Assessed Coursework - Statistical analysis of network data – MATH-448

November 2024

*Solutions to the assessed coursework should be handed in at the lecture **22nd of November at 10am in Ma B1 427**. You should not ask for help during the tutorial from the TA and should only hand in your own workings. Group work is not allowed, and please sign a statement saying "This is all my own work" on the first page of the assessed coursework. **Using LaTeX is highly advised for the clarity of your answers.** The mark allocation to each question will depend on how difficult it is, and you should give solutions to all questions. **Please note that there are 5 questions.** The assessed coursework is worth 15% of the course mark.*

Instructions to hand back your coursework:

- Number all your pages in the following format: page number/total number of pages.
- Put your name and SCIPER number on all pages.
- Staple together all answer sheets you want to hand back.

Exercise 1

Assume you observe a network $\mathcal{G}$ corresponding to Figure 1.

(i) Write down the adjacency matrix of the network represented in Figure 1. Calculate the degrees of this network.

(ii) Write down all cycles present in the network with four nodes or less represented in Figure 1, and write down how many copies of these cycles are found in the network.

(iii) Decide which node in the network is most "important"; justify your choice of "important".

(iv) Assume you add and delete two edges. The new edges you add are 15 and 57, and the ones you delete are 23 and 34. Compute the new degrees in the new network. Does this network correspond to one connected component? Explain the rationale to your answer.

Solution 1

(i) The adjacency matrix of the network is:

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

The degrees of each node are:

$$\deg(1) = 4, \quad \deg(2) = 4, \quad \deg(3) = 2, \quad \deg(4) = 4, \quad \deg(5) = 2, \quad \deg(6) = 3, \quad \deg(7) = 3$$

(ii) **Triangles (3-cycles):**

- Cycle 1: $1 \to 2 \to 7 \to 1$
- Cycle 2: $1 \to 4 \to 6 \to 1$

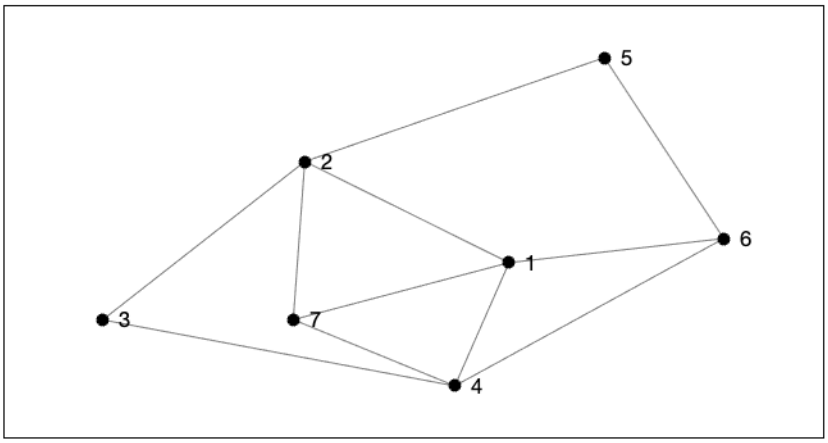

Figure 1: A planar representation of a seven node network.

- Cycle 3: $1 \to 4 \to 7 \to 1$

Quadrilaterals (4-cycles):

- Cycle 1: $1 \to 6 \to 5 \to 2 \to 1$
- Cycle 2: $1 \to 6 \to 4 \to 7 \to 1$
- Cycle 3: $1 \to 2 \to 7 \to 4 \to 1$
- Cycle 4: $2 \to 3 \to 4 \to 7 \to 2$
- Cycle 5: $2 \to 3 \to 4 \to 1 \to 2$

(iii) Node 1 is the most "important" because:

- It has the highest degree ($\deg(1) = 4$) alongside nodes 2 and 4.
- Further compute centralities to show that It connects to other high-degree nodes (nodes 2, 4, 6, and 7) and it participates in multiple cycles, indicating a central position in the network structure.

(iv) After adding edges 1-5 and 5-7, and deleting edges 2-3 and 3-4, the new degrees are:

$$\deg(1) = 5$$
$$\deg(2) = 3$$
$$\deg(3) = 0$$
$$\deg(4) = 3$$
$$\deg(5) = 4$$
$$\deg(6) = 3$$
$$\deg(7) = 4$$

Node 3 becomes isolated ($\deg(3) = 0$), so the network consists of two connected components:

- A main component with nodes $1, 2, 4, 5, 6, 7$.
- An isolated node 3.

Therefore, the network does not correspond to one connected component.

Exercise 2

A *directed* network makes a difference between an edge between i and j versus a directed edge starting at i and ending in j. Assume the directed network has an adjacency matrix A which has binary independent entries, and is defined on n nodes. Introduce the two parameters $\boldsymbol{\pi}$ and $\boldsymbol{\psi}$ that are $n \times 1$ vectors with entries between zero and unity and assume that the edges are binary and independent with expectation

$$\mathrm{E}\{A_{ij}\} = \pi_i \psi_j, \quad 1 \leq i \neq j \leq n. \tag{1}$$

(i) Calculate the expectation of the 'in' and 'out' degrees of the network. These are defined as

$$d_i^{(\mathrm{in})} = \sum_{j \neq i} A_{ji},$$

and

$$d_i^{(\mathrm{out})} = \sum_{j \neq i} A_{ij},$$

respectively.

(ii) What constraints need to be placed on the vectors $\boldsymbol{\pi}$ and $\boldsymbol{\psi}$ given the total number of edges are fixed to $\mathcal{E} = \sum_{i \neq j} A_{ij}$?

(iii) Assume we propose the estimators $\widehat{\pi}_i = d_i^{(\mathrm{out})}/C_1$ and $\widehat{\psi}_j = d_j^{(\mathrm{in})}/C_2$, what values should we choose as C_1 and C_2 ? **Hints:** You may note that as all directed edges need to start and end somewhere

$$\sum_i \mathrm{E}\{d_i^{in}\} = \sum_i \mathrm{E}\{d_i^{out}\},$$

and you may assume that as n is large

$$\|\boldsymbol{\pi}\|_1 >> \pi_i, \quad \|\boldsymbol{\psi}\|_1 >> \psi_i,$$

as well as since $0 < \pi_i < 1$ and $0 < \psi_i < 1$, it is reasonable to assume

$$\|\boldsymbol{\pi}\|_1 >> \|\boldsymbol{\pi}\|_2, \quad \|\boldsymbol{\psi}\|_1 >> \|\boldsymbol{\psi}\|_2.$$

Solution 2

(i) The **in-degree** of node i is defined as:

$$d_i^{(\mathrm{in})} = \sum_{j \neq i} A_{ji}.$$

The expected value is:

$$\mathrm{E}\{d_i^{(\mathrm{in})}\} = \mathrm{E}\left\{\sum_{j \neq i} A_{ji}\right\} = \sum_{j \neq i} \mathrm{E}\{A_{ji}\} = \sum_{j \neq i} \pi_j \psi_i = \psi_i \sum_{j \neq i} \pi_j.$$

The **out-degree** of node i is defined as:

$$d_i^{(\mathrm{out})} = \sum_{j \neq i} A_{ij}.$$

The expected value is:

$$\mathrm{E}\{d_i^{(\mathrm{out})}\} = \mathrm{E}\left\{\sum_{j \neq i} A_{ij}\right\} = \sum_{j \neq i} \mathrm{E}\{A_{ij}\} = \sum_{j \neq i} \pi_i \psi_j = \pi_i \sum_{j \neq i} \psi_j.$$

(ii) The total number of edges is:

$$\mathcal{E} = \sum_{i \neq j} A_{ij}.$$

Taking the expected value:

$$\mathrm{E}\{\mathcal{E}\} = \mathrm{E}\left\{\sum_{i \neq j} A_{ij}\right\} = \sum_{i \neq j} \mathrm{E}\{A_{ij}\} = \sum_{i \neq j} \pi_i \psi_j.$$

Thus, the constraint is:

$$\sum_{i \neq j} \pi_i \psi_j = \mathcal{E}.$$

(iii) We wish to estimate π_i and ψ_j. We wish to chose estimators that are unbiased. We shall take estimators as suggested to be

$$\hat{\pi}_i = \frac{d_i^{out}}{C_1}. \tag{2}$$

We shall calculate the expectation of this:

$$\mathrm{E}\hat{\pi}_i = \frac{\mathrm{E}\{d_i^{out}\}}{C_1}. \tag{3}$$

We first need to calculate the rhs of that equation and find:

$$\begin{aligned}
\mathrm{E}\{d_i^{out}\} &= \sum_{j \neq i} \pi_i \psi_j / C_1 && (4)\\
&= \pi_i \{\|\boldsymbol{\psi}\|_1 - \psi_i\}/C_1 && (5)\\
&= \frac{\pi_i \|\boldsymbol{\psi}\|_1}{C_1}(1 - \psi_i/\|\boldsymbol{\psi}\|_1) && (6)\\
&\approx \frac{\pi_i \|\boldsymbol{\psi}\|_1}{C_1}. && (7)
\end{aligned}$$

Thus we should take $C_1 = \|\boldsymbol{\psi}\|_1$.

$$\hat{\psi}_i = \frac{d_i^{in}}{C_2}. \tag{8}$$

We shall calculate the expectation of this:

$$\mathrm{E}\hat{\psi}_i = \frac{\mathrm{E}\{d_i^{in}\}}{C_2}. \tag{9}$$

We first need to calculate the rhs of that equation and find:

$$\begin{aligned}
\mathrm{E}\{d_i^{in}\} &= \sum_{j \neq i} \psi_i \pi_j / C_2 && (10)\\
&= \psi_i \{\|\boldsymbol{\pi}\|_1 - \pi_i\}/C_2 && (11)\\
&= \frac{\psi_i \|\boldsymbol{\pi}\|_1}{C_2}(1 - \pi_i/\|\boldsymbol{\pi}\|_1) && (12)\\
&\approx \frac{\psi_i \|\boldsymbol{\pi}\|_1}{C_2}. && (13)
\end{aligned}$$

Thus we should take $C_2 = \|\boldsymbol{\pi}\|_1$.

But we prefer to estimate this from the data. Note that from (ii)

$$\mathcal{E} = \sum \mathrm{E}A_{ij} \approx \|\boldsymbol{\pi}\|_1 \|\boldsymbol{\psi}\|_1.$$

Because all edges need to start and end somewhere

$$\sum_i \mathrm{E}\{d_i^{in}\} = \sum_i \mathrm{E}\{d_i^{out}\} = \|\boldsymbol{\pi}\|_1 \|\boldsymbol{\psi}\|_1 = \|\boldsymbol{\pi}\|_1^2 = \|\boldsymbol{\psi}\|_1^2,$$

and so we can deduce

$$\|\boldsymbol{\pi}\|_1 = \|\boldsymbol{\psi}\|_1 = \sqrt{\mathcal{E}}.$$

Thus if we want an estimator use method of moments and we take

$$\hat{\pi}_i = \frac{d_i^{out}}{\sqrt{\sum_i \mathrm{E}\{d_i^{out}\}}}, \tag{14}$$

and

$$\hat{\psi}_i = \frac{d_i^{in}}{\sqrt{\sum_i \mathrm{E}\{d_i^{in}\}}}. \tag{15}$$

Exercise 3

Assume that we generate the network $\mathcal{G}$ with adjacency matrix A from the generating mechanism

$$A_{ij} \,|\, \boldsymbol{\alpha}, \boldsymbol{\xi} \sim \text{Bernoulli}\left(\alpha_1 + \alpha_2 \xi_i \xi_j\right), \quad 1 \le i < j \le n. \tag{16}$$

Assume that $\alpha_1 \sim U(0, 1/2)$ and $\alpha_2 \,|\alpha_1 \sim U(0, 1-\alpha_1)$, and as usual $\xi_i \sim U(0,1)$. The variable α is independent of the vector $\boldsymbol{\xi}$.

(i) Calculate the mean degree, conditional on the latent variable $\boldsymbol{\xi}$.

(ii) Calculate the degree variance given $\boldsymbol{\xi}$.

Solution 3

(i) The degree of node i in an undirected network is:

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

The expected value of d_i is:

$$\mathrm{E}[d_i] = \sum_{j \neq i} \mathrm{E}[A_{ij}] = \sum_{j \neq i} \mathrm{E}_{\alpha_1,\alpha_2}\left[\mathrm{E}[A_{ij} \mid \alpha_1, \alpha_2, \xi]\right] = \sum_{j \neq i} \mathrm{E}_{\alpha_1,\alpha_2}\left[\alpha_1 + \alpha_2 \xi_i \xi_j\right] = \sum_{j \neq i} \left(\mathrm{E}[\alpha_1] + \mathrm{E}[\alpha_2] \xi_i \xi_j\right).$$

Since $\alpha_1 \sim U(0, 1/2)$,

$$\mathrm{E}[\alpha_1] = \frac{0 + \frac{1}{2}}{2} = \frac{1}{4}.$$

Given α_1, $\alpha_2 \mid \alpha_1 \sim U(0, 1-\alpha_1)$,

$$\mathrm{E}[\alpha_2 \mid \alpha_1] = \frac{0 + (1-\alpha_1)}{2} = \frac{1-\alpha_1}{2}.$$

Therefore,

$$\mathrm{E}[\alpha_2] = \mathrm{E}_{\alpha_1}\left[\mathrm{E}[\alpha_2 \mid \alpha_1]\right] = \mathrm{E}_{\alpha_1}\left[\frac{1-\alpha_1}{2}\right] = \frac{1}{2}\left(1 - \mathrm{E}[\alpha_1]\right) = \frac{1}{2}\left(1 - \frac{1}{4}\right) = \frac{3}{8}.$$

Now, compute $\mathrm{E}[d_i]$:

$$\mathrm{E}[d_i] = \sum_{j \neq i} \left(\frac{1}{4} + \frac{3}{8}\xi_i \xi_j\right) = (n-1) \cdot \frac{1}{4} + \frac{3}{8}\xi_i \sum_{j \neq i} \xi_j = \frac{n-1}{4} + \frac{3}{8}\xi_i S_i,$$

where $S_i = \sum_{j \neq i} \xi_j$. **Thus,**

$$\mathrm{E}[d_i] = \frac{n-1}{4} + \frac{3}{8}\xi_i \sum_{j \neq i} \xi_j.$$

(ii) We will use the law of total variance:

$$\mathrm{Var}[d_i \mid \xi] = \mathrm{E}\left[\mathrm{Var}[d_i \mid \alpha_1, \alpha_2, \xi]\right] + \mathrm{Var}\left[\mathrm{E}[d_i \mid \alpha_1, \alpha_2, \xi]\right].$$

First, compute $\mathrm{E}[d_i \mid \alpha_1, \alpha_2, \xi]$:

$$\mathrm{E}[d_i \mid \alpha_1, \alpha_2, \xi] = \sum_{j \neq i} \mathrm{E}[A_{ij} \mid \alpha_1, \alpha_2, \xi] = \sum_{j \neq i} (\alpha_1 + \alpha_2 \xi_i \xi_j) = (n-1)\alpha_1 + \alpha_2 \xi_i S_i.$$

Next, compute $\mathrm{Var}[d_i \mid \alpha_1, \alpha_2, \xi]$:

$$\mathrm{Var}[d_i \mid \alpha_1, \alpha_2, \xi] = \sum_{j \neq i} \mathrm{Var}[A_{ij} \mid \alpha_1, \alpha_2, \xi] = \sum_{j \neq i} p_{ij}(1 - p_{ij}),$$

where $p_{ij} = \alpha_1 + \alpha_2 \xi_i \xi_j$ and as they are conditionally independent. Now, compute $\mathrm{E}\left[\mathrm{Var}[d_i \mid \alpha_1, \alpha_2, \xi]\right]$ and $\mathrm{Var}\left[\mathrm{E}[d_i \mid \alpha_1, \alpha_2, \xi]\right]$. First, compute $\mathrm{E}\left[\mathrm{Var}[d_i \mid \alpha_1, \alpha_2, \xi]\right]$:

$$\mathrm{E}\left[\mathrm{Var}[d_i \mid \alpha_1, \alpha_2, \xi]\right] = \sum_{j \neq i} \mathrm{E}\left[p_{ij}(1 - p_{ij})\right].$$

We expand $p_{ij}(1 - p_{ij})$:

$$p_{ij}(1 - p_{ij}) = (\alpha_1 + \alpha_2 \xi_i \xi_j)(1 - \alpha_1 - \alpha_2 \xi_i \xi_j).$$

Simplify the expression:

$$\begin{aligned} p_{ij}(1 - p_{ij}) &= (\alpha_1 + \alpha_2 \xi_i \xi_j)\,(1 - \alpha_1 - \alpha_2 \xi_i \xi_j) \\ &= \alpha_1(1 - \alpha_1) - \alpha_1 \alpha_2 \xi_i \xi_j + \alpha_2 \xi_i \xi_j (1 - \alpha_1) - \alpha_2^2 \xi_i^2 \xi_j^2. \end{aligned}$$

Take expectations:

$$\mathrm{E}\left[p_{ij}(1 - p_{ij})\right] = \mathrm{E}[\alpha_1(1 - \alpha_1)] - 2\mathrm{E}[\alpha_1 \alpha_2]\xi_i \xi_j + \mathrm{E}[\alpha_2]\xi_i \xi_j - \mathrm{E}[\alpha_2^2](\xi_i \xi_j)^2.$$

We have $\mathrm{E}[\alpha_1(1 - \alpha_1)] = \mathrm{E}[\alpha_1] - \mathrm{E}[\alpha_1^2] = \frac{1}{4} - \frac{1}{12} = \frac{1}{6}$. $\mathrm{E}[\alpha_1 \alpha_2] = \frac{1}{12}$ (easy check via integration). $\mathrm{E}[\alpha_2] = \frac{3}{8}$. $\mathrm{E}[\alpha_2^2] = \frac{7}{36}$ (computed via integration).

Therefore,

$$\mathrm{E}\left[p_{ij}(1 - p_{ij})\right] = \frac{1}{6} - \frac{1}{6}\xi_i \xi_j + \frac{3}{8}\xi_i \xi_j - \frac{7}{36}(\xi_i \xi_j)^2.$$

Simplify:

$$\mathrm{E}\left[p_{ij}(1 - p_{ij})\right] = \frac{1}{6} + \frac{5}{24}\xi_i \xi_j - \frac{7}{36}(\xi_i \xi_j)^2.$$

Now, sum over $j \neq i$:

$$\mathrm{E}\left[\mathrm{Var}[d_i \mid \alpha_1, \alpha_2, \xi]\right] = \sum_{j \neq i} \left(\frac{1}{6} + \frac{5}{24}\xi_i \xi_j - \frac{7}{36}(\xi_i \xi_j)^2\right).$$

Define $S_i = \sum_{j \neq i} \xi_j$ and $Q_i = \sum_{j \neq i} (\xi_j)^2$:

$$\mathrm{E}\left[\mathrm{Var}[d_i \mid \alpha_1, \alpha_2, \xi]\right] = \frac{n-1}{6} + \frac{5}{24}\xi_i S_i - \frac{7}{36}\xi_i^2 Q_i.$$

Next, compute $\mathrm{Var}\left[\mathrm{E}[d_i \mid \alpha_1, \alpha_2, \xi]\right]$:

$$\mathrm{Var}\left[\mathrm{E}[d_i \mid \alpha_1, \alpha_2, \xi]\right] = \mathrm{Var}\left[(n-1)\alpha_1 + \alpha_2 \xi_i S_i\right].$$

Compute the variance:

$$\mathrm{Var}\left[(n-1)\alpha_1 + \alpha_2 \xi_i S_i\right] = (n-1)^2 \mathrm{Var}[\alpha_1] + (\xi_i S_i)^2 \mathrm{Var}[\alpha_2] + 2(n-1)(\xi_i S_i)\,\mathrm{Cov}[\alpha_1, \alpha_2].$$

where

- $\mathrm{Var}[\alpha_1] = \frac{1}{48}$ (using variance of uniforms formula)

- $\mathrm{Var}[\alpha_2] = \frac{31}{576}$, using computations done previously for $\mathrm{E}[\alpha_2]$ and $\mathrm{E}[\alpha_2^2]$.
- $\mathrm{Cov}[\alpha_1, \alpha_2] = -\frac{1}{96}$ follows using $\mathrm{Cov}\left[\alpha_1, \alpha_2\right] = E\left[\alpha_1 \alpha_2\right] - E\left[\alpha_1\right] E\left[\alpha_2\right]$ and the computations we have done previously.

Therefore,

$$\mathrm{Var}\left[\mathrm{E}[d_i \mid \alpha_1, \alpha_2, \xi]\right] = \frac{(n-1)^2}{48} + \frac{31(\xi_i S_i)^2}{576} - \frac{(n-1)(\xi_i S_i)}{48}.$$

Finally, the degree variance given ξ is:

$$\mathrm{Var}[d_i \mid \xi] = \left(\frac{n-1}{6} + \frac{5}{24}\xi_i S_i - \frac{7}{36}\xi_i^2 Q_i\right) + \left(\frac{(n-1)^2}{48} + \frac{31(\xi_i S_i)^2}{576} - \frac{(n-1)(\xi_i S_i)}{48}\right).$$

Thus,

$$\mathrm{Var}[d_i \mid \xi] = \frac{n-1}{6} + \frac{(n-1)^2}{48} + \frac{5}{24}\xi_i S_i - \frac{(n-1)(\xi_i S_i)}{48} - \frac{7}{36}\xi_i^2 Q_i + \frac{31(\xi_i S_i)^2}{576}.$$

Exercise 4

Starting from lecture 6, slide 11, Aldous Hoover's theorem can be stated as: An infinite symmetric array $\mathbf{A} = (A_{ij})_{i<j}$ is exchangeable if and only if there is a function $f_\alpha : [0,1]^2 \to [0,1]$ for uniform random variable α, with the sequence of independent random variables (ξ_i) and the array with independent entries (χ_{ij}) are $U[0,1]$ random variables and deterministic function f_α, such that

$$A_{ij} \mid \alpha, \xi \sim \mathrm{I}\left(\chi_{ij} \leq f_\alpha\left(\xi_i, \xi_j\right)\right).$$

Another representation is as follows, assuming F_α is deterministic: There is a function $F_\alpha : [0,1]^3 \to [0,1]$, that is symmetric in its first two arguments, with the sequence (ξ_i) and the array (χ_{ij}) are $U[0,1]$ independent random variables independent such that

$$A_{ij} | \alpha, \xi \sim F_\alpha\left(\xi_i, \xi_j, \chi_{i,j}\right)$$

Show that the two representations are equivalent. Further show that they are equivalent to

$$A_{ij} \mid \alpha, \xi \sim \text{Bernoulli}(f_\alpha\left(\xi_i, \xi_j\right)).$$

Solution 4

The main point is the following: For this Bernoulli representation, we need to show that there is a symmetric function $W : [0,1]^2 \to [0,1]$ such that

$$F\left(\xi_i, \xi_j, \chi_{ij}\right) \stackrel{d}{=} I_{\{(x,y,z) | z \leq f(x,y)\}}\left(\xi_i, \xi_j, \chi_{ij}\right).$$

To prove it, we define

$$f(x,y) := \mathbb{E}[F(x,y,\chi)], \quad \chi \sim U[0,1].$$

Then,

$$\begin{aligned}\mathbb{P}\left[F\left(\xi_i, \xi_j, \chi_{ij}\right) = 1\right] = \mathbb{E}\left[f\left(\xi_i, \xi_j, \chi_{ij}\right)\right] =& \mathbb{E}\left[\mathbb{E}\left[F\left(\xi_i, \xi_j, \chi_{ij}\right) \mid \xi_i, \xi_j\right]\right] \\ &= \mathbb{E}\left[f\left(\xi_i, \xi_j\right)\right] = \mathbb{P}\left[\chi_{i,j} \leq f\left(\xi_i, \xi_j\right)\right].\end{aligned}$$

Thus,

$$A_{ij} | \xi_i, \xi_j \sim \text{Bernoulli}(f(\xi_i, \xi_j)). \tag{17}$$

Exercise 5

The empirical modularity of a network G is defined for label vector z, edge variables $\{A_{ij}\}$, and degrees $\{d_i\}$ to be

$$\widehat{Q}_G(z) = \sum_{i<j}\left\{A_{ij} - \frac{d_i d_j}{\sum_l d_l}\right\}\delta_{z_i z_j}.$$

Let us study the properties of a related object by replacing A_{ij} and d_i by their expectations $\mathrm{E}\{A_{ij}\}$ $\mathrm{E}\{d_i\}$, defining this theoretical modularity to be for deterministic z

$$Q_G(z) = \sum_{i<j}\left\{\mathrm{E}\{A_{ij}\} - \frac{\mathrm{E}\{d_i\}\mathrm{E}\{d_j\}}{\sum_l \mathrm{E}\{d_l\}}\right\}\delta_{z_i z_j}. \tag{18}$$

Assume that for label vector z

$$\mathrm{E}\{A_{ij} \mid z\} = \theta_{z_i z_j}, \quad i,j = 1, \ldots, n. \tag{19}$$

(i) Assume that first $\theta_{11} = 1/2 = \theta_{22}$ but $\theta_{12} = 1/4$: there are $\lfloor n/2 \rfloor$ nodes in each of the two classes of nodes, so $z_i = 1$ if $i \leq \lfloor n/2 \rfloor$ and $z_i = 2$ if $i > \lfloor n/2 \rfloor$. Compute the expected modularity, defined in (18), $Q_G(z)$ in this instance ?

(ii) Assume that first $\theta_{12} = 1/2$ but $\theta_{11} = 1/4 = \theta_{22}$: there are $\lfloor n/2 \rfloor$ nodes in each of the two classes of nodes, so $z_i = 1$ if $i \leq \lfloor n/2 \rfloor$ and $z_i = 2$ if $i > \lfloor n/2 \rfloor$. Compute the expected modularity, defined in (18), $Q_G(z)$ in this instance ?

(iii) As a last choice assume that z_i is assigned to one or two independently at random with equal probability. What value does the modularity $Q_G(z)$ then take on average?

(iv) Please provide interpretation of the results you have just derived.

Solution 5

(i) First note that, $\theta_{11} = \theta_{22} = \dfrac{1}{2}$, $\theta_{12} = \dfrac{1}{4}$, we have equal-sized classes: $n_1 = n_2 = \dfrac{n}{2}$ and $z_i = 1$ for $i \leq \dfrac{n}{2}$, $z_i = 2$ otherwise. As the two classes are of equal size and equal intra probability, their expected degree is the same. We first begin by computing the expected degree of nodes from the first class, The expected degree within Class 1, thus the connection with the remaining $n-1$ nodes from its class, is as follows

$$d_{\text{in}}^{(1)} = (n_1 - 1)\,\theta_{11} = \left(\frac{n}{2} - 1\right) \times \frac{1}{2}$$

With the other class, its degree is

$$d_{\text{out}}^{(1)} = n_2 \theta_{12} = \frac{n}{2} \times \frac{1}{4}$$

As such, the total expected degree for nodes in Class 1:

$$\mathbb{E}\{d_i\} = d_{\text{in}}^{(1)} + d_{\text{out}}^{(1)} = \left(\frac{n}{2} - 1\right) \times \frac{1}{2} + \frac{n}{2} \times \frac{1}{4} = \frac{3n}{8} - \frac{1}{2}.$$

This yields the following total expected degree

$$\sum_l \mathbb{E}\{d_l\} = n \times \mathbb{E}\{d_i\} = n\left(\frac{3n}{8} - \frac{1}{2}\right) = \frac{3n^2}{8} - \frac{n}{2}$$

Now we compute the full expression $\theta_{11} - \dfrac{(\mathbb{E}\{d_i\})^2}{\sum_l \mathbb{E}\{d_l\}}$. We first have

Calculate $(\mathbb{E}\{d_i\})^2$:

$$(\mathbb{E}\{d_i\})^2 = \left(\frac{3n}{8} - \frac{1}{2}\right)^2 = \left(\frac{3n-4}{8}\right)^2 = \frac{(3n-4)^2}{64}, \quad \sum_l \mathbb{E}\{d_l\} = \frac{3n^2}{8} - \frac{n}{2} = \frac{3n^2 - 4n}{8}$$

Thus,

$$\frac{(3n-4)^2}{8(3n^2-4n)} = \frac{(3n-4)^2}{8n(3n-4)} = \frac{3n-4}{8n}$$

And we have

$$\theta_{11} - \frac{(\mathbb{E}\{d_i\})^2}{\sum_l \mathbb{E}\{d_l\}} = \frac{1}{2} - \frac{3n-4}{8n} = \frac{4n - (3n-4)}{8n} = \frac{n+4}{8n}$$

Thus, the inside of the sum is constant. As such, we now need to compute the number of within-class pairs to simplify the expression.

$$N_{\text{within}} = \binom{n_1}{2} + \binom{n_2}{2} = 2\binom{n/2}{2} = \frac{n(n-2)}{4}$$

The final expression is

$$\boxed{\mathbb{E}\{\widehat{Q}_G(z)\} = N_{\text{within}} \left(\theta_{11} - \frac{(\mathbb{E}\{d_i\})^2}{\sum_l \mathbb{E}\{d_l\}}\right) = \frac{n(n-2)}{4} \times \frac{n+4}{8n} = \frac{(n-2)(n+4)}{32}}$$

(ii) We perform the same computations as before and only replacing with the new parameters. First, we have

$$\mathbb{E}\{d_i\} = \left(\frac{n}{2} - 1\right) \times \frac{1}{4} + \frac{n}{2} \times \frac{1}{2} = \left(\frac{n}{8} - \frac{1}{4}\right) + \frac{n}{4} = \frac{3n-2}{8}$$

which gives

$$\sum_l \mathbb{E}\{d_l\} = n\left(\frac{n}{2} - \frac{1}{4}\right) = n\frac{3n-2}{8},$$

which gives

$$\frac{E\left[d_i\right] E\left[d_j\right]}{\sum_l E\left[d_l\right]} = \frac{\left(\frac{3n-2}{8}\right)^2}{\frac{(3n-2)n}{8}} = \frac{3n-2}{8n}$$

Combining with the number of pairs, we get

$$\boxed{\mathbb{E}\{\widehat{Q}_G(z)\} = N_{\text{within}}\left(\frac{1}{4} - \frac{3n-2}{8n}\right) = \frac{-(n-2)^2}{32} < 0}$$

which is always negative for $n > 2$.

(iii) When labels z_i are assigned randomly with equal probability, the expected modularity is zero because the labels are uncorrelated with the network structure. Thus,

$$\boxed{\mathbb{E}\{\widehat{Q}_G(z)\} = 0}$$

(iv) (i) The positive modularity indicates strong community structure. Nodes are more likely to connect within their own group than to nodes in the other group.

(ii) The negative modularity suggests that nodes are more connected to nodes in the other group, indicating disassortative mixing.

(iii) A modularity of zero implies that the connections are random with respect to the assigned labels; there is no community structure.