Introduction to Exponential-family Random Graph Models (ERGMs)

Mark S. Handcock

handcock@ucla.edu

Statistical Analysis of Networks

November 2, 2024

Reprise: What is a statistical model?

The word "model" means different things in different subfields A *statistical* model is a

- formal representation of a
- stochastic process
- specified at one level (e.g., person, dyad) that
- aggregates to a higher level (e.g., population, network)

A well specified stochastic model allows us to understand the uncertainty associated with observed outcomes.

Why take a statistical approach?

Descriptive vs. generative goals

- Descriptive: numerical summary measures
    - Nodal level: e.g., centrality
    - Configuration level: e.g., triad census
    - Network level: e.g., centralization, clustering
- Generative: micro foundations for macro patterns
    - Recover underlying dynamic processes from cross-sectional data
    - Test hypotheses
    - Extrapolate and simulate from model

Substantive considerations

However, different processes can lead to similar macro signatures

- For example: “clustering” typically observed in social nets can be a result of
  - Sociality - highly active persons create clusters
  - Homophily - assortative mixing by attribute creates clusters
  - Transitive triad closure - triangles create clusters

  Want to be able to fit these terms simultaneously, and identify the independent effects of each process on the overall outcome.

Example: Friend of a friend, or birds of a feather?

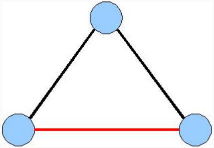
Two theories about the process that generates 3-cycles in an undirected graph

1. Homophily: People tend to chose friends who are like them, in grade, race, etc. (“birds of a feather”), triad closure is a by-product
2. Transitivity: People who have friends in common tend to become friends (“friend of a friend”), closure is the key process

Example: Friend of a friend, or birds of a feather?

Two theories about the process that generates 3-cycles in an undirected graph: Homophily verses Transitivity?

So, for three actors of the same type:



Cycle-closing tie may form because of *transitivity* but also *homophily*

Modeling endogenous network processes

- Guiding principles:
    - Network ties are the outcome of (unobserved) social processes that tend to be local and interactive
    - There are both regularities and variability in these local interactive processes
    - The observed network is only one realization from a set of possible networks with similar characteristics: the outcome of some unknown stochastic process
    - We do not know what is the stochastic process, and the goal is to propose a plausible and theoretically principled hypothesis for the process

Modeling endogenous network processes

- We construct statistical models in which:
    - network ties are variable and network nodes are fixed
    - assumptions about the form of local interactions (i.e., dependence) among tie variables are explicit
    - regularities are represented by model parameters and estimated from data

The logic behind ERG models - example

Friendship network in a classroom, size n

- Focus on structural characteristic of interest: are there more reciprocated ties than would be expected by chance?
- Posit a stochastic network model that includes a density parameter (ties occur at random) and a reciprocation parameter (tendency for reciprocation to occur)
- The probability distribution on the set of all possible graphs of size n is constructed such that graphs with a lot of reciprocation are more probable

The logic behind random graph models

- Once we have estimated the parameters of the probability distribution, we can draw graphs at random and compare their characteristics with those of the observed network
- If the model is good, then sampled graphs will resemble the observed network
- Then we may hypothesize that the modeled structural effects could explain the emergence of the network

Four generations of dependence hypotheses

1. Bernoulli (Erdős-Rényi) models
2. p_1 models
3. Markov random graphs
4. Realization-dependent models

Each of these hypotheses gives rise to certain classes of statistics that can be included in the model

Generative Theory for Network Structure

Frank and Strauss (1986)

Represent dependence structure of a model by its *dependence graph*, D:

- vertices: labels of the tie variables in Y.
- edges: if the tie variables are conditionally dependent given the rest of the network

Result: Any, and all, models for Y have the ERGM form with sufficient statistics

$$\prod_{i,j \in A} Y_{ij} \qquad A \in \mathcal{A}$$

where $\mathcal{A}$ is the set of cliques of D.

First generation - Bernoulli (Erdős-Rényi) models

Tie variables Y_{ij} and Y_{kl} are conditionally independent unless they are incident on the same nodes

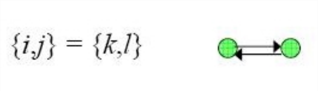
In this case the sufficient statistics are the tie variables Y_{ij} themselves, typically reduced to counts of tie variables by homogeneity.

Second generation - p_1 models

Tie variables Y_{ij} and Y_{kl} are conditionally independent unless they involve the same dyad

Holland & Leinhardt (1981)
Wasserman & Galaskiewicz (1985)

In this case the sufficient statistics are the tie variables Y_{ij} and $Y_{ij}Y_{ji}$ (“mutual ties”) typically reduced to counts of variables by homogeneity



Homogeneity of model parameters

If we assume that isomorphic configurations have equal parameters, then:

- There is one parameter for each class of network configurations
- The corresponding statistic is the number of configurations in y
- For example, for a p_1 model:

The

$$\sum_{i,j} y_{ij} \quad \{\sum_{j} y_{ij}\}_{i=1}^{n} \quad \{\sum_{i} y_{ij}\}_{j=1}^{n} \quad \sum_{i<j} y_{ij} y_{ji}$$

Generative Theory for Network Structure

Frank and Strauss (1986)

Represent dependence structure of a model by its *dependence graph*, D:

- vertices: labels of the tie variables in Y.
- edges: if the tie variables are conditionally dependent given the rest of the network

Result: Any, and all, models for Y have the ERGM form with sufficient statistics

$$\prod_{i,j \in A} Y_{ij} \qquad A \in \mathcal{A}$$

where $\mathcal{A}$ is the set of cliques of D.

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

$\Rightarrow$ 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

Generative Theory for Network Structure

Actor Markov statistics

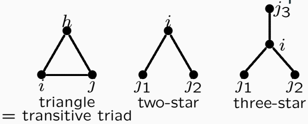
$\Rightarrow$ Frank and Strauss (1986)

– 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.
- triangles: $\mathrm{triangle}(y) =$
  number of triads that form a complete sub-graph in y.

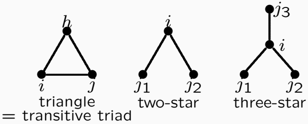


Generative Theory for Network Structure

Actor Markov statistics

$\Rightarrow$ Frank and Strauss (1986)

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

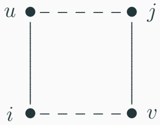


More General mechanisms motivated by conditional independence

⇒ 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



Partial conditional dependence when four-cycle is created

This produces features on configurations of the form:

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

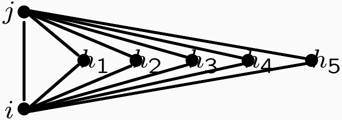
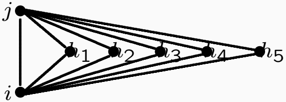


Figure 1: The actors in the non-directed (i, j) edge have 5 shared partners

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

- k-triangle distribution:
  $\mathrm{t}_k(y) =$ proportion of k-triangles in the graph 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*

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

Or generalizations with better statistical properties:
Clustering degrees - the *shared degree* statistics

– $esp_k(y)$ = the proportion of dyads that are tied and have exactly k neighbors in common

Combine into *geometrically weighted edgewise shared partners*

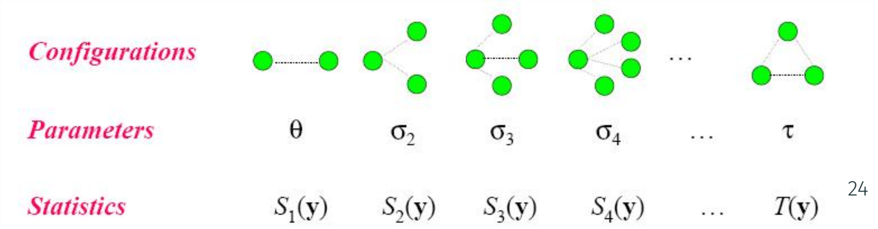
$$gwesp(y) = \sum_{k=1}^{g-2} e^{-\eta k} esp_k(y)$$

⇒ Snijders, Pattison, Robbins and Handcock (2006)
⇒ Handcock and Hunter (2006)

Homogeneity of model parameters

If we assume that isomorphic configurations have equal parameters, then:

- There is one parameter for each class of network configurations
- The corresponding statistic is the number of configurations in *y*
- For example, for an Actor Markov model:



The Model

Probability distribution of the set of possible graphs

$$P(Y = y) = \frac{\exp\left\{\sum_{k=1}^{K} \theta_k g_k(y)\right\}}{c(\theta)}$$

where $\theta_k k = 1, 2...K$ are parameters $g_k(y) k = 1, 2...K$ are statistics, and $c(\theta)$ is a normalizing constant:

$$c(\theta) = \sum_{y \in \mathcal{Y}} \exp\left\{\sum_{k=1}^{K} \theta_k g_k(y)\right\}$$

In other words,

$$P(Y = y) \propto \theta_1 g_1(y) + \theta_2 g_2(y) + \theta_3 g_3(y) + ... + \theta_k g_k(y)$$

The Exponential-family Random Graph Model

We can also re-express it in terms of the odds of tie y_{ij} conditional on the rest of a graph

$$\frac{Pr(Y_{ij} = 1|y_{ij}^C)}{Pr(y_{ij} = 0|y_{ij}^C)} = \exp\left\{\sum_{k=1}^{K} \theta_k(g_k(y_{ij}^+) - g_k(y_{ij}^-))\right\}$$

where y_{ij}^+ is the graph with $y_{ij} = 1$,
y_{ij}^- is the graph with $y_{ij} = 0$ and
y_{ij}^C is the graph excluding y_{ij}.
Implications:

- Log-odds only depend on the "change-score", $\Delta_{ij}^k = g_k(y_{ij}^+) - g_k(y_{ij}^-)$

The Exponential-family Random Graph Model

We can also re-express it in terms of the odds of tie y_{ij} conditional on the rest of a graph

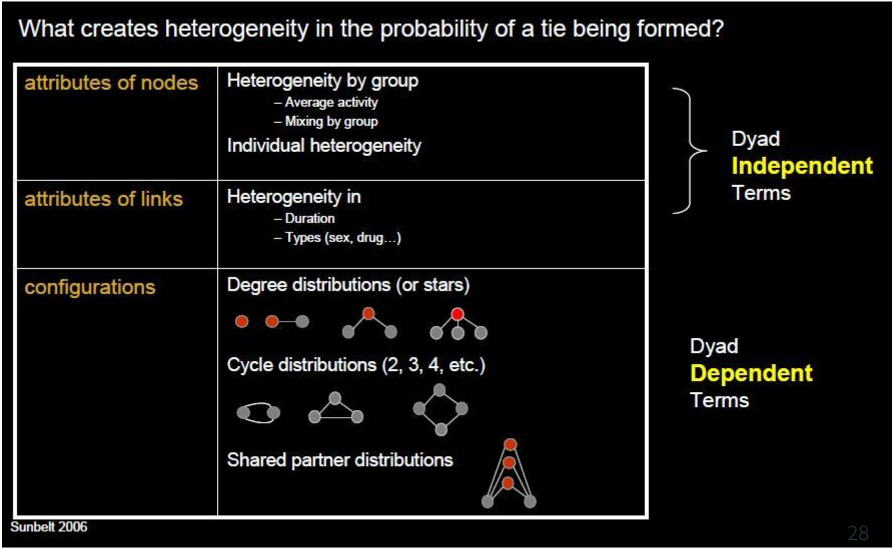
$$\frac{Pr(Y_{ij} = 1|y_{ij}^C)}{Pr(Y_{ij} = 0|y_{ij}^C)} = \exp\left\{\sum_{k=1}^{K} \theta_k(g_k(y_{ij}^+) - g_k(y_{ij}^-))\right\}$$

where y_{ij}^+ is the graph with $y_{ij} = 1$,
y_{ij}^- is the graph with $y_{ij} = 0$ and
y_{ij}^C is the graph excluding y_{ij}.

Implications:

- Each unit change in Δ_{ij}^k increases the conditional log-odds of $Y_{ij} = 1$ by θ_k
- θ_k is the impact of the k^{th}covariate on the log-odds of a tie

What kinds of Covariates?


What creates heterogeneity in the probability of a tie being formed?

attributes of nodes	Heterogeneity by group – Average activity – Mixing by group Individual heterogeneity	Dyad **Independent** Terms
attributes of links	Heterogeneity in – Duration – Types (sex, drug…)	
configurations	Degree distributions (or stars) Cycle distributions (2, 3, 4, etc.) Shared partner distributions	Dyad **Dependent** Terms

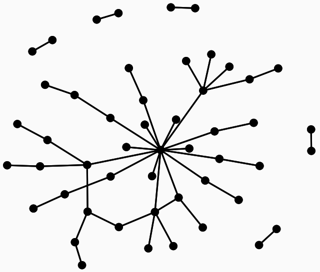
Sunbelt 2006

Illustrations of good models within this model-class

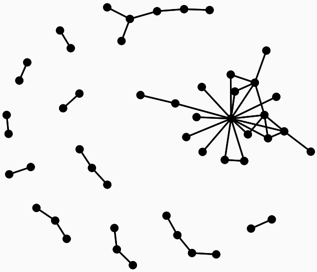
- village-level structure
  - $n = 50$
  - mean clustering coefficient = 15%
- larger-level structure
  - $n = 1000$
  - mean clustering coefficient = 15%
- Attribute mixing
  - Two-sex populations
  - mean clustering coefficient = 15%

village-level structure

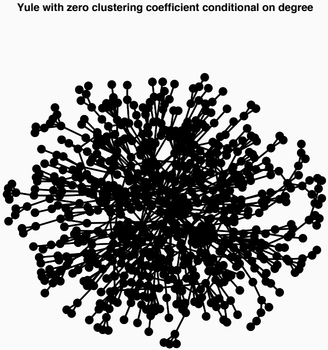
Yule with zero clustering coefficient conditional on degree

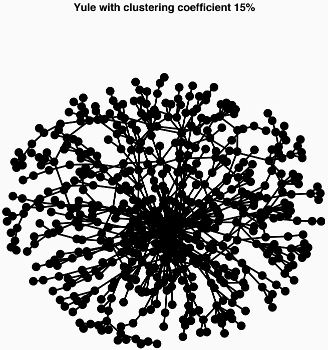


Yule with clustering coefficient 15%

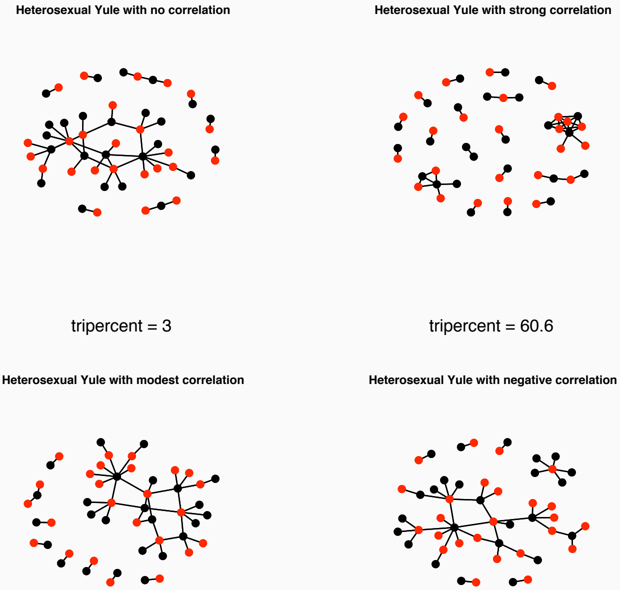


larger-level structure





Heterosexual population



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