Social Network Analysis ERGM 1

Robin Burke
DePaul University
Chicago, IL

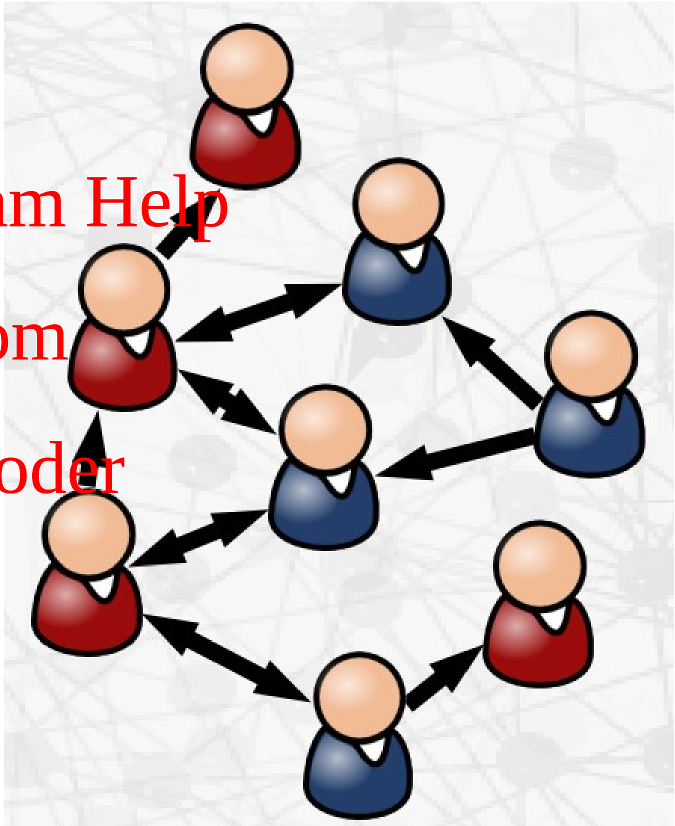
Thanks to Carter Butts for some slides and illustrations

ERGM

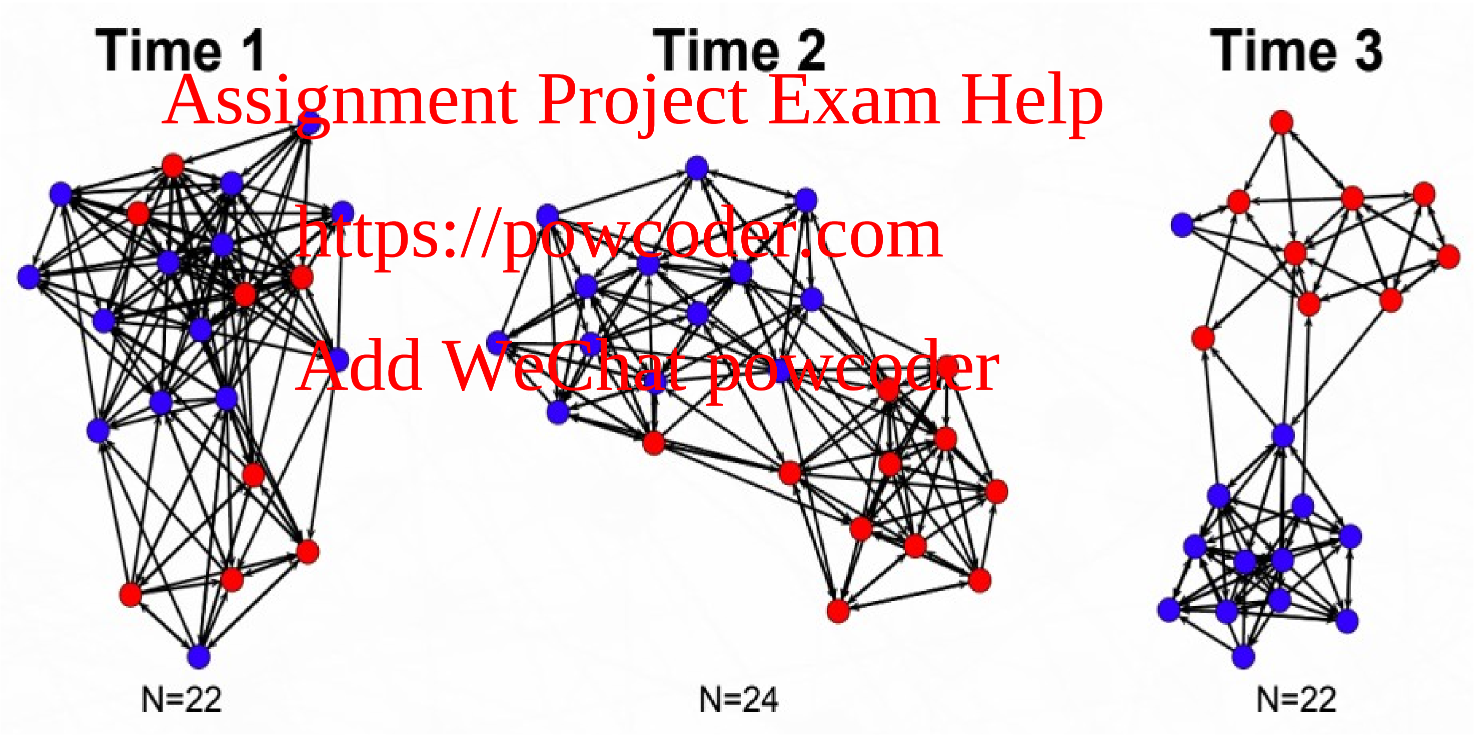
- Fit network formation models to data
- What factors are responsible for the observed effects?

Example: The Reds and The Blues

- A community w/two groups - the "Reds" and the "Blues"
- Question
  - exploring cooperation and trust in the community
  - during a period of upheaval
- We can observe networks of trust/friendship within representative subgroups....



Polarization: Why?



Why does this happen?

- We want to find low-level mechanisms
  - that explain the macro-level changes

- Example

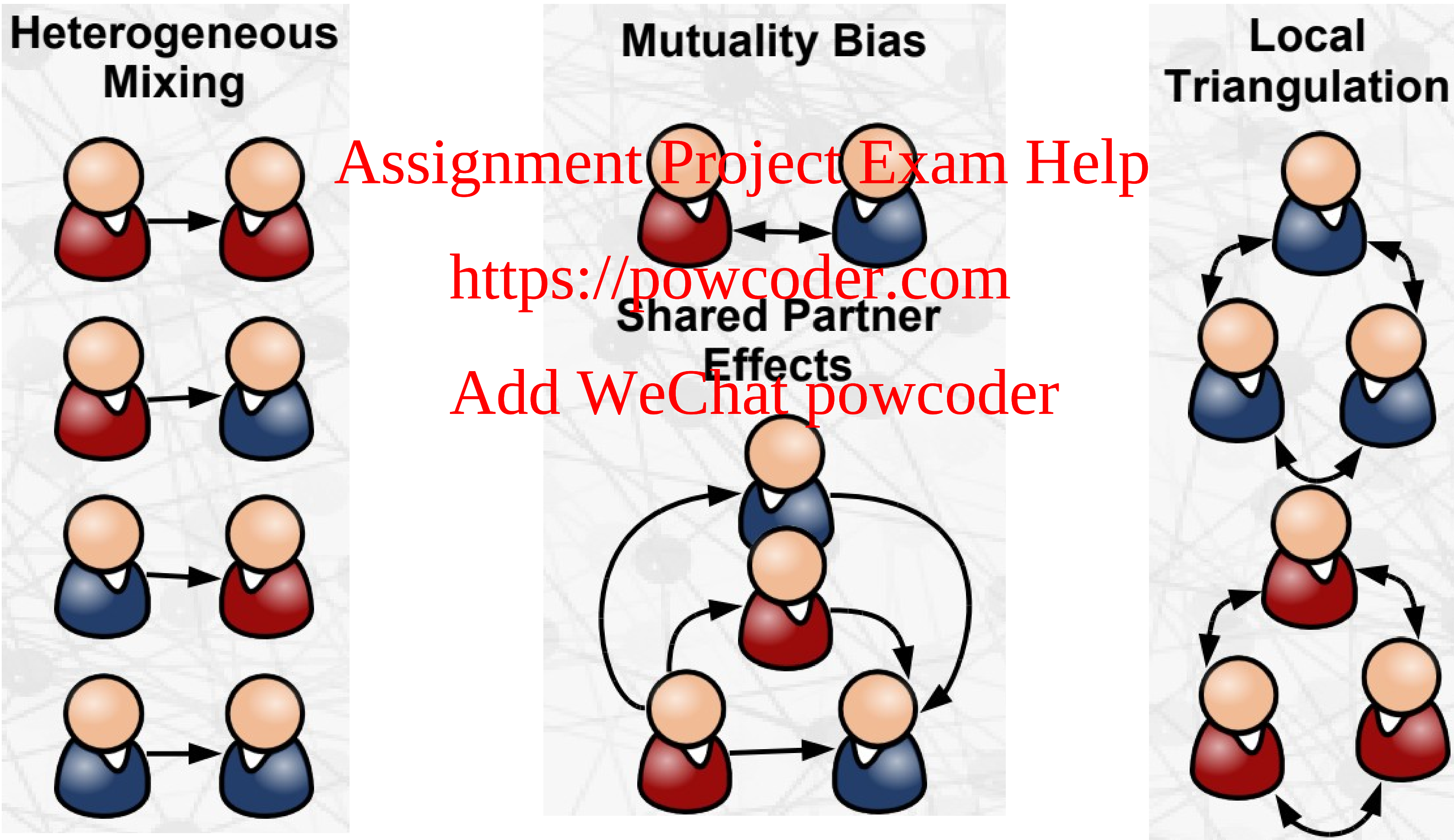
  - Increased tendency to closed triangles?

    - if density is constant
    - Leads to greater clustering
  - Decreased tendency to mutuality
    - combined with increased homophily?
- Can we quantify the effects?

Modeling

- Solution: parametric models
  - Identify candidate structural mechanisms
  - Parameterize using graph statistics
  - Fit models to data
    - Compare alternatives
    - Interpret parameter estimates
    - Assess adequacy

Possible Mechanisms



Factors in Tie Formation

- All ties are not equally probable
  - Chance of an (i,j) edge may depend on properties of i and j
  - Can also depend on other (i,j) relationships
- Examples
  - Homophily
  - Preferential attachment
  - Transitive closure

Graph model

- ER random graph
  - Assumes n nodes
  - Adds edges with uniform random probability
- What we want
  - a model where the edge probability varies
  - based on the nodes themselves
  - based on the characteristics of the network as whole

Logistic Network Regression

- Why not treat edges as independent binary variables?
  - log-odds as a linear function of covariates?
- A (very) special case of standard logistic regression
  - Dependent variable is a network adjacency matrix
- We could do this

$$\log\left(\frac{\Pr(M_{ij}=1)}{\Pr(M_{ij}=0)}\right) = \theta_1 X_{ij1} + \theta_2 X_{ij2} + \ldots + \theta_l X_{ijl} = \theta^T X_{ij}$$

log(p/(1-p) = logit(p) maps (0,1) to (-∞,∞)

$\theta_1..\theta_l$ are the parameters we want to estimate

X_{ijk} is the value of the kth predictor on the i,j pair

But

- The logistic model can be quite powerful, but very limiting
  - No way to model conditional dependence among edges
    - E.g., true triad closure bias, reciprocity
  - Cannot handle exotic support constraints
- Not a good match to certain networks
  - Example: network that must be transitive
- A more general framework: discrete exponential families
  - Very general way of representing discrete distributions
  - Turns up frequently in statistics, physics, etc.
  - Subsumes many common distributions
    - Bernoulli, gamma, Poisson, normal, etc.

Exponential Random Graph Model

θ = [p1, p2, p3 … pk]
A vector of parameters
What we are trying to fit

t(g) = [t1(g), t2(g), t3(g), … tk(g)]
t(g) = a vector of computed properties of g
What we are measuring about g

Dot product of two vectors = scalar value

$$P(g \mid \theta, t, \gamma) = \frac{e^{\boxed{\theta^T t(g)}}}{\sum_{g' \in \gamma} e^{\theta^T t(g')}}$$

g = the graph we observed

γ = the set of all possible graphs of interest

Exponential Random Graph

$$P(g \mid \theta, t, \gamma) = \frac{e^{\theta^T t(g)}}{\sum_{g' \in \gamma} e^{\theta^T t(g')}}$$

- We model the probability of drawing the observed graph (g) from the set of all random graphs of the same size γ
  - as a function of
    - t = functions that compute the characteristics we are interested in
    - θ = the weights associated with those characteristics
- We want to find the θ values that maximize the probability of the observed graph
- Computing the denominator is crazy expensive
  - sum of all possible graphs of a given size!

Equivalent for Adjacency Matrix

$$P(m \mid \theta, t, \mu) = \frac{e^{\theta^T t(m)}}{\sum_{m' \in \mu} e^{\theta^T t(m')}}$$

- Now we are modeling the matrix

- In particular, we can talk about modeling

  - the presence or absence of an edge
  - m_{ij}^{+} represents the presence of edge <i,j>
  - m_{ij}^{-} represents its absence

Conditional odds of an edge

- Ratio of
  - probability of the network with the edge
  - probability of the network without the edge

$$\frac{P(M = m_{i,j}^{+} \mid \theta, t, \mu)}{P(M = m_{i,j}^{-} \mid \theta, t, \mu)}$$

- If this is high
  - then the edge is likely
  - given θ,t,μ

Conditional odds

$$\frac{P(M = m^{+}_{i,j} \mid \theta, t, \mu)}{P(M = m^{-}_{i,j} \mid \theta, t, \mu)} = \frac{e^{\theta^T t(m^{+}_{i,j})}}{\cancel{\sum_{m' \in \mu} e^{\theta^T t(m')}}} \cdot \frac{\cancel{\sum_{m' \in \mu} e^{\theta^T t(m')}}}{e^{\theta^T t(m^{-}_{i,j})}}$$

- Which means...
  - we can avoid the ugly denominator

Simplifying

$$\frac{P(M = m^+_{i,j} \mid \theta, t, \mu)}{P(M = m^-_{i,j} \mid \theta, t, \mu)} = \frac{e^{\theta^T t(m^+_{i,j})}}{e^{\theta^T t(m^-_{i,j})}} = e^{\theta^T [t(m^+_{i,j}) - t(m^-_{i,j})]}$$

- So the conditional odds
  - is a function of the “change score”
  - how the value of each t changes when an edge is added
- Remember that
  - θ is a vector of parameters
  - t is a vector of graph measurements
    - e.g. the # of asymmetric dyads

Conditional Log-odds

$$\log\left[\frac{P(M=m^{+}_{i,j}\mid\theta,t,\mu)}{P(M=m^{-}_{i,j}\mid\theta,t,\mu)}\right]=\theta^{T}\left[t(m^{+}_{i,j})-t(m^{-}_{i,j})\right]=\theta^{T}\Delta_{ij}$$

$$\Delta_{ij}=t(m^{+}_{ij})-t(m^{-}_{ij})$$

- Useful implication
  - each unit change in the measurement t_k when (i,j) edge is present
  - increases the conditional log-odds of (i,j) by θ_k
- This is only conditionally true!
  - The marginal log-odds of an (i,j) edge is allowed to depend on the whole adjacency matrix

Conditional edge probability

- prob = odds/(1+odds)

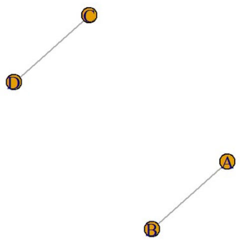
$$P(m_{ij} \mid \theta, t, \mu) = \frac{e^{\theta^T \Delta_{ij}}}{1+e^{\theta^T \Delta_{ij}}} = \frac{1}{1+e^{-\theta^T \Delta_{ij}}} = \text{logit}^{-1}(\theta^T \Delta_{ij})$$

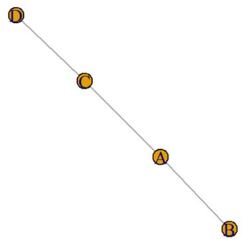
- So, the conditional probability of an edge
  - is the inverse logit of $\theta^T \Delta_{ij}$
- It would be nice if we could perform regression on this
  - “autologistic regression”
- The problem is that it is only conditional probability
  - and edge existence is not independent
  - and t() values are not independent

Estimation

- Perform maximum likelihood estimation
- We want θ^* such that
  - $Pr(M=m_{obs}|t, \theta^*, \mu)$ is maximized, and
  - $E_{\theta^*}(t(M))=t(m_{obs})$
- Such parameters exist
  - as long as data is not “too extreme”
- Can calculate approximate standard errors
  - again, some assumptions, but all indications are that these are usually met

Example





- Simple networks
- What does the ERGM model fit mean?

Break