STA6800 - Statistical Analysis of Network

Introduction to ERGM

Ick Hoon Jin

Yonsei University, Department of Statistics and Data Science

1 Exponential Random Graph Models

2 Difficulty in Parameter Estimation

What is a Network?

- Definition: A representation of "relational data" in the form of a mathematical graph: A set of nodes along with a set of edges connecting some pairs of nodes.
- Adjacency Matrix

$$X_{ij} = \begin{cases} 1 & \text{node } i \text{ and } j \text{ are connected,} \\ 0 & \text{node } i \text{ and } j \text{ are not connected.} \end{cases}$$

Network Examples: Florentine Business Network

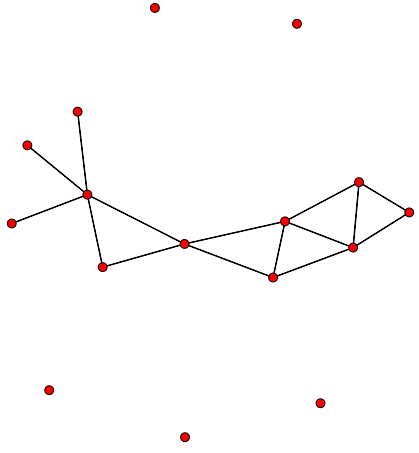


Figure: This network represents a set of business ties among Renaissance Florentine families. The network consists of 16 families who were locked in a struggle for political control of the city of Florence around 1430.

Network Examples: AddHealth Network

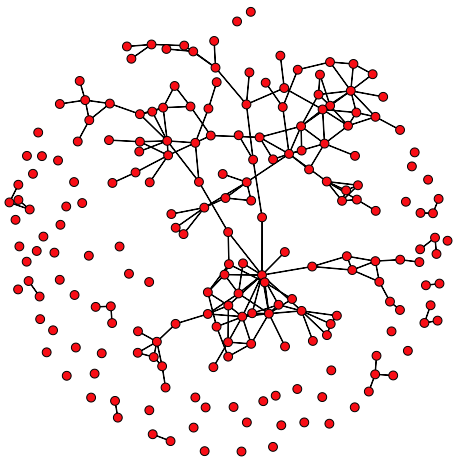


Figure: This network represents a set of friendship ties among high school students. It was collected during the first wave (1994-1995) of National Longitudinal Study of Adolescent Health (AddHealth).

Exponential Random Graph Model

Define the class of exponential random graph models (ERGMs) as

$$P_\theta(X = x) = \frac{\exp(\theta^t g(x))}{\kappa(\theta)} \tag{1}$$

where

- X: A random network written as an adjacency Matrix, X_{ij} is an indicator of an edge from node i to node j.
- $g(x)$: A vector of network statistics of interest.
- θ: The vector of parameters measuring the strengths of the effects of the corresponding entries in the vector $g(x)$.
- $\kappa(\theta)$: A normalizing constant

Exponential Random Graph Model

- Explain parsimoniously the local selection forces that shapes the global structure of a network.
- A network dataset may be considered like the response in a regression model, where the predictors are things such as “propensity for Indiv. to form triangles of partnerships.”
  $\Rightarrow$ ERGM help us quantify the strength of local transitivity.
- The information from the use of an ERGM can be used to understand a particular phenomenon or to simulate new random realizations to networks that retain the essential properties of the original.

Interpretation of ERGM

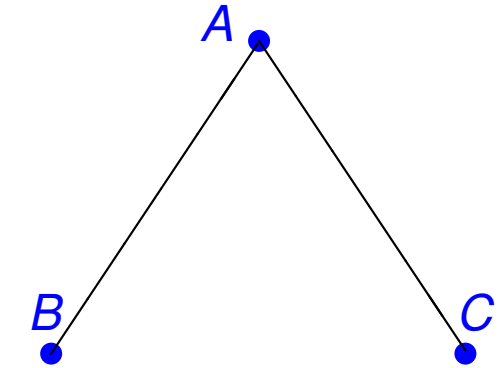
Exponential Random Graph Models (ERGMs):

$$P_\theta(X = x) = \frac{\exp(\theta^t g(x))}{\kappa(\theta)}$$

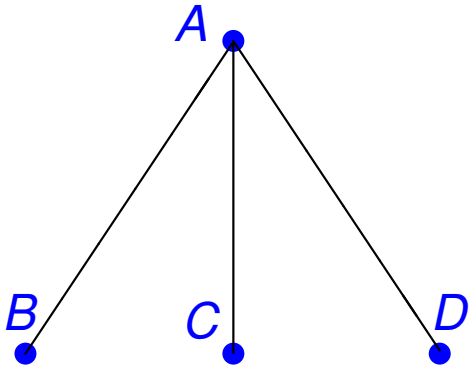
- $\theta > 0$: There exists a tendency to form $g(x)$ when changing X_{ij} value from 0 to 1.
- $\theta < 0$: There exists a tendency not to form $g(x)$ when changing X_{ij} value from 0 to 1.

Basic Markov Network Statistics

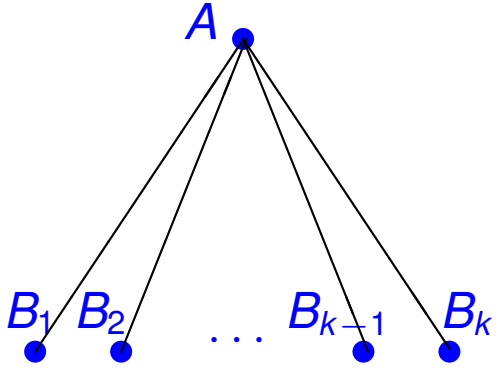
2-star ($S_2(x) = K_2$-star)



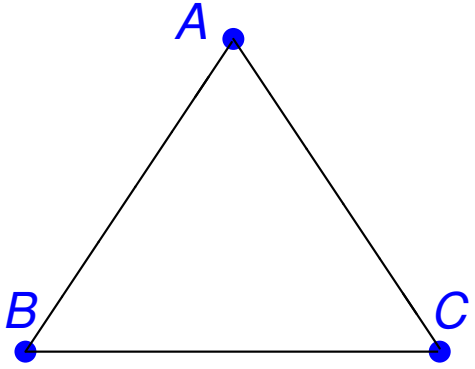
3-star ($S_3(x) = K_3$-star)



k-star ($S_k(x) = K_k$-star)



Triangle ($T(x)$)

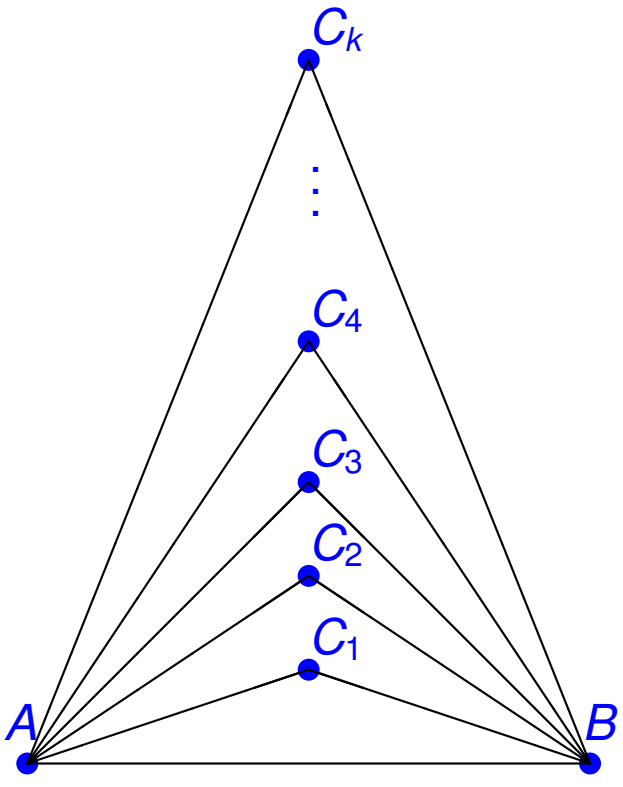


Degree and Shared Partnership Distribution

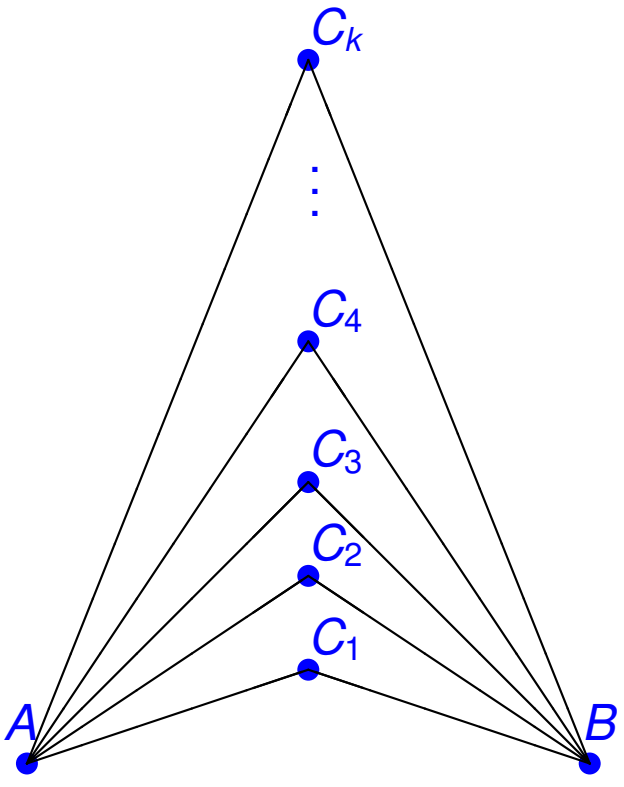
- Degree: The number of edges the node has to other nodes.
- $D_k(x)$: The number of nodes with degree k. $\sum D_k(x) = n$.
- Shared Partnership Distribution
  - $EP_k(x)$: The number of unordered pairs (i, j) for which i and j have exactly share k common neighbors and $X_{ij} = 1$.
  - $NP_k(x)$: The number of unordered pairs (i, j) for which i and j have exactly share k common neighbors and $X_{ij} = 0$
  - $DP_k(x)$: The number of unordered pairs (i, j) for which i and j have exactly share k common neighbors regardless of value X_{ij}.
  - $\sum EP_k(x) = S_1(x)$ (edge counts) and $\sum DP_k(x) = \binom{n}{2}$ (dyad counts).

Degree and Shared Partnership Distribution

$EP_k(x)$



$NP_k(x)$



Degree and Shared Partnership Distribution

The geometrically weighted statistics for degree and shared partnership distribution are defined by

$$u(x|\tau) = e^{\tau} \sum_{i=1}^{n-2} \left\{ 1 - \left(1 - e^{-\tau} \right)^{i} \right\} D_i(x),$$

$$v(x|\tau) = e^{\tau} \sum_{i=1}^{n-2} \left\{ 1 - \left(1 - e^{-\tau} \right)^{i} \right\} EP_i(x),$$

$$w(x|\tau) = e^{\tau} \sum_{i=1}^{n-2} \left\{ 1 - \left(1 - e^{-\tau} \right)^{i} \right\} DP_i(x),$$

where the additional parameter τ specifies the decreasing rate of the weights put on the higher order terms.

Intractable Normalizing Constants

- The normalizing constant of ERGMs is

$$\kappa(\theta) = \sum_{\text{all possible } x} \exp\left\{\theta^t g(\mathbf{x})\right\}.$$

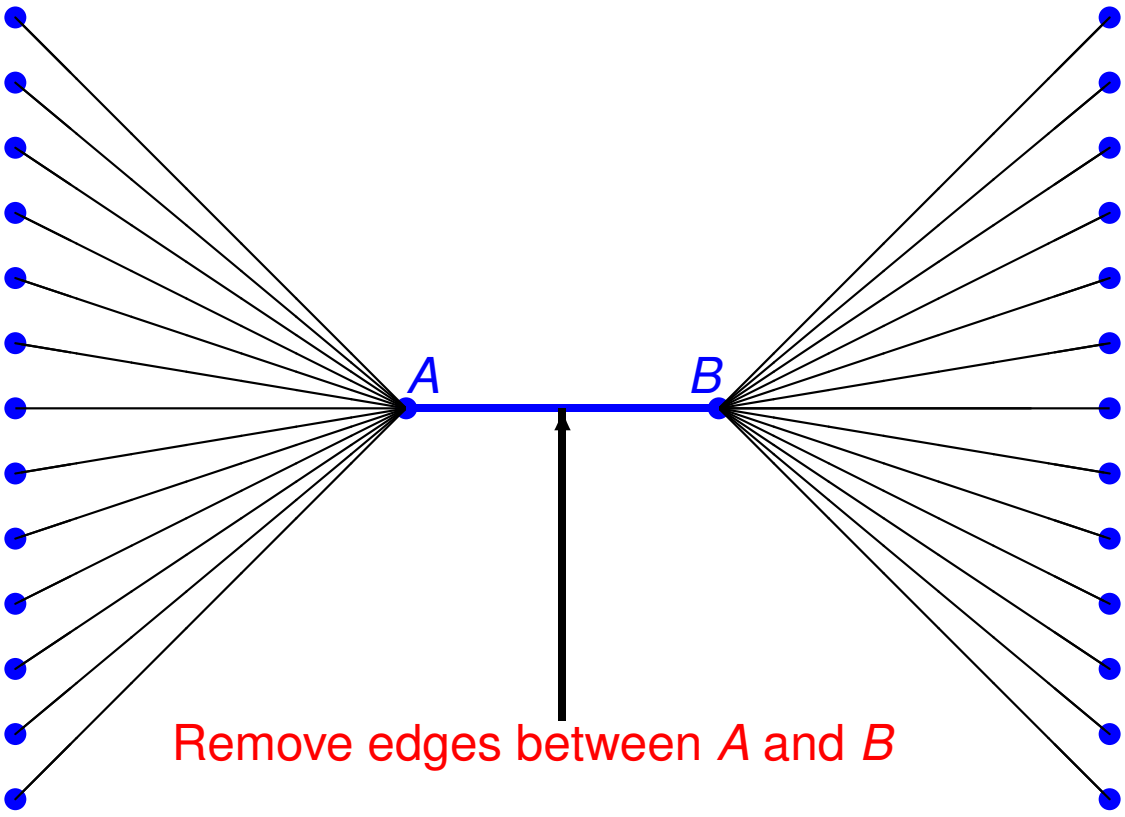
  Since there exist $2^{\binom{n}{2}}$ networks even in the undirected case, $\kappa(\theta)$ is not directly computable.

- Due to the intractable normalizing constant, MCMC is key to both simulation and statistical inference.
- However, for the general MH algorithm, the acceptance probability involve an unknown normalizing constant ratio $\kappa(\theta)/\kappa(\theta')$, where θ' denotes the proposed value, and it renders its failure.

Model Degeneracy

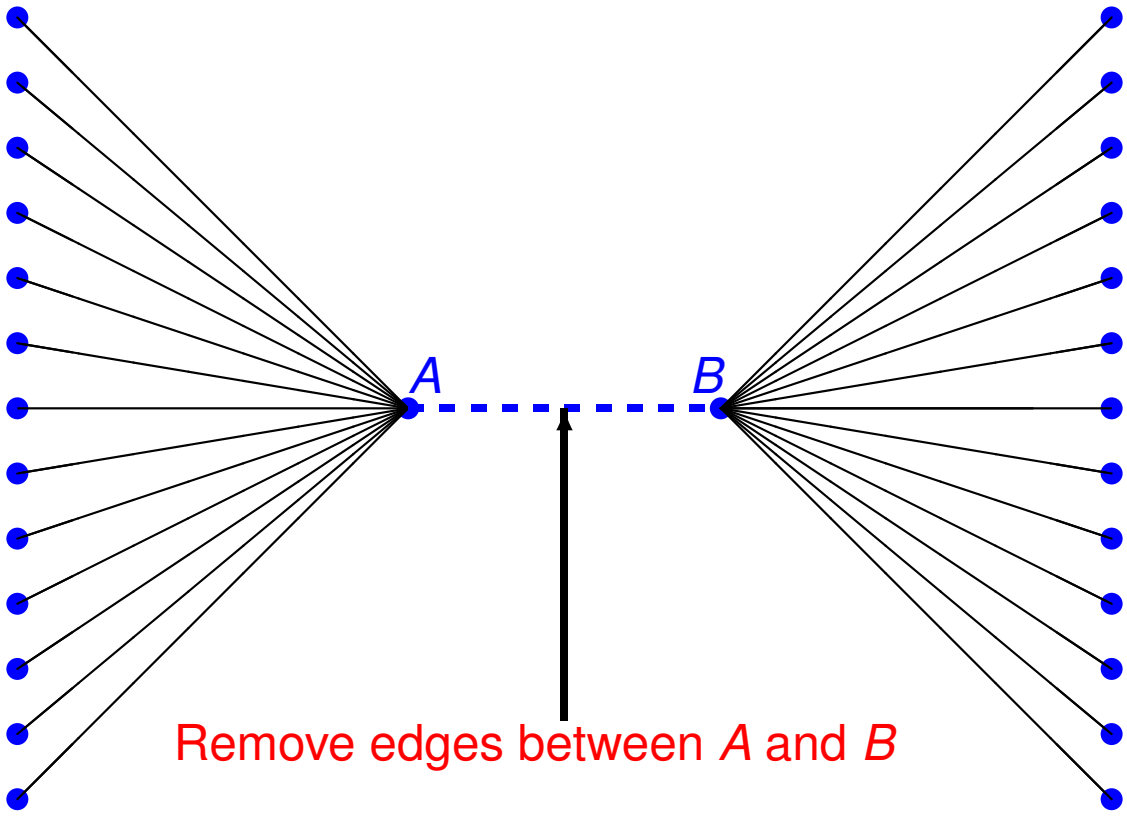
- For some configurations of θ, the ERGMs produces networks that are either full (every tie exists) or empty (no ties exist) with probability close to one.
- Example: Basic Markovian Statistics
  - When one edge is added to or removed from the network, the values of the basic Markovian statistics can change a lot while the values of other statistics do not change proportionally, so the dyadic dependence effects amplify quickly and the model tend to be degenerated.
- Current methods, MCMLE and stochastic approximation, sometimes produce degenerate estimates of θ if the starting value is in or close to a degeneracy region. $\Rightarrow$ Local convergence property.

Model Degeneracy for Basic Markov Statistics



$S_2(x) = 182$, $S_3(x) = 728$, $S_4(x) = 2002$.

Model Degeneracy for Basic Markov Statistics



$S_2(x) = 156$, $S_3(x) = 572$, $S_4(x) = 1430$.

Visualization of the Model Degeneracy Region

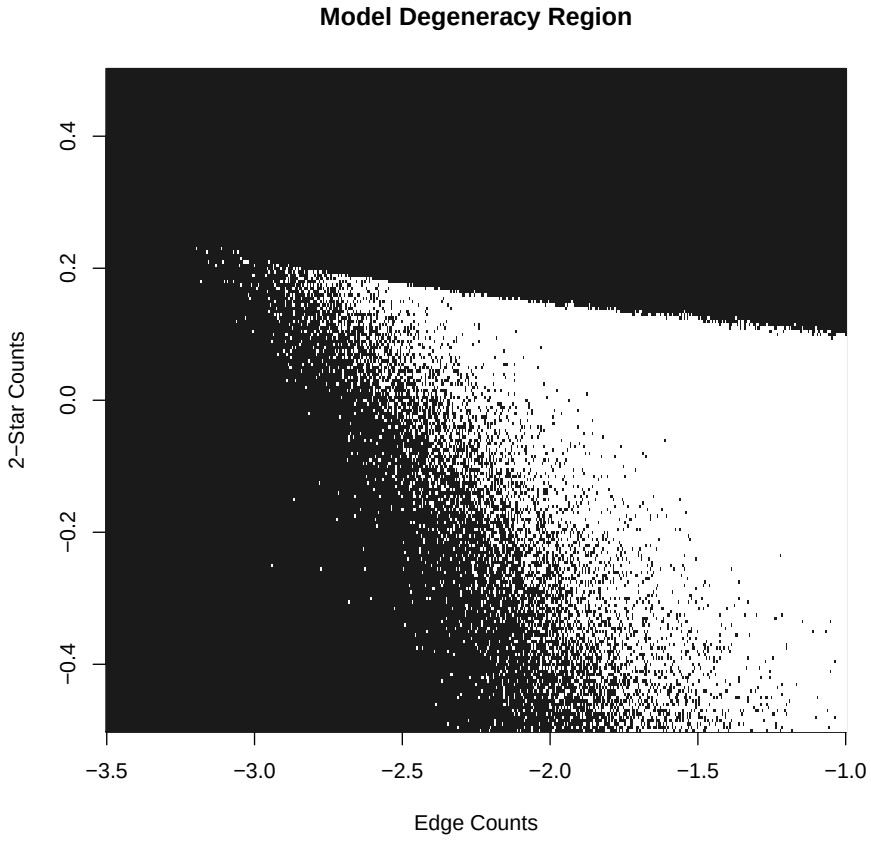


Figure: Visualization of the degeneracy (black) and non-degeneracy (white) region of an ERGM with edge counts and K_2-star.