Exponential Random Graph Models (ERGM)

Jeffrey A. Smith
University of Nebraska-Lincoln
Department of Sociology
Social Networks & Health Workshop
May 19, 2016

The Macro Structure of the Session(s)

- Introduction to ERGMs (1 hour and 15 minutes)
  - Introduction to the model
  - Estimation
- Hands on approach to modeling networks using ERGM (1 hour and 30 minutes)
  - A quick review of R (data management, network objects…)
  - Describing and visualizing the network
  - Fitting the network using statnet
    - Model terms and constraints
    - Interpretation
  - Checking the diagnostics of the model
  - Checking model fit

Goals

- Have a basic sense of ERGMs
  - What they are and what you can do with them
- Be able to fit an ERGM to your data
- Be able to tell if the algorithm “worked”
- Be able to tell if the model was a good one

- Run the model, know what it means, know if there are any problems

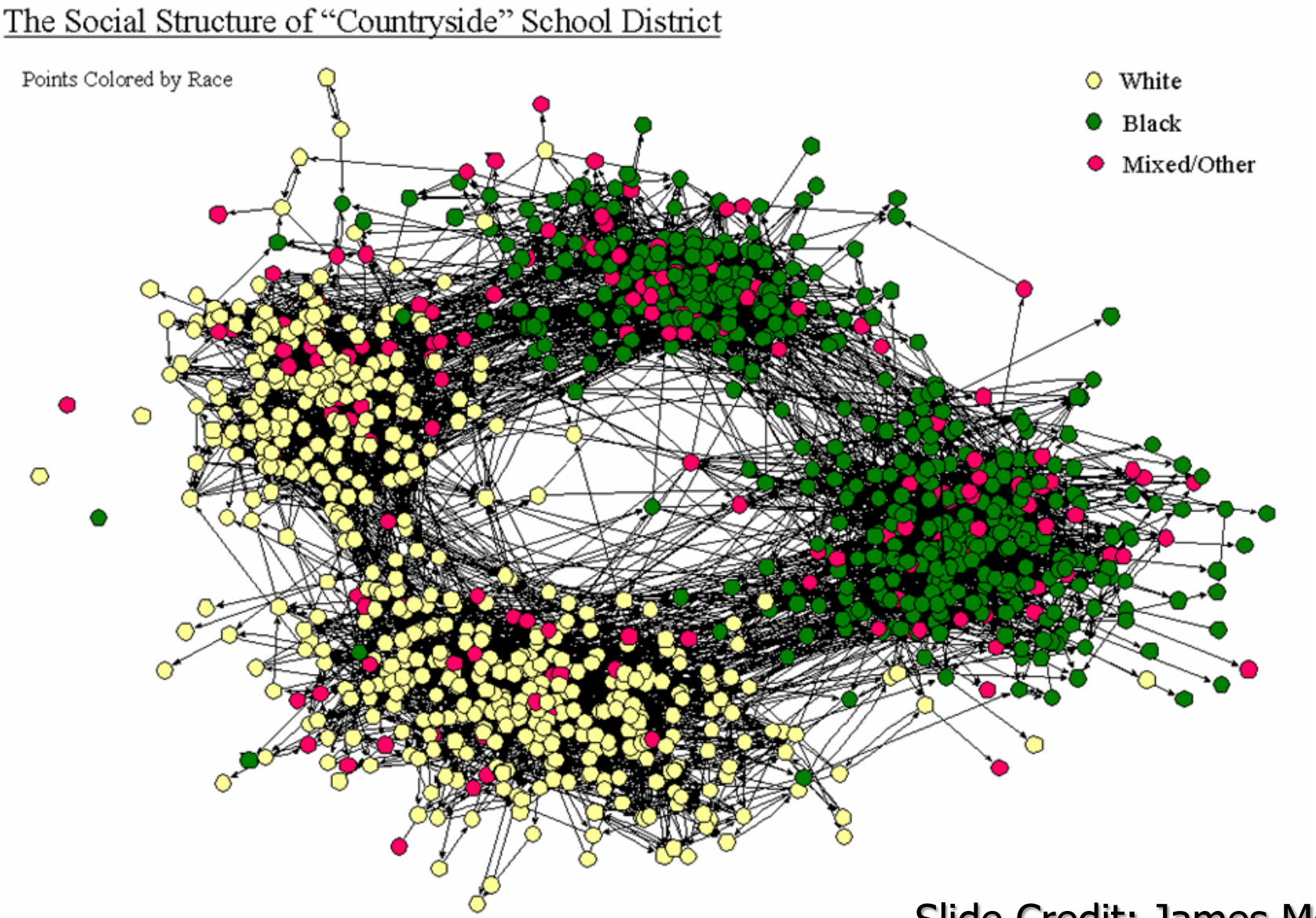
Introduction to ERGM

- Introduction
  - Why statistically model a network?
- Introduce the exponential family
- Some example models: simple to complex
- Estimation

Some Preliminaries

- Formally: a network is a set of n nodes and the social relationships between each pair of nodes.
  - For each pair of actors, or nodes, i,j in the set N (N=1,2…n), let Y_{ij}= 1 if there exists a tie from i to j and Y_{ij}=0 if no tie exists
- Assume here that we are working with complete network data
  - But can run ERGMs using sample data (Morris and Krivistky et al. 2011; Smith 2012)

Why Statistically Model a Network?



Slide Credit: James Moody

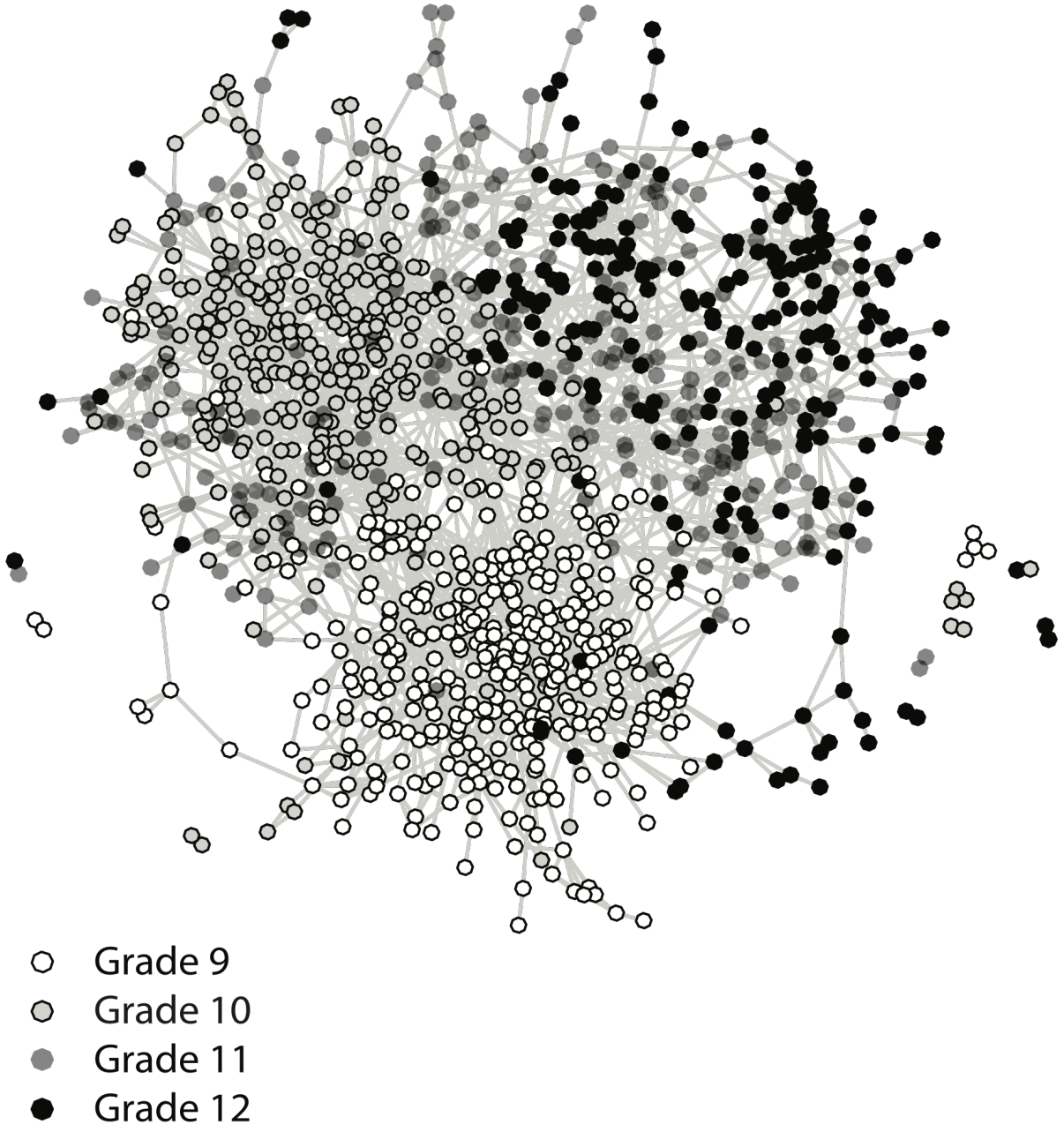
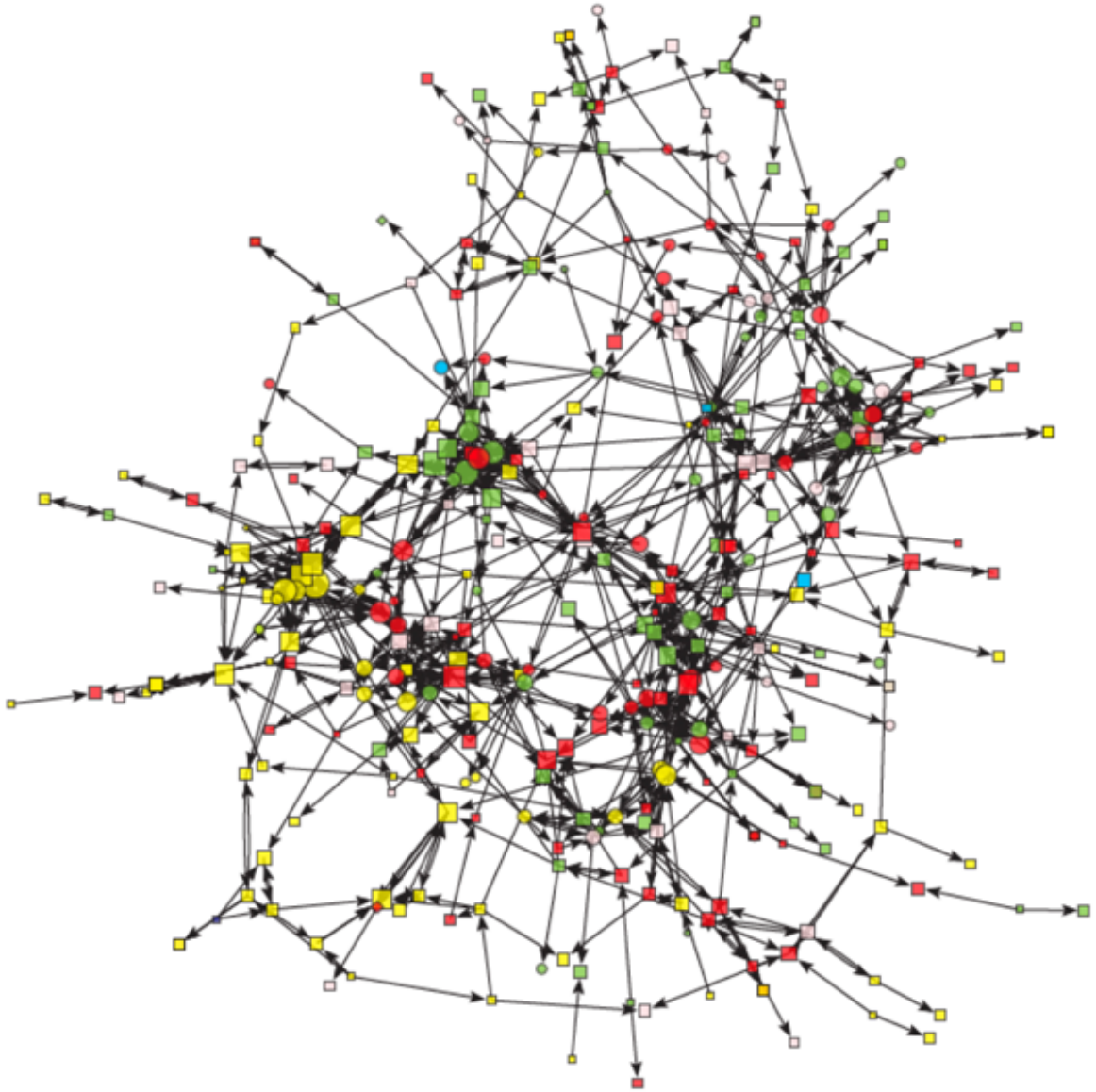
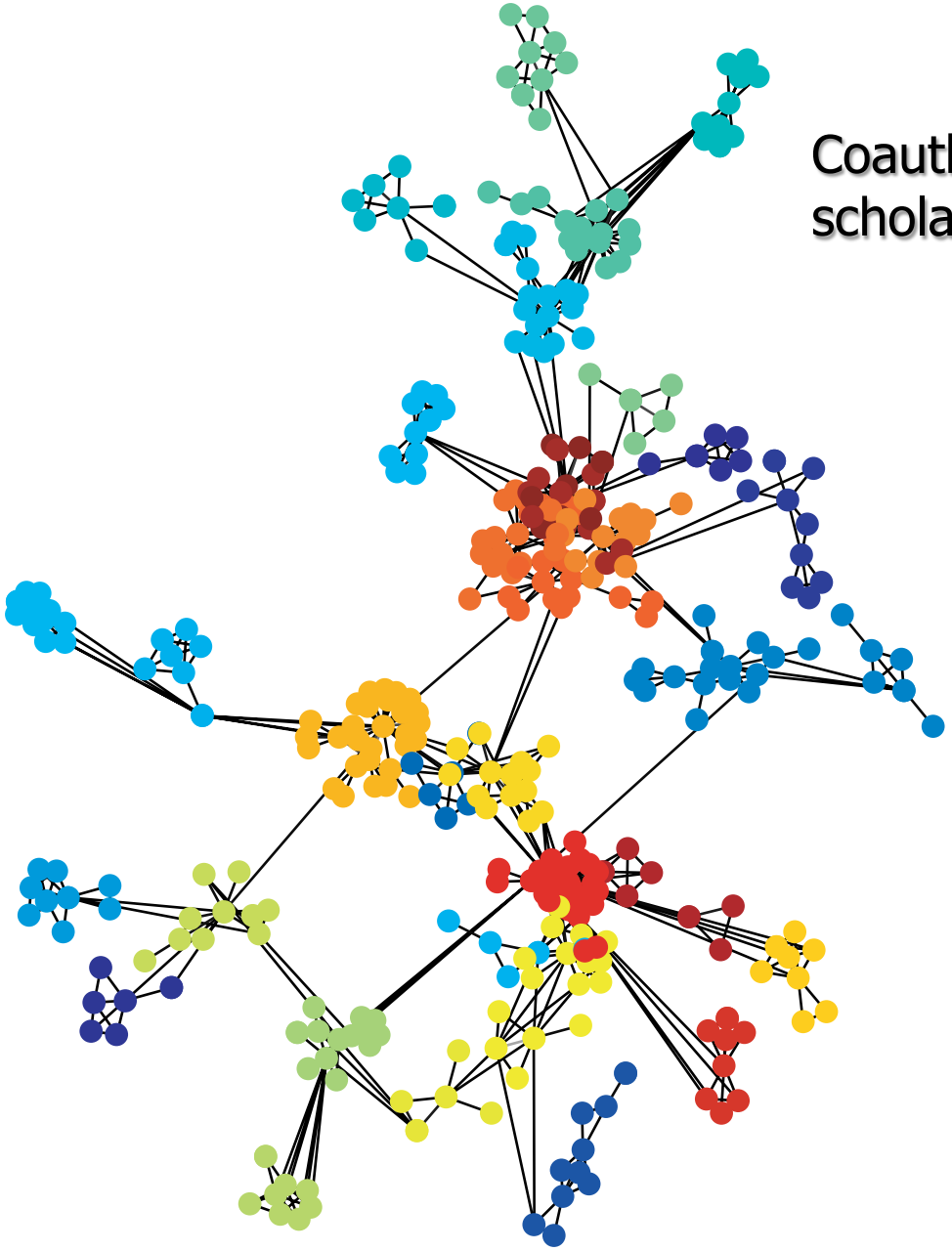

Grade 9
Grade 10
Grade 11
Grade 12

FIGURE 4. School 39 – Clustering and Hierarchy



Key: Colors reflect race, where yellow = White, blue = Asian, green = African American, red = Hispanic, and pink = other / mixed. Shape reflects gender, where circles are female, squares are male. Size reflects popularity (# of received friendship selections).



Coauthorship between scholars in physics literature

Slide Credit: Porter et al. 2009

Why statistically model a network? (see Robins et al. 2007)

- 1. Capture regularities in the network while recognizing the uncertainties surrounding the network (and the modeling process)
- 2. Test hypothesis about structural features (relative to chance)
  - Is the amount of cross race contact less than that expected by chance?

Why statistically model a network? (see Robins et al. 2007)

- 3. Test competing hypothesis about process of network generation
  - transitivity versus homophily in generating group structure
- 4. Macro-micro link
  - Can we specify the local processes that generate the features of the whole network?

Example Health Related Questions We Could Answer

- Are popular (high status) adolescents more likely to drink, smoke, etc?
- Are people who smoke/drink, etc. more likely to be friends with others who smoke/ drink?
  - How much of this tendency is due to transitive closure and reciprocity?

Example Health Related Questions We Could Answer

- What processes determine whom one goes to for health advice?
- How prone is the network to epidemics?
  - spread of HIV, HCV
- How effective are different interventions likely to be?

A word of warning

- 1. Fancy network models are not a substitute for asking good, interesting questions
- 2. The following discussion has many formulas and some technical details. The details are not critical for actually running the models. The key is to have some sense of what the models do and the terms you can use in them. We will get to the hands on portion in short order.

The Exponential Family

$$p(X = x) = \frac{\exp\{\theta' z(x)\}}{\kappa(\theta)}$$

Where:
X is a random network on n nodes
x is the observed network
θ is a vector of parameters (like regression coefficients)
z(x) is a vector of network statistics
κ is a normalizing constant, to ensure the probabilities sum to 1:
$\kappa(\theta) = \sum_{\substack{all possible \\ graphs \\ x}} \exp\{\theta' z(x)\}$

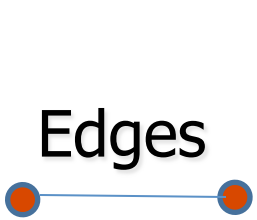
p(X=x)

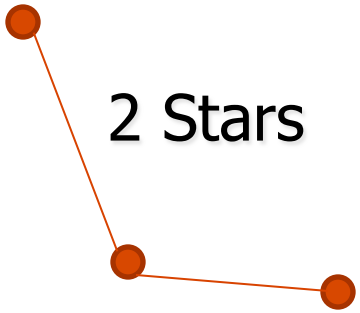
- The dependent variable is all ij pairs in the network
- Trying to predict the presence or absence of ties between pairs of nodes in x
- In other words:
  - Each possible network tie is a random variable
  - Predicting ties between i and j probabilistically

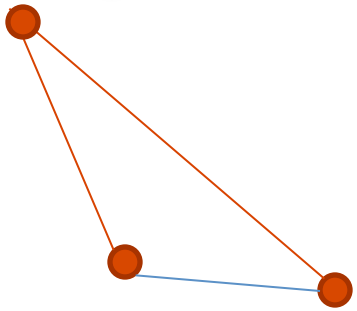
$$\sum \exp\{\theta' z(x)\}$$

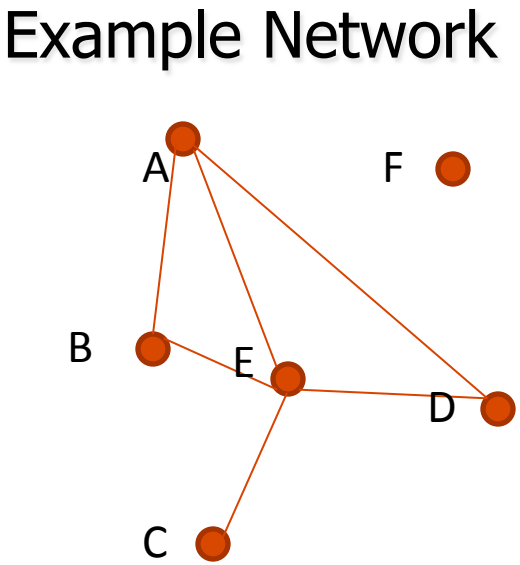
- The independent variables are counts of structural configurations
  - Ultimately change statistics (for the sake of estimation)
- You specify the network statistics to be counted
  - Edges, 2-star, triangle, homophily terms…

Network Statistic Counts for Example Network









6 Edges {A-B, A-E, B-E, A-D, C-E, E-D}

2 Triangles {A-B-E, A-E-D}

11 2-Stars {A-B-E,B-E-A,B-E-D,B-E-C,A-E-D,A-E-C,C-E-D,B-A-E,B-A-D,E-A-D,E-D-A}

- You want model terms that predict ties
  - Model terms that describe the local process generating the network
- When making decisions about the model terms consider
  - Dependence assumptions: for example, reciprocity term means that ij variable (or tie) is dependent on the variable ji.
  - Homogeneity constraints: for example, one parameter for all isomorphic two stars:

θ

- The parameters indicate whether there is a large or small amount of that configuration (edges, triangles) relative to a random network, conditioned on the rest of the model

$K(\theta)$

- Normalizing constant
- More on this when we get to estimation

Some example Models: Simple to Complex

- Simplest: Bernoulli Random Graph
- Dependence and Homogeneity Assumptions
  - Y_{ij} are independent and all equally likely
- Equates to a model with just an edge term (the number of ties in the network)

$$p(X = x) = \frac{\exp\{\theta' z(x)\}}{\kappa(\theta)}$$

Written as:

$$p(X = x) = \frac{\exp\theta\{\sum_{i,j} x_{ij}\}}{\kappa(\theta)}$$

Dyadic Independent Models

- Somewhat more complicated
- Dyads, but not edges, are independent of one another
- Model with edges (θ) and reciprocity (ρ) (mutual) terms

$$p(X = x) = \frac{\exp(\theta \sum_{i,j} x_{ij} + \rho \sum_{i,j} x_{ij} x_{ji})}{k(\theta, \rho)}$$

Dyadic Independent Models: p1 (Holland and Leinhardt-1981)

- Somewhat more complicated
- Dyads, but not edges, are independent of one another
- Model with edges (θ) and reciprocity (ρ) (mutual) terms
- But now add terms for out-degree (α) and in-degree (β) of each node (i.e. attractiveness)

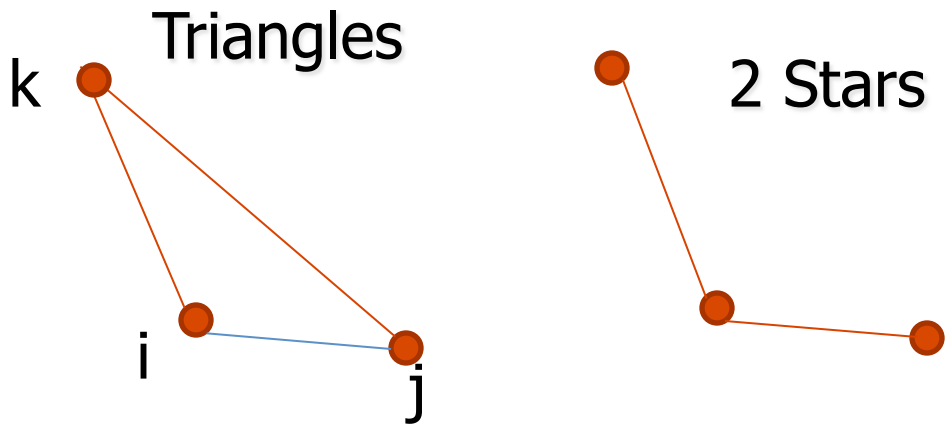
$$p(X = x) = \frac{\exp(\theta \sum_{i,j} x_{ij} + \sum_{i} x_{i+} \alpha_i + \sum_{j} x_{+j} \beta_j + \rho \sum_{i,j} x_{ij} x_{ji})}{k(\theta, \rho, \alpha, \beta)}$$

Dyadic Independent Models

- Can add terms for homophily
  - Ties more likely if i and j share some characteristic
  - Drop homogeneity assumption=differential homophily (ties are more likely if i and j share a characteristic, but more true for some groups than others)
- Nodal characteristics instead of dummies for each person (for example, do girls have fewer ties than boys?)

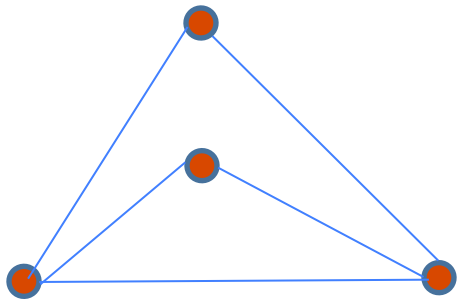
Markov random graphs (Frank and Strauss 1986)

- X_{ij} are independent if they do not share a common actor
- Looking at "local neighborhoods"
- Terms like k-star and triangle



Moving beyond Markov Neighborhoods

- Terms:
- Edgewise shared partner distribution
- Geometrically weighted edgewise shared partner distribution (GWESP)
- For example: 2-triangle



Moving beyond Markov Neighborhoods

- Higher order terms like GWESP often lead to better model fit (more on this later)
- But are they easily interpretable?

List of Typical Terms

- Edges
- Mutuality
- Attractiveness/expansiveness
- Homophily (match, differential, mixing)
- Nodefactor
- Edge-covariates
- Degree distribution
- Triangles
- 2-stars
- Cycles
- GWESP

Example output from an ERGM

```
==========================
Summary of model fit
==========================

Formula:   mesa ~ edges + nodematch("Grade", diff = T) + nodematch("Race",
    diff = F)

Iterations:  8 out of 20

Monte Carlo MLE Results:
                  Estimate Std. Error MCMC % p-value
edges              -6.2228     0.1738      0 < 1e-04 ***
nodematch.Grade.7   2.8256     0.1975      0 < 1e-04 ***
nodematch.Grade.8   2.9148     0.2382      0 < 1e-04 ***
nodematch.Grade.9   2.4474     0.2642      0 < 1e-04 ***
nodematch.Grade.10  2.6080     0.3743      0 < 1e-04 ***
nodematch.Grade.11  3.3376     0.2967      0 < 1e-04 ***
nodematch.Grade.12  3.7041     0.4573      0 < 1e-04 ***
nodematch.Race      0.4214     0.1435      0 0.00333 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     Null Deviance: 28987  on 20910  degrees of freedom
 Residual Deviance:  1919  on 20902  degrees of freedom

AIC: 1935    BIC: 1999    (Smaller is better.)
>
```

But how we do get those coefficients?

Estimation

- Ideally use MLE
  - Find the estimates of θ that make the observed network x most likely
- The difficulty of estimating the vector of coefficients lies in the normalizing constant
- Recall:

$$p(X = x) = \frac{\exp\{\theta' z(x)\}}{\kappa(\theta)}$$

- To estimate coefficient values by MLE, we need to know something about the distribution of graph statistics across “all possible networks”
- But this is difficult as there are a very large number of possible networks-can’t enumerate them all

MLE Estimation

- Likelihood formula:

$$\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0) = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{g}(\mathbf{y}_{\text{obs}}) - \log\left[\frac{\kappa(\boldsymbol{\theta}, \mathcal{Y})}{\kappa(\boldsymbol{\theta}_0, \mathcal{Y})}\right],$$

- Want to estimate unknown parameters ,θ, given the observed network
- Can't get ML estimates directly (because the normalizing constant can't be enumerated) so approximate by simulation

Markov Chain Monte Carlo MLE (MCMCMLE)

- A somewhat complicated process so we won't go into detail, just the basics so we can tell if something hasn't gone well

Markov Chain Monte Carlo MLE (MCMCMLE)

- Basic Idea:
- When model includes dyad dependent terms, use MCMC estimation
- Simulate a set of networks to use as sample
- Necessary to approximate unknown part of likelihood ($k(\theta)$)
  - Gibbs Sampling (or Metropolis)

The Problem of Degeneracy

- Some parameter values will generate a sample of networks (the MCMC sample) with probability mass on 1 or only a few networks
- Often complete or empty network
- Can't get MLE estimates under such conditions
- Problems with triangle and other markov neighborhood terms
- GWESP and similar terms less likely to have such problems (Snidjer et al. 2006)

Goodness of Fit

- How do we know if our model is a good one?
- Simulate networks from resultant ERGM and compare properties of simulated network to observed network
- Distance, degree distributions