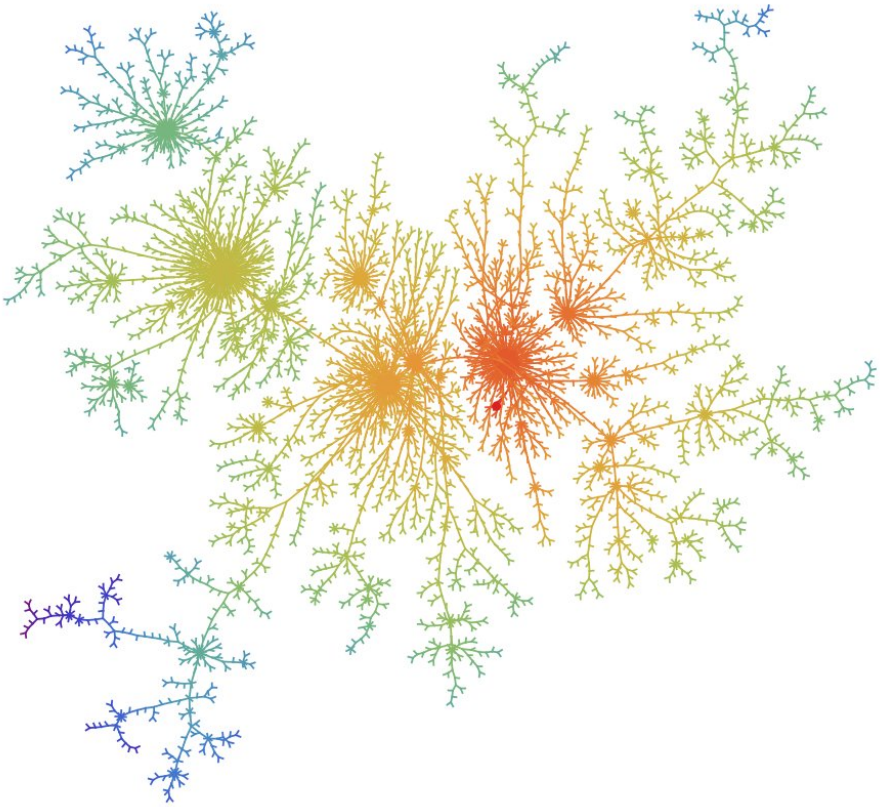
EA MAP572 2020-21

Projet final : Graphes (modèles et algorithmes)



Crédit : Igor Kortchemski (CNRS et École Polytechnique).

Les graphes interviennent de façon essentielle dans de nombreux problèmes de mathématiques appliquées. Ils interviennent en particulier lorsque l'on souhaite modéliser le graphe des pages internet (les connexions étant les liens entre les pages), les maillages ou des études statistiques. Leur étude soulève de nombreuses questions et nous allons en aborder certaines dans ce travail. Les objectifs sont de

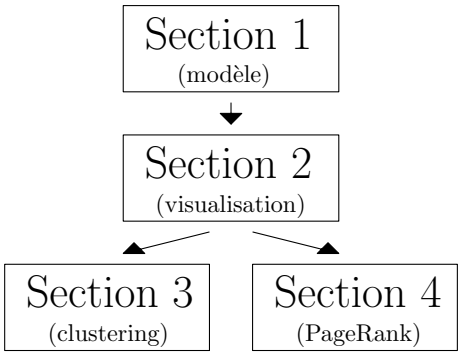
— définir et simuler des modèles de graphes,
— les dessiner/représenter de façon pertinente,
— faire de l'analyse de données et de l'algorithmique sur ces graphes.

À la fin de la dernière séance, vous devrez rendre un rapport individuel de 15 pages environ au format [1] `pdf`. Vous pouvez y inclure tout ce qui vous semble pertinent sur la résolution des problèmes posés : les sorties graphiques et leurs interprétations, des morceaux de votre code `python`, des remarques sur la modélisation, des résultats théoriques et expérimentaux, etc. Les questions ci-dessous sont des suggestions, et vous n'êtes pas obligés de toutes les traiter. Toute prise d'initiative est la bienvenue !

Pour enrichir le rapport vous êtes aussi tout à fait autorisé à chercher sur le web des références supplémentaires sur les problèmes et algorithmes évoqués ici (et à citer vos sources !).

Les différentes parties du projet proposées ci-dessous sont très largement indépendantes, selon la hiérarchie suivante :

1. Vous pouvez travailler sur un notebook mais il faudra nous transmettre un pdf, ce qui nous intéresse est avant tout les sorties graphiques et l'interprétation des résultats.



1 Simulation de graphes aléatoires : graphe par attachement préférentiel

Nous allons commencer par le graphe à *attachement préférentiel*[2], introduit pour modéliser de façon dynamique la création de grands réseaux d'interactions (typiquement, le graphe d'internet : les sommets sont des pages web et une arête entre 2 sommets correspond à un lien).

Le modèle

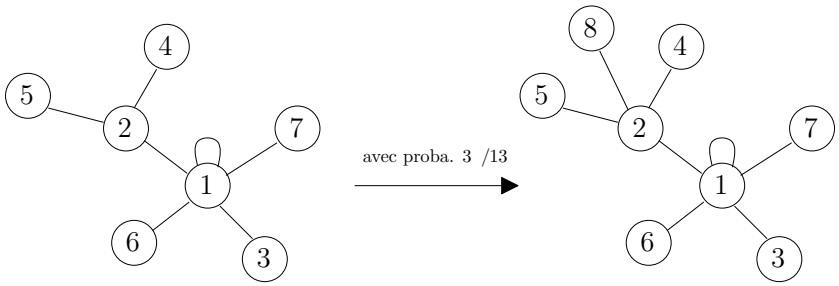
Soit $n \geq 1$ un entier, nous allons construire de façon dynamique un graphe aléatoire[3] G_n à n sommets $\{1, \dots, n\}$ et n arêtes.

1. On part du graphe G_1 qui contient un unique sommet $\{1\}$, et dont l'unique arête est une boucle $1 \to 1$.
2. Connaissant G_k, on rajoute un sommet $k+1$ et une nouvelle arête $k+1 \to v_{k+1}$, où $v_{k+1} \in \{1, 2, \dots, k\}$ est aléatoire, indépendant de G_k, et tiré de la façon suivante :

$$\mathbb{P}(v_{k+1} = i) = \frac{\text{degré}(i)}{\sum_j \text{degré}(j)} = \frac{\text{degré}(i)}{2k-1}, \qquad (\star)$$

où degré(i) est le nombre d'arêtes **distinctes** qui touchent le sommet i.

Voici un exemple avec $k = 7$ (dans ce cas-là $v_8 = 2$) :



Question 1.1. Écrire un script qui simule le graphe G_n. Pour l'instant on représente un graphe par sa *matrice d'adjacence* $n \times n$ à valeurs dans $\{0, 1\}$ définie par `Adjacence(i,j)`$= 1$ si `i` et `j` sont voisins dans le graphe.

Question 1.2. L'une des caractéristiques de ce modèle est que les degrés des sommets suivent approximativement une loi de puissance (*power-law*) : si s est un sommet tiré uniformément dans G_n,

$$\mathbb{P}(\text{ degré de } s = k) \overset{k \text{ grand}}{\approx} ck^{-\alpha},$$

2. Le modèle a été introduit de façon implicite par A.-L. Barabási, R.Albert, "Emergence of scaling in random networks", *Science* (1999).

3. Ce sera en réalité un arbre, et pour l'instant les arêtes sont *non* orientées.

où c, α sont des constantes positives. Cette asymptotique est à comprendre lorsque n est grand, et k petit devant n. Pouvez-vous illustrer ce phénomène (et évaluer le coefficient α) sur une simulation de G_n ?

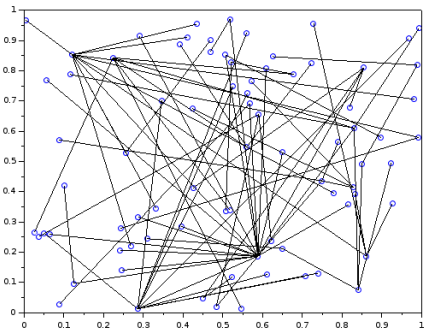
Question 1.3. **[Théorique]** Dans l'article original, les auteurs prétendent que leur modèle reproduit *"a rich-gets-richer phenomenon that can be easily detected in real networks"*. Pour illustrer le fait que G_n est extrêmement non-homogène, estimer l'espérance du degré du sommet 1 dans G_n, et comparer avec le degré moyen d'un sommet.
Indication : On pourra calculer $\mathbb{E}[X_{\ell+1}|X_\ell]$, où X_ℓ est le degré du sommet 1 lorsque le ℓ-ème sommet apparaît dans le graphe.

2 Visualisation d'un graphe : un problème d'optimisation

Pour n grand, on souhaite représenter un graphe G_n plus visuellement qu'avec sa matrice d'adjacence. Ce genre de méthode peut aussi servir à visualiser une matrice creuse qui peut alors être vue comme la matrice d'adjacence d'un graphe. De tels dessins sont visibles sur `http://www.cise.ufl.edu/research/sparse/matrices/list_by_nnz.html`.

Question 2.1. Tracer le graphe de votre choix de la façon suivante :
- Tirer au sort n points uniformes indépendants $(X_1, Y_1), \dots, (X_n, Y_n)$ dans le carré $[0,1]^2$.
- Si i, j sont voisins dans G_n, tracer un segment entre (X_i, Y_i) et (X_j, Y_j).

Pour un graphe à attachement préférentiel on obtient quelque chose d'un peu fouillis :



On cherche à représenter ce graphe de façon plus élégante, de façon à représenter visuellement la topologie du graphe. Soit `Distance` la matrice $n \times n$ dans laquelle le coefficient (i,j) est la longueur du plus court chemin entre i et j dans le graphe. L'idée est que les coordonnées $M_i = (X_i, Y_i)$ et $M_j = (X_j, Y_j)$ des points i et j doivent être d'autant plus éloignées que $\texttt{Distance}(i,j)$ est grand :

$$\frac{1}{\sqrt{2}} \parallel M_i - M_j \parallel \approx \frac{\texttt{Distance}(i,j)}{\max_{i,j} \texttt{Distance}(i,j)}$$

(le terme $\sqrt{2}$ correspond à la plus grande distance euclidienne dans le carré $[0,1]^2$, et est donc à comparer avec $\max_{i,j} \texttt{Distance}(i,j)$). Pour une configuration $M_1, \dots M_n$ donnée, on introduit donc l'*énergie*

$$E = \sum_{i,j} \frac{\left(\frac{1}{\sqrt{2}} \parallel M_i - M_j \parallel - D^\star_{i,j}\right)^2}{(D^\star_{i,j})^2},$$

où

$$D^\star_{i,j} = \frac{\texttt{Distance}(i,j)}{\max_{i,j} \texttt{Distance}(i,j)}.$$

Question 2.2. Tracer le graphe G_n de façon à minimiser le plus possible l'énergie E. On pourra utiliser des méthodes de type gradient. Vous pouvez aussi modifier le choix de E pour obtenir des graphes plus harmonieux. Enfin, on peut se poser la question de représenter le graphe en dimension 3, voire d'utiliser des couleurs.

Une variante : le δ-attachement préférentiel

On peut généraliser le modèle du graphe à attachement préférentiel de la façon suivante. Soit $\delta > -1$ un paramètre fixé, on remplace l'équation $(\star)$ par

$$\mathbb{P}(v_{k+1} = i) = \frac{\text{degré}(i) + \delta}{\sum_j (\text{degré}(j) + \delta)} = \frac{\text{degré}(i) + \delta}{2k - 1 + k\delta}.$$

Pour $\delta = 0$, on retrouve le modèle initial.

Question 2.3. Tracer des graphes G_n pour différentes valeurs de δ. Comment peut-on décrire l'influence de δ sur G_n ?

3 Clustering et détection de communautés : utilisation de la décomposition spectrale

Dans cette partie on cherche à faire le *clustering* (ou segmentation) d'un graphe donné en un certain nombre k de classes, selon la proximité dans le graphe.

Nous allons utiliser l'algorithme *Spectral Clustering* qui est détaillé dans l'article (téléchargeable sur le moodle du cours)

U.Von Luxburg. A tutorial on spectral clustering.
Statistics and computing, vol.17 (2007) n.4, p.395-416.

Question 3.1. Implémenter l'algorithme sur le graphe par attachement préférentiel pour différentes valeurs de k.

On va maintenant chercher à tester *Spectral Clustering* sur un autre modèle de graphe : le *Stochastic Block Model* (SBM). Ce graphe aléatoire est utilisé pour modéliser les communautés dans des réseaux sociaux. Les paramètres du modèle sont les suivants :

n	Nombre de sommets
K	Nombre de classes (ou communautés)
$E_1, \dots, E_K$	Partition de $\{1, 2, \dots, n\}$ en K classes
$(q_{r,s})_{r,s \leq K}$	Probabilités inter-classes

Le modèle est construit de la façon suivante : si i, j appartiennent respectivement aux classes E_r, E_s, alors on met une arête $i \sim j$ avec probabilité $q_{r,s}$, et ce indépendamment pour chaque paire $\{i, j\}$.

Question 3.2. Vous pouvez simuler quelques graphes obtenus selon le *Stochastic Block Model*, les tracer et essayer de retrouver les différentes classes avec l'algorithme *Spectral Clustering*. (Essayer avec $n \approx 200$, et $2 \leq K \leq 5$ classes.)

Question 3.3. Ne pas hésiter à comparer *Spectral Clustering* avec le clustering obtenu en effectuant simplement k-means sur les graphes dessinés dans la partie précédente.

4 Popularité dans un graphe : une approche par valeurs propres (PageRank)

L'algorithme PageRank [4], introduit en 1998 et à la base de Google, attribue à chaque sommet d'un graphe un score censé évaluer sa popularité dans le graphe.

Le principe est le suivant. Soit G un graphe à n sommets, et $A \in \mathcal{M}_{n\times n}(\{0,1\})$ sa matrice d'adjacence. On note $\tilde{A}$ la matrice d'adjacence renormalisée :

$$\tilde{A}_{i,j} = \frac{A_{i,j}}{\sum_j A_{i,j}}.$$

Pour $\varepsilon > 0$, on note P_ε la matrice

$$P_\varepsilon = (1-\varepsilon)\tilde{A} + \frac{\varepsilon}{n}\begin{pmatrix} 1 & \dots & 1 \\ & \ddots & \\ 1 & \dots & 1 \end{pmatrix}.$$

Alors le vecteur des scores X de PageRank est le vecteur propre (à gauche) associé à la valeur propre 1 de la matrice P_ε, renormalisé par $\sum X_i = 1$.

Remarque : *Pour que les résultats de* PageRank *soient plus pertinents, il est préférable d'utiliser un modèle de graphe* orienté. *C'est très facile avec l'attachement préférentiel : on oriente chaque nouvelle arête dans le sens $k+1 \to v_k$.*

Question 4.1. En évaluant les scores PageRank sur quelques graphes de petite taille, discuter du rôle de ε (pour donner une idée, Brin & Page proposaient initialement de prendre $\varepsilon = 0,15$). Discuter de l'implémentation de l'algorithme pour des graphes de grande taille. Quels algorithmes de recherche de valeurs/vecteurs propres peut-on proposer qui soient utilisables dans ce cas (on pourra commencer par étudier la méthode de la puissance) ?

Question 4.2. Implémenter PageRank pour un grand graphe à attachement préférentiel, est-ce que cela semble un bon critère pour la popularité d'un site ?

Question 4.3. On se demande si on peut "tricher" pour augmenter son PageRank. Pour répondre à cette question, estimer ce qu'il se passe pour X si l'on créé artificiellement quelques sommets qui ne pointent que vers un sommet fixé.

4. S.Brin, L.Page. The anatomy of a large-scale hypertextual web search engine. *Proceedings of the Seventh World Wide Web Conference* (1998).