
“Ask 10% of employees about their collaborations”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole company, but someone is out sick”
- **Boundary Specification Problem:**
“Should a contractor be considered a part of the company?”



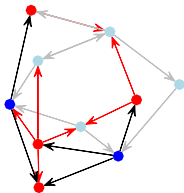
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
“Ask 10% of employees about their collaborations”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole company, but someone is out sick”
- **Boundary Specification Problem:**
“Should a contractor be considered a part of the company?”



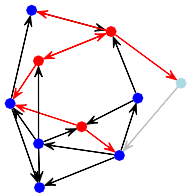
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
“Ask 10% of employees about their collaborations”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole company, but someone is out sick”
- **Boundary Specification Problem:**
“Should a contractor be considered a part of the company?”



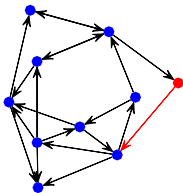
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
“Ask 10% of employees about their collaborations”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole company, but someone is out sick”
- **Boundary Specification Problem:**
“Should a contractor be considered a part of the company?”



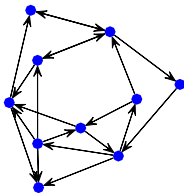
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
“Ask 10% of employees about their collaborations”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole company, but someone is out sick”
- **Boundary Specification Problem:**
“Should a contractor be considered a part of the company?”



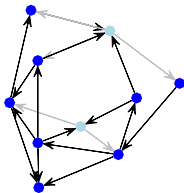
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
“Ask 10% of employees about their collaborations”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole company, but someone is out sick”
- **Boundary Specification Problem:**
“Should a contractor be considered a part of the company?”



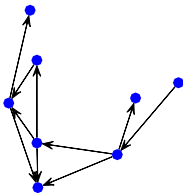
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
“Ask 10% of employees about their collaborations”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole company, but someone is out sick”
- **Boundary Specification Problem:**
“Should a contractor be considered a part of the company?”



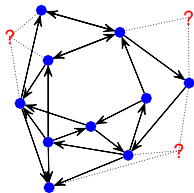
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
“Ask 10% of employees about their collaborations”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole company, but someone is out sick”
- **Boundary Specification Problem:**
“Should a contractor be considered a part of the company?”



Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
“Ask 10% of employees about their collaborations”
 - Egocentric
 - Adaptive
- **Out-of-design Missing Data:**
“Try to survey the whole company, but someone is out sick”
- **Boundary Specification Problem:**
“Should a contractor be considered a part of the company?”



Frameworks for Statistical Analysis

	Network Specific	Population Process
Fully Observed Data	Description	Modeling (Statistical)
Partially Observed Data	Design-Based Inference	Likelihood Inference

statnet capabilities

Required packages: `ergm` and `network` available on CRAN

- `ergm` is a collection of functions to fit, simulate from, plot and evaluate exponential-family random graph models.

The main functions within the `ergm` package are:

- `ergm`, a function to fit exponential-family random graph models in which the probability of a network is dependent upon a vector of network statistics specified by the user
- `simulate`, a function to simulate random networks using an ERGM
- `gof`, a function to evaluate the goodness of fit of an ERGM to the data.

`ergm` contains many other functions as well.

- `network` is a package to create, store, modify and plot the data in network objects.

The network object class, defined in the `network` package, can represent a range of relational data types and it supports arbitrary vertex / edge / network attributes.

Data stored as network objects can then be analyzed using all of the component packages in the `statnet` suite.

statnet capabilities

Optional packages

The optional packages `sna`, `degreenet`, `latentnet`, and `networksis` are all available on CRAN:

- `sna`: A set of tools for traditional social network analysis .
- `degreenet`: This package was developed for the degree distributions of networks. It implements likelihood-based inference, bootstrapping, and model selection, and it includes power-law models such as the Yule and Waring as well as a range of alternative models that have been proposed in the literature. .
- `latentnet`: A package to fit and evaluate latent position and cluster models for statistical networks.
- `networksis`: A package to simulate bipartite networks with fixed marginals through sequential importance sampling .

statnet capabilities

Additional optional packages are available on request, as described below.

- **dynamicnetwork**: A set of tools for visualizing dynamically changing networks .
- **netperm**: A package for permutation Models for relational data . It provides simulation and inference tools for exponential families of permutation models on relational structures.
- **rSoNIA**: Provides a set of methods to facilitate exporting data and parameter settings and launching SoNIA, which stands for Social Network Image Animator . SoNIA facilitates interactive browsing of dynamic network data and exporting animations as a QuickTime movies.

Additional capabilities

- statnet can efficiently deal with large networks (it handles data natively in edgelist form (within the backend)).
- In terms of data representations, it can generally support networks on the order of 10^8 edges and/or nodes.
- missing data on relations are handled
- dynamic models has been developed and coded, but is not yet on CRAN (Krivitsky 2009)

Statistical Challenges and Opportunities

- massive and varied types of data
 - incorporation of these into the model is sometimes difficult
- networks fundamentally relational
 - traditional notions based on independence flawed
- noise in the relations and attributes
- partially observed networks
 - almost always (non-ignorable) missing values
 - ⇒ Handcock and Gile (2008)
 - often the boundary of the network is endogenous
- measuring goodness-of-fit of network models
 - ⇒ Hunter, Goodreau and Handcock (2007)
- representing uncertainty in the inference
- visualization of complex models and networks

- In some disciplines the basic question of inference is ignored
- understanding properties of sparse representations
 - e.g., concept of “model degeneracy” \Rightarrow Handcock (2003)
 - MLE, maximum pseudo-likelihood
- improve estimation methods
 - technology transfer of approximate likelihood methods and ideas developed in Genetics and Computer Science
 - Variational methods (Jordan et al 1998, ...)

Opportunities

- Sciences should make better use of network sampling techniques
 - adaptive network designs (e.g., link tracing)
 - ⇒ Handcock and Gile (2008)
 - respondent-driven sampling for hard-to-reach populations
 - ⇒ Gile and Handcock (2008)
- Dynamic and longitudinal models (harder and easier)
- Most models condition on the number of nodes
 - models “generating” the number of nodes are important

Summary

- Network representations intersect with most sciences
- Sparse models are being used to capture structural properties
- The models must depend on the scientific objective.
- Some seemingly simple models are not so.
- The inclusion of attributes is very important
 - actor attributes
 - dyad attributes e.g. homophily, race, location
 - structural terms e.g. transitive homophily