Estimation Methods for Statistical Network Modeling

David Hunter

Department of Statistics
Penn State University

Supported by ONR MURI Award Number N00014-08-1-1015

Outline

Outline

What is a network model?

- For a network observed at a single instant: any probability distribution on the set of all possible networks (say, binary networks on a fixed set of n nodes).

What is a network model?

- For a network observed at a single instant: any probability distribution on the set of all possible networks (say, binary networks on a fixed set of n nodes).
- Thus, we assign each possible network a probability, e.g.,

$$P\left[Y = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}\right] = \frac{1}{64}, P\left[Y = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}\right] = \frac{1}{64},$$

and so on.

What is a network model?

- For a network observed at a single instant: any probability distribution on the set of all possible networks (say, binary networks on a fixed set of n nodes).
- Thus, we assign each possible network a probability, e.g.,

$$P\left[Y = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}\right] = \frac{1}{64}, P\left[Y = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}\right] = \frac{1}{64},$$

  and so on.
- Fortunately, there are better ways than explicit enumeration!

What is a network model?

- For a network observed at a single instant: any probability distribution on the set of all possible networks (say, binary networks on a fixed set of n nodes).
- Thus, we assign each possible network a probability, e.g.,

$$P\left[Y = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}\right] = \frac{1}{64}, P\left[Y = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}\right] = \frac{1}{64},$$

  and so on.
- Fortunately, there are better ways than explicit enumeration!
- This notion can be generalized to more general situations (time-varying networks, non-binary edges, etc.)

Exponential family, or p-star, models

Exponential-Family Random Graph Model (ERGM)

$$P_\theta(Y = y) \propto \exp\{\theta^\top g(y)\}$$

or

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)},$$

where

- Y is a random network on n nodes (e.g., a 0–1 matrix)
- θ is a vector of parameters
- $g(y)$ is a known vector of network statistics on y
- $\kappa(\theta)$ makes all the probabilities sum to 1

Ultimately, we care about what <u>data</u> (y) tell us about θ.

The Erdős-Rényi model for random networks

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}$$

- The normalizing “constant” $\kappa(\theta)$ can be troublesome, but not always.

The Erdős-Rényi model for random networks

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}$$

- The normalizing "constant" $\kappa(\theta)$ can be troublesome, but not always.

Example: The Erdős-Rényi model

Let p be some fixed constant between 0 and 1. Let $P(Y = y)$ be equal to $p^{E(y)}(1 - p)^{\overline{E}(y)}$, where $E(y)$ is the number of edges in y and $\overline{E}(y)$ is the number of non-edges in y.

The Erdős-Rényi model for random networks

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}$$

- The normalizing "constant" $\kappa(\theta)$ can be troublesome, but not always.

Example: The Erdős-Rényi model

Let p be some fixed constant between 0 and 1. Let $P(Y = y)$ be equal to $p^{E(y)}(1-p)^{\overline{E}(y)}$, where $E(y)$ is the number of edges in y and $\overline{E}(y)$ is the number of non-edges in y.

Rewriting, we get

$$P(Y = y) \quad = \quad \left(\frac{p}{1-p}\right)^{\text{\# of edges}} \times (1-p)^N$$

The Erdős-Rényi model for random networks

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}$$

- The normalizing "constant" $\kappa(\theta)$ can be troublesome, but not always.

Example: The Erdős-Rényi model

Let p be some fixed constant between 0 and 1. Let $P(Y = y)$ be equal to $p^{E(y)}(1-p)^{\overline{E}(y)}$, where $E(y)$ is the number of edges in y and $\overline{E}(y)$ is the number of non-edges in y.

Rewriting, we get

$$\begin{aligned} P(Y = y) &= \left(\frac{p}{1-p}\right)^{\text{\# of edges}} \times (1-p)^N \\ &= e^{\theta(\text{\# of edges})} \times \frac{1}{\kappa(\theta)} \end{aligned}$$

ERGMs as a starting point

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}$$

- We will use the basic ERGM framework as a jumping-off point for discussing much of our work on estimation, including topics such as:
  - Various methods for intractable normalizing constants $\kappa(\theta)$
  - Latent space models
  - Mixtures models of simple ERGMs
  - Relational event models

Outline

The goal of estimation

Exponential-family Random Graph Model (ERGM)

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}$$

- When θ is known, this is a probability model describing the random behavior of Y.

The goal of estimation

Exponential-family Random Graph Model (ERGM)

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}$$

- When θ is known, this is a probability model describing the random behavior of Y.
- Statistical estimation is “probability in reverse”: We don’t know θ but instead we observe $Y = y^{\text{obs}}$.

Goal:

Use observed data to select from the given ERGM class — i.e., to learn about θ.

The goal of estimation

Exponential-family Random Graph Model (ERGM)

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}$$

- When θ is known, this is a probability model describing the random behavior of Y.
- Statistical estimation is “probability in reverse”: We don’t know θ but instead we observe $Y = y^{\text{obs}}$.

Goal:

Use observed data to select from the given ERGM class — i.e., to learn about θ.

We might search for a “best” θ (MLE) or a density $p(\theta \mid \text{data})$.

The loglikelihood function

The model class:

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}, \text{ where } \kappa(\theta) = \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\theta^\top g(z)\}$$

- The likelihood is just $L(\theta) = P_\theta(Y = y^{\text{obs}})$, viewed as a function of θ.

The loglikelihood function

The model class:

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}, \text{ where } \kappa(\theta) = \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\theta^\top g(z)\}$$

- The likelihood is just $L(\theta) = P_\theta(Y = y^{\text{obs}})$, viewed as a function of θ.
- To choose a θ, we might try to search for a "best" theta by maximizing $L(\theta)$ or

$$\ell(\theta) = \log L(\theta) = \theta^\top g(y^{\text{obs}}) - \log \kappa(\theta)$$

The loglikelihood function

The model class:

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}, \text{ where } \kappa(\theta) = \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\theta^\top g(z)\}$$

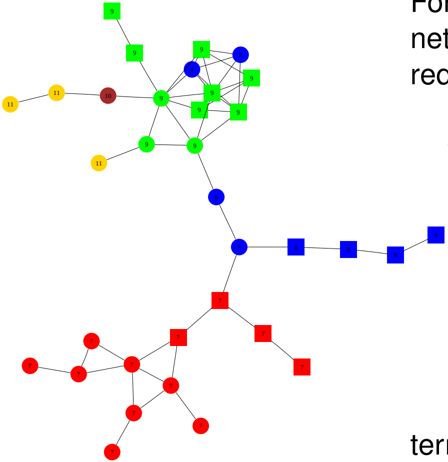
- The likelihood is just $L(\theta) = P_\theta(Y = y^{\mathrm{obs}})$, viewed as a function of θ.
- To choose a θ, we might try to search for a “best” theta by maximizing $L(\theta)$ or

$$\ell(\theta) = \log L(\theta) = \theta^\top g(y^{\mathrm{obs}}) - \log \kappa(\theta)$$

- Alternatively, a Bayesian approach tries to describe an entire distribution over θ values, the posterior:

$$p(\theta \mid Y = y^{\mathrm{obs}}) \propto L(\theta) \times \pi(\theta).$$

Computing the likelihood is sometimes very difficult



For this undirected, 34-node network, computing $\ell(\theta)$ directly requires summation of

7,547,924,849,643,082,704,483, 109,161,976,537,781,833,842, 440,832,880,856,752,412,600, 491,248,324,784,297,704,172, 253,450,355,317,535,082,936, 750,061,527,689,799,541,169, 259,849,585,265,122,868,502, 865,392,087,298,790,653,952

terms.

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $y_{ij} = 0$ or 1, depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i, j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $y_{ij} = 0$ or 1, depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i, j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

Conditional on $Y_{ij}^c = y_{ij}^c$, Y has only two possible states, depending on whether $Y_{ij} = 0$ or $Y_{ij} = 1$.

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $y_{ij} = 0$ or 1, depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i, j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

Conditional on $Y_{ij}^c = y_{ij}^c$, Y has only two possible states, depending on whether $Y_{ij} = 0$ or $Y_{ij} = 1$.
Let's calculate the ratio of the two respective probabilities.

[We'll use $P_\theta(Y = y) = \exp\{\theta^\top g(y)\}/\kappa(\theta)$.]

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $y_{ij} = 0$ or 1, depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i, j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

$$\frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \frac{\exp\{\theta^\top g(y_{ij}^+)\}}{\exp\{\theta^\top g(y_{ij}^-)\}}$$

A lot of cancellation happened on the right hand side!

Conditional log-odds of an edge

Notation: For a network y and a pair (i,j) of nodes,

- $y_{ij} = 0$ or 1, depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i,j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

$$\frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \exp\{\theta^\top [g(y_{ij}^+) - g(y_{ij}^-)]\}$$

A lot of cancellation happened on the right hand side!

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $y_{ij} = 0$ or 1, depending on whether there is an edge
- y_{ij}^c denotes the status of all pairs in y other than (i, j)
- y_{ij}^+ denotes the same network as y but with $y_{ij} = 1$
- y_{ij}^- denotes the same network as y but with $y_{ij} = 0$

$$\log \frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \theta^\top [g(y_{ij}^+) - g(y_{ij}^-)]$$

Conditional log-odds of an edge

Notation: For a network y and a pair (i, j) of nodes,

- $\delta(y)_{ij}$ denotes the vector of change statistics,

$$\delta(y)_{ij} = g(y_{ij}^{+}) - g(y_{ij}^{-}).$$

So $\delta(y)_{ij}$ is the conditional log-odds of edge (i, j).

$$\log \frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \theta^\top \delta(y)_{ij}$$

This simple formula can serve as the basis for a Markov chain Monte Carlo (MCMC) scheme for simulating random networks.

Maximum Pseudolikelihood: Intuition

- What if we assume that there is no dependence (or very weak dependence) among the Y_{ij}?

Maximum Pseudolikelihood: Intuition

- What if we assume that there is no dependence (or very weak dependence) among the Y_{ij}?
- In other words, what if we approximate the marginal $P(Y_{ij} = 1)$ by the conditional $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$?

Maximum Pseudolikelihood: Intuition

- What if we assume that there is no dependence (or very weak dependence) among the Y_{ij}?
- In other words, what if we approximate the marginal $P(Y_{ij} = 1)$ by the conditional $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$?
- Then the Y_{ij} are independent with

$$\log \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \theta^\top \delta(y^{\text{obs}})_{ij},$$

so we obtain an estimate of θ using straightforward logistic regression.

Maximum Pseudolikelihood: Intuition

- What if we assume that there is no dependence (or very weak dependence) among the Y_{ij}?
- In other words, what if we approximate the marginal $P(Y_{ij} = 1)$ by the conditional $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$?
- Then the Y_{ij} are independent with

$$\log \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \theta^\top \delta(y^{\mathrm{obs}})_{ij},$$

  so we obtain an estimate of θ using straightforward logistic regression.
- Result: The maximum pseudolikelihood estimate.

Maximum Pseudolikelihood: Intuition

- What if we assume that there is no dependence (or very weak dependence) among the Y_{ij}?
- In other words, what if we approximate the marginal $P(Y_{ij} = 1)$ by the conditional $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$?
- Then the Y_{ij} are independent with

$$\log \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \theta^\top \delta(y^{\text{obs}})_{ij},$$

  so we obtain an estimate of θ using straightforward logistic regression.
- Result: The maximum pseudolikelihood estimate.
- For independence models, MPLE = MLE!

MLE vs. MPLE

> *Far better an approximate answer to the 'right' question, which is often vague, than an 'exact' answer to the wrong question, which can always be made precise.*
>
> — *John W. Tukey*

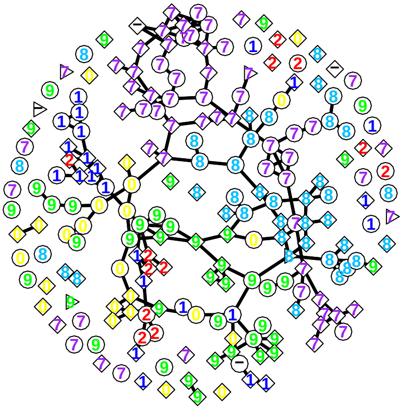
- MLE (maximum likelihood estimation): Well-established method but very hard because the normalizing constant $\kappa(\alpha)$ is difficult to evaluate, so we approximate it instead.
- MPLE (maximum pseudo-likelihood estimation): Easy to do using logistic regression, but based on an independence assumption that is often not justified.

Several authors, notably van Duijn et al. (2009), argue forcefully against the use of MPLE (except when MLE=MPLE!).

Outline

Example Network: High School Friendship Data

School 10: 205 Students



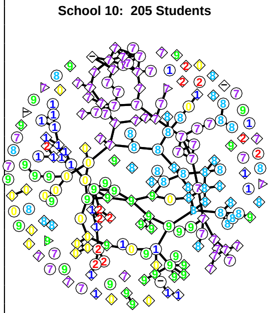
- An edge indicates a mutual friendship.
- Colored labels give grade level, 7 through 12.
- Circles = female, squares = male, triangles = unknown.
- N.B.: Missing data ignored here, though this could be altered.

Fitting an ERGM to the high school dataset

- ERGM parameter estimates from Hunter et al (2008):

Coefficient		Coefficient	
edges	−3.49(1.92)	AD (Gr.) = 1	3.41(1.42)*
GWESP	0.83(0.13)***	AD (Gr.) = 2	2.42(1.48)
GWD	−2.01(0.35)***	AD (Gr.) = 3	1.43(1.62)
GWDSP	0.50(0.09)***		
		DH (Gr. 7)	6.00(1.56)***
NF (Gr. 8)	−0.34(0.78)	DH (Gr. 8)	6.48(1.64)***
NF (Gr. 9)	0.64(0.59)	DH (Gr. 9)	4.52(1.58)**
NF (Gr. 10)	0.55(0.59)	DH (Gr. 10)	4.96(1.59)**
NF (Gr. 11)	0.97(0.60)	DH (Gr. 11)	4.32(1.54)**
NF (Gr. 12)	1.23(0.60)*	DH (Gr. 12)	4.11(1.58)**
NF (Gr. NA)	3.86(1.30)**		
		DH (White)	1.55(0.68)*
NF (Black)	0.51(0.42)	DH (Black)	0.92(1.55)
NF (Hisp)	−0.23(0.33)	DH (Hisp)	0.87(0.43)*
NF (Nat Am)	−0.21(0.32)	DH (Nat Am)	1.31(0.43)**
NF (Other)	−0.61(0.69)		
NF (Race NA)	1.53(0.89)		
NF (Female)	0.09(0.10)	UH (Sex)	0.67(0.16)***
NF (Sex NA)	−0.18(0.47)		
NF stands for Node Factor.		AD stands for Absolute Difference. DH stands for Differential Homophily. UH stands for Uniform Homophily.	

* Significant at 0.05 level ** Significant at 0.01 level *** Significant at 0.001 level



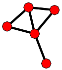
Goodness of fit intuition

ERGM
class
$\exp\{\theta^\top g(y)\}$

Goodness of fit intuition

ERGM
class
$\exp\{\theta^\top g(y)\}$ $\longrightarrow$

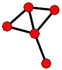
$\uparrow$
y^{obs}



Goodness of fit intuition

ERGM class		(approx) MLE
$\exp\{\theta^\top g(y)\}$	$\longrightarrow$	$\widehat{\theta}$

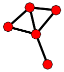
$$\uparrow$$
$$y^{\text{obs}}$$



Goodness of fit intuition

ERGM class		(approx) MLE		Fitted ERGM
$\exp\{\theta^\top g(y)\}$	$\longrightarrow$	$\widehat{\theta}$	$\longrightarrow$	$\exp\{\widehat{\theta^\top} g(y)\}$

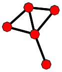
$$\uparrow$$
$$y^{\mathrm{obs}}$$



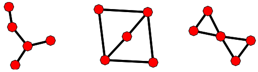
Goodness of fit intuition

ERGM class $\exp\{\theta^\top g(y)\}$ $\longrightarrow$ (approx) MLE $\hat{\theta}$ $\longrightarrow$ Fitted ERGM $\exp\{\widehat{\theta^\top} g(y)\}$

$\uparrow$
y^{obs}



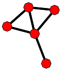
$\downarrow$
Randomly generated networks $\tilde{Y}_1, \tilde{Y}_2, \ldots$



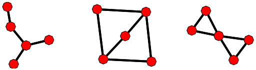
Goodness of fit intuition

ERGM class $\exp\{\theta^\top g(y)\}$ $\longrightarrow$ (approx) MLE $\widehat{\theta}$ $\longrightarrow$ Fitted ERGM $\exp\{\widehat{\theta^\top} g(y)\}$

$\uparrow$ y^{obs}



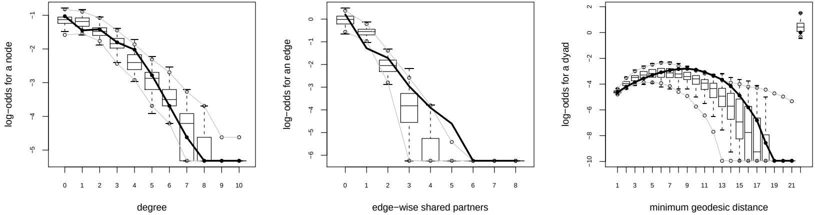
$\downarrow$ Randomly generated networks $\tilde{Y}_1, \tilde{Y}_2, \ldots$



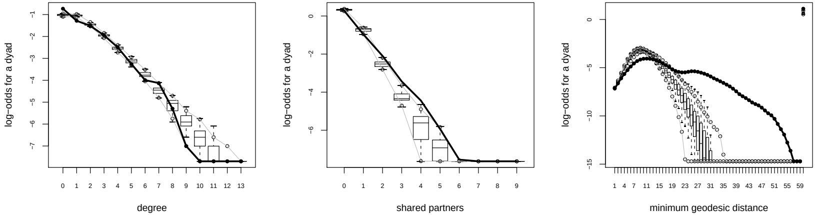
- Question: How does y^{obs} "look" as a representative of the sample $\tilde{Y}_1, \tilde{Y}_2, \ldots$?

Graphical GOF check (from Hunter et al, 2008)

n=205 (dataset shown)



n=2209 (different dataset but same model)



Outline

Intractable normalizing constants

Exponential-family Random Graph Model (ERGM)

$$P_\theta(Y = y) = \frac{\exp\{\theta^\top g(y)\}}{\kappa(\theta)}$$

- In ERGMs for which $\kappa(\theta)$ is intractable, we are working on improved MCMC-based maximum likelihood schemes
- In addition, by considering MPLE and MLE to be at either end of a spectrum of algorithms, it may be possible to balance the computational benefits of MPLE with the accuracy and precision of MLE.

Latent space models

Latent space network model (Handock et al, 2007)

$$P_\theta(Y = y \mid z) \propto \prod_{i \neq j} \left[\theta_0^\top x_{ij} + \theta_1 \| z_i - z_j \| \right],$$

where z_i and z_j are (unobserved) positions in latent space of the ith and jth nodes.

- Conditional on the z's, the normalizing constant is simple; but there are many z parameters!
- Additional structure may be assumed on the z_i as in Handcock et al (2007).
- Implementation of an estimation algorithm (θ and the z_i) may be dramatically aided through improved algorithms and data structures.

Mixtures of ERGMs

Stochastic blockstructure model (Nowicki and Snijders, 2001)

$$P_\theta(Y = y \mid z, \lambda) = \prod_{i \neq j} \left[\theta_{z_i, z_j}\right]^{y_{ij}},$$

where z_i and z_j are (unobserved) latent categories and the z_i are independent with $P(z_i = k) = \lambda_k$.

- Like the latent position model, the normalizing constant is simple conditional on the z's.
- However, the full (joint) likelihood is too complicated for direct methods.
- MCMC methods (e.g., Nowicki and Snijders, 2001) are possible but do not scale well to large networks.
- An alternative (Daudin et al, 2008) is a variational method.

Relational Event Models

In the relational events model, “events” happen at particular moments in time according to:

- a particular hazard function $\lambda(t)$, which may involve various
- parameters and
- statistics defined on a network determined by the cumulative sequence of events.

Depending on the choice of parameterization, estimation may or may not be challenging numerically; but typically one may avoid the difficult normalizing constant issue.

Cited References

- Butts CT (2008, *Soc. Meth.*), A Relational Event Framework for Social Action.
- Daudin JJ, Picard F, Robin S (2008, *Stat. & Comp.*), A Mixture Model for Random Graphs.
- Handcock MS, Raftery AE, Tantrum JM (2007, *J Roy Stat Soc A*), Model-Based Clustering for Social Networks.
- Hunter DR, Goodreau SM, and Handcock MS (2008, *J. Am. Stat. Assoc.*) Goodness of fit for social network models.
- Nowicki K and Snijders TAB (2001, *J. Am. Stat. Assoc*) Estimation and Prediction for Stochastic Blockstructures.
- van Duijn MAJ, Gile K, and Handcock MS (2009, *Social Networks*), A Framework for the Comparison of Maximum Pseudo-Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models.