Exponential Random Graph Models

University of Washington Network Modeling Group	University of Melbourne	Other contributors
Mark Handcock Steven Goodreau Skye Bender-deMoll Martina Morris (UW) Carter Butts (UCI) Dave Hunter (PSU) Jim Moody (Duke)	Garry Robins Pip Pattison Peng Wang	Tom Snijders Marijtje von Duijn (Groningen)

Outline of Workshop

Introduction	Why statistical models for networks?
Dyad *in*dependence	P1, loglinear models
Dyad dependence	Neighborhoods, conditional independence, Markov models, latent space, general form
ERGM framework	Model specification, degeneracy, diagnostics
Estimation	MCMC, MLE vs. MPLE, simulation
Advanced model specification	New terms, goodness of fit

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)

Statistical analysis, given a model

- Estimate parameters of the process
  - Joint estimation of multiple, possibly correlated, effects
- Inference from sampled data to population
  - Uncertainty in parameter estimates
- Goodness of fit
  - Traditional diagnostics
    - Model fit (BIC, AIC)
    - Estimation diagnostics (MCMC performance)
  - Network-specific GOF
    - Network statistics in the model as covariates
    - Network properties not in the model

Why take a statistical approach?

Descriptive vs. generative goals

- ***Descriptive:*** numerical summary measures
  - Nodal level: e.g., centrality, geodesic distribution
  - Configuration level: e.g., cycle census
  - Network level: e.g., centralization, clustering, small world, core/periphery
- ***Generative***: micro foundations for macro patterns
  - Recover underlying dynamic process from x-sectional data
  - Test alternative hypotheses
  - Extrapolate and simulate from model

ERGMs are a hybrid:

Traditional (generalized) linear models (statistical)

- *but*
  - Unit of analysis: relation (dyad)
  - Observations may be dependent (like a complex system)
  - Complex nonlinear and threshold effects
  - Estimation is different

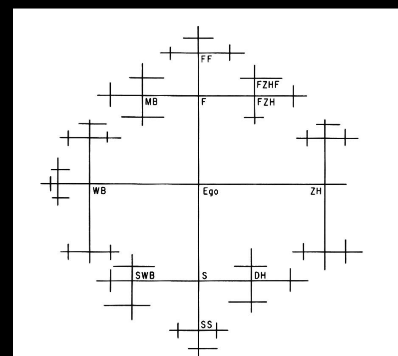
Agent-based models (mathematical)

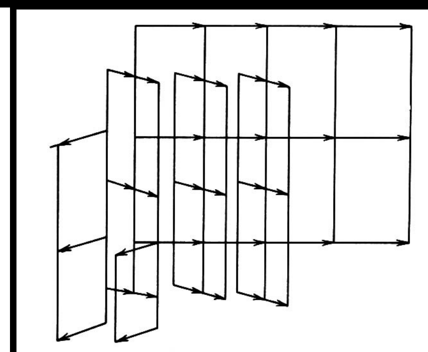
- *but*
  - Can estimate model parameters from data
  - Can test model goodness of fit

Substantive considerations

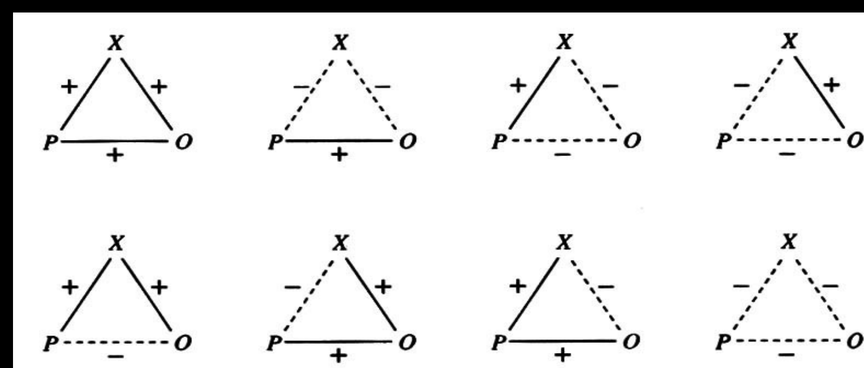
- Original focus of network analysis was generative

Kinship exchange
(Levi-Strauss, White)





Balance theory
(Heider)



Substantive considerations

But different processes can lead to similar macro signatures

For example: “clustering” typically observed in social nets

- Sociality – highly active persons create clusters
- Homophily – assortative mixing by attribute creates clusters
- Triad closure – triangles create clusters

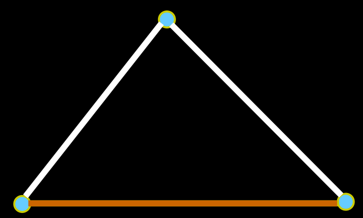
Example: Friend of a friend, or birds of a feather?

Two theories about the process that generates 3-cycles:

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

So, for three actors of the same type:



Cycle-closing tie forms because of transitivity

but also homophily

Basic statistical model

$$P(Y) \propto \theta_1 g_1(y) + \theta_2 g_2(y) + \ldots + \theta_k g_k(y)$$

Probability of the graph

coefficient*covariate

Simplest model: Bernoulli random graph (Erdös-Rènyi)

All ties x_{ij} are equally probable and independent (iid)

So the probability of the graph just depends on the cumulative probability of each tie:

$$\theta \sum_{i=1}^{n} y_{ij}$$

Re-expressed in terms of p(tie)

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

Probability of the graph

$$P(y_{ij}=1|Y^{(ij)})=P(Y^{+})/\{P(Y^{+})+P(Y^{-})\}$$

Probability of *ij* tie, conditional on the rest of the graph

$$\text{logit}[P(y_{ij}=1)|Y^{(ij)}]=\theta_1\partial_1(y^{(ij)})+\theta_2\partial_2(y^{(ij)})+\ldots+\theta_k\partial_k(y^{(ij)})$$

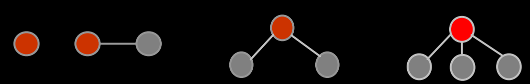
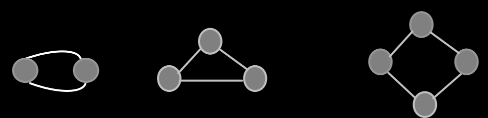
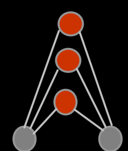
Conditional log odds of the tie.

δ is the “change statistic”, the change in the value of the covariate g(y) when the *ij* tie changes from 1 to 0

So θ is the impact of the 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
attributes of links	Heterogeneity in – Duration – Types (sex, drug…)
configurations	Degree distributions (or stars)  Cycle distributions (2, 3, 4, etc.)  Shared partner distributions 

Dyad **Independent** Terms (attributes of nodes, attributes of links)

Dyad **Dependent** Terms (configurations)

Two modeling components

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

θ $g(y)$

Covariates: what terms in the model?

Coefficients: ***Homogeneity constraints*** for coefficients on terms?

For example:

covariate: edges, etc.

coefficient: each edge has the same probability vs.
each edge has different probabilities or
some edges have different probabilities

Example: homogeneous Bernoulli random graph

Every tie has equal probability, log-odds of a tie = θ

$$P(Y_{ij} = y_{ij}) = e^{\theta y_{ij}} / (1 + e^{\theta})$$

And because the ties are independent,

the joint probability is simply the product of the individual probabilities:

$$P(Y = y) = \exp\left\{ \theta \sum_{i,j} y_{ij} \right\} / \left[1 + \exp\{\theta\} \right]^{\mathbf{E}}$$

$g(y)$ $\quad c(\theta_1, \ldots, \theta_q), \quad q = 1$

P_1 Model Holland and Leinhardt (1981)

Attractiveness, expansiveness & mutuality

$$\sum_{i=1}^{n} \alpha_i y_{i+} + \sum_{j=1}^{n} \beta_j y_{+j} + \rho \sum_{i<j} y_{ij} y_{ji}$$

Marginal effects for each actor (indegree and outdegree)

Homogeneity constraint on mutuality

Dyadic independence

Stochastic block-model Fienberg, Wasserman et al. (1981,5)

Add exogenous attributes as covariates

group specific indegree outdegree, and mixing

$$\sum \alpha_k \sum_{i\in k} y_{i+} + \sum \beta_l \sum_{j\in l} y_{+j} + \sum \phi_{kl} \sum_{i\in k,\, j\in l} y_{ij} + \rho \sum y_{ij} y_{ji}$$

Similar to loglinear models for attribute based mixing

Still dyadic independence

What happens when ties are dependent?