Analyse statistique de graphes — Année 2022-2023
Catherine Matias — Sorbonne Université

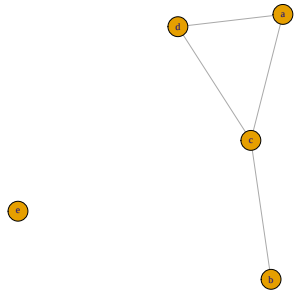
Examen - 1ère session - Durée : 3h

Les documents sont autorisés. Toutes les communications sont interdites. L'énoncé contient 3 parties : QCM de cours ; exercice d'application pratique (sur `R` ou Python) et problème théorique. Chacune compte pour environ 1/3 de la note, mais la difficulté de ces 3 parties est croissante.

NOM : PRÉNOM :

Questions cours - QCM (6 points).
Répondre directement sur l'énoncé. **ATTENTION : Cocher une seule réponse par question. Les mauvaises réponses sont comptées négativement. Il vaut mieux ne pas répondre au hasard !**

Dans les 4 questions suivantes, on considère le graphe :



Question 1. Le graphe ci-dessus est :
☐ simple et non dirigé
☐ acyclique et connexe
☐ à boucles (self-loops) et possède 2 composantes connexes

Question 2. Ce graphe contient :
☐ Un triangle et un nœud isolé
☐ Un cycle de longueur 4
☐ Une étoile à 4 branches (4-star)

Question 3. On s'intéresse au spectre d'une matrice Laplacienne de ce graphe :
☐ 0 n'est pas valeur propre de cette matrice
☐ On ne peut pas savoir si 0 est valeur propre sans faire le calcul
☐ 0 est valeur propre, de multiplicité 2

Question 4. Si on ajuste un modèle $\mathcal{G}(n,p)$ sur ce graphe alors on obtient :
☐ $n = 5$ et $\hat{p} = 2/5$
☐ $n = \{a, b, c, d, e\}$ et $\hat{p} = 4/10$
☐ $n = 4$ et $\hat{p} = 4/5$

Question 5. Pour mettre en œuvre un algorithme de re-branchement, il suffit d'avoir :
☐ Une suite de degrés observés
☐ Le nombre d'étoiles à 3 branches
☐ Un générateur de loi de puissance

Question 6. Dans un modèle à blocs stochastiques, l'estimation des paramètres repose sur une approximation variationnelle de l'algorithme `EM`, parce que :
☐ C'est une version beaucoup plus rapide en temps de calcul
☐ C'est une version plus simple à coder qui donne les mêmes résultats
☐ L'algorithme `EM` exact ne peut pas être mis en œuvre

Question 7. Lorsqu'un graphe est connexe, un algorithme de clustering spectral :
☐ permet de retrouver des groupes de nœuds de degrés similaires
☐ permet de retrouver le groupe des nœuds rouges et le groupe des nœuds bleus
☐ permet de retrouver des groupes de nœuds fortement connectés entre eux

Question 8. Pour faire du clustering de nœuds sur un graphe on peut :
☐ construire la matrice de similarité de ce graphe sur laquelle on applique le clustering spectral
☐ ajuster un modèle à blocs stochastiques sur ce graphe
☐ utiliser au choix le clustering spectral ou un modèle à blocs stochastiques dont les résultats sont identiques

Question 0. Dans un algorithme de clustering spectral, le nombre de clusters :
☐ Peut être estimé par un critère pénalisé analogue au BIC
☐ Peut être estimé par l'heuristique du 'trou des valeurs propres'
☐ C'est un paramètre qui doit être choisi a priori et pas estimé

Question 10. Le modèle d'attachement préférentiel permet de générer un graphe :
☐ Qui possède la propriété petit monde
☐ Qui est un arbre (donc un graphe acyclique)
☐ Qui a exactement la suite des degrés que l'on s'est fixé à l'avance

Question 11. La matrice d'adjacence d'un graphe :
☐ Est toujours une matrice symétrique
☐ Est toujours une matrice carrée
☐ Est toujours constituée de 0 et de 1

Question 12. Un graphe régulier :
☐ A une suite des degrés constante
☐ Est constitué de groupes de nœuds de taille constante
☐ Possède toutes les arêtes possibles

Problème I (partie pratique, 7 points). Écrire vos réponses un fichier RMarkdown, R script ou Jupyter notebook (noyau `R` ou Python) **bien commenté** dans lequel vous donnerez toutes les réponses **explicitement**, en identifiant bien le numéro de la question. Ce fichier est à déposer sur Moodle avant l'heure de fin de l'examen.

On souhaite analyser un jeu de données de relations de type Facebook-like entre des pages de shows télévisés. Vous pouvez bien sûr utiliser `igraph` pour répondre aux questions.

1. Charger le graphe disponible dans le fichier `tvshow_edges.csv` à télécharger sur Moodle (dossier 'Examen - partie pratique'). Le graphe est-il dirigé ? Simple ? Valué ? Connexe ? Donner sa taille et son ordre. Visualiser le graphe.
2. Identifier les nœuds qui occupent une position centrale dans le graphe selon les trois notions de centralité définies en cours.

Dans la suite :
— Si le graphe est simple et non dirigé, gardez le tel quel ;
— sinon, supprimez les directions, les boucles et la multiplicité des arêtes, et travaillez avec ce nouveau graphe (simple et non dirigé).

3. Ajuster un modèle $G(n, p)$ sur ce graphe.
4. Simuler 100 graphes $G(n, p)$ avec les paramètres estimés sur le graphe observé et évaluer :
   — la moyenne (sur les 100 graphes simulés) des degrés moyens des nœuds,
   — le nombre moyen (sur les 100 graphes simulés) d'arêtes par graphe,
   — le nombre moyen (sur les 100 graphes simulés) de triangles.
   Comparer ces valeurs aux valeurs correspondantes observées sur le graphe. **Interprétez les résultats.**
5. Reprendre la question précédente avec un modèle 'fixed-degree' FD$(d_1, \ldots, d_n)$ où $d_1, \ldots, d_n$ est la suite des degrés du graphe observé. **Commentez.**

Problème II (partie théorique, 7 points + bonus). Rédiger vos réponses sur la copie fournie.

On note $G = (V = \{1, \dots, n\}, E)$ un graphe simple, non dirigé, binaire ou à valeurs positives et connexe. Soit A sa matrice d'adjacence et $d_i = \sum_{j=1}^{n} A_{ij}$ le degré (éventuellement valué) de chaque noeud ($1 \leq i \leq n$).

Pour tout entier $K \geq 1$, une partition $V_1, \dots, V_K$ de taille K de l'ensemble V vérifie $V_1 \cup \dots \cup V_K = V$ et $V_k \cap V_\ell = \emptyset$ pour tout $k \neq \ell$. Pour tout graphe et toute partition $V_1, \dots, V_K$, on définit

$$\mathrm{Cut}(A, V_1, \dots, V_K) = \sum_{k=1}^{K} \mathrm{Cut}_k(A, V_k) = \sum_{k=1}^{K} \sum_{i \in V_k} \sum_{j \in V_k^c} A_{ij},$$

où V_k^c est le complémentaire de V_k dans V.

On s'intéresse au problème de minimisation sur l'ensemble des partitions de V de cette quantité Cut pour le graphe G de matrice d'adjacence A.

1. Que compte la quantité $\mathrm{Cut}(A, V_1, \dots, V_K)$ pour une partition $V_1, \dots, V_K$ fixée ? Lorsque K est fixé, qu'est-ce qui caractérise une partition $V_1, \dots, V_K$ de V qui minimise la quantité $\mathrm{Cut}(A, V_1, \dots, V_K)$?
2. Lorsque K varie, donnez un minorant et un majorant de $\mathrm{Cut}(A, V_1, \dots, V_K)$ valable pour tout K. Sont-ils atteints et si oui pour quelles partitions ?

Dans la suite, on suppose toujours que $K \geq 2$ (parce que $K = 1$ n'est pas intéressant).

3. Lorsque $K = 2$ et G est un chemin, que vaut $\min_{V_1, V_2} \mathrm{Cut}(A, V_1, V_2)$?
4. Lorsque $K = 2$ et G est quelconque (mais vérifiant les hypothèses de l'énoncé), donner un minorant strictement positif de $\mathrm{Cut}(A, V_1, V_2)$. Est-ce que cette valeur peut toujours être atteinte ? Vous donnerez soit une preuve soit un contre-exemple.
5. Lorsque G est un graphe complet, calculer $\mathrm{Cut}(A, V_1, \dots, V_K)$ en fonction des caractéristiques de la partition $V_1, \dots, V_K$. Que se passe-t-il si on choisit une partition équilibrée, i.e. telle que $|V_1| = \dots = |V_K| = |V|/K$?

On introduit à présent une version normalisée de la quantité précédente. Pour chaque ensemble V_k de la partition, on définit son volume

$$\mathrm{Vol}(V_k) = \sum_{i \in V_k} d_i$$

et

$$\mathrm{Ncut}(A, V_1, \ldots, V_K) = \sum_{k=1}^{K} \frac{\mathrm{Cut}_k(A, V_k)}{\mathrm{Vol}(V_k)}.$$

On s'intéresse à nouveau au problème de minimisation de cette quantité normalisée. Cette normalisation a pour objectif d'obtenir des partitions non déséquilibrées, c'est-à-dire dont toutes les parties ont des tailles comparables ou similaires.

On note $D = \mathrm{diag}(d_1, \ldots, d_n), L = D - A, I_n$ la matrice identité de taille $n \times n$ et $\mathcal{L} = I_n - D^{-1/2} A D^{-1/2}$ et pour toute partition $V_1, \ldots, V_K$ on introduit $N = (n_{ik})_{1 \leq i \leq n, 1 \leq k \leq K}$ la matrice définie par

$$\forall 1 \leq i \leq n, 1 \leq k \leq K, \quad n_{ik} = \begin{cases} \frac{1}{\sqrt{\mathrm{Vol}(V_k)}} & \text{si } i \in V_k, \\ 0 & \text{sinon.} \end{cases}$$

6. Calculez $\mathrm{Tr}(N^\intercal L N)$. Que constatez-vous ?
7. En déduire que pour $X = D^{1/2} N$, on a

$$\mathrm{Tr}(X^\intercal \mathcal{L} X) = \mathrm{Ncut}(A, V_1, \ldots, V_K) \quad \text{et} \quad X^\intercal X = I_K.$$

On a donc montré que :

$$\min_{(V_1, \ldots, V_K)} \mathrm{Ncut}(A, V_1, \ldots, V_K) = \min_{X = D^{1/2} N} \mathrm{Tr}(X^\intercal \mathcal{L} X) \quad \text{avec } X \in \mathbb{R}^{n \times K}; X^\intercal X = I_K.$$

Comme le problème de droite est difficile à résoudre, une approche classique consiste à considérer le problème dit *relaxé* suivant :

$$\min_{X \in \mathbb{R}^{n \times K}; X^\intercal X = I_K} \mathrm{Tr}(X^\intercal \mathcal{L} X).$$

(Le problème est relaxé car on a enlevé la contrainte $X = D^{1/2} N$). On admet (c'est une conséquence du théorème min-max de Courant-Fischer) que la solution de ce problème est

$$\min_{X \in \mathbb{R}^{n \times K}; X^\intercal X = I_K} \mathrm{Tr}(X^\intercal \mathcal{L} X) = \sum_{k=1}^{K} \lambda_k$$

où $\lambda_i, 1 \leq i \leq n$ sont les valeurs propres de $\mathcal{L}$, classées dans l'ordre croissant ; et le min est atteint pour X la matrice colonne des K plus petits vecteurs propres de $\mathcal{L}$.

8. Question Bonus : Qu'en déduisez-vous ?