TP n°2 : Détection de Communautés dans les Graphes

Le compte-rendu doit être rédigé en LaTeX et être le plus concis possible. Toutes les figures doivent être commentées.

Nous allons étudier dans ce TP les performances de l'algorithme spectral de détection de communautés sur des graphes. Nous nous plaçons pour cela dans le contexte suivant : on définit G un graphe **symétrique** de n nœuds, répartis en K classes $\mathcal{C}_1, \dots, \mathcal{C}_K$ de cardinalités $|\mathcal{C}_i|/n \to c_i > 0$. La matrice d'adjacence de G est donnée par A avec

$$A_{ij} \sim \text{Bernoulli}(q_i q_j C_{ab})$$

lorsque $i \in \mathcal{C}_a$ et $j \in \mathcal{C}_b$ (à symétrie près). Le paramètre $q_i \in (0,1)$ est la probabilité intrinsèque du nœud i à se connecter à tout autre nœud du graphe. Le paramètre C_{ab} est un facteur de pondération définissant le lien entre les classes $\mathcal{C}_a$ et $\mathcal{C}_b$. On prendra pour hypothèse que

$$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}$$

avec $M_{ab} = O(1)$ vis-à-vis de n. Par ailleurs, on supposera que les q_i sont i.i.d. issus d'une loi indépendante des M_{ab}. On définit enfin B, matrice de *modularité* de G, comme étant

$$B = A - qq^*$$

avec $q \in \mathbb{R}^n$ le vecteur des q_i.

Observations préliminaires

Nous allons observer ici plusieurs résultats élémentaires sur A et B.

1. Montrez que, conditionnellement aux q_i, $\frac{1}{\sqrt{n}}A$ est la somme d'une matrice de rang au plus K et d'une matrice aléatoire à entrées indépendantes de moyenne nulle avec un profil de variances.
2. Qu'en est-il de la matrice $\frac{1}{\sqrt{n}}B$?
3. Représentez graphiquement le spectre de $\frac{1}{\sqrt{n}}B$ pour $K = 3$ et :
   - $q_i = q_0$ pour tout i
   - q_i uniforme autour d'une valeur q_0, avec un étalement plus ou moins large
   - $q_i \in \{q^{(1)}, q^{(2)}\}$ pour différentes valeurs de $q^{(1)}, q^{(2)}$
   - Dans les trois cas précédents, prendre différentes valeurs pour M.

   Commentez vos observations, et justifier en particulier la forme du spectre ainsi que le nombre de valeurs propres isolées.
4. Observez et commentez l'allure des vecteurs propres associés aux valeurs propres extrêmes :
   - Conclure sur un algorithme spectral de détection de communautés
   - Que peut-il se passer lorsque les q_i ne sont pas égaux ? Justifiez. What can happen ?

Cas Homogène

[redacted] considère dans cette partie que $q_i = q_0 \in (0,1)$ pour tout i et on prendra M diagonale.

Nous admettrons les résultats suivants. Pour $X_n \in \mathbb{R}^{n \times n}$ une matrice à entrées indépendantes de moyenne nulle, variance (égale ou convergeant uniformément vers) $1/n$ et de moment d'ordre quatre fini et indépendant de n, la mesure spectrale $\frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i(X_n)}$ converge vers la loi du demi-cercle $\mathbb{P}_{sc}$ de densité

$$\mathbb{P}_{sc}(dx) = \frac{1}{2\pi}\sqrt{(4-x^2)^+}dx.$$

Par ailleurs, pour $z \in \mathbb{C}^+$, la transformée de Stieltjes $g_{sc}(z)$ est l'unique solution dans $\mathbb{C}^+$ de l'équation en g

$$g = -\frac{1}{z+g}$$

qui s'étend naturellement sur $\mathbb{R} \setminus [-2,2]$. On sait également que, pour tous vecteurs $u, v \in \mathbb{R}^n$ déterministes, avec probabilité un

$$u^*(X_n - zI_n)^{-1}v - g_{sc}(z)u^*v \to 0.$$

De plus, $g'_{sc}(z) = \frac{g_{sc}^2(z)}{1-g_{sc}^2(z)}$.

1. En utilisant les données précédentes ainsi que les résultats vus en cours concernant les perturbations de petits rangs, déterminez une condition sur les c_i et M_{ii} pour l'existence asymptotique de valeurs propres isolées dans le spectre de $\frac{1}{\sqrt{n}}B$.
   Indice : Pour cette question et la suivante, on pourra utiliser, comme vu en cours, l'approche par résolution de $\det(\frac{1}{\sqrt{n}}B - \lambda I_n)$. Par ailleurs, il sera pratique d'utiliser la matrice $J = [j_1, \ldots, j_k] \in \mathbb{R}^{n \times k}$ avec $j_i \in \mathbb{R}^n$ le vecteur canonique de la classe $\mathcal{C}_i$.
2. Déterminez les valeurs asymptotiques des valeurs propres isolées et vérifiez par simulations.
3. Déterminez les valeurs asymptotiques des alignements entre le vecteur propre isolé et les vecteurs canoniques des classes $\mathcal{C}_1$ et $\mathcal{C}_2$.
   Indice : Pour cela, on pourra évaluer $\frac{1}{2\pi i}\oint_\Gamma \frac{1}{n_a} j_a^*(\frac{1}{\sqrt{n}}B - zI_n)^{-1}j_a dz$ pour Γ un contour complexe bien choisi. On rappelle par ailleurs la formule d'inversion de Woodbury $(A + UV^*)^{-1} = A^{-1} - A^{-1}U(I + VA^{-1}U)^{-1}VA^{-1}$.
4. Confirmez l'adéquation entre théorie et pratique par simulations pour des valeurs grandissantes de n.
5. Proposez un algorithme de détection de communautés basé sur une représentation dans un espace à K dimensions des vecteurs propres. Que nous manque-t-il pour évaluer les performances de cet algorithme ? Proposez une approche.

Cas Hétérogène

Reprenons ici l'hypothèse q_i quelconque.

1. En se référant à la Question 4 de la partie "Observations préliminaires", générez un tracé où, pour un choix judicieux des q_i l'algorithme proposé précédemment ne fonctionne plus.
2. Proposez deux algorithmes : (i) l'un basé sur une renormalisation de la matrice B, (ii) l'autre basé sur une renormalisation des vecteurs propres de B, permettant de corriger le problème. Effectuez des simulations rendant compte du résultat et conclure.