ST5225: Statistical Analysis of Networks
Lecture 8: Statistical Modelling of Networks: Random Graph

WANG Wanjie
staww@nus.edu.sg

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

Saturday 24 March, 2018

Outline

- Review: World Wide Web, Part I
- World Wide Web, Part II
- Statistical Modelling of Networks

Review

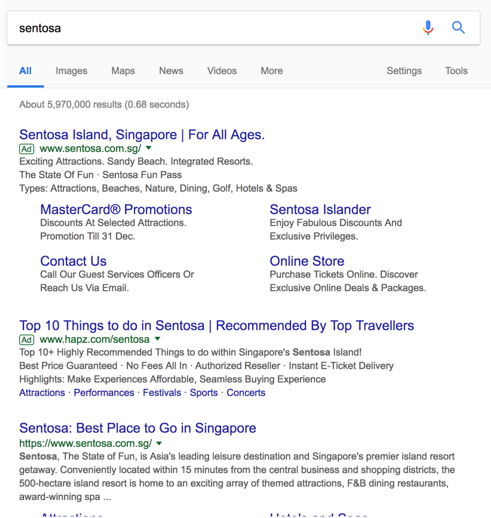
- Introduction of World Wide Web. Review of strongly connected and bowtie structure
- Methods for Web Search
  - Authority and Hub Algorithm
  - Page Rank
  - Mathematics

Overview

- Advertisement
  - Have the idea what is the problem for advertisements
  - Relate the idea of advertisement price with the matching market problem
  - Construct the price for advertisements
- Review of Statistical notions: model, PDF, likelihood function, MLE, simulation, etc.
- Random Graph
  - Know how to simulate a random graph
  - With a given network, know how to calculate the likelihood
  - Problems with random graph
- If time permits, stochastic block model

Advertisement

- If you search “Sentosa” on google, the results will have two advertisement, followed by the results



Advertisement

- The companies, say Sentosa trip agents, pay Google to post an advertisement about them
- When does google display these ads?
- People who search for “Sentosa” would be more interested in the Sentosa trips
- When people search for “Sentosa”, Google displays the ads according to the money paid
- People who are interested in the ads may click to the link

- It happens for every search engine.
- Without the search engines, companies pay to display ads for everyone. Now they show ads to those who have intent

Advertisement

- The companies, say Sentosa trip agents, pay Google to post an advertisement about them
- When does google display these ads?
- People who search for “Sentosa” would be more interested in the Sentosa trips
- When people search for “Sentosa”, Google displays the ads according to the money paid
- People who are interested in the ads may click to the link

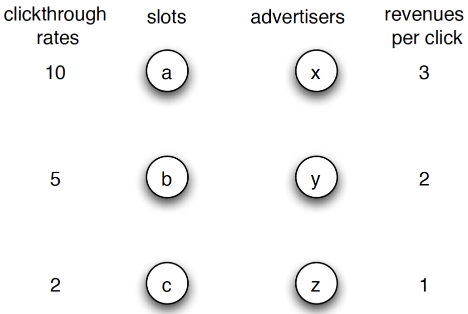
- It happens for every search engine.
- Without the search engines, companies pay to display ads for everyone. Now they show ads to those who have intent

Advertisement

Settings and Problems:

- The company creates an ad that shows every time when the user enters "Sentosa", and that ad links to their website. *The company pay the search engine only when some one clicks through the link*. This strategy is called *paying per click*.
  - Examples in the textbook:
  - "Calligraphy pens": $1.70 per click
  - "mortgage refinancing": $50 per click

- How to set the price for each company?
  - Too many keywords, and it is hard to set price for each search
  - Auction the slots would help, yet how to set the auction?
  - Vickrey-Clarke-Groves (VCG) mechanism



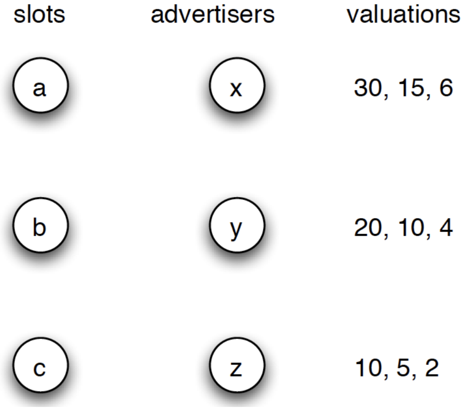
- Several slots for one keyword. Top slots indicate high clichthrough rates.
- Several advertisers. Each advertiser gets different revenue from clicks
- **Target**: Charge on slots properly

Matching Market

- For one advertiser, the valuation of a slot is

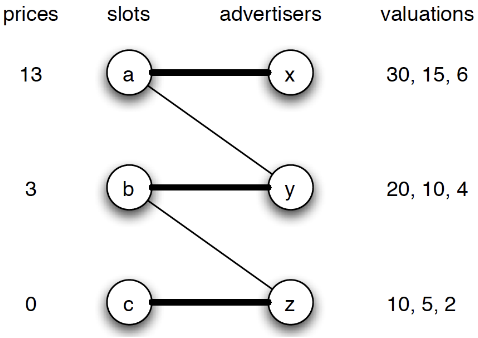
$$\text{valuation} = \text{clickthrough rate} \times \text{ revenue per click}$$

- Matching market problem!



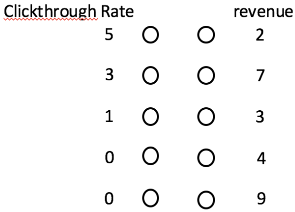
Matching Market

- Find a series of market-clearing prices, and the corresponding matching.



Matching Market

- More slots than advertises, or more advertisers than slots?
- Create "fake" slots and "fake" advertisers.
  - Advertisers > slots: create "fake" slots of clickthrough rate 0
  - Advertisers < slots: create "fake" advertisers of revenue 0.



Summary

- The advertisement problem for WWW network
- How to formulate the problem
- Solution: matching markets
- In reality, the advertisers may lie on the revenue. One solution is to apply VCG mechanism. Check the textbook if you are interested.
- More topics of interest. e.g., the “hidden” clickthrough rates.

With the observed data, what we can do?

- Design-based approach: there is a population, our observed data are randomly chosen from the population, which is a sample. We use the descriptive statistics of the sample as the estimator of the population
  - totally random
  - for the networks, we have discussed possible problems for this approach
- Model-based approach: we assume the population follow some model. The observed data are randomly chosen from the population, which can be regarded as the sample of the model. With the statistics, we estimate the parameters of the model, and then estimate the population
  - Assumption: population follows some model

Example: Flip a coin for 10 times, and get the result THHHTTHTTH (denoted as 0111001001).

- Design-based approach: with the observed data, we have the sample mean 0.5, and sample variance 0.278.
- Model-based approach:
  - assume the outcome follows Bernoulli dist with parameter p
  - Estimate p by some estimator (e.g., MOM, MLE, etc.)
  - With MLE, we estimate $\hat{p}$ = mean of the outcomes, which is 0.5
  - Therefore, our estimation is that the outcome follows Bernoulli(0.5) distribution, with mean 0.5, and variance 0.25.
  - The distribution can be used for further estimations, say, kurtosis, tail probabilities, etc.

Example: Flip a coin for 10 times, and get the result THHHTTHTTH (denoted as 0111001001).

- Design-based approach: with the observed data, we have the sample mean 0.5, and sample variance 0.278.
- Model-based approach:
  - assume the outcome follows Bernoulli dist with parameter p
  - Estimate p by some estimator (e.g., MOM, MLE, etc.)
  - With MLE, we estimate $\hat{p} =$ mean of the outcomes, which is 0.5
  - Therefore, our estimation is that the outcome follows Bernoulli(0.5) distribution, with mean 0.5, and variance 0.25.
  - The distribution can be used for further estimations, say, kurtosis, tail probabilities, etc.

Review: Models

Example: Flip a coin for 10 times, and get the result THHHTTHTTH (denoted as 0111001001).

- Design-based approach: with the observed data, we have the sample mean 0.5, and sample variance 0.278.
- Model-based approach:
  - assume the outcome follows Bernoulli dist with parameter p
  - Estimate p by some estimator (e.g., MOM, MLE, etc.)
  - With MLE, we estimate $\hat{p} =$ mean of the outcomes, which is 0.5
  - Therefore, our estimation is that the outcome follows Bernoulli(0.5) distribution, with mean 0.5, and variance 0.25.
  - The distribution can be used for further estimations, say, kurtosis, tail probabilities, etc.

Review: Models

- When we say "models", we mean a collection

$$\{P_\theta : \theta \in \Theta\},$$

  where Θ is the parameter space (in the previous example, $\theta = p$, $\Theta = (0,1)$). P_θ gives the probability function.
- We estimate the parameter $\hat{\theta}$, and give the likelihood for our observed data as $P_{\hat{\theta}}(data)$

Note:

- The models can be viewed as our understanding of the experiment (prior information)
- θ can be a vector, containing multiple unknown parameters
- We can also check whether the model fits the data or not

- When we say "models", we mean a collection

$$\{P_\theta : \theta \in \Theta\},$$

where Θ is the parameter space (in the previous example, $\theta = p$, $\Theta = (0, 1)$). P_θ gives the probability function.

- We estimate the parameter $\hat{\theta}$, and give the likelihood for our observed data as $P_{\hat{\theta}}(data)$

Note:

- The models can be viewed as our understanding of the experiment (prior information)
- θ can be a vector, containing multiple unknown parameters
- We can also check whether the model fits the data or not

Review: Models: estimators

Multiple ways to estimate θ. Some popular choices: MLE, MOM, etc.

- Maximum Likelihood Estimate (MLE)
  - Given θ, the prob. density function of a single observation: $f_\theta(x)$
  - Given θ, the joint density function for the observed data is $f_\theta(x_1, x_2, \cdots, x_n)$
  - Fixed $x_1, x_2, \cdots, x_n$, then the density function is a function of θ:

$$L(\theta|x_1, x_2, \cdots, x_n) = f_\theta(x_1, x_2, \cdots, x_n),$$

    for which we call *likelihood function*.
  - MLE is $\hat{\theta}$ that maximizes $L(\theta)$.
  - Example: for the coin toss problem, the prob. density function is

$$f_p(x) = p^x(1-p)^{1-x}, \qquad x = 0, 1.$$

    The observed data is 0111001001, so the joint density function is

$$f_p(0111001001) = p^5(1-p)^5 = L(p; 0111001001).$$

    Take the derivative of $L(p)$ and let it equal to 0. It gives $p = 0.5$. Check the original function to make sure $p = 0.5$ maximizes $L(p)$. Therefore, MLE here is $\hat{p} = 0.5$

Review: Models: estimators

Multiple ways to estimate θ. Some popular choices: MLE, MOM, etc.

- Maximum Likelihood Estimate (MLE)
  - Given θ, the prob. density function of a single observation: $f_\theta(x)$
  - Given θ, the joint density function for the observed data is $f_\theta(x_1, x_2, \cdots, x_n)$
  - Fixed $x_1, x_2, \cdots, x_n$, then the density function is a function of θ:

    $$L(\theta | x_1, x_2, \cdots, x_n) = f_\theta(x_1, x_2, \cdots, x_n),$$

    for which we call *likelihood function*.
  - MLE is $\hat{\theta}$ that maximizes $L(\theta)$.
  - Example: for the coin toss problem, the prob. density function is

    $$f_p(x) = p^x (1-p)^{1-x}, \qquad x = 0, 1.$$

    The observed data is 0111001001, so the joint density function is

    $$f_p(0111001001) = p^5 (1-p)^5 = L(p; 0111001001).$$

    Take the derivative of $L(p)$ and let it equal to 0. It gives $p = 0.5$. Check the original function to make sure $p = 0.5$ maximizes $L(p)$. Therefore, MLE here is $\hat{p} = 0.5$

Goodness-of-fit of model:

- For the linear regression problem, we consider multiple models
- For example, we consider a dataset about students, with inputs: height, weight, age, GPA, citizenship, and the output is the salary of them after they graduated for 3 years.
- The model can be $salary \sim height + weight + age + GPA + citizenship$, and it can also be $salary \sim age + GPA + citizenship$
  For both model, there are parameters to estimate
- Model selection: which model is better?
  The goodness-of-fit statistic is such a measurement for the model
- Model selection tells us whether the variable is important or not. It is much more than then descriptive statistics.

- For networks, we want to propose some models to help us explore

$$\{P_\theta(G); \theta \in \Theta\},$$

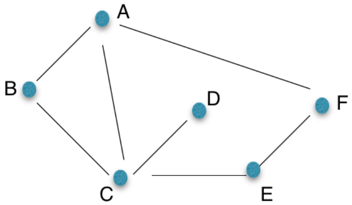
where G is the observed network, and θ are the parameters

- What is the PDF P? Hopefully, P fits the data, and can be estimated from the data
- In our module, I will introduce the following models:
  - Random Graph Model
  - Stochastic block model
  - Exponential random graph model
  - Latent space model

Graph $G = (V, E)$ consists of the set of nodes V and the set of edges $E \subset V \times V$. We have the adjacency matrix A to denote whether there is an edge or not.

- We want to figure out $P_\theta(G)$ for possible graphs G
- We consider the set of graphs where $|V|$ is given
- Now the set of nodes is given. For i and j, we flip an (unbalanced) coin to decide whether there is an edge or not.
  Assumption: $A_{ij} \overset{i.i.d}{\sim} Bernoulli(p)$, p is a constant for any i, j.
  - $P(A_{ij} = 1) = p$. With prob. p, i and j are connected
  - *i.i.d.*: identical and independently distributed. All the edges follow the same distribution, independent with each other.
  - For any two nodes, the $P(connected)$ is the same
  - $A_{ij} = 1$ and $A_{ik} = 1$ does not change the prob. that $A_{jk} = 1$
  - For directed graphs, A_{ij} and A_{ji} are independent. For undirected graphs, $A_{ij} = A_{ji} \sim Bernoulli(p)$

Example.



What is the probability for this network, if it is generated as a random graph model with $p = 0.4$?

Solution. In there network, there are 6 nodes, and $\binom{6}{2} = 15$ possible pairs of these nodes. For these pairs, 7 of them are connected (7 edges).
Therefore, the prob. for this observed network is

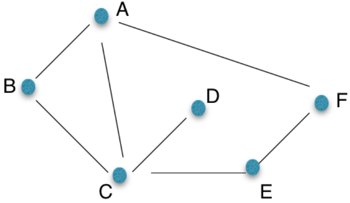
$$P_{p=0.5}(G) = (0.4)^7 * (0.6)^{15-7} = 2.75e - 05$$

For an undirected graph $G = (V, E)$, what is the prob. of it under the random graph model with parameter p? What is the maximum likelihood estimate (MLE) of p?

- $$P_p(G) = \prod_{i<j} p^{A_{ij}}(1-p)^{1-A_{ij}} = p^{|E|}(1-p)^{\binom{|V|}{2}-|E|},$$
  which is also the likelihood function $L(p)$
- The log-likelihood function is
  $$\log P_p(G) = |E| \log(p) + (\binom{|V|}{2} - |E|) \log(1-p)$$
- Maximize $P_p(G)$ is equivalent with maximizing $\log P_p(G)$. Therefore, we take derivative of $\log P_p(G)$ w.r.t. p, and let it equal to 0.
  $$\frac{d \log P_p(G)}{dp} = \frac{|E|}{p} - \frac{\binom{|V|}{2} - |E|}{1-p} = 0, \qquad \hat{p} = \frac{|E|}{\binom{|V|}{2}}$$
- Check it achieves the maximum

- The estimation $\hat{p}$ is the density of the graph
- For G follows random graph model, the expectation of degree for each node is
$$(n-1)p$$
- Parameterization I: Fixed p, then when $n \to \infty$, the expected degree goes to infinity. We call this case as *dense graph*.
- Parameterization II: Fixed the mean degree $\lambda = (n-1)p$, then when $n \to \infty$, p decreases to 0. The graph gets sparser and sparser, and we call it as *sparse graph*.
- In the mean degree parameterization, the distribution of the degree follows $Poisson(\lambda)$.
This result follows from the property of Binomial distribution, that $Binomial(n, p)$ converges to $Poisson(np)$ for large n, if np is a constant.

Example.



Assume this network follows the random graph model, what is MLE for p? What is the prob. of this network with the MLE?

Solution. The network consists of 6 nodes and 7 edges. According to the property of random graph models, the MLE is

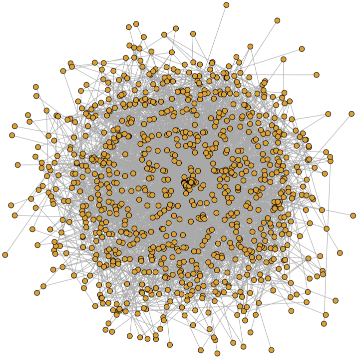
$$\hat{p} = \frac{7}{\binom{6}{2}} = 7/15.$$

The prob. with $\hat{p} = 7/15$ is

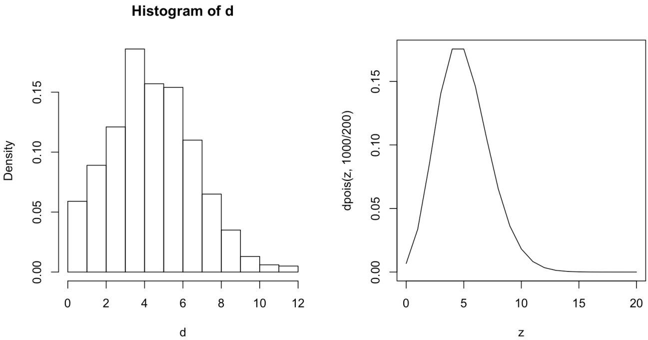
$$P_{\hat{p}}(G) = (7/15)^7 * (8/15)^{15-7} = 3.16e - 05$$

Random Graph Model: Simulation

- Now, given n and p, we can generate a random graph
  1. Create an $n \times n$ matrix A
  2. Let $A_{ij} \sim Bernoulli(p)$, $i < j$
  3. Make A to be symmetric
  4. Generate a graph according to the adjacency matrix A
- Example: $n = 1000$ nodes, $p = 1/200$, $np = 1000/200 = 5$



Random Graph Model: Degree distribution



Left: Histogram of the example random graph in previous slide
Right: Poisson density with parameter $\lambda = 5$

Problems

- Degree distributions in real networks are not binomial/Poisson (Recall the power-law)
- Too few triangles in this model, since the edges are totally independent

$$\text{Proportion of triangles} = p^3$$

- It cannot describe the clustering structure

Model criticism: find things where

- the model makes predictions
- You didn't fit to them
- check against data

- Random graph model is not perfect

Summary

- A simple model that describes the graphs with two kinds of parameterizations
- Likelihood function/MLE
- It helps us to simulate/generate graphs.
- Problems with the random graph model

We try to model the communities.

- Assume there are K communities in total, which are community $1, 2, 3, \cdots K$
- Each node has a label ℓ_i, which indicates the community it belongs to ($\ell_i = 1, 2, \cdots, K$)
- For node i and j, the prob. that they are connected is b_{ℓ_i,ℓ_j}, i.e., $A_{ij} \sim Bernoulli(b_{\ell_i,\ell_j})$
- The prob. that i and j are connected are related to the community labels of nodes, not the nodes themselves
- It allows different prob. of connections between/within the communities/groups. Usually, we would assume higher prob. for the connections within one group, than those between groups.
- The model allows the community structure.

Stochastic Block Model, II

- For each node, we denote its label as a $K \times 1$ vector

$$\pi_i = (0\ 0\ 0\ \cdots\ 1\ 0\ \cdots\ 0)',$$

  which contains one and only one "1", and all the other elements are 0. The location of "1" indicates the community label of node i
- The labels of n nodes can be denoted as an $n \times K$ matrix, Π, where each row contains only one nonzero element, as "1"
- The prob. of connections between the groups can be written as an $K \times K$ matrix B
- The prob. of connections between nodes can be denoted as $P = \Pi B \Pi'$. For Bernoulli distribution, the prob. equals to the expectation of the random variable, therefore

$$E[A] = \Pi B \Pi' - diag(\Pi B \Pi'),$$

  where $diag(\Pi B \Pi')$ is the diagonal matrix formed by the diagonals of $\Pi B \Pi'$. It's to exclude self-loops.

Matrix form for $\Pi B \Pi'$ in SBM ($K = 2$)

Consider a simple example with only 2 communities.

- Note that B is 2×2 matrix, where $B = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$
- The result expectation of A is

$$\Pi B \Pi' = \begin{bmatrix} \pi_1' \\ \vdots \\ \pi_n' \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} [\pi_1, \ldots, \pi_n] =$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 1 \end{bmatrix} = \begin{bmatrix} a & b & a & b & a & b & a \\ b & c & b & c & b & c & b \\ a & b & a & b & a & b & a \\ b & c & b & c & b & c & b \\ a & b & a & b & a & b & a \\ b & c & b & c & b & c & b \\ a & b & a & b & a & b & a \end{bmatrix}$$

Matrix form for $\Pi B \Pi'$ in SBM ($K = 2$), II

$$(\Pi B \Pi')_{i,j} = P(\text{edge between nodes } i \text{ and } j), \quad i \neq j$$

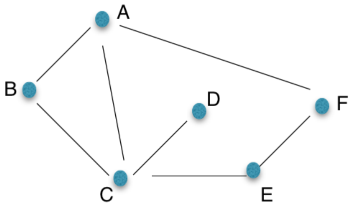
$$(\Pi B \Pi')(i,j) = \begin{cases} a, & i \text{ and } j \text{ belong to community 1,} \\ c, & i \text{ and } j \text{ belong to community 2,} \\ b, & \text{otherwise,} \end{cases}$$

$$\Pi B \Pi' = \begin{bmatrix} a & b & a & b & a & b & a \\ b & c & b & c & b & c & b \\ a & b & a & b & a & b & a \\ b & c & b & c & b & c & b \\ a & b & a & b & a & b & a \\ b & c & b & c & b & c & b \\ a & b & a & b & a & b & a \end{bmatrix} \xrightarrow{permute} \begin{bmatrix} a & a & a & a & b & b & b \\ a & a & a & a & b & b & b \\ a & a & a & a & b & b & b \\ a & a & a & a & b & b & b \\ b & b & b & b & c & c & c \\ b & b & b & b & c & c & c \\ b & b & b & b & c & c & c \end{bmatrix}$$

Note: because of the community partition, there is block structure in the probability matrix.

SBM: example

Example.



Assume that nodes A, B, C, E and F belong to community 1, node D belongs to community 2, and the prob. matrix is

$$B = \begin{pmatrix} 0.7 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}$$

What is the prob. of this network under this setting?

Solution. The prob. function is

$$P(G) = (0.7)^6(1-0.7)^4 \times 0.1^1(1-0.1)^4 = 6.25e-05$$

SBM: Example II

Note

- we should also consider the prob. of no edges.
- Proper partition will increase the prob. of the network

For this problem, the matrices are

$$\Pi = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 0.7 & 0.1 \\ 0.1 & 0.5 \end{pmatrix} \quad \Pi B \Pi' = \begin{pmatrix} 0.7 & 0.7 & 0.7 & 0.1 & 0.7 & 0.7 \\ 0.7 & 0.7 & 0.7 & 0.1 & 0.7 & 0.7 \\ 0.1 & 0.1 & 0.1 & 0.5 & 0.1 & 0.1 \\ 0.7 & 0.7 & 0.7 & 0.1 & 0.7 & 0.7 \\ 0.7 & 0.7 & 0.7 & 0.1 & 0.7 & 0.7 \\ 0.7 & 0.7 & 0.7 & 0.1 & 0.7 & 0.7 \end{pmatrix}$$

- For each matrix, the labels should be written down (community 1, community 2, node A, B, C, etc.)
- Rearrage $\Pi B \Pi'$ here, would result a block matrix.

SBM: Properties

- Degree: say n_k nodes in block k ($\sum_{k=1}^{K} n_k = n = |V|$), then for node i, the degree distribution is

$$\sum_{k=1}^{K} Binomial(n_k - \delta_{\ell_i,k}, b_{\ell_i,k}).$$

- Probability of an edge:

$$\begin{aligned}
Pr(A_{ij}=1) &= \sum_{j=1}^{K}\sum_{k=1}^{K} P(A_{ij}=1, \ell_i=j, \ell_j=k) \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K} P(A_{ij}=1|\ell_i=j, \ell_j=k)P(\ell_i=j, \ell_j=k) \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K} b_{jk}P(\ell_i=j, \ell_j=k) = \sum_{j=1}^{K}\sum_{k=1}^{K} b_{jk}\frac{n_j n_k}{n^2}
\end{aligned}$$

SBM: Properties II

- Parameters: Π, B
- Likelihood

$$L = \prod_{i,j} b_{\ell_i,\ell_j}^{A_{ij}} (1 - b_{\ell_i,\ell_j})^{1-A_{ij}}.$$

Assume we have n_r nodes in block r and e_{rs} edges between block r and block s. Then,

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

- Log-likelihood:

$$l = \log(L) = \sum_{r,s} e_{rs} \log \frac{b_{rs}}{1 - b_{rs}} + \sum_{r \neq s} n_r n_s \log(1 - b_{rs})] + \sum_r \binom{n_r}{2} \log(1 - b_{rs})$$

- If we know the partition, then the MLE is

$$\hat{b}_{rs} = \frac{e_{rs}}{n_r n_s}, \quad \hat{b}_{rr} = \frac{e_{rs}}{\binom{n_r}{2}},$$

Comparison between random graph and SBM

Random Graph	**Block model**
Edges are independent	Edges are independent given the labels
All edges have equal prob.	All edges between 2 groups have equal prob.
All nodes have the same binomial degree dist.	All nodes in the same group have the same dist.
No preference on triangles	Large prob. for triangles if all the three nodes are in the same group
MLE: $\hat{p}$ =density of the graph	MLE: $\hat{b}_{rs} = \frac{e_{rs}}{n_r n_s}$, $1 \leq r, s \leq K$

SBM: unknown labels

Usually, the community labels are unknown to us

- Assume that $\ell_i \sim Multinomial(\rho)$, where ρ is an $K \times 1$ vector, indicates the prob. that a node belongs to each community
- Likelihood

$$L = \sum_{\ell \in \{1:k\}^n} \prod_{i,j} b_{\ell_i,\ell_j}^{A_{ij}} (1 - b_{\ell_i,\ell_j})^{1-A_{ij}} \prod_{i=1}^{n} \rho_{\ell_i}.$$

- Log-likelihood:

$$l = \log(L) = \log \sum_{\ell \in \{1:k\}^n} \prod_{i,j} b_{\ell_i,\ell_j}^{A_{ij}} (1 - b_{\ell_i,\ell_j})^{1-A_{ij}} \prod_{i=1}^{n} \rho_{\ell_i}$$

- Note that $\log \sum \neq \sum \log \cdots$
- NO explicit solution!

SBM: unknown labels

Methods

- Expectation-Maximization algorithm:
- Spectral method
- Modularity