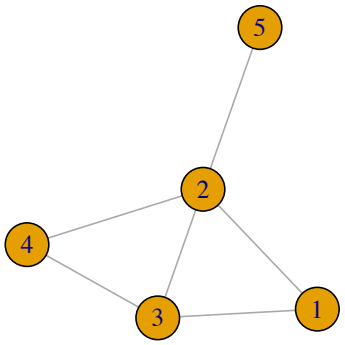
Mock exam for Statistical analysis of network data - solutions

1. (a)



(b)

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

(c)

$$I = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

(d) By definition, $d_i = \sum_j A_{ij}$. Substituting, $d_1 = 2$, $d_2 = 4$, $d_3 = 3$, $d_4 = 2$, $d_5 = 2$.

(e) $N_{C_3} = \frac{1}{6} \sum_{i,j,k} A_{ij} A_{jk} A_{ki} = 2$.

2. We define

$$d_i = \sum_{j \neq i} A_{ij}. \tag{1}$$

(a) To calculate the moments of this random variable we use the law of iterated expectation

$$\begin{aligned} \mathrm{E}\{d_i\} &= \mathrm{E}_\xi \mathrm{E}_{A|\xi}\{d_i\} = \sum_{j \neq i} \mathrm{E}_\xi g(\xi_i) g(\xi_j) \\ &= (n-1)\mathrm{E}^2\{g(\xi_i)\} = (n-1)(\int g(x) \cdot 1 \, dx)^2 = (n-1)\|g\|_1^2. \end{aligned}$$

We define the function p-norm to be

$$\|g\|_p^p = \int_0^1 g^p(x) \, dx.$$

(b) To compute the variance we use the law of total variance as seen in class:

$$\begin{aligned}
\mathrm{Var}\{d_i\} &= \mathrm{E}_\xi \mathrm{Var}_{A|\xi}\{d_i\} + \mathrm{Var}_\xi \mathrm{E}_{A|\xi}\{d_i\} \\
&= \mathrm{E}_\xi \sum_{j\neq i} g(\xi_i)g(\xi_j)(1-g(\xi_i)g(\xi_j)) + \mathrm{Var}_\xi \sum_{j\neq i} g(\xi_i)g(\xi_j) \\
&= (n-1)\{\|g\|_1^2 - \|g\|_2^4\} + \mathrm{E}\sum_{j\neq i}\sum_{k\neq i} g(\xi_i)g(\xi_j)g(\xi_k)g(\xi_i) - (n-1)^2\|g\|_1^4 \\
&= (n-1)\{\|g\|_1^2 - \|g\|_2^4\} + \mathrm{E}\sum_{j\neq i}\sum_{k\neq i} g(\xi_i)g(\xi_j)g(\xi_k)g(\xi_i) - (n-1)^2\|g\|_1^4 \\
&= (n-1)\{\|g\|_1^2 - \|g\|_2^4\} + \mathrm{E}\sum_{j\neq i}\sum_{k\neq i,j} g^2(\xi_i)g(\xi_j)g(\xi_k) + \mathrm{E}\sum_{j\neq i} g^2(\xi_i)g^2(\xi_j) \\
&- (n-1)^2\|g\|_1^4 \\
&= (n-1)\{\|g\|_1^2 - \|g\|_2^4\} + (n-1)(n-2)\|g\|_2^2\|g\|_1^2 + (n-1)\|g\|_2^4 - (n-1)^2\|g\|_1^4 \\
&= (n-1)\|g\|_1^2 + (n-1)(n-2)\|g\|_2^2\|g\|_1^2 - (n-1)^2\|g\|_1^4.
\end{aligned}$$

For the special case of an Erdos Renyi network this becomes

$$\begin{aligned}
\mathrm{Var}\{d_i\} &= (n-1)\rho_n + (n-1)(n-2)\rho_n^2 - (n-1)^2\rho_n^2 \\
&= (n-1)\rho_n + (n-1)\{n-2-(n-1)\}\rho_n^2 = (n-1)\rho_n + (n-1)(-1)\rho_n^2.
\end{aligned}$$

We can also note that if A is an Erdos–Renyi graph with edge probability ρ_n then $\|g\|_1 = \sqrt{\rho_n}$, then $d_i = \sum_{j\neq i} A_{ij} \sim \mathrm{Bin}\{(n-1), \rho_n\}$ which has mean $(n-1)\rho_n$ and variance $(n-1)\rho_n(1-\rho_n)$, which matches.

We can also compute this directly. We already have $\mathrm{E}\{d_i\}$ above. Then it follows

$$\begin{aligned}
\mathrm{E}\{d_i^2\} &= \mathrm{E}\sum_{j\neq i}\sum_{k\neq i} A_{ij}A_{ik} \\
&= \mathrm{E}\sum_{j\neq i}\sum_{k\neq i,j} A_{ij}A_{ik} + \mathrm{E}\sum_{j\neq i} A_{ij} \\
&= \|g\|_2^2\|g\|_1^2(n-1)(n-2) + \|g\|_1^2(n-1). \qquad (2)
\end{aligned}$$

We now note that:

$$\begin{aligned}
\mathrm{Var}\{d_i\} &= \mathrm{E}\{d_i^2\} - \mathrm{E}^2\{d_i\} \\
&= \|g\|_2^2\|g\|_1^2(n-1)(n-2) + \|g\|_1^2(n-1) - (n-1)^2\|g\|_1^4. \qquad (3)
\end{aligned}$$

This gives the same result as the law of total variance.

(c) We can calculate this by brute force for $i \neq j$:

$$\begin{aligned}
\mathrm{Cov}\{d_i, d_j\} &= \mathrm{E}\{d_i d_j\} - \mathrm{E}\{d_i\}\mathrm{E}\{d_j\} \\
&= \mathrm{E}\{\sum_{k\neq i} A_{ik}\sum_{l\neq j} A_{jl}\} - \mathrm{E}\{d_i\}\mathrm{E}\{d_j\} \\
&= \mathrm{E}\{\sum_{k\neq i,j,l} A_{ik}\sum_{l\neq j,i} A_{jl}\} + \mathrm{E}\{\sum_{k=j}\sum_{l=i} A_{ik}A_{jl}\} + \mathrm{E}\{\sum_{l\neq j} A_{ij}A_{jl}\} + \mathrm{E}\{\sum_{l\neq j} A_{il}A_{jl}\} \\
&+ \mathrm{E}\{\sum_{l\neq j} A_{il}A_{ji}\} - (n-1)^2\|g\|_1^4 \\
&= \sum_{k\neq i,j,l}\sum_{l\neq j,i} \|g\|_1^4 + \|g\|_2^2 + 3(n-2)\|g\|_2^2\|g\|_1^2 - (n-1)^2\|g\|_1^4 \\
&= (n-3)(n-2)\|g\|_1^4 - (n-1)^2\|g\|_1^4 + \|g\|_2^2 + 3(n-2)\|g\|_2^2\|g\|_1^2 \\
&= (n^2 - 5n + 6 - n^2 + 2n - 1)\|g\|_1^4 + \|g\|_2^2 + 3(n-2)\|g\|_2^2\|g\|_1^2 \\
&= (-3n+5)\|g\|_1^4 + 3(n-2)\|g\|_2^2\|g\|_1^2 + \|g\|_2^2. \qquad (4)
\end{aligned}$$

Therefore in the special case of an Erdos–Renyi graph we get as $\|g\|_p = \sqrt{\rho_n}$ for $p = 1, 2, 3, \ldots$ the simple form of

$$\mathrm{Cov}\{d_i, d_j\} = \rho_n(1 - \rho_n),$$

and this concurs with our expectation, as one edge will be in common between the two sums.

Part (ii):

(a) A graph model is finitely exchangeable if for all permutations $\pi : [n] \to [n]$, $\boldsymbol{A} \stackrel{d}{=} \boldsymbol{A}^{(\pi)}$, where $\boldsymbol{A}^{(\pi)}$ is the permuted adjacency matrix with elements $A_{ij}^{(\pi)} = A_{\pi_i, \pi_j}$.

(b) The joint probability is

$$\Pr\{\boldsymbol{A} = \boldsymbol{a}\} = \Pr\{\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{\xi}\}\Pr\{\boldsymbol{\xi}\} = \prod_{i<j} g(\xi_i)g(\xi_j)\Pr\{\xi_i\}\Pr\{\xi_j\},$$

which is obviously invariant to permutations of the indices.

3. (i)

$$D = \begin{pmatrix} 0 & 2 & 1 & 2 & 2 \\ 2 & 0 & 1 & 2 & 2 \\ 1 & 1 & 0 & 1 & 1 \\ 2 & 2 & 1 & 0 & 1 \\ 2 & 2 & 1 & 1 & 0 \end{pmatrix}$$

(ii) Closeness centrality:

$$C_i = \frac{n}{\sum_{j \neq i} D_{ij}} = \frac{5}{\sum_{j \neq i} D_{ij}},$$

from which $C_1 = 5/7$, $C_2 = 5/7$, $C_3 = 5/4$, $C_4 = 5/6$, $C_5 = 5/6$.

(iii) Harmonic centrality:

$$C_i^{(H)} = \sum_{j \neq i} \frac{1}{D_{ij}},$$

from which $C_1^{(H)} = 2.5$, $C_2^{(H)} = 2.5$, $C_3^{(H)} = 4$, $C_4^{(H)} = 3$, $C_1^{(H)} = 3$.

4. (i)

$$\Pr\{\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{z}\} = \prod_{i<j} \theta_{z_i z_j}^{a_{ij}} (1 - \theta_{z_i z_j})^{1 - a_{ij}},$$

from which the log-likelihood is

$$\ell(\theta; \boldsymbol{a}, \boldsymbol{z}) = \sum_{i<j} \{a_{ij} \log(\theta_{z_i z_j}) + (1 - a_{ij}) \log(1 - \theta_{z_i z_j})\}.$$

(ii) Described in slides 14-17 of Lecture 8 (Latent space models, April 7).

(iii) Described in slides 4-5 of Week 6 (Network clustering, March 25).

5. (i) The degree-corrected stochastic blockmodel is written as

$$A_i j \mid z_i, z_j, \xi_i, \xi_j \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta_{z_i z_j} + \pi(\xi_i)\pi(\xi_j)).$$

From this, using the shorthand $P_{ij} = \pi(\xi_i)\pi(\xi_j)$, the log-likelihood is

$$\ell(\theta; \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{\xi}) = \sum_{i<j} \{a_{ij} \log(\theta_{z_i z_j} + P_{ij}) + (1 - a_{ij}) \log(1 - \theta_{z_i z_j} - P_{ij})\}.$$

Assuming all latent variables and the degree structure π known, we can separate out the likelihood component pertaining to each of the blocks:

$$\ell(\theta_{ab}; \boldsymbol{a}, \boldsymbol{z}, \boldsymbol{\xi}) = \sum_{i<j} \{a_{ij} \log(\theta_{ab} + P_{ij}(\boldsymbol{\xi})) + (1 - a_{ij}) \log(1 - \theta_{ab} - P_{ij}(\boldsymbol{\xi}))\} I(z_i = a, z_j = b).$$

We can maximize this likelihood with respect to θ_{ab}. The solution will depend on both $\boldsymbol{z}$ and $\boldsymbol{\xi}$. In principle, we can obtain a global solution by considering the estimator of θ_{ab} for all possible configurations of $\boldsymbol{z}$ and $\boldsymbol{\xi}$, and find the optimum of the likelihood, but this is very impractical because of the high number of possible cluster membership configurations. Moreover, the presence of the uniform $\boldsymbol{\xi}$ requires using Bayesian methods, marginalising over $\boldsymbol{\xi}$. The EM algorithm may also be used.

(ii) Introducing the notations $\bar{\theta} = \sum_{a,b} \frac{h_a h_b}{n^2} \theta_{ab}$ and $P = \int_{[0,1]} \pi(x) dx$,

$$\begin{aligned}
\mathrm{E}(d_i) &= \sum_j \mathrm{E}_{\boldsymbol{\xi},\boldsymbol{z}}[\mathrm{E}(A_{ij} \mid \boldsymbol{\xi}, \boldsymbol{z})] = \sum_j \mathrm{E}_{\boldsymbol{\xi},\boldsymbol{z}}[\theta_{z_i z_j} + \pi(\xi_i)\pi(\xi_j)] \\
&= \sum_j \mathrm{E}_{\boldsymbol{z}}(\theta_{z_i z_j}) + \mathrm{E}_{\xi_i}[\pi(\xi_i)] \sum_j \mathrm{E}_{\xi_j}[\pi(\xi_j)] \\
&= \sum_j \bar{\theta} + \sum_j P^2 = (n-1)(\bar{\theta} + P^2).
\end{aligned}$$

(iii) A_{ij} and A_{ji} represent now edges between the same nodes but with opposite direction. Their probability may differ, so the generating model is not necessarily symmetric, neither the interaction matrix θ_{ab}, nor the degree correction. Moreover, even one single node can belong to different clusters when being a donor and when being a receptor, and we might also specify different uniform latent variables for these roles. The model becomes

$$A_i j \mid \boldsymbol{\xi}^{(\mathrm{in})}, \boldsymbol{z}^{(\mathrm{in})}, \boldsymbol{\xi}^{(\mathrm{out})}, \boldsymbol{z}^{(\mathrm{out})} \overset{\mathrm{iid}}{\sim} \mathrm{Bernoulli}\{\theta_{z_i^{(\mathrm{in})} z_j^{(\mathrm{out})}} + \pi^{(\mathrm{in})}(\xi_i^{(\mathrm{in})})\pi^{(\mathrm{out})}(\xi_j^{(\mathrm{out})})\}.$$

The corresponding log-likelihood is

$$\begin{aligned}
\ell(\theta; \boldsymbol{\xi}^{(\mathrm{in})}, \boldsymbol{z}^{(\mathrm{in})}, \boldsymbol{\xi}^{(\mathrm{out})}, \boldsymbol{z}^{(\mathrm{out})}) = \sum_{i,j} \{ & a_{ij} \log(\theta_{z_i^{(\mathrm{in})} z_j^{(\mathrm{out})}} + \pi^{(\mathrm{in})}(\xi_i^{(\mathrm{in})})\pi^{(\mathrm{out})}(\xi_j^{(\mathrm{out})})) \\
& + (1 - a_{ij}) \log(1 - \theta_{z_i^{(\mathrm{in})} z_j^{(\mathrm{out})}} - \pi^{(\mathrm{in})}(\xi_i^{(\mathrm{in})})\pi^{(\mathrm{out})}(\xi_j^{(\mathrm{out})}))\}.
\end{aligned}$$