Inference for a Longitudinal Model of Network Formation: Heider's theory of Balance vs Simmel's triadic formation

Mark S. Handcock

Department of Statistics
University of Washington

Joint work with

Martina Morris
David Krackhardt

and the

U. Washington Network Modeling Group

Supported by NIH NIDA Grant DA012831 and NICHD Grant HD041877

Inference for a Longitudinal Model of Network Formation: Heider's theory of Balance vs Simmel's triadic formation

Mark S. Handcock

Department of Statistics
University of Washington

Joint work with

Martina Morris
David Krackhardt

and the

U. Washington Network Modeling Group

Supported by NIH NIDA Grant DA012831 and NICHD Grant HD041877

Working Papers available at

http://www.csss.washington.edu/Papers

SFI Workshop "Is there a Physics of Society? January 10-12 2008

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

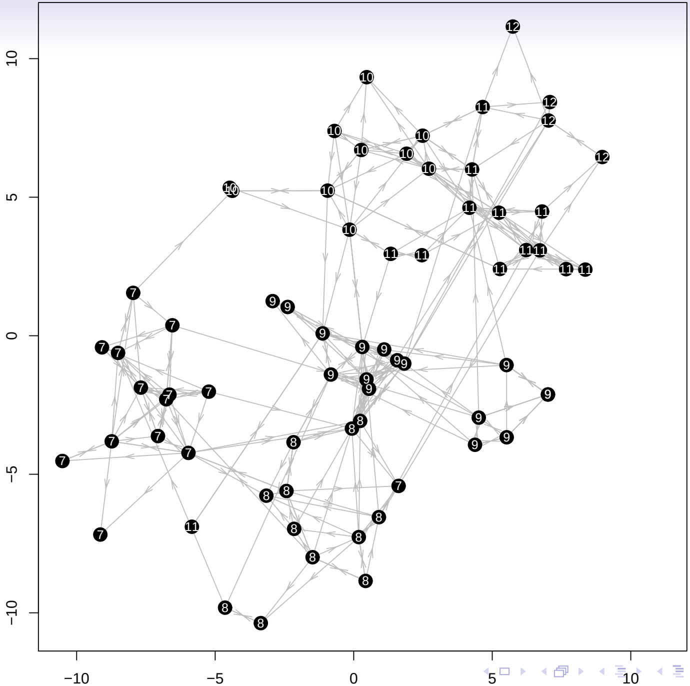
Deep literatures available

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

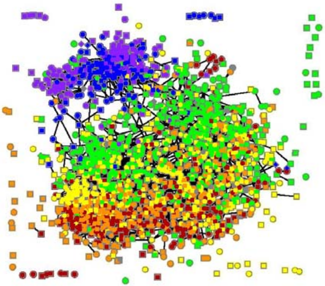
Examples of Friendship Relationships

Examples of Friendship Relationships

- The National Longitudinal Study of Adolescent Health
  $\Rightarrow$ `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

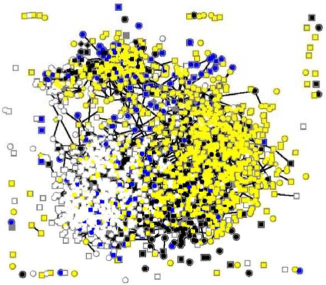

10
5
0
−5
−10
−10
−5
0
5
10

School Community Stratum 44
mutual friendships by Grade



2209 Students

School Community Stratum 44
mutual friendships by Race



2209 Students

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

- **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

  $\Rightarrow$ 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

  ⇒ 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

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.

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

- 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 (Erdős-Rényi model)

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

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

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,

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

Models for the Degree Distribution in Isolation

Let $P(K = k)$ be the probability mass function of the degree of a randomly chosen actor.

$P(K = k)$ has *power-law behavior* with scaling $\rho > 1 \iff \exists c_1, c_2,$ and $M : 0 < c_1 \leq P(K = k)k^{\rho} \leq c_2 < \infty$ for $k > M$.

Models for the Degree Distribution in Isolation

Let $P(K = k)$ be the probability mass function of the degree of a randomly chosen actor.

$P(K = k)$ has *power-law behavior* with scaling $\rho > 1 \iff$ $\exists c_1, c_2$, and $M : 0 < c_1 \leq P(K = k)k^{\rho} \leq c_2 < \infty$ for $k > M$.

- Preferential Attachment (Albert and Barabasi 2000)

$$\textit{Yule Model}: \quad P(K = k | K > 0) = \frac{(\rho - 1)\Gamma(k)\Gamma(\rho)}{\Gamma(k + \rho)} \quad k = 1, 2, \ldots$$

$\Rightarrow$ Simon (1955), Jones and Handcock (2003c)

Models for the Degree Distribution in Isolation

Let $P(K = k)$ be the probability mass function of the degree of a randomly chosen actor.

$P(K = k)$ has *power-law behavior* with scaling $\rho > 1 \iff$
$\exists c_1, c_2$, and $M : 0 < c_1 \leq P(K = k)k^\rho \leq c_2 < \infty$ for $k > M$.

- Preferential Attachment (Albert and Barabasi 2000)

$$\textit{Yule Model}: \quad P(K = k | K > 0) = \frac{(\rho - 1)\Gamma(k)\Gamma(\rho)}{\Gamma(k + \rho)} \quad k = 1, 2, \ldots$$

$\Rightarrow$ Simon (1955), Jones and Handcock (2003c)

$$\textit{Waring Model}: \quad P(K = k | K > 0) = \frac{(\rho - 1)\Gamma(\rho + \rho_0)}{\Gamma(\rho_0 + 1)} \cdot \frac{\Gamma(k + \rho_0)}{\Gamma(k + \rho_0 + \rho)}$$

$$\textit{probability of a new actor}: \quad p = \frac{\rho - 2}{\rho + \rho_0 - 1} \quad \rho_0 > -1$$

$\Rightarrow$ Irwin (1963), Jones and Handcock (2003c, ...)

Models for the Degree Distribution in Isolation

Let $P(K = k)$ be the probability mass function of the degree of a randomly chosen actor.

$P(K = k)$ has *power-law behavior* with scaling $\rho > 1 \iff \exists c_1, c_2$, and $M : 0 < c_1 \leq P(K = k)k^\rho \leq c_2 < \infty$ for $k > M$.

- Preferential Attachment (Albert and Barabasi 2000)

$$\textit{Yule Model}: \quad P(K = k | K > 0) = \frac{(\rho - 1)\Gamma(k)\Gamma(\rho)}{\Gamma(k + \rho)} \quad k = 1, 2, \ldots$$

$\Rightarrow$ Simon (1955), Jones and Handcock (2003c)

$$\textit{Waring Model}: \quad P(K = k | K > 0) = \frac{(\rho - 1)\Gamma(\rho + \rho_0)}{\Gamma(\rho_0 + 1)} \cdot \frac{\Gamma(k + \rho_0)}{\Gamma(k + \rho_0 + \rho)}$$

$$\textit{probability of a new actor}: \quad p = \frac{\rho - 2}{\rho + \rho_0 - 1} \quad \rho_0 > -1$$

$\Rightarrow$ Irwin (1963), Jones and Handcock (2003c, ...)

- Vetting Models $\Rightarrow$ Handcock and Jones (2004)
  - A partnership network as a subset of an underlying acquaintance network
- Decoupling Models (tail behavior) $\Rightarrow$ Handcock and Jones (2004)

- Failure of log-log plot "curve fitting" approaches
  $\Rightarrow$ Jones and Handcock (2002a,2002b,2003a,2003b,2003c, ...)

- Failure of log-log plot "curve fitting" approaches

  $\Rightarrow$ Jones and Handcock (2002a,2002b,2003a,2003b,2003c, ...)
- Likelihood-based inference for the parameters (arbitrary model)

  $\Rightarrow$ Jones and Handcock (2002a,2002b,2003a,2003b,2003c, ...),
  - continuous models (Pareto) $\Rightarrow$ Jones and Handcock (2002a)
  - discrete models (Discrete Pareto) $\Rightarrow$ Jones and Handcock (2002b)
  - random generation of degrees $\Rightarrow$ Jones and Handcock (2002b)
- Model selection
  - likelihood-based
  - AIC (best for large samples)
  - BIC (good for small samples)

- Failure of log-log plot "curve fitting" approaches

  $\Rightarrow$ Jones and Handcock (2002a,2002b,2003a,2003b,2003c, ...)
- Likelihood-based inference for the parameters (arbitrary model)

  $\Rightarrow$ Jones and Handcock (2002a,2002b,2003a,2003b,2003c, ...),
  - continuous models (Pareto) $\Rightarrow$ Jones and Handcock (2002a)
  - discrete models (Discrete Pareto) $\Rightarrow$ Jones and Handcock (2002b)
  - random generation of degrees $\Rightarrow$ Jones and Handcock (2002b)
- Model selection
  - likelihood-based
  - AIC (best for large samples)
  - BIC (good for small samples)
- Quantifying uncertainty
  - confidence intervals for the parameters
  - Anderson-Darling Statistics $\Rightarrow$ (Handcock, Jones, Morris 2003)
  - Bootstrap confidence intervals (for model selection)
- Fit to real network data (primarily sexual partnership data)
  - Sweden, US (many), Rakai, UK, etc

    $\Rightarrow$ Hamilton, Handcock, Morris (2008)
- Confidence intervals for epidemic potential R_0 for networks

  $\Rightarrow$ Handcock and Jones (2006)

Some References on Inference and Models for Degree Distributions

1. Jones JH, Handcock MS (2002a). "Statistical Evidence Tells Tails of Human Sexual Contacts." *Working Paper 21*, Center for Statistics and the Social Sciences. URL `http://www.csss.washington.edu/Papers`.
2. Jones JH, Handcock MS (2002b). "Epidemic thresholds exist in human sexual contact networks." *Working Paper 23*, Center for Statistics and the Social Sciences. URL `http://www.csss.washington.edu/Papers`.
3. Jones JH, Handcock MS (2003a). "Sexual contacts and epidemic thresholds." *Nature*, **423**(6940), 605–606.
4. Handcock MS, Jones JH, Morris M (2003b). "On 'Sexual contacts and epidemic thresholds,' models and inference for Sexual partnership distributions." *Working Paper 31*, Center for Statistics and the Social Sciences. URL `http://www.csss.washington.edu/Papers`.
5. Jones JH, Handcock MS (2003c). "An assessment of preferential attachment as a mechanism for human sexual network formation." *Proceedings of the Royal Society of London, B*, **270**, 1123–1128.
6. Handcock MS, Jones JH (2004). "Likelihood-Based Inference for Stochastic Models of Sexual Network Formation." *Theoretical Population Biology*, **65**, 413–422.
7. Handcock MS, Jones JH (2006). "Interval estimates for epidemic thresholds in two-sex network models." *Theoretical Population Biology*, **70**, 125–134.
8. Hamilton DT, Handcock MS, Morris M (2008). "Degree distributions in sexual networks: A framework for evaluating evidence." *Sexually Transmitted Diseases*, **35**, 30–40.

All methodology implemented in the `R` package `degreenet`, available as part of `statnet` at `http://statnetproject.org`

Generative Theory for Network Structure

Actor Markov statistics

$\Rightarrow$ Frank and Strauss (1986)

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

Generative Theory for Network Structure

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

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

Generative Theory for Network Structure

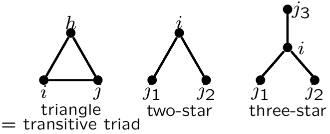
Actor Markov statistics

$\Rightarrow$ 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 modeling

- 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, $\mathrm{s}_2 =$ 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.



More General mechanisms motivated by conditional independence

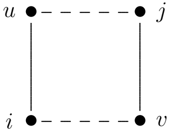
⇒ Pattison and Robins (2002), Butts (2005)
⇒ Snijders, Pattison, Robins and Handcock (2004)

– 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

More General mechanisms motivated by conditional independence

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

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

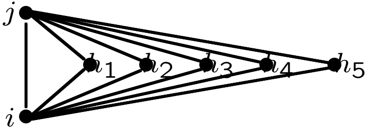


Figure: The actors in the non-directed (i, j) edge have 5 shared partners

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

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

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

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%

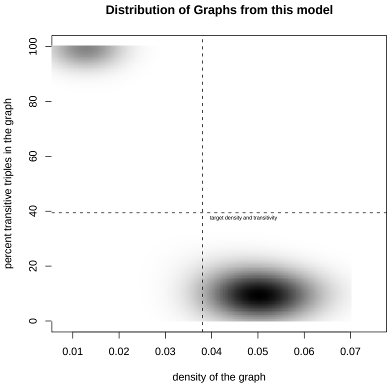
How can we tell if the model is useful?

- 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,

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

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

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

  – $n = 1000$

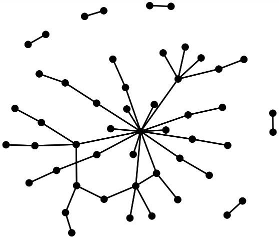
  – mean clustering coefficient $= 15\%$ – degree distribution: Yule with scaling exponent 3.

- Attribute mixing

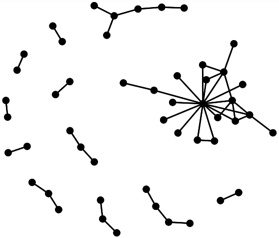
  – Two-sex populations

  – mean clustering coefficient $= 15\%$ – degree distribution: Yule with scaling exponent 3.

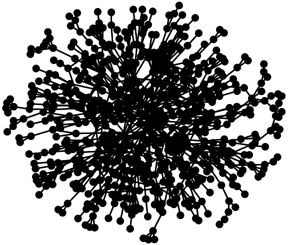
Yule with zero clustering coefficient conditional on degree



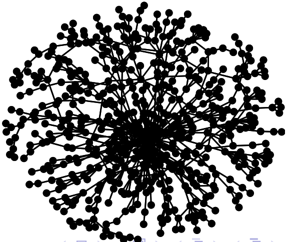
Yule with clustering coefficient 15%



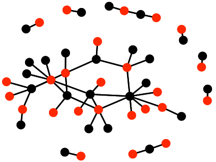
Yule with zero clustering coefficient conditional on degree



Yule with clustering coefficient 15%

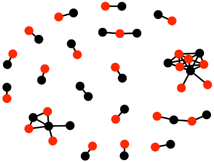


Heterosexual Yule with no correlation



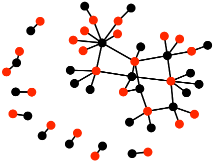
tripercent = 3

Heterosexual Yule with strong correlation

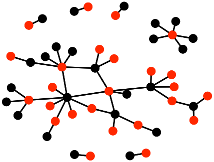


tripercent = 60.6

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.

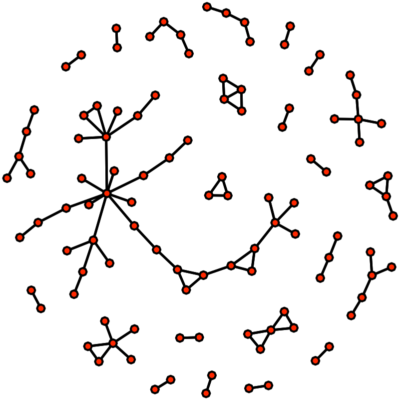


Figure: A protein - protein interaction network for *E. Coli*. The nodes represent proteins and the ties indicate that the two proteins are known to interact with each other.

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.

Preferred Friends within a Fraternity over Time

Preferred Friends within a Fraternity over Time

- In 1956, 17 men were recruited to live in a fraternity house
  $\Rightarrow$ Newcomb (1961)
- Each week in the Fall semester asked to rank their peers is terms of how much he liked them
  - Longitudinal data over 15 weeks (except week 9)
  - Rank order of all 16 peers
  - Present a tie when peer is ranked in the top half

Theories of Social Structure

- Fritz Heider's Theory of Balance
  - a person is motivated to establish and maintain balance in their relationships
- Heider (1946. p. 110) simplified the predictions of the theory

> "In the case of two entities, a balanced state exists if
> the relation between them is [mutually] positive (or [mutually]
> (or [mutually] negative....
> In the case of three entities, a balanced state exists if
> all three relations [among the three entities] are positive...,
> or if two are negative and one positive"

- balance is a state of equilibrium
- balance predicts dynamics: networks tend to balance

Georg Simmel's Theory

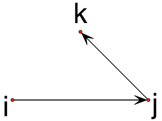
- consideration of the dyad is not enough
    - triad, not the dyad, is the fundamental social unit
- Simmel (1908. p. 136): Members of a dyad experience an

> "intensification of relation by [the addition of] a third element, or by a social framework that transcends both members of the dyad"

- *Simmelian tie*: a tie that was embedded in a clique
- In a dynamic context, Simmelian ties are hard to break

Quantifying the Theories of Heider and Simmel

- Heiderian Theory
  - dyads: balance (symmetric pair) vs. imbalance (asymmetric)
  - triads: balance (transitive) vs. imbalance (pre-transitive triple)



- Simmelian Theory
  - triads: complete sub-graphs of size three

Statistical Signatures for Heider and Simmel

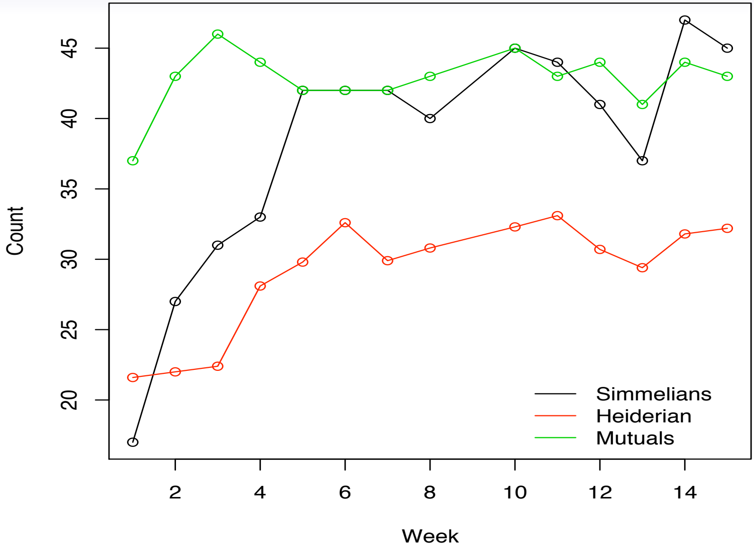
To capture the propensity for a network to have Heiderian dyads and triads we use:

$$g_1(y) = \text{number of symmetric dyads in } y$$

$$g_2(y) = \text{number of Heiderian (i.e., transitive) triads in } y$$

To capture the propensity for the network to have Simmelian triads we use the statistic:

$$g_3(y) = \text{number of Simmelian triads in } y,$$

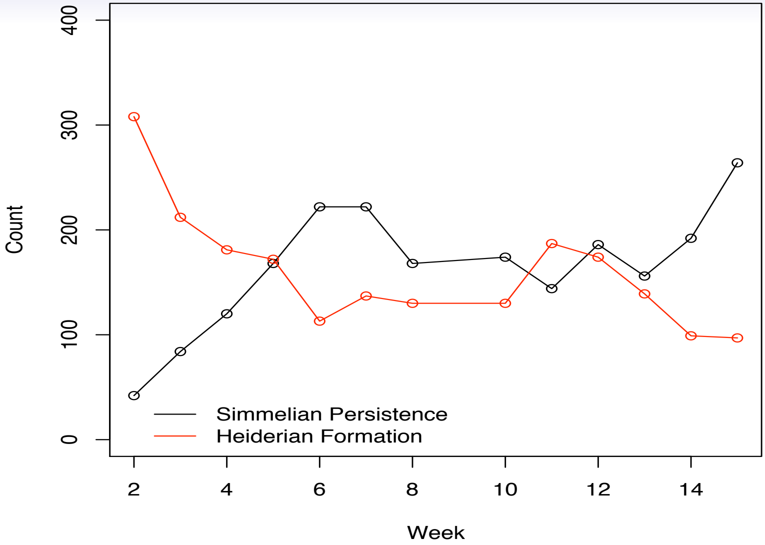


Simmelian and Heiderian Statistics for the Networks over Time.

Cross-sectional Models to Represent the Theories

$$P_\eta(Y = y) = \frac{e^{\sum_{k=1}^{3} \eta_k g_k(y)}}{\kappa(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

where Y=$\{y$: each node in y has exactly 8 out-ties$\}$



The persistence of Simmelian triads and the formation of Heiderian triads for the Newcomb Networks over Time.

Modeling Longitudinal Networks

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

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

Modeling Longitudinal Networks

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

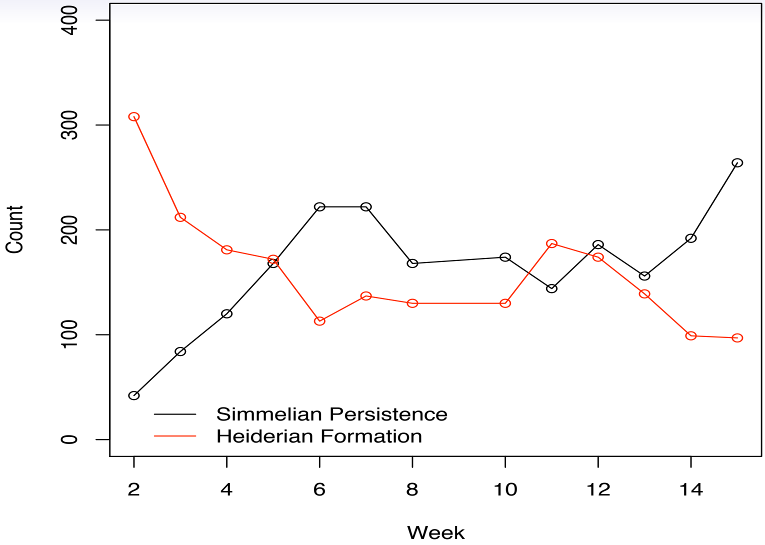
$$Y_{ij}^{(t)} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

Consider a dynamic variant of the above model:

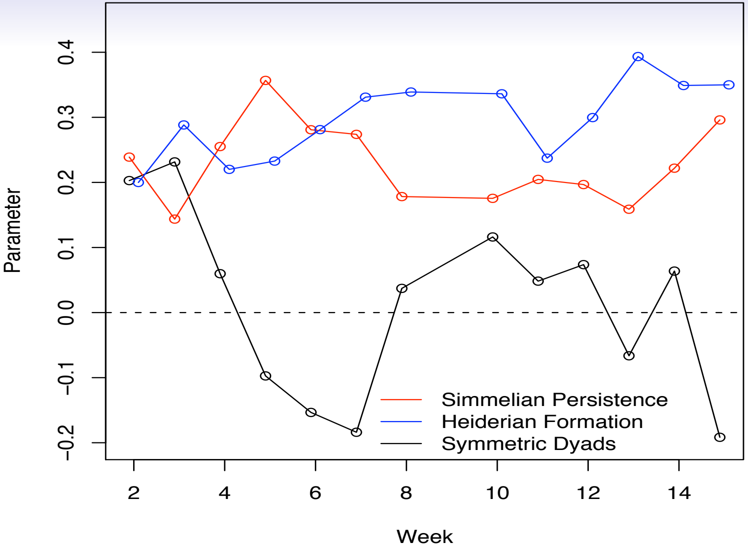
$$P_\eta(Y^{(t+1)} = y^{(t+1)} | Y^{(t)} = y^{(t)}) = \frac{\exp\left(\eta^{(t+1)} \cdot g\left(y^{(t+1)}; y^{(t)}\right)\right)}{\sum_{s \in \mathcal{Y}} \exp\left(\eta^{(t+1)} \cdot g\left(x; y^{(t)}\right)\right)} \quad t = 2, \ldots, T$$

We add two additional statistics dynamic statistics:

$$\begin{aligned} g_4(y^{(t+1)}; y^{(t)}) &= \text{number of pre-Heiderian triads in } y^{(t)} \\ &\quad\ \text{that are Heiderian in } y^{(t+1)} \\ g_5(y^{(t+1)}; y^{(t)}) &= \text{number of Simmelian triads in } y^{(t)} \\ &\quad\ \text{that persist in } y^{(t+1)} \end{aligned}$$



The persistence of Simmelian triads and the formation of Heiderian triads for the Newcomb Networks over Time.



The joint effects of the persistence of Simmelian triads, Heiderian dyadic balance, and the formation of Heiderian triads for the Newcomb networks over time. The values plotted are the MLEs of the parameters of model for $t = 2, \ldots, 15$.

Conclusions and Challenges

- Network models are a very constructive way to represent (social) theory
- The models can be used to compare the predictions of social theory
- Simple models are being used to capture structural properties
- Homogeneity is a foundation to build models on

Conclusions and Challenges

- Network models are a very constructive way to represent (social) theory
- The models can be used to compare the predictions of social theory
- Simple models are being used to capture structural properties
- Homogeneity is a foundation to build models on
- Some seemingly simple models are not so.

Conclusions and Challenges

- Network models are a very constructive way to represent (social) theory
- The models can be used to compare the predictions of social theory
- Simple models are being used to capture structural properties
- Homogeneity is a foundation to build models on
- Some seemingly simple models are not so.
- Large and deep literatures exist are often ignored or not cited

Conclusions and Challenges

- Network models are a very constructive way to represent (social) theory
- The models can be used to compare the predictions of social theory
- Simple models are being used to capture structural properties
- Homogeneity is a foundation to build models on
- Some seemingly simple models are not so.
- Large and deep literatures exist are often ignored or not cited
- 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

Conclusions and Challenges

- Network models are a very constructive way to represent (social) theory
- The models can be used to compare the predictions of social theory
- Simple models are being used to capture structural properties
- Homogeneity is a foundation to build models on
- Some seemingly simple models are not so.
- Large and deep literatures exist are often ignored or not cited
- 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
- The predominant social theory of balance alone can not explain the evolution of the Fraternity network.
- Both effects appear to exist simultaneously.

Conclusions and Challenges

- Network models are a very constructive way to represent (social) theory
- The models can be used to compare the predictions of social theory
- Simple models are being used to capture structural properties
- Homogeneity is a foundation to build models on
- Some seemingly simple models are not so.
- Large and deep literatures exist are often ignored or not cited
- 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
- The predominant social theory of balance alone can not explain the evolution of the Fraternity network.
- Both effects appear to exist simultaneously.
- Software: A suite of `R` packages to implement this `statnetproject.org`
- See all of Volume 24 of the *Journal of Statistical Software*