Exponential Random Graph Models for Social Network Analysis

Danny Wyatt
590AI
March 6, 2009

Traditional Social Network Analysis

- Covered by Eytan
- Traditional SNA uses descriptive statistics
  - Path lengths
  - Degree distributions
  - Thousands of different centrality metrics

Stochastic Social Network Analysis

- Treat networks as realizations of random variables
- Propose a model for the distribution of those variables
- Fit the model to some observed data
- With the learned model
  - Interpret the parameters to gain insight into the properties of the network
  - Use the model to predict properties of the network

This Tutorial

- Exponential Random Graph Models
  - EGRMs, *p**, *p*-star
- How they're applied in sociological research
- How they related to techniques in machine learning
- Some work I've done with them

Exponential Random Graph Models

- Exponential family distribution over networks

$$p(\mathbf{Y} = \mathbf{y} | \boldsymbol{\theta}) = \frac{1}{Z} e^{\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\phi}(\mathbf{y})}$$

$\mathbf{y}$ Observed network adjacency matrix

y_{ij} Binary indicator for edge (i,j)

$\boldsymbol{\phi}(\mathbf{y})$ Features

- Properties of the network considered important
- Independence assumptions

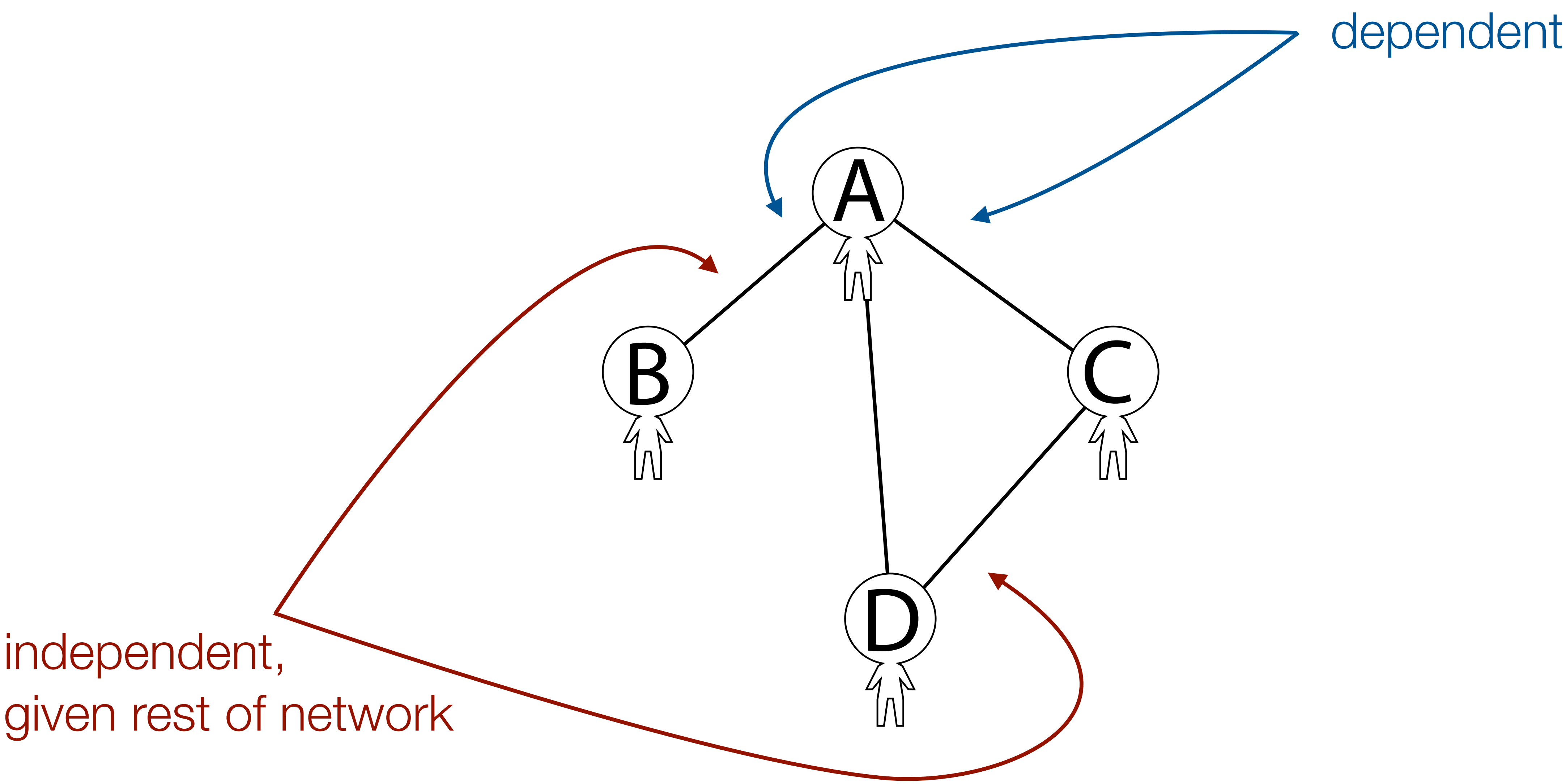
$\boldsymbol{\theta}$ Parameters to be learned

Z Normalizing constant: $\sum_{\mathbf{y}} e^{\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\phi}(\mathbf{y})}$

Markov Random Graphs [Frank86]

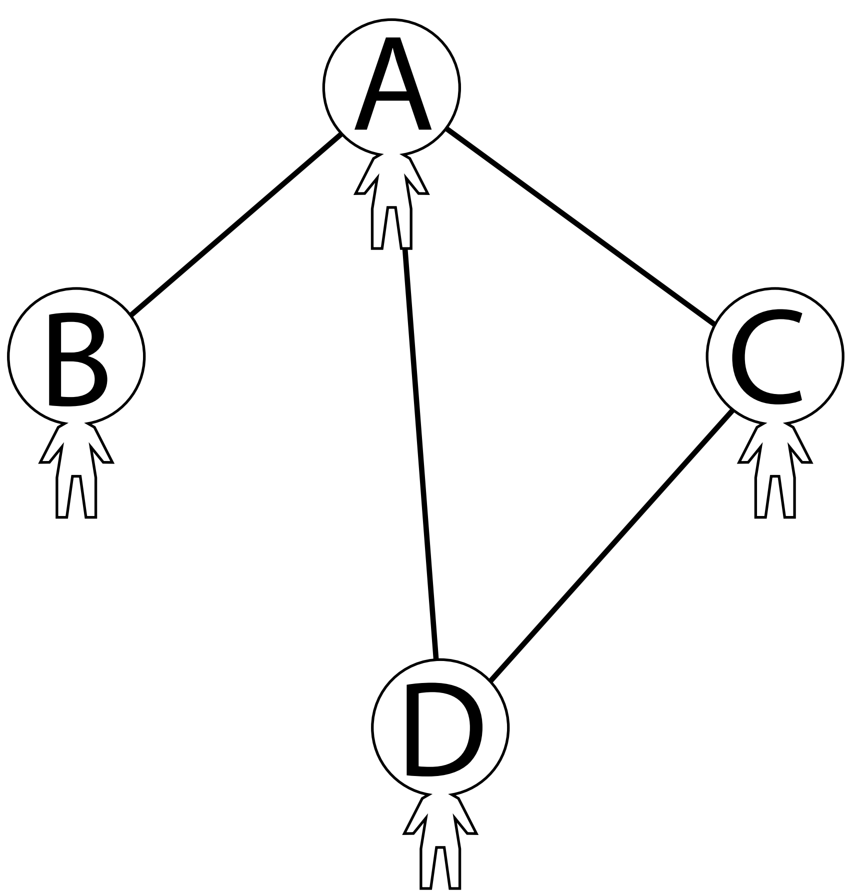
- Edges considered conditionally independent if they don't share a node

- Social phenomena are local

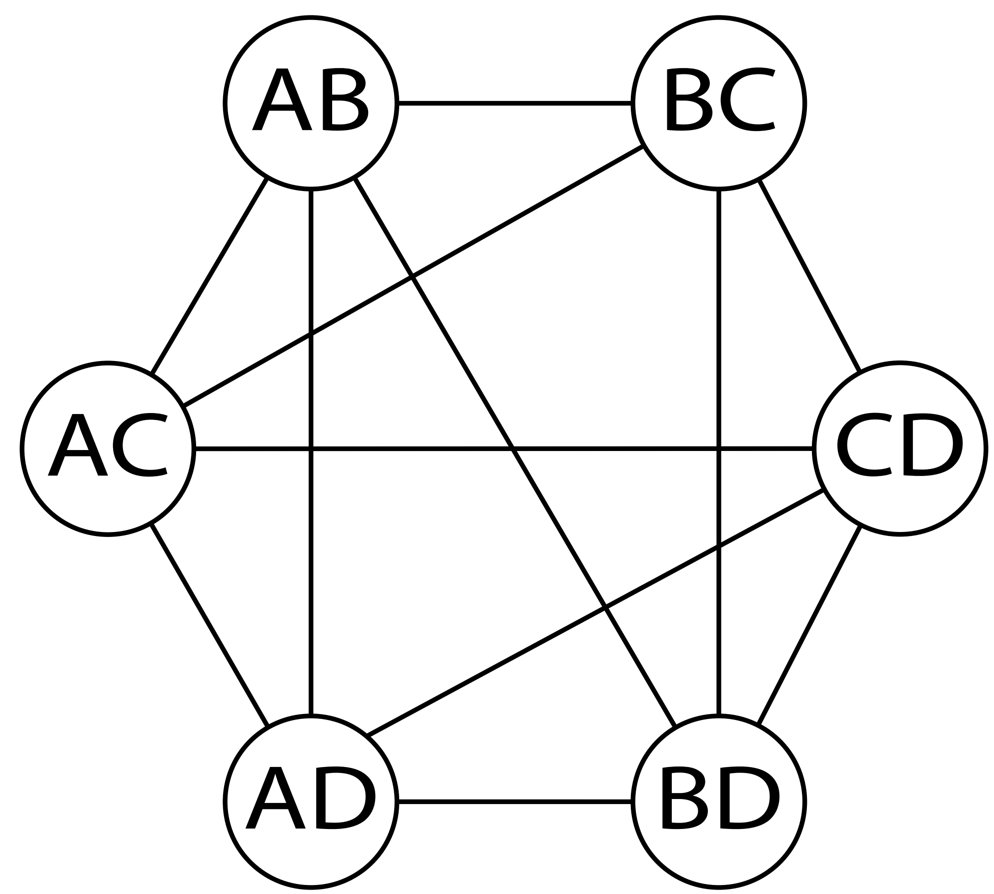


Graphical Model

- Nodes in the graphical model are *edges* in the social network
- Edges in graphical model indicate conditional dependency between edges in the social network



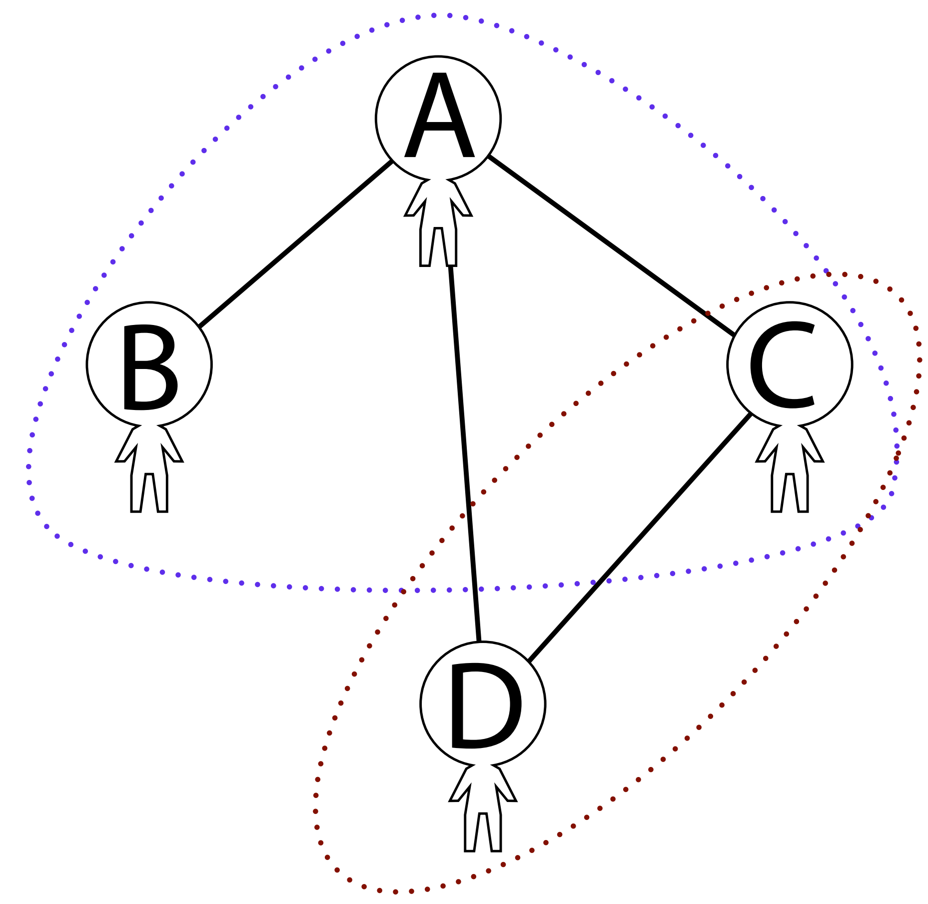
Social Network

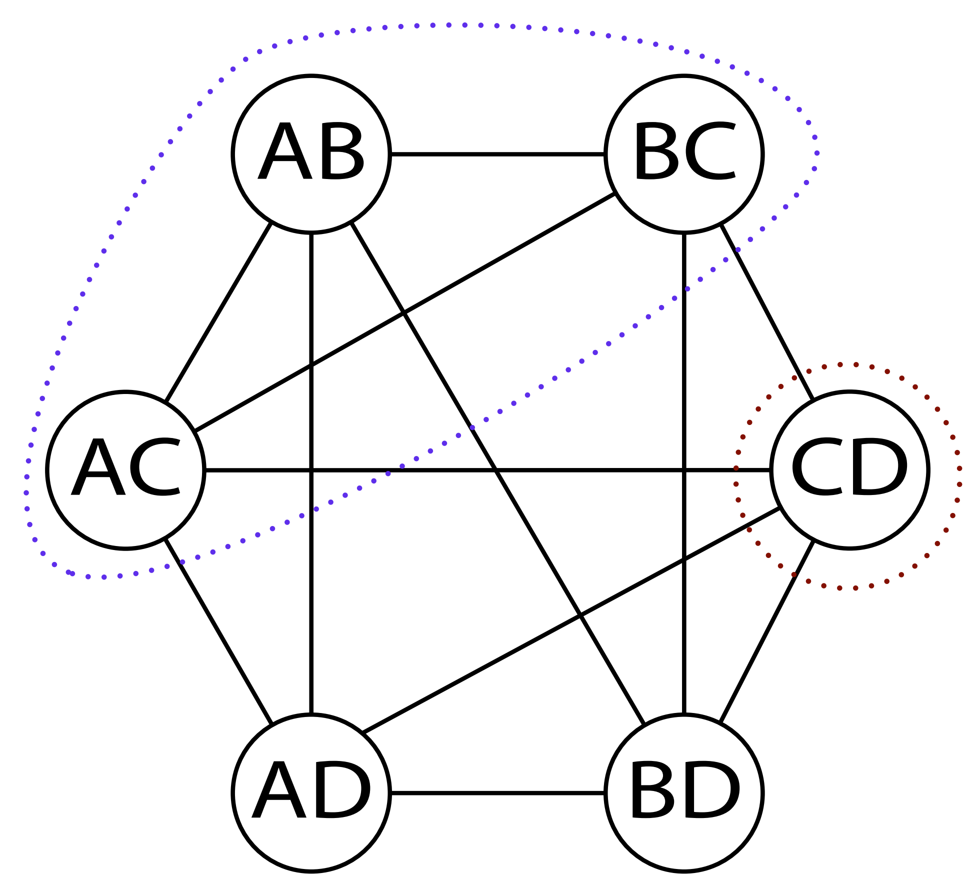


Graphical Model

A Simple Example

- Two repeated features
  - Edge indicator per pair in the social network
    - Singleton potential on each node in the graphical model
  - Triangle indicator per triad in the social network
    - 3 variable clique potentials in graphical model



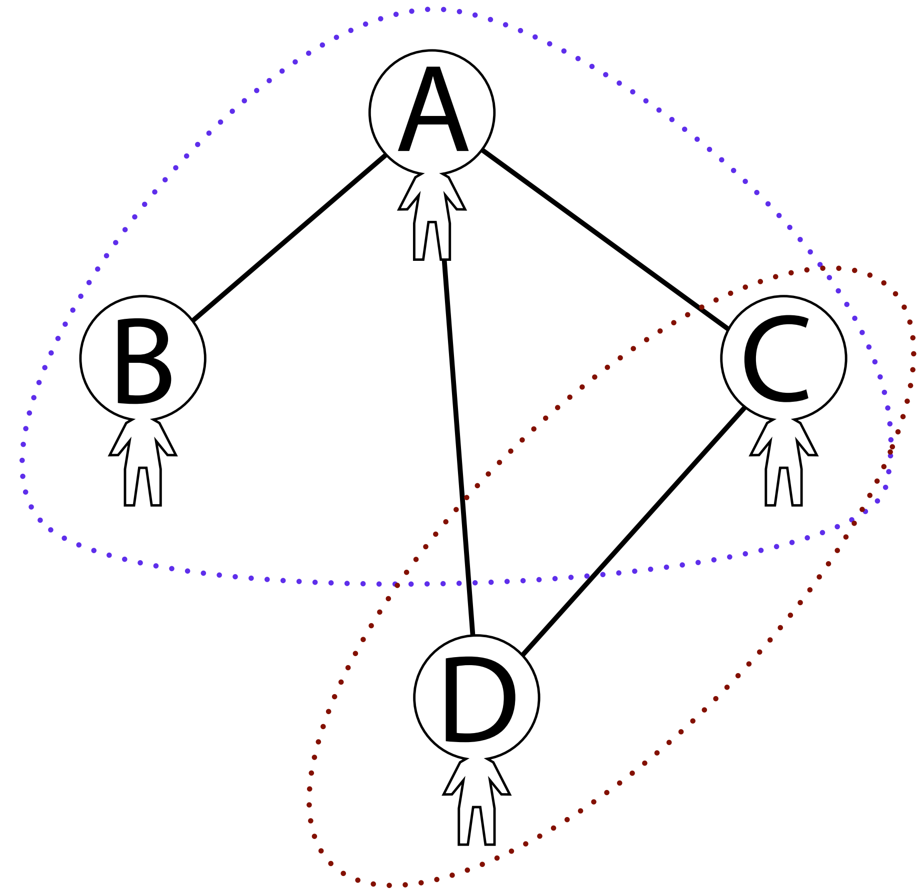


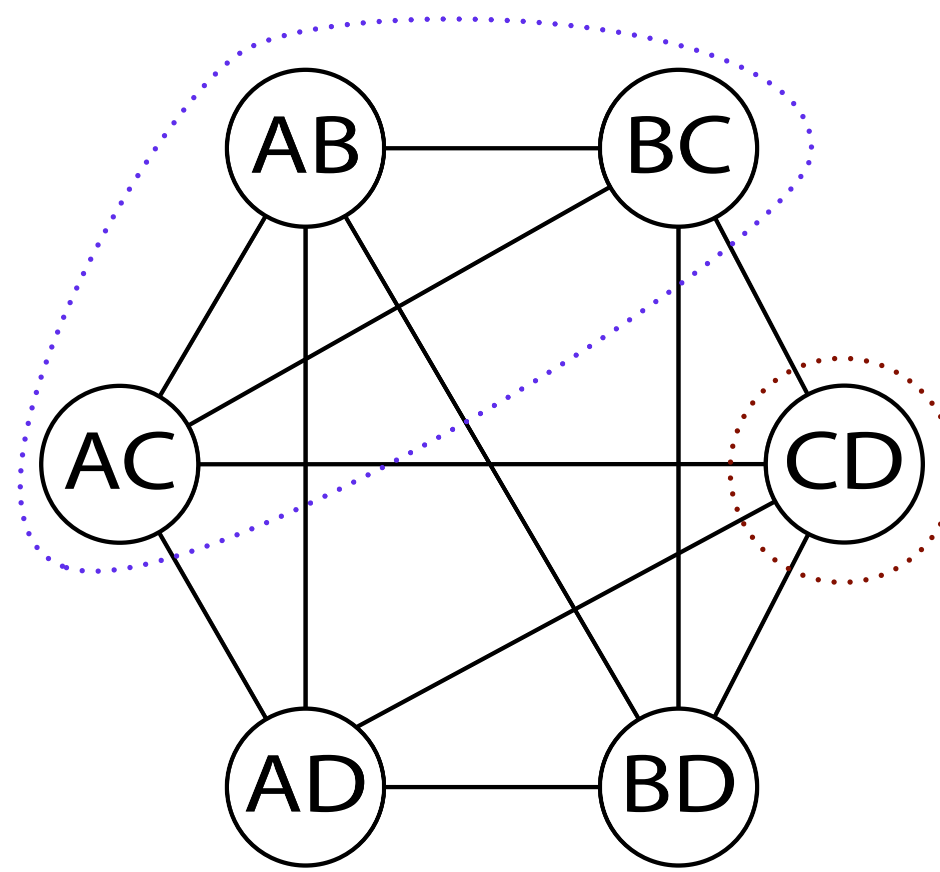
Homogeneity

- Parameters are tied for repeated features
    - Relational model
- Nodes are equivalent
- Isomorphic networks have the same probability
- Larger networks provide more information about the parameters

A Simple Example

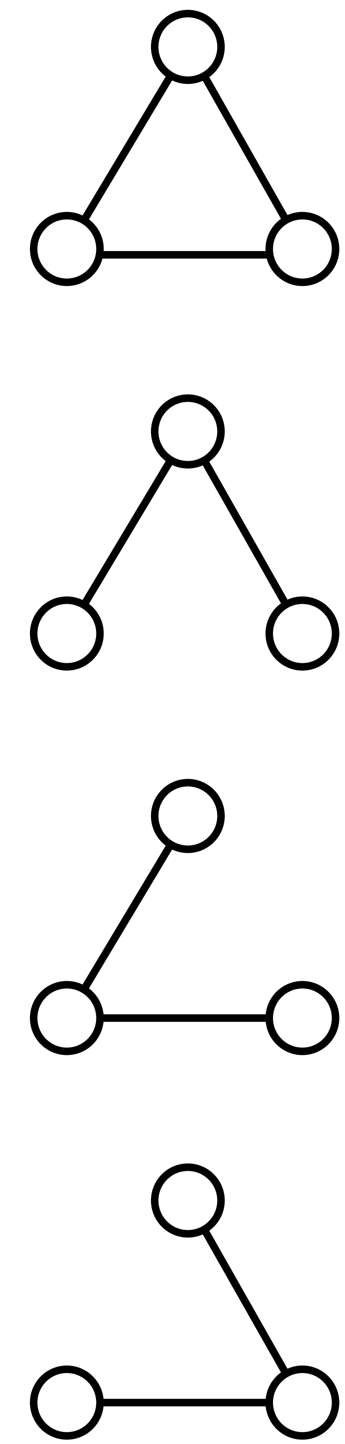
- Repeated tied, features are replaced with counts
- Tied edge indicators → edge count
  - Density
- Tied triangle indicators → triangle count
  - Transitivity





ERGM Network Features

- Usually subgraph counts
- Nested, for interpretability
  - Include all sub-subgraphs
  - e.g. All triangles also include three 2-stars



Nodal Covariates

$$p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z} e^{\boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathbf{y}, \mathbf{x})}$$

- $\mathbf{X}$ – variables with information about people
- Exogenous
  - Sex, age
- Possibly Nonexgenous
  - Religion, political affiliation, smoker
- $\boldsymbol{\phi}(\mathbf{x}, \mathbf{y})$ now also captures information about relationship between ties and covariates

Parameter Learning

- Maximum Likelihood Estimation

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{y}) - \log Z$$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathcal{L}(\boldsymbol{\theta}, \mathbf{y})$$

- Second order gradient ascent

$$\frac{\partial}{\partial \theta_i} \mathcal{L}(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\phi}(\mathbf{y}) - \underset{\mathbf{y}}{\mathrm{E}} \left[\boldsymbol{\phi}(\mathbf{y}) \right]$$

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} \mathcal{L}(\boldsymbol{\theta}, \mathbf{y}) = -\mathrm{cov}_{\boldsymbol{\theta}} \left[\boldsymbol{\phi}(\mathbf{y}) \right]$$

- Both approximated with MCMC

An Optimization [Geyer92, Hunter06]

$$\mathcal{L}(\boldsymbol{\theta}_b, \mathbf{y}) - \mathcal{L}(\boldsymbol{\theta}_a, \mathbf{y}) = \left(\boldsymbol{\theta}_b{}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{Y}) - \log Z_{\boldsymbol{\theta}_b}\right) - \left(\boldsymbol{\theta}_a{}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{Y}) - \log Z_{\boldsymbol{\theta}_a}\right)$$
$$= (\boldsymbol{\theta}_b - \boldsymbol{\theta}_a)^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{Y}) - \log \frac{Z_{\boldsymbol{\theta}_a}}{Z_{\boldsymbol{\theta}_b}}$$

- Change in loglikelihood can be approximated with a single sample drawn at one setting of theta
- Gradient can be approximated as well by re-weighting samples and recomputing expectation

$$w_k^{\boldsymbol{\theta}_b} = \frac{\exp([\boldsymbol{\theta}_b - \boldsymbol{\theta}_a]^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{y}_k))}{\sum_j \exp([\boldsymbol{\theta}_b - \boldsymbol{\theta}_a]^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{y}_j))}$$

- Approximation degrades with distance
- But can take many steps with a single sample

ERGMs in Practice

- Devise the model to capture phenomena of interest
  - Carefully include nested/confounding features
  - Account for all nodal covariates of interest
- Learn the parameters
- See what the parameters tell you about your network
- (Not yet much work on using ERGMs for prediction)

Interpreting Parameters

- Weight is log odds of unit increase in feature value, everything else kept equal
- Positive weight means probability increases with feature value
- Negative weight means probability decreases with feature value
- Zero weight means feature has no effect
- If you've accounted for nodal covariates
    - Network feature weights tell you importance of network structure

Confidence Intervals

$$\mathrm{cov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) \geq \mathbf{I}(\boldsymbol{\theta})^{-1}$$

$$\mathbf{I}(\boldsymbol{\theta}) = -\operatorname*{E}_{\mathbf{Y}}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\mathcal{L}(\boldsymbol{\theta}, \mathbf{y})\middle|\boldsymbol{\theta}\right]$$

$$\hat{\mathbf{I}}(\boldsymbol{\theta}) = -\operatorname*{E}_{\mathbf{Y}}\left[\frac{\partial^2}{\partial \hat{\boldsymbol{\theta}}^2}\mathcal{L}(\hat{\boldsymbol{\theta}}, \mathbf{y})\middle|\hat{\boldsymbol{\theta}}\right]$$

- How reliable are your parameter estimates?
- Use inverse Fisher information to estimate sampling covariance
- Divide by square root of sample size to get standard error
- But what is the sample size?

Model Degeneracy

- As described, models work very poorly
- Learned parameters do not generate data that resembles the input
    - Tend toward wholly connected or completely empty graphs

Model Degeneracy

- Most parameter values place all of the probability on unrealistic networks

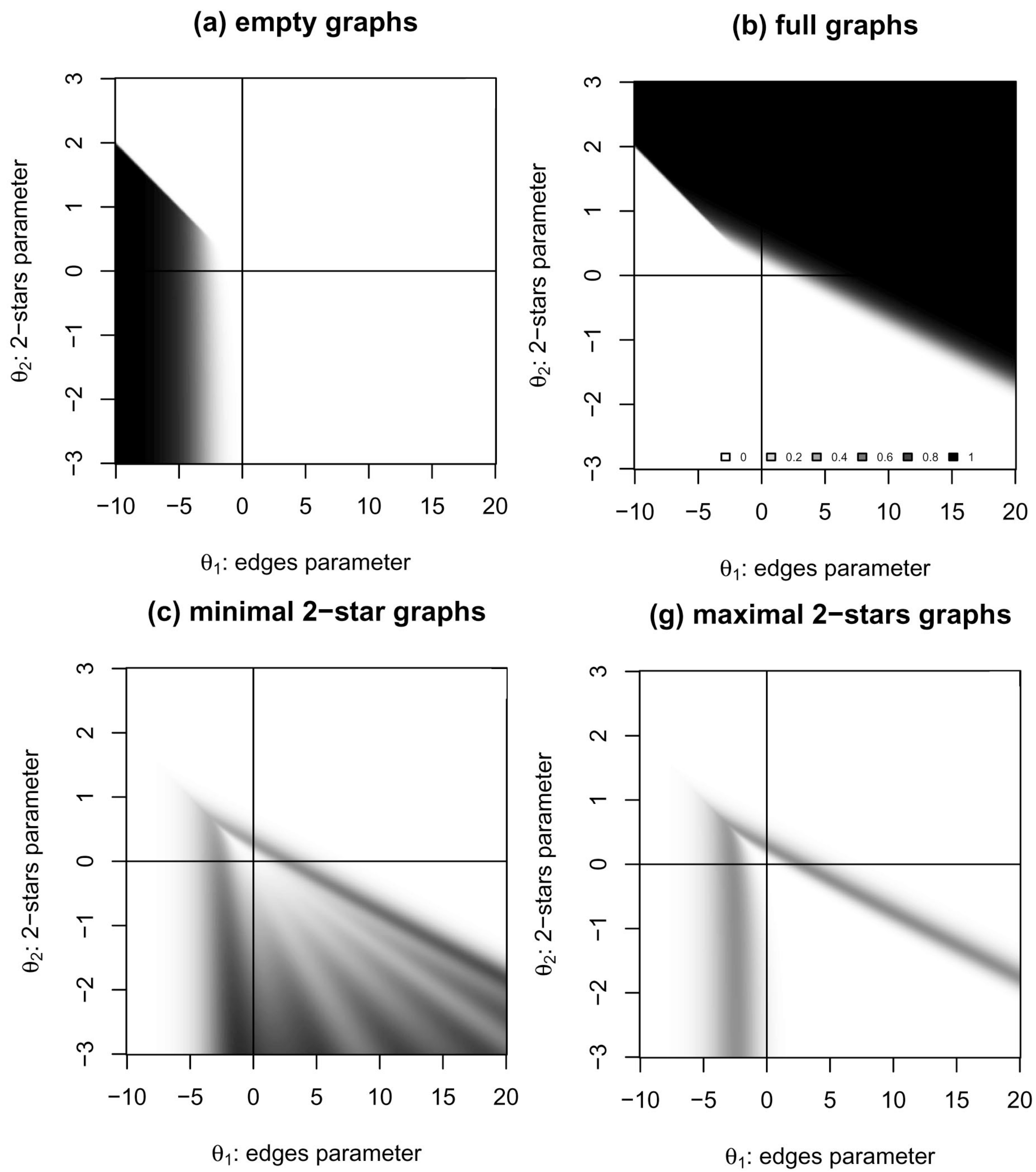


figure from [Handcock03]

Model Degeneracy

- Most parameter values place all of the probability on unrealistic networks

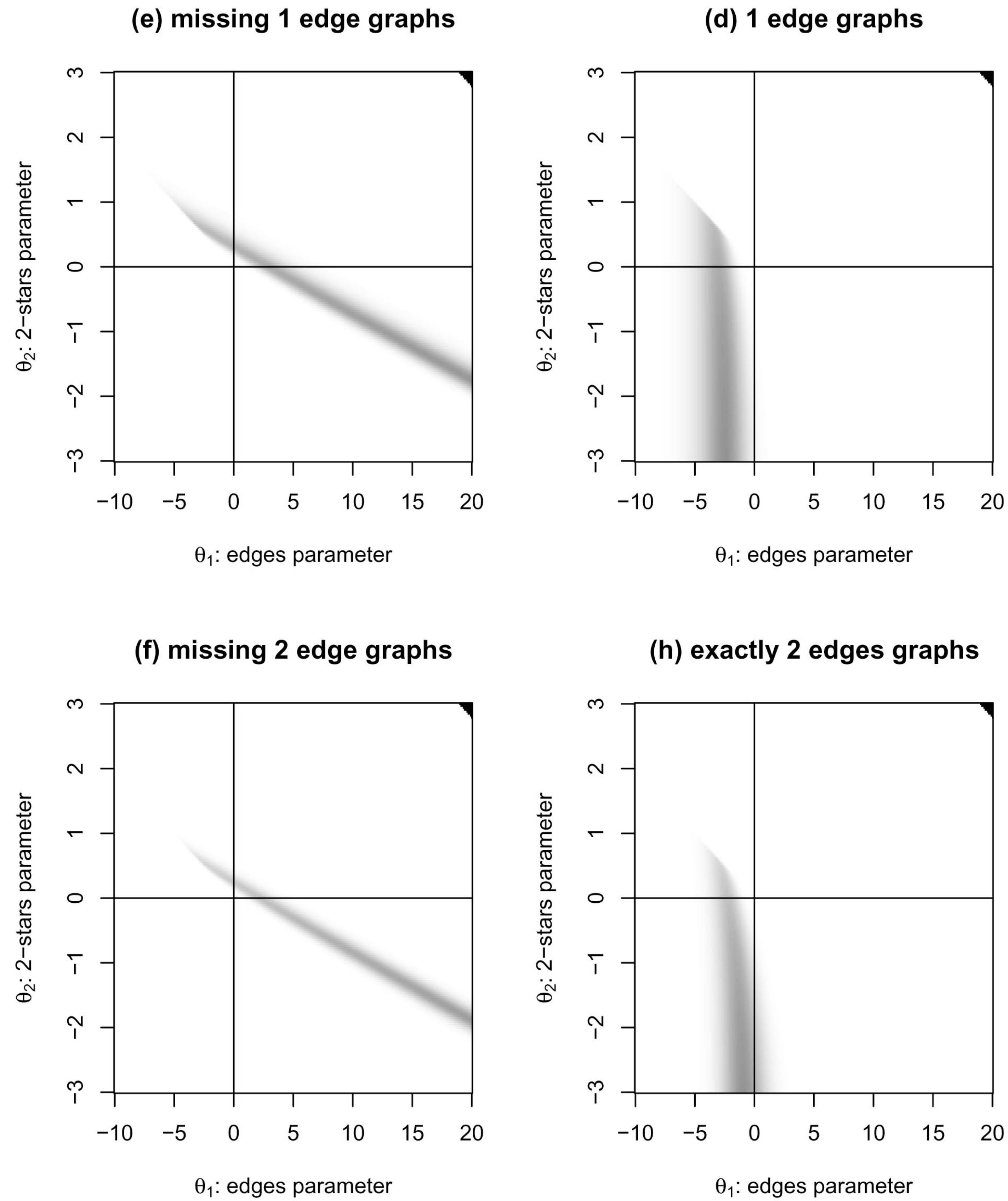


figure from [Handcock03]

Model Degeneracy

- Putting them all together
- Region of “realistic” parameters is small and unfriendly in shape
- Difficult to reach using gradient methods and MCMC

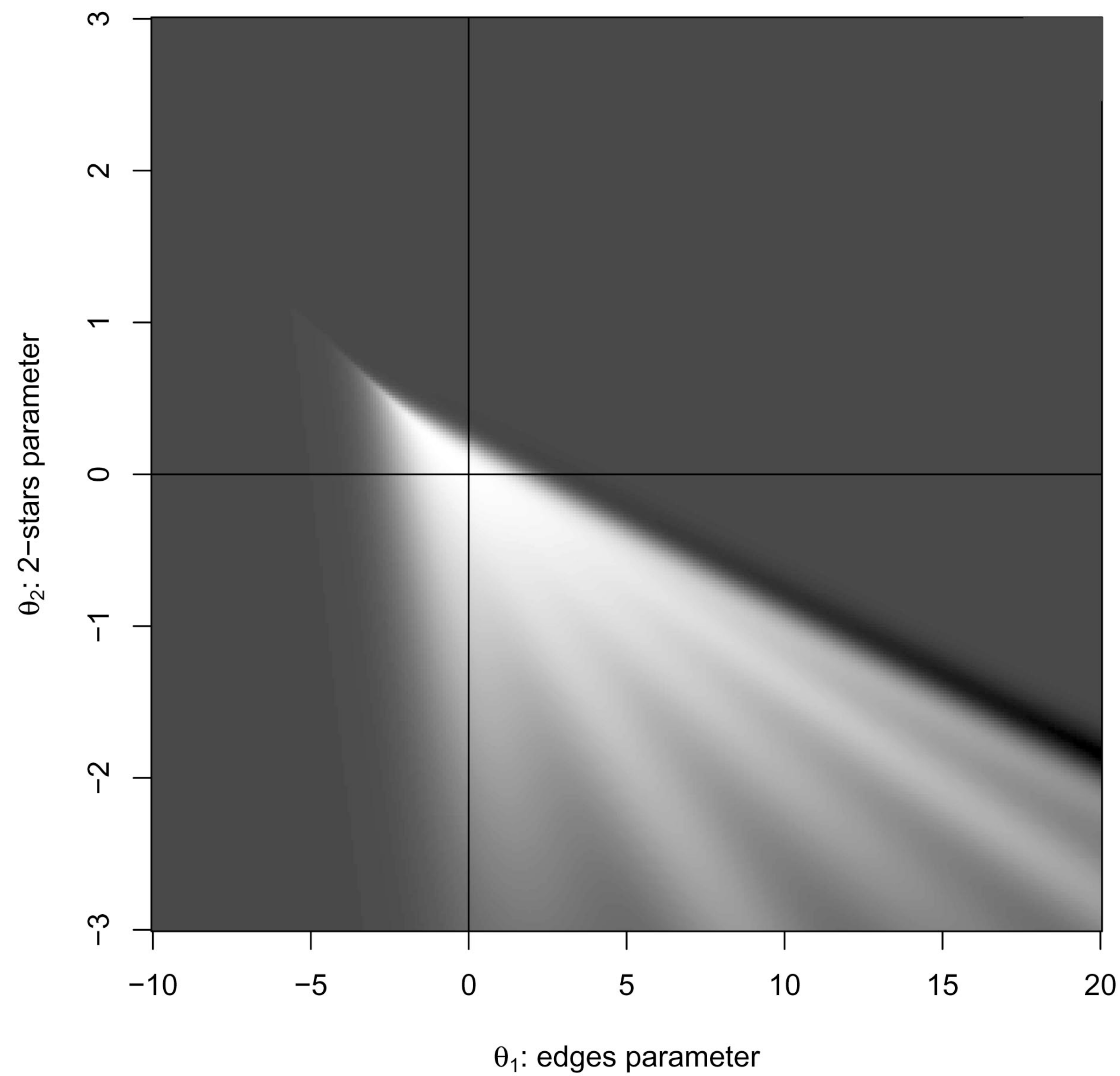
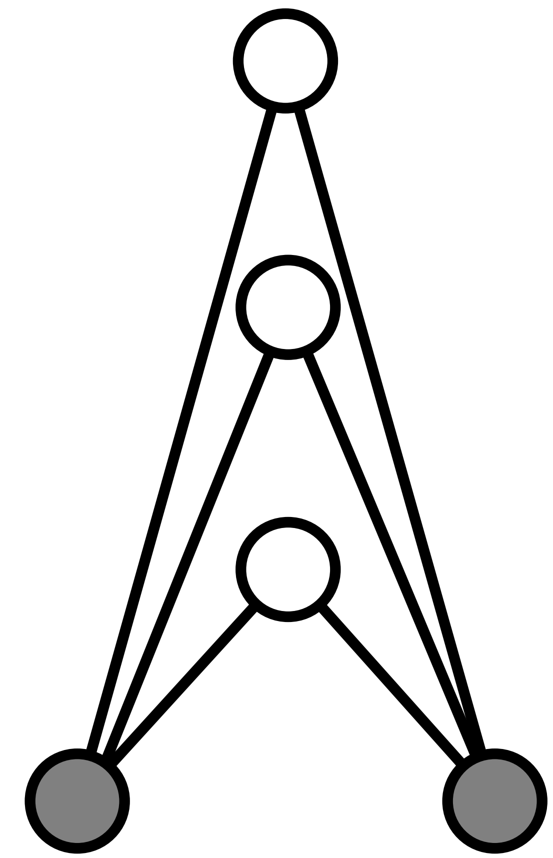


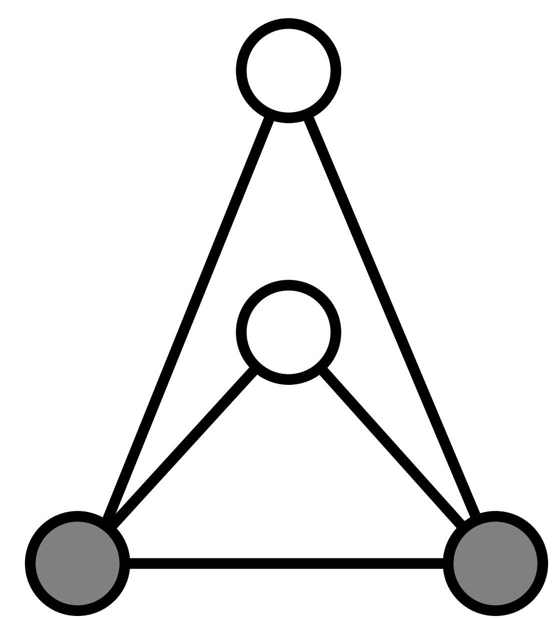
figure from [Handcock03]

New Features for ERGMs

- More nuanced notions of structure
- Look at many more features of the networks
  - Degree histogram
  - Shared partner histograms
- Too many features!
  - Nonparametric



3 shared partners



2 edgewise shared partners

Parameter Constraints [Hunter06]

- Constrain weights to reduce number of parameters
- Exploit ordinality of histogram features
- Maintain socially intuitive parameter values
- Diminishing returns constraint

“cost” parameter

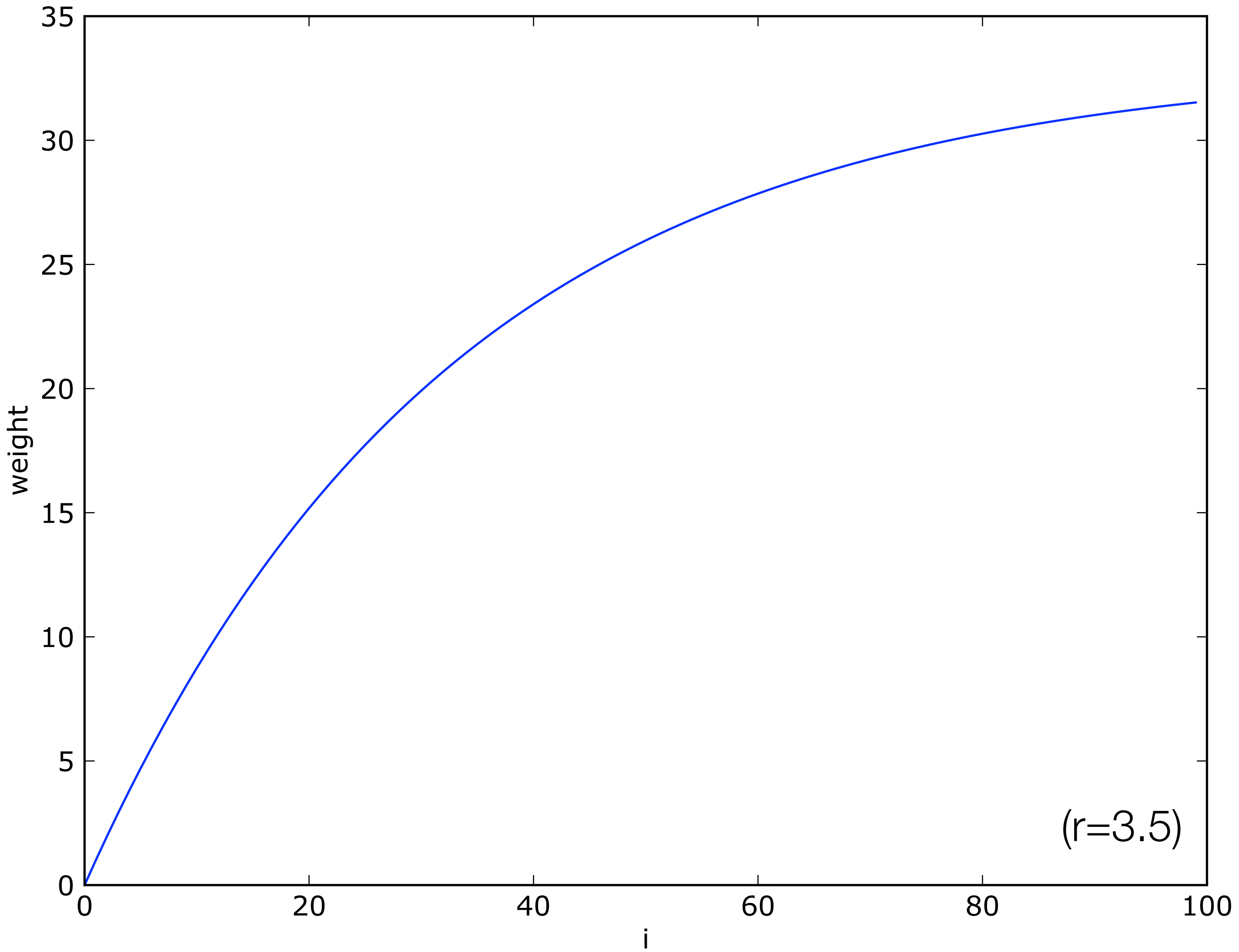
$$\theta_i = w\left(ci \cdot e^r(1 - (1 - e^{-r})^i)\right)$$

usual multiplicative weight

geometric rate parameter

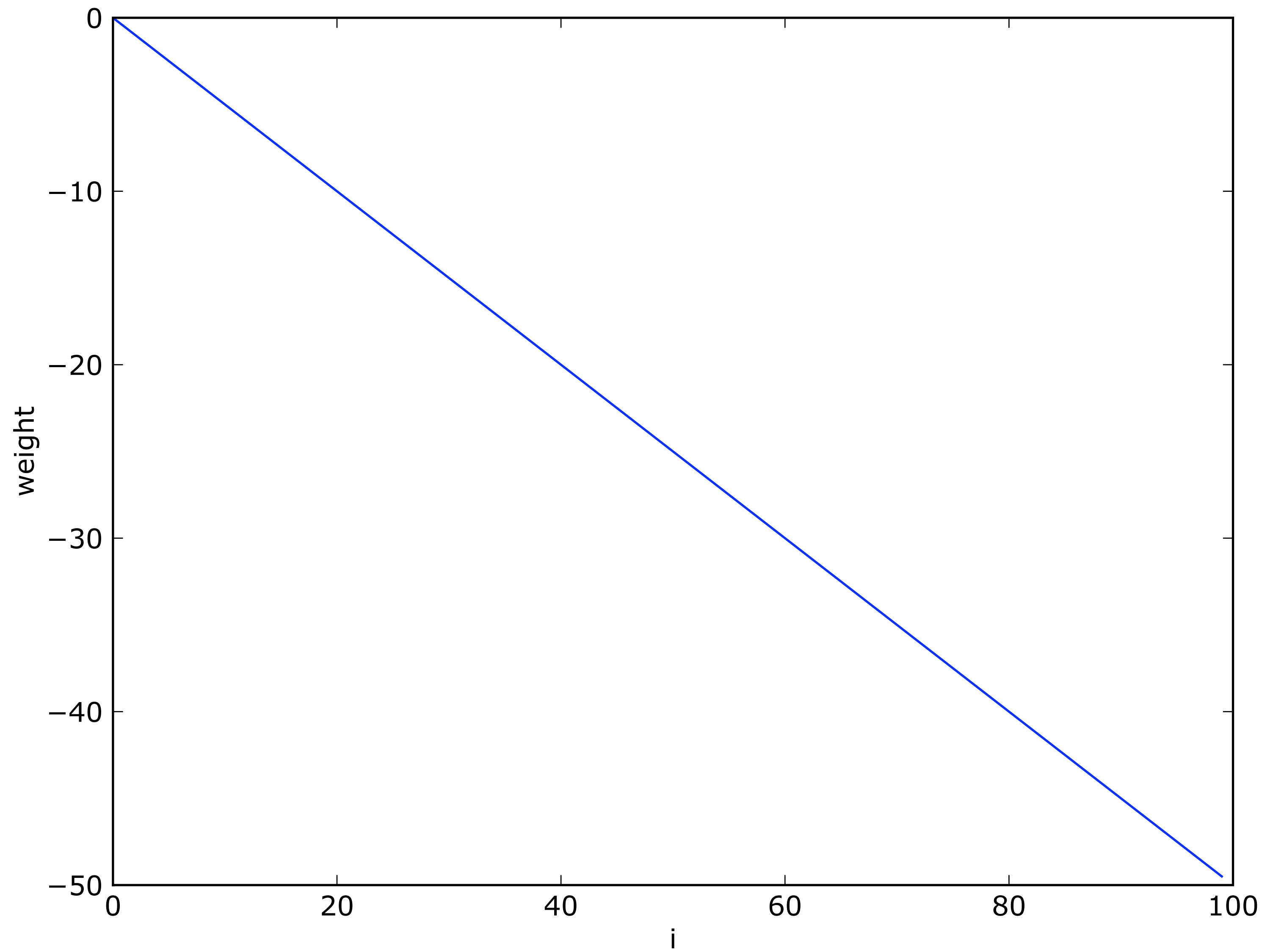
Geometric Weight Constraints

$$\theta_i = w\left(ci \cdot e^r(1-(1-e^{-r})^i)\right)$$



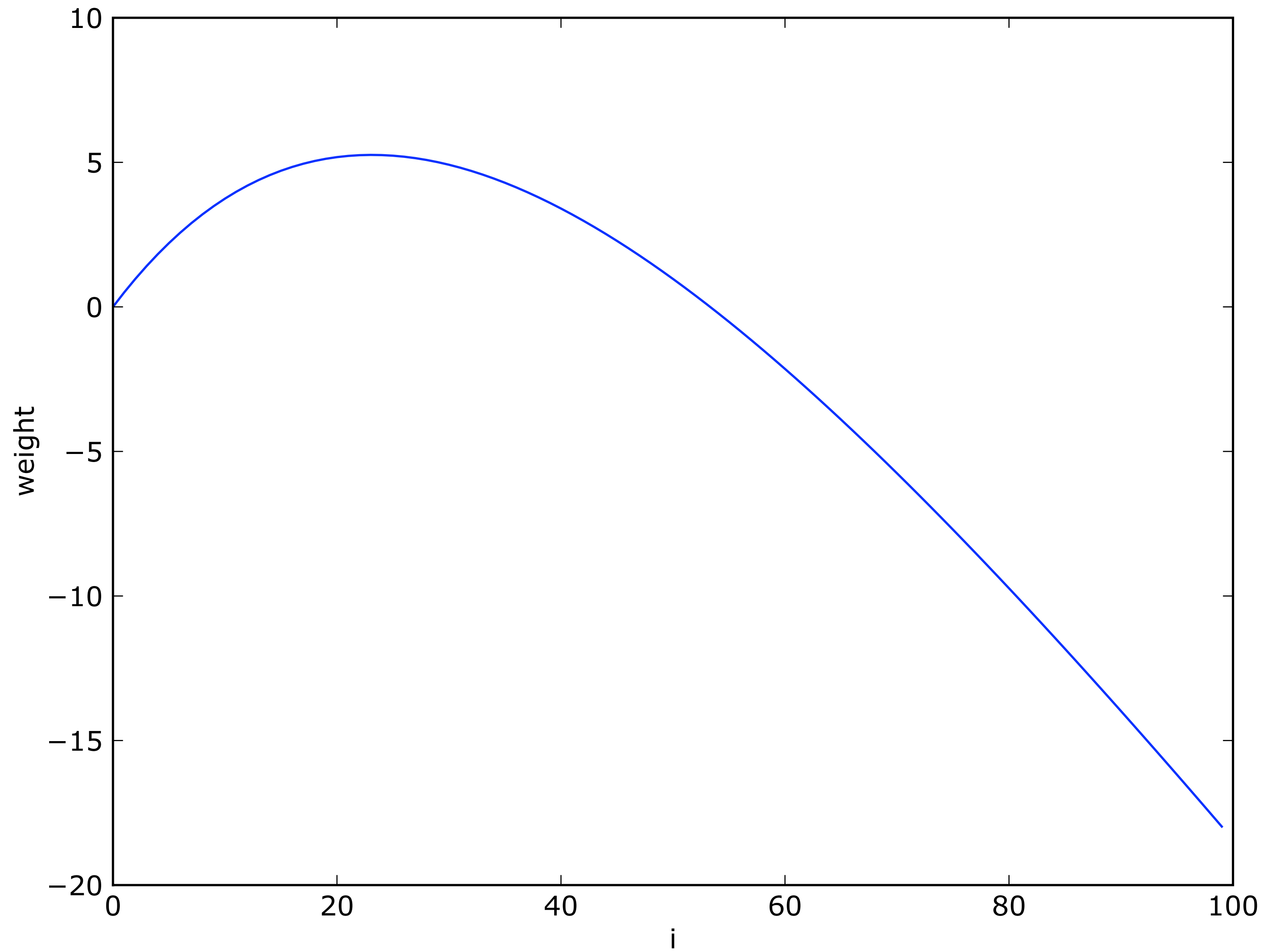
Geometric Weight Constraints

$$\theta_i = w\left(ci \cdot e^r(1 - (1 - e^{-r})^i)\right)$$



Geometric Weight Constraints

$$\theta_i = w\left(ci \cdot e^r(1-(1-e^{-r})^i)\right)$$



Curved Exponential Families

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z} e^{\boldsymbol{\eta}(\boldsymbol{\theta})^\mathsf{T} \boldsymbol{\phi}(\mathbf{y})}$$

$$\boldsymbol{\theta} \in \mathbb{R}^n$$

$$\boldsymbol{\eta} : \mathbb{R}^n \to \mathbb{R}^n$$

$$n < m$$

- Fewer parameters than features
- Non-linear mapping from low dimensional parameter space to high dimensional feature space
    - Linear mapping (e.g. tied parameters) are ordinary exponential family
- Parameters lie on curved p-dimensional manifold in q-dimensional space

Learning Curved Exponential Families

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{y}) = \boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{y}) - \log Z$$

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{y}) &= \nabla_{\boldsymbol{\theta}} \boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \nabla_{\boldsymbol{\eta}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{y}) \\
&= \nabla_{\boldsymbol{\theta}} \boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \left[\boldsymbol{\phi}(\mathbf{y}) - \mathop{\mathrm{E}}_{\mathbf{Y}} [\boldsymbol{\phi}(\mathbf{y})] \right]
\end{aligned}$$

$$[\nabla_{\boldsymbol{\theta}} \boldsymbol{\eta}(\boldsymbol{\theta})]_{ij} = \frac{\partial \eta_i}{\partial \theta_j}$$

- Use Jacobian to project high dimensional gradient onto manifold
- Not convex, in general

Curved Exponential Families and Graphical Models

- Bayes net with k binary variables and n CPT entries is a CEF [Geiger98]
  - $m = 2^k$
  - Generalizes to any discrete number of states
- Bayes nets with hidden variables are not, in general, CEFs [Geiger01]
- Has implications for model selection

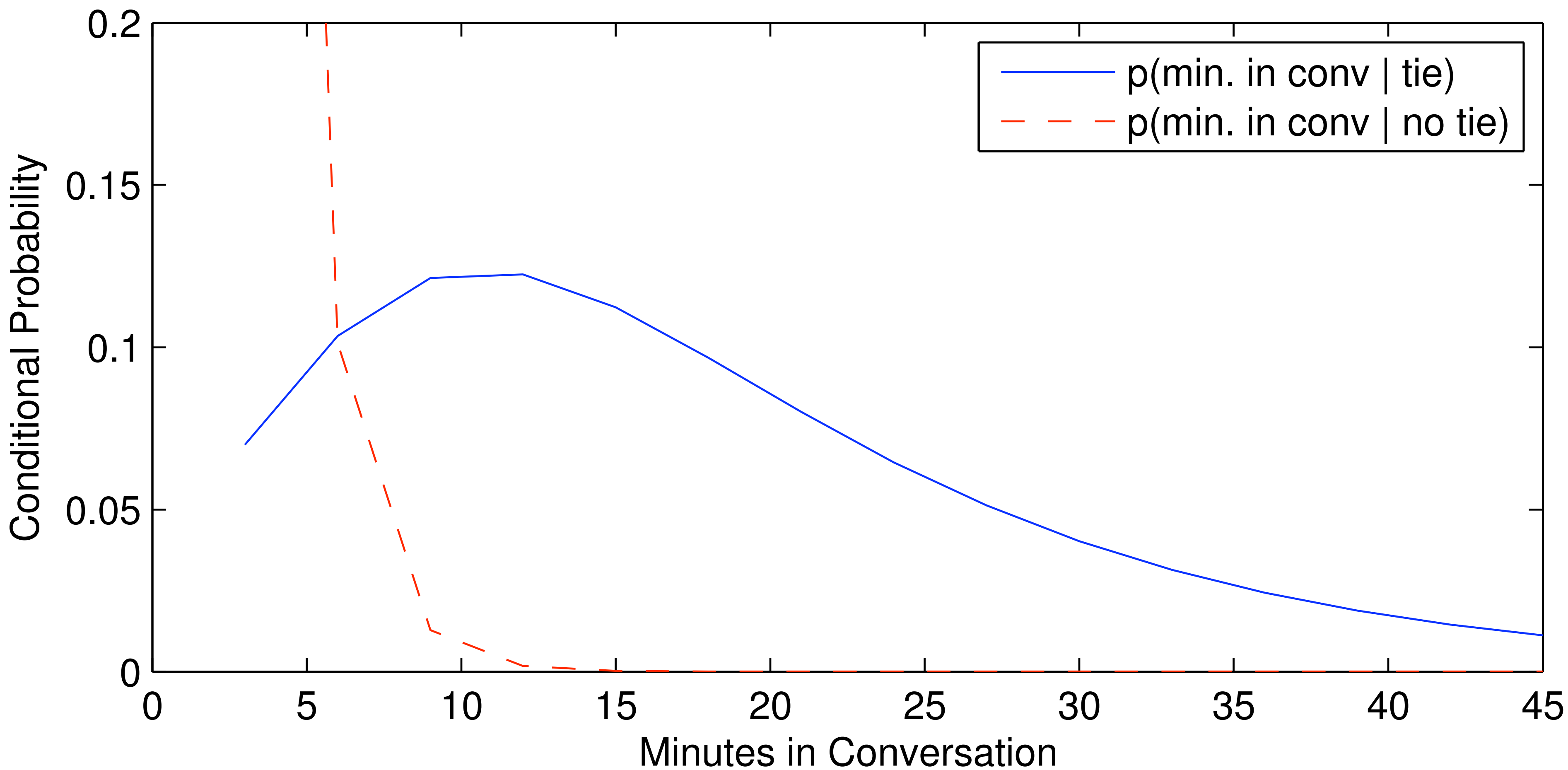
CERGMs for Latent Social Networks [Wyatt08]

$$p(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z} \sum_{\mathbf{y}} e^{\boldsymbol{\eta}(\boldsymbol{\theta})^\top \boldsymbol{\phi}(\mathbf{x},\mathbf{y})}$$

- Have discrete-but-ordinal observations of time spent in conversation: $\mathbf{x}$
- Social network $\mathbf{y}$ is now hidden
- Use curved model to express “diminishing returns” on time in conversation
- Simultaneously learn (unsupervised) parameters governing
  - Latent social structure
  - Relationship between time in conversation and latent social tie
    - Different parameters for “edge on” and “edge off” time in conversation curves

Interpretable Parameters

- Compute conditional probabilities of time in conversation given edge / no edge
- Look for
  - Threshold for “socially significant” time in conversation
  - Point of maximum “socially useful” time in conversation



References

[Frank86] Markov Graphs. Ove Frank and David Strauss. *J. American Statistical Association*, 81(395), 1986.

[Geiger98] Graphical Models and Exponential Families. Dan Geiger and Christopher Meek. *Proc. of UAI*. 1998.

[Geiger01] Stratified Exponential Families: Graphical Models and Model Selection. Dan Geiger, David Heckerman, Henry King and Christopher Meek. *The Annals of Statistics*, 29(2), 2001.

[Geyer92] Constrained Monte Carlo Maximum Likelihood for Dependent Data. Charles J. Geyer and Elizabeth Thompson. *J. Royal Statistical Society*, 54(3), 1992.

[Handcock03] Assessing degeneracy in statistical models of social networks. Mark Handcock. UW CSSS, TR No. 39. 2003.

[Hunter06] Inference in Curved Exponential Family Models for Networks. David R. Hunter and Mark Handcock. *J. Computational and Graphical Statistics*, 15(3), 2006.

[Hunter07] Curved Exponential Family Models for Social Networks. David Hunter. *Social Networks*, 29, 2007.

[Hunter08] Goodness of Fit of Social Network Models. David Hunter, Steven Goodreau and Mark Handcock. *J. American Statistical Association*, 103(481), 2008.

[Wyatt08] Learning Hidden Curved Exponential Random Graph Models to Infer Face-to-Face Interaction Networks from Situated Speech Data. Danny Wyatt, Tanzeem Choudhury, and Jeff Bilmes. *Proc. of AAAI*. 2008.

- `statnet` software: http://statnet.org/
    - `R` based
    - Developed here at UW