ST5225: Statistical Analysis of Networks
Lecture 9: Exponential Random Graph Models

WANG Wanjie
staww@nus.edu.sg

Department of Statistics and Applied Probability
National University of Singapore (NUS)

Saturday 31 March, 2018

Outline

- Review: World Wide Web, Part I

- Exponential Random Graph Models

Review

- Advertisement
  - How to set the price for advertisements in search engines
  - Formulation of the problem: clickthrough rate, revenue per click, valuation, matching market
- Review of Statistical notions: model, PDF, likelihood function, MLE.
- Random Graph
  - $|V|$ is given, $(i,j) \overset{i.i.d.}{\sim} Bernoulli(p)$
  - Likelihood, MLE, and an example
  - More properties: degree dist., prob. of edge, parameterization
  - drawbacks of the model: few triangles, no clustering structure, degree dist.
- Stochastic block model
  - $|V|$ is given
  - Each node has a label ℓ_i, indicating which community it belongs to. The prob. of an edge depends on ℓ_i and ℓ_j
  - Likelihood, MLE, and an example
  - More properties: degree dist., prob. of an edge
  - If the labels are unknown, we model the labels as multinomial dist., and have new likelihood function
  - MLE does not have explicit solution in this case

Overview

- Generalizations of SBM (revisit in the future)
  - Degree Corrected SBM
  - Mixed membership SBM
- Exponential Random Graph Model
  - Motivation
  - Sufficient Statistics
  - Exponential family distributions
  - Model
  - Edge prob., MLE
  - Example: $p1$ model

Exponential Random Graph Model

Recall the likelihood for the RGM and the SBM for graph $G = (V, E)$

- Random Graph Model:

$$L(p) = p^{|E|}(1-p)^{\binom{|V|}{2}-|E|} = \exp\left\{|E|\log p + (\binom{|V|}{2} - |E|)\log(1-p)\right\}$$

$$= \exp\left\{|E|\log\frac{p}{1-p} + \binom{|V|}{2}\log(1-p)\right\}$$

- SBM

$$L(B) = \prod_{r\neq s} b_{rs}^{e_{rs}}(1-b_{rs})^{n_r n_s - e_{rs}} \times \prod_{r} b_{rr}^{e_{rr}}(1-b_{rr})^{\binom{n_r}{2}-e_{rr}}$$

$$= \exp\left\{\sum_{r,s} e_{rs}\log\frac{b_{rs}}{1-b_{rs}} + \sum_{r\neq s} n_r n_s \log(1-b_{rs}) + \sum_{r}\binom{n_r}{2}\log(1-b_{rs})\right\}$$

Both can be written in the form of $\exp\{\sum_{i=1}^{L} f_i(\theta)S_i(data)\}$, where $S_i(data)$ is some statistic of the data.

How to make the model more flexible?

- Generalise the model in a similar form

$$\exp\{\sum \text{parameter} \times \text{statistic}\}$$

- Why do we select this form?
- What does it mean?
- How to select the parameters and the statistics?

With data points $(x_1, x_2, x_3, \cdots, x_n)$, we may calculate many many statistics:

Statistic

Statistic is a function of the random variables.

Examples:

- Mean of the data: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Variance of the data $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Minimum/Maximum of the data: $\min_i x_i$, $\max_i x_i$
- The first observation the data: x_1
- Statistics of interest depend on the case

One common problem for the data is to get estimate of the parameters

- Say that the sample $X_i \overset{i.i.d}{\sim} Unif(0,b)$, $1 \leq i \leq n$. What is the MLE for b?

  Solution. Note that for the uniform dist., the PDF is

  $$f(x) = \frac{1}{b} I_{0 \leq x \leq b}.$$

  Therefore, the likelihood function is

  $$L(b) = \prod_{i=1}^{n} [\frac{1}{b} I_{0 \leq x \leq b}] = \frac{1}{b^n} I_{0 \leq \min x_i \leq \max x_i \leq b}.$$

  To figure out the likelihood function, we only need $\max x_i$.

  - Therefore, knowing $\max x_i$ is *sufficient* to figure out the MLE.
  - Further, if we have an estimate $\hat{b}$, then knowing $\max x_i$ is *sufficient* to figure out the density of the data points.

Sufficient statistic, II

Now we consider the normal distribution.

- Say that the sample $X_i \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, $1 \leq i \leq n$. What is the likelihood function for (μ, σ^2)?

  Solution. Note that for the normal dist., the PDF is

  $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}.$$

  Therefore, the likelihood function is

  $$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} [\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}] \\ &= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-\frac{\sum_{i=1}^{n}(x-\mu)^2}{2\sigma^2}\} \\ &= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-\frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2} + \frac{2\mu\sum_{i=1}^{n} x_i}{2\sigma^2} - \frac{n\mu^2}{2\sigma^2}\} \end{aligned}$$

  To figure out the likelihood function, we only need $\sum x_i^2$ and $\sum x_i$

  - Again, knowing $\sum_{i=1}^{n} x_i^2$ and $\sum_{i=1}^{n} x_i$ is *sufficient* to figure out the MLE and calculate the likelihood function
  - We do not need the details of the data

Sufficient statistic, III

In the following two slides, X denotes the data vector $(X_1, X_2, \cdots, X_n)$

Sufficient Statistic

With respect to a model P_θ, a statistic $T(X)$ is *sufficient* for underlying parameter θ if the conditional probability distribution of the data X, given the statistic $T(X)$, does not depend on the parameter θ, i.e.

$$P(X|T(X), \theta) = P(X|T(X)).$$

- The relationship between X and the parameter θ are totally expressed by the relationship between X and $T(X)$
- Instead of storing all the data, we may store $T(X)$ only
- Note: $T(X)$ is a function of X only, which does not include any parameter
- The model parameters are decided by $T(X)$. It can be viewed that *the model targets on* $T(X)$

Sufficient statistics, IV

Factorization Theorem

T is sufficient for θ if and only if nonnegative functions g and h can be found such that:

$$P_\theta(x) = h(x)g(\theta, T(x)).$$

- In other words, the data only interacts with parameter θ via $T(X)$.
- Proof. (sufficiency)

$$\begin{aligned}
P(X|T(X),\theta) &= P_\theta(X|T(X)) = \frac{P_\theta(X, T(X))}{P_\theta(T(X))} \\
&= \frac{h(X)g(\theta, T(X))}{\sum_{x:T(x)=T} h(x)g(\theta, T(x))} \\
&= \frac{h(X)g(\theta, T(X))}{g(\theta, T(x))\sum_{x:T(x)=T} h(x)} \\
&= \frac{h(X)}{\sum_{x:T(x)=T} h(x)} = P(X|T(X))
\end{aligned}$$

Sufficient statistics, Examples

- Uniform dist. $X_i \overset{i.i.d}{\sim} Unif(0,\theta)$

$$f_\theta(x_1,\cdots,x_n) = \prod_{i=1}^{n}[\frac{1}{\theta} I_{0\le x\le\theta}] = \frac{1}{\theta^n} I_{0\le \min x_i \le \max x_i \le b}.$$

So the sufficient stat is $\max x_i$.

The uniform dist. model is interested in the range of the data.

- Suppose that $X_i \overset{i.i.d}{\sim} f_\alpha$, where $f_\alpha(x) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}[x(1-x)]^{\alpha-1}$, $\alpha > 0$. Then,

$$f_\alpha(x_1,x_2,\cdots,x_n) = \prod_{i=1}^{n}\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}[x_i(1-x_i)]^{\alpha-1} = \Big(\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}\Big)^n[\prod_{i=1}^{n}x_i(1-x_i)]^{\alpha-1},$$

where the sufficient statistic is $T = \prod_{i=1}^{n} X_i(1-X_i)$.

- *MLE is always sufficient stat.*

Exponential Family Distributions

According to the factorization theorem, if the parametric model has a distribution with the form

$$P_\theta(x) = h(x)g(\theta_1, \theta_2, \cdots, \theta_d)\exp\{\sum_{i=1}^{d} T_i(x)\theta_i\},$$

then the sufficient statistics are

$$T_1(x), T_2(x), \cdots, T_d(x).$$

Exponential Family Distribution

We call $f_\theta(x)$ as an *exponential family distribution*, if it satisfies

$$f_\theta(x) = h(x)g(\theta)\exp\left\{\sum_{i=1}^{d} \theta_i T_i(x)\right\}$$

where $T_i(x)$, $h(x)$, and $g(\theta)$ are known functions.

Exponential Family Distributions, II

Exponential Family Distribution

We call $f_\theta(x)$ as an *exponential family distribution*, if it satisfies

$$f_\theta(x) = h(x)g(\theta)\exp\left\{\sum_{i=1}^{d} \theta_i T_i(x)\right\}$$

where $T_i(x)$, $h(x)$, and $g(\theta)$ are known functions.

Remarks:

- The data X and the parameter interacts *through* $T_i(x)$ *only*. $h(x)$ is a function about the data only, and $g(\theta)$ is a function about the parameter only.
- $g(\theta)$ is to normalize the density function, so that the integration is 1.
- The part θ_i can be generalized to be $\eta(\theta) = (\eta_1(\theta), \eta_2(\theta), \cdots, \eta_d(\theta))$, where $\eta(\theta)$ is a one-to-one mapping.

Exponential Family Distributions, Examples

- Normal dist.

$$f_{\mu,\sigma}(x_1,\cdots,x_n) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-\frac{n\mu^2}{2\sigma^2}\} \exp\{-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2} + \frac{2\mu\sum_{i=1}^n x_i}{2\sigma^2}\}.$$

Define the sufficient stat as $T_1(X) = \sum_{i=1}^n x_i^2$, $T_2(X) = \sum_{i=1}^n x_i$, and define $\theta_1 = -\frac{1}{2\sigma^2}$, $\theta_2 = \frac{2\mu}{2\sigma^2}$. Then the density function can be rewritten as

$$f_{\mu,\sigma}(x_1,\cdots,x_n) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{-\frac{n\mu^2}{2\sigma^2}} \exp\{\theta_1 T_1(X) + \theta_2 T_2(X)\},$$

which belongs to the exponential family

- Bernoulli dist.

$$\begin{aligned} f_p(x_1,\cdots,x_n) &= p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)} \\ &= \exp\{\sum_{i=1}^n x_i \log p + \sum_{i=1}^n (1-x_i)\log(1-p)\} \\ &= \exp\{\sum_{i=1}^n x_i \log\frac{p}{1-p} + n\log(1-p)\} \end{aligned}$$

Define $\theta = \log\frac{p}{1-p}$, and $T(X) = \sum x_i$. The Bernoulli dist. also belongs to the exponential family.

Exponential Family Distributions, Examples

- Normal dist.

$$f_{\mu,\sigma}(x_1, \cdots, x_n) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-\frac{n\mu^2}{2\sigma^2}\} \exp\{-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2} + \frac{2\mu \sum_{i=1}^n x_i}{2\sigma^2}\}.$$

Define the sufficient stat as $T_1(X) = \sum_{i=1}^n x_i^2$, $T_2(X) = \sum_{i=1}^n x_i$, and define $\theta_1 = -\frac{1}{2\sigma^2}$, $\theta_2 = \frac{2\mu}{2\sigma^2}$. Then the density function can be rewritten as

$$f_{\mu,\sigma}(x_1, \cdots, x_n) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{-\frac{n\mu^2}{2\sigma^2}} \exp\{\theta_1 T_1(X) + \theta_2 T_2(X)\},$$

which belongs to the exponential family

- Bernoulli dist.

$$\begin{aligned} f_p(x_1, \cdots, x_n) &= p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)} \\ &= \exp\{\sum_{i=1}^n x_i \log p + \sum_{i=1}^n (1-x_i)\log(1-p)\} \\ &= \exp\{\sum_{i=1}^n x_i \log\frac{p}{1-p} + n\log(1-p)\} \end{aligned}$$

Define $\theta = \log\frac{p}{1-p}$, and $T(X) = \sum x_i$. The Bernoulli dist. also belongs to the exponential family.

Exponential Family Distributions, Property

- Many distributions belong to the exponential family, say, normal, exponential, gamma, chi-squared, beta, Bernoulli, Poisson, Wishart, geometric, etc.
- There are some exceptions, such as uniform dist.
- If a distribution belongs to the exponential family, it is easy to figure out the sufficient statistics $(T_i(X))$
- On the other hand, if we can find a finite set of sufficient statistics, then very possibly it belongs to the exponential family.

Fisher-Pitman-Koopman-Darmois Theorem

Let $T = (T_1, T_2, ..., T_d)$ be a finite set of sufficient statistics for a model $p_\theta(x)$ with support that does not depend on θ. Then, $p_\theta(x)$ must either be an exponential family distribution, or a uniform distribution.

- We may also define a model to be with the form $\exp\{\sum_{i=1}^{d} \theta_i T_i(X)\}$, so that the statistic of interest, $T_i(X)$, would be considered in the model

Random Graph Model

Recall the Random graph model with parameter p,

$$L(p) = p^{|E|}(1-p)^{\binom{|V|}{2}-|E|} = \exp\left\{|E| \log \frac{p}{1-p} + \binom{|V|}{2} \log(1-p)\right\}$$

- The distribution belongs to exponential family
- The sufficient statistic is $|E|$, number of edges
- $\theta = \log \frac{p}{1-p}$, which projects the interval $(0,1)$ to $\mathcal{R}$
- Since we already assumed $|V|$ is given, so the part $\exp\{\binom{|V|}{2} \log(1-p)\}$ does not depend on data, regarded as $h(p)$
- For this model, the sufficient statistic of interest is *the number of edges*

Degree Correction

For the random graph model, the degree distribution for any node is the same. Allow degree heterogeneity, we assume the model follows:

$$P(A_{ij} = 0) = c_{ij}, \qquad P(A_{ij} = 1) = c_{ij} e^{a_i}.$$

Recall that $P(A_{ij} = 0) + P(A_{ij} = 1) = 1$, so

$$c_{ij} = \frac{1}{1 + e^{a_i}}, \qquad P(A_{ij} = 1) = \frac{e^{a_i}}{1 + e^{a_i}}.$$

Define the logit function as $\mathrm{logit}(p) = \log \frac{p}{1-p}$, then

$$\mathrm{logit} P(A_{ij} = 1) = \log \frac{P(A_{ij} = 1)}{1 - P(A_{ij} = 1)} = \log \frac{c_{ij} e^{a_i}}{c_{ij}} = a_i, \quad i \in V$$

- In this model, we allow the edge connection probability p differs according to the node it starts with
- It targets on the directed graph.

Degree Correction, II

The likelihood function for the above model is

$$L(a) = \prod_{i,j} P(A_{ij} = 1)^{A_{ij}} (1 - P(A_{ij} = 1))^{1 - A_{ij}}.$$

The log-likelihood function is

$$\begin{aligned}
l(a) = \log L(a) &= \sum_{i,j} A_{ij} \log P(A_{ij} = 1) + (1 - A_{ij}) \log(1 - P(A_{ij} = 1)) \\
&= \sum_{i,j} A_{ij} \log \frac{P(A_{ij} = 1)}{1 - P(A_{ij} = 1)} + \log(1 - P(A_{ij} = 1)) \\
&= \sum_{i,j} A_{ij} a_i + \log(1 - P(A_{ij} = 1)) \\
&= \sum_{i} A_{i+} a_i + \sum_{i,j} \log(1 - P(A_{ij} = 1)),
\end{aligned}$$

where $A_{i+} = \sum_j A_{ij}$, and the second part $\sum_{i,j} \log(1 - P(A_{ij} = 1))$ does not depend on the data

Degree Correction, III

- The model still belongs to the exponential family
- The sufficient statistic is A_{i+}, $i \in V$, i.e., it is the out-degree for each node
- Take the partial derivative of the log-likelihood function and let it equal to 0. The solutions suggests that

$$\hat{P}(A_{ij} = 1) = \frac{e^{\hat{a}_i}}{1 + e^{\hat{a}_i}} = \frac{A_{i+}}{n-1},$$

  which is the standardized out-degree of node i
- Conclusion: it models the out-degree for each node by the parameter a_i. The density function can be written as a function of the out-degree. The sufficient statistic is the out-degree of each node i. The MLE can be represented by the sufficient statistics.

- If we are interested in other graphical structure, such as reciprocal edges ($(i,j) \in E$ and $(j,i) \in E$), complete structures (say, triangles), we can decide a distribution with sufficient statistics as the number of these structures.
- With the sufficient statistics, we may decide an exponential family distribution on the graph.

Exponential Random Graph Model (ERGM)

Exponential-family Random Graph Models (ERGMs) are exponential families over graphs,

$$P_\theta(G) = h(\theta) \sum_{i=1}^{d} T_i(G)\theta_i,$$

where $T_i(G)$ are functions of the graph/adjacency matrix.

ERGM: construction

To create an ERGM of interest, the following procedure works:

- Pick d (distinct) functions of the graph; they might be chosen through appeals to theory, experience, guesswork, tradition, referee pressure, trial and error, etc.
- Build the model based on these functions.

Examples:

- Random graph model: the function is the number of all the edges
- Model we just discussed: $|V|$ functions in total, each is the out-degree for one node
- Block model: The number of nodes in each community, n_r, and the number of edges between communities, e_{rs}, for $1 \leq r, s \leq K$.
- Not all the models are ERGMs!

ERGM: Properties

Consider the model as

$$P_\theta(G) = g(G)h(\theta)\sum_{i=1}^{d} T_i(G)\theta_i,$$

what is the probability for $(i,j) \in E$?

Solution. Let A_{+ij} denotes the adjacency matrix with $A_{ij} = 1$, and A_{-ij} denotes the adjacency matrix with $A_{ij} = 0$. We have two sets of statistics, $T(A_{+ij})$ and $T(A_{-ij})$.
According to the definition of ERGM,

$$P_\theta(A_{+ij}) = e^{T(A_{+ij})\theta}h(\theta) \qquad P_\theta(A_{-ij}) = e^{T(A_{-ij})\theta}h(\theta)$$

So, given all the other edges,

$$P((i,j) \in E|\text{the other edges}) = \frac{P_\theta(A_{+ij})}{P_\theta(A_{-ij})} = e^{(T(A_{+ij})-T(A_{-ij}))\theta}.$$

Therefore, the edge prob. for (i,j) is concluded as a logistic regression problem:

$$\log\frac{P(A_{ij}=1)}{1-P(A_{ij}=1)} = (T(A_{+ij}) - T(A_{-ij}))\theta.$$

NUS National University of Singapore

$G = (V, E)$ with adjacency matrix A follows ERGM with joint density

$$P_\theta(A) = e^{T(A)\theta} h(\theta) = e^{\sum_i T_i(A)\theta_i} h(\theta) = e^{\sum_i T_i(A)\theta_i} / Z(\theta),$$

where $Z(\theta) = \sum_x \exp\{\sum_i T_i(x)\theta_i\}$ since it is the normalizing function.

Take the partial derivative of $Z(\theta)$, we have

$$\begin{aligned} \frac{\partial Z(\theta)}{\partial \theta_i} &= \sum_x \exp\{\sum_j T_j(x)\theta_j\} T_i(x) \\ &= \sum_x T_i(x) \frac{\exp\{\sum_j T_j(x)\theta_j\}}{Z(\theta)} \times Z(\theta) \\ &= \sum_x T_i(x) Z(\theta) p_\theta(x) = Z(\theta) \sum_x T_i(x) p_\theta(x) = Z(\theta) E_\theta[T_i]. \end{aligned}$$

Therefore, for the sufficient statistics T_i, the expectation is

$$E_\theta[T_i] = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta_i} Z(\theta) = \frac{\partial}{\partial \theta_i} \log Z(\theta)$$

$G = (V, E)$ with adjacency matrix A follows ERGM with joint density

$$P_\theta(A) = e^{T(A)\theta} h(\theta) = e^{\sum_i T_i(A)\theta_i} h(\theta) = e^{\sum_i T_i(A)\theta_i} / Z(\theta),$$

where $Z(\theta) = \sum_x \exp\{\sum_i T_i(x)\theta_i\}$ since it is the normalizing function.

Take the partial derivative of $Z(\theta)$, we have

$$\begin{aligned}
\frac{\partial Z(\theta)}{\partial \theta_i} &= \sum_x \exp\{\sum_j T_j(x)\theta_j\} T_i(x) \\
&= \sum_x T_i(x) \frac{\exp\{\sum_j T_j(x)\theta_j\}}{Z(\theta)} \times Z(\theta) \\
&= \sum_x T_i(x) Z(\theta) p_\theta(x) = Z(\theta) \sum_x T_i(x) p_\theta(x) = Z(\theta) E_\theta[T_i].
\end{aligned}$$

Therefore, for the sufficient statistics T_i, the expectation is

$$E_\theta[T_i] = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta_i} Z(\theta) = \frac{\partial}{\partial \theta_i} \log Z(\theta)$$

Recall that the likelihood function is the density function,

$$L(\theta) = P_\theta(A) = e^{T(A)\theta}/Z(\theta).$$

The log-likelihood function is

$$l(\theta) = T(A)\theta - \log Z(\theta).$$

Take the derivative of it and let it equal to 0,

$$\frac{\partial l(\theta)}{\partial \theta_i}\Big|_{\theta=\hat{\theta}} = T_i(A) - \frac{\partial \log Z(\theta)}{\partial \theta_i}\Big|_{\theta=\hat{\theta}} = 0,$$

and the MLE satisfies

$$T_i(A) = \frac{\partial \log Z(\theta)}{\partial \theta_i}\Big|_{\theta=\hat{\theta}} = E_{\hat{\theta}}[T_i]$$

- With MLE, *the expectation of the sufficient stat. equals to the observed sufficient stat.*

ERGM: MLE

Recall that the likelihood function is the density function,

$$L(\theta) = P_\theta(A) = e^{T(A)\theta}/Z(\theta).$$

The log-likelihood function is

$$l(\theta) = T(A)\theta - \log Z(\theta).$$

Take the derivative of it and let it equal to 0,

$$\frac{\partial l(\theta)}{\partial \theta_i}\Big|_{\theta=\hat{\theta}} = T_i(A) - \frac{\partial \log Z(\theta)}{\partial \theta_i}\Big|_{\theta=\hat{\theta}} = 0,$$

and the MLE satisfies

$$T_i(A) = \frac{\partial \log Z(\theta)}{\partial \theta_i}\Big|_{\theta=\hat{\theta}} = E_{\hat{\theta}}[T_i]$$

- With MLE, *the expectation of the sufficient stat. equals to the observed sufficient stat.*

- Random graph model: The sufficient statistic is $|E|$, and the parameter is $\theta = \log \frac{p}{1-p}$.

  Given θ, $p = \frac{e^\theta}{1+e^\theta}$, and the expectation of $|E|$ is $\binom{|V|}{2} \times p = \binom{|V|}{2} \times \frac{e^\theta}{1+e^\theta}$. Therefore, MLE satisfies

$$|E| = \binom{|V|}{2} \times \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} \Longleftrightarrow \hat{p} = \frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}} = \frac{|E|}{\binom{|V|}{2}}$$

- Block model. The sufficient stat. are e_{rs}. The expectation of e_{rs} for $r \neq s$ is $n_r n_s b_{rs}$, where the parameter $\theta_{rs} = \log \frac{b_{rs}}{1-b_{rs}}$. Therefore, MLE satisfies

$$e_{rs} = n_r n_s \hat{b}_{rs} \Longleftrightarrow \hat{b}_{rs} = \frac{e^{\hat{\theta}_{rs}}}{1+e^{\hat{\theta}_{rs}}} = \frac{e_{rs}}{n_r n_s}$$

Remark 1. Since $T_i(A)$ can be calculated from the graph, the conclusion builds the equations for MLE

Remark 2. Yet, $E_\theta[T_i]$ may be hard to calculate

ERGM: Example

Question. Consider the politics blog dataset. Whether there is a link from A to B depends on the number of edges (base parameter), popularity of B (whether other blogs refer to it or not), the expansiveness of A (whether A refers to other blogs), and the probability of reciprocal edges if there is a link from B to A. Build an ERGM for this data set, which includes these information.

Solution. Let A denote the adjacency matrix. Mathematically, we represent the information with some statistics

$$A_{++}, \quad A_{i+}, \quad A_{+i}, \quad \sum_{i,j} A_{ij}A_{ji}.$$

Therefore, we build an ERGM with all these stats are sufficient stats. The density function would be

$$P_\theta(A) = \exp\{A_{++}\theta_0 + \sum_{i\in V} A_{i+}\theta_i^{(1)} + \sum_{j\in V} A_{+j}\theta_j^{(2)} + \theta_n \sum_{i,j} A_{ij}A_{ji}\}/Z(\theta),$$

where

$$Z(\theta) = \sum_{A} \exp\{A_{++}\theta_0 + \sum_{i\in V} A_{i+}\theta_i^{(1)} + \sum_{j\in V} A_{+j}\theta_j^{(2)} + \theta_n \sum_{i,j} A_{ij}A_{ji}\}$$

ERGM: Example

Question. Consider the politics blog dataset. Whether there is a link from A to B depends on the number of edges (base parameter), popularity of B (whether other blogs refer to it or not), the expansiveness of A (whether A refers to other blogs), and the probability of reciprocal edges if there is a link from B to A. Build an ERGM for this data set, which includes these information.

Solution. Let A denote the adjacency matrix. Mathematically, we represent the information with some statistics

$$A_{++}, \quad A_{i+}, \quad A_{+i}, \quad \sum_{i,j} A_{ij}A_{ji}.$$

Therefore, we build an ERGM with all these stats are sufficient stats. The density function would be

$$P_\theta(A) = \exp\{A_{++}\theta_0 + \sum_{i\in V} A_{i+}\theta_i^{(1)} + \sum_{j\in V} A_{+j}\theta_j^{(2)} + \theta_n \sum_{i,j} A_{ij}A_{ji}\}/Z(\theta),$$

where

$$Z(\theta) = \sum_A \exp\{A_{++}\theta_0 + \sum_{i\in V} A_{i+}\theta_i^{(1)} + \sum_{j\in V} A_{+j}\theta_j^{(2)} + \theta_n \sum_{i,j} A_{ij}A_{ji}\}$$

The notation is confusing here. To avoid misunderstanding, let

$$\theta = \theta_0, \quad \alpha_i = \theta_i^{(1)}, \quad \beta_j = \theta_j^{(2)}, \quad \rho = \theta_n.$$

So the density function is

$$P_\theta(A) = \exp\{A_{++}\theta + \sum_{i \in V} A_{i+}\alpha_i + \sum_{j \in V} A_{+j}\beta_j + \rho \sum_{i,j} A_{ij}A_{ji}\}/Z(\theta),$$

where $Z(\theta) = \sum_A \exp\{A_{++}\theta + \sum_{i \in V} A_{i+}\alpha_i + \sum_{j \in V} A_{+j}\beta_j + \rho \sum_{i,j} A_{ij}A_{ji}\}$. Note that Z is hard to calculate.

Now we consider the prob. for edges. Since Z is hard to calculate, we cannot calculate it directly. Given the probability of the other edges, we consider the following conditions:

$P_{ij}(0,0)$: probability of no edge between nodes i and j
$P_{ij}(1,0)$: probability of existence of $i \to j$ but absence of $j \to i$
$P_{ij}(0,1)$: probability of existence of $j \to i$ but absence of $i \to j$
$P_{ij}(1,1)$: probability of existence of both $i \to i$ and $j \to i$

ERGM: Example, II

The notation is confusing here. To avoid misunderstanding, let

$$\theta = \theta_0, \quad \alpha_i = \theta_i^{(1)}, \quad \beta_j = \theta_j^{(2)}, \quad \rho = \theta_n.$$

So the density function is

$$P_\theta(A) = \exp\{A_{++}\theta + \sum_{i\in V} A_{i+}\alpha_i + \sum_{j\in V} A_{+j}\beta_j + \rho \sum_{i,j} A_{ij}A_{ji}\}/Z(\theta),$$

where $Z(\theta) = \sum_A \exp\{A_{++}\theta + \sum_{i\in V} A_{i+}\alpha_i + \sum_{j\in V} A_{+j}\beta_j + \rho \sum_{i,j} A_{ij}A_{ji}\}$. Note that Z is hard to calculate.

Now we consider the prob. for edges. Since Z is hard to calculate, we cannot calculate it directly. Given the probability of the other edges, we consider the following conditions:
$P_{ij}(0,0)$: probability of no edge between nodes i and j
$P_{ij}(1,0)$: probability of existence of $i \to j$ but absence of $j \to i$
$P_{ij}(0,1)$: probability of existence of $j \to i$ but absence of $i \to j$
$P_{ij}(1,1)$: probability of existence of both $i \to i$ and $j \to i$

ERGM: Example, III

Compared to the case $P_{ij}(0,0)$, note that $P_{ij}(1,0)$ means the number of edges A_{++} increases by 1, A_{i+} increases by 1, and A_{+j} increases by 1. Therefore,

$$\begin{aligned} P_\theta(A \text{ with } i \to j) \quad &= \quad \exp\{(A_{++}+1)\theta + \sum_{k \neq i} A_{i+}\alpha_i + (A_{i+}+1)\alpha_i \\ &\quad + \sum_{j \neq i} A_{+j}\beta_j + (A_{+j}+1)\beta_j + \rho \sum_{i,j} A_{ij}A_{ji}\}/Z(\theta) \end{aligned}$$

$$P_\theta(A \text{ without } (i,j) \text{ or } (j,i)) = \exp\{A_{++}\theta + \sum_{i \in V} A_{i+}\alpha_i + \sum_{j \in V} A_{+j}\beta_j + \rho \sum_{i,j} A_{ij}A_{ji}\}/Z($$

$$\Longrightarrow \frac{P_\theta(A \text{ with } i \to j)}{P_\theta(A \text{ without } (i,j) \text{ or } (j,i))} = \exp\{\theta + \alpha_i + \beta_j\}$$

If we say the prob. for $P_{ij}(0,0) = c_{ij}$, then

$$P_{ij}(1,0) = c_{ij}\exp\{\theta + \alpha_i + \beta_j\}.$$

Similarly,

$$P_{ij}(0,1) = c_{ij}\exp\{\theta + \alpha_j + \beta_i\},$$
$$P_{ij}(1,1) = c_{ij}\exp\{\theta + \alpha_i + \beta_j + \alpha_j + \beta_i + \rho\}.$$

Since $P_{ij}(0,0) + P_{ij}(1,0) + P_{ij}(0,1) + P_{ij}(1,1) = 1$,

$$\begin{aligned} c_{ij} &= 1/\big[1 + \exp\{\theta + \alpha_i + \beta_j\} + \exp\{\theta + \alpha_j + \beta_i\} \\ &\quad + \exp\{\theta + \alpha_i + \beta_j + \alpha_j + \beta_i + \rho\}\big]. \end{aligned}$$

The density can be written as

$$P_\theta(A_{ij}, A_{ji}) = \frac{e^{\mu_{ij}A_{ij} + \mu_{ji}A_{ji} + \rho A_{ij}A_{ji}}}{1 + e^{\mu_{ij}} + e^{\mu_{ji}} + e^{\mu_{ij} + \mu_{ji} + \rho}},$$

where $\mu_{ij} = \theta + \alpha_i + \beta_j$.

- The above model is called $p1$ model, which is the origin of ERGM
- ERGM is also called p^* model

Similarly,

$$P_{ij}(0,1) = c_{ij}\exp\{\theta + \alpha_j + \beta_i\},$$
$$P_{ij}(1,1) = c_{ij}\exp\{\theta + \alpha_i + \beta_j + \alpha_j + \beta_i + \rho\}.$$

Since $P_{ij}(0,0) + P_{ij}(1,0) + P_{ij}(0,1) + P_{ij}(1,1) = 1$,

$$\begin{aligned} c_{ij} &= 1/\big[1 + \exp\{\theta + \alpha_i + \beta_j\} + \exp\{\theta + \alpha_j + \beta_i\} \\ &\quad + \exp\{\theta + \alpha_i + \beta_j + \alpha_j + \beta_i + \rho\}\big]. \end{aligned}$$

The density can be written as

$$P_\theta(A_{ij}, A_{ji}) = \frac{e^{\mu_{ij}A_{ij} + \mu_{ji}A_{ji} + \rho A_{ij}A_{ji}}}{1 + e^{\mu_{ij}} + e^{\mu_{ji}} + e^{\mu_{ij} + \mu_{ji} + \rho}},$$

where $\mu_{ij} = \theta + \alpha_i + \beta_j$.

- The above model is called $p1$ model, which is the origin of ERGM
- ERGM is also called p^* model

ERGM: MLE Calculation

For $p1$ model, the expectation of sufficient stat is hard to calculate, since it is edges are dependent. Not to mention the cases when we consider more complicated structures (k-cliques, k-stars, ect.)

To solve it, here are some alternative methods:

- Stochastic Approximation (short introduction)
- Pseudo MLE
- MCMC

Stochastic Approximation

If we draw many many graphs with that distribution under $\hat{\theta}$, then the average of the sufficient stats from these graphs are close to the expectation. Relate that to $T(A)$ from data, and update the estimate.

- Start with a guess $\hat{\theta}^{(0)}$
- Generate many graphs from $\hat{\theta}^{(0)}$
- Approximate $E_{\hat{\theta}}[T]$ by sample averages
- Adjust $\hat{\theta}^{(i)}$ to $\hat{\theta}^{(i+1)}$ to bring $E_{\hat{\theta}}[T]$ closer to $T(x)$
- Repeat the procedure to get better and better estimation, until it converges

Stochastic Approximation

In the approximation, we need to generate the

- Start with an initial graph configuration $A^{(0)}$
- Pick an edge (i, j) at random
- Flip the edge with probability

$$\frac{p_\theta(A^{(0)}_{+ij})}{p_\theta(A^{(0)}_{-ij})},$$

  which does not involve $Z(\theta)$.
- Repeat the procedure a few times, and the result graph is a graph follows the distribution.

This is a Gibbs sampling procedure.