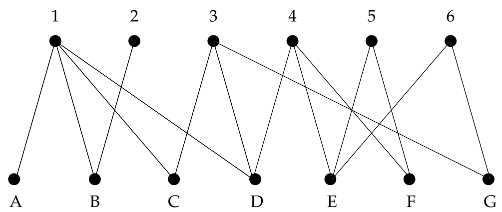
Graphe du Web – modélisation par les graphes bipartis

Une des méthodes vues en cours pour modéliser le graphe du Web s'appuie sur les graphes bipartis. Nous développons ici quelques aspects de cette méthode.

1 Rappels

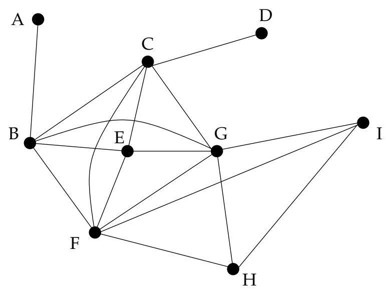
On rappelle qu'un graphe biparti $G = (\bot, \top, E)$ est un graphe composé de deux ensemble de sommets distincts $\bot$,$\top$ et où $E \subset \bot \times \top$. On considère alors la *projection sur* $\bot$ du graphe G, qui est le graphe $G_p = (\bot, E_p)$ avec $E_p = \{(u, v) \in \bot{\times}\bot$ tels qu'il existe un sommet $w \in \top$ avec $(u, w) \in E$ et $(v, w) \in E\}$.

Question 1.1. *Construisez la projection sur $\bot$ du graphe biparti ci-dessous.*



2 Retrouver la structure bipartie

Question 2.1. *Donner un graphe biparti dont la projection donne le graphe ci-dessous, d'abord de façon simple, puis en minimisant le nombre de sommets de $\top$.*



Question 2.2. *Que représente un sommet de $\top$ dans le graphe projeté ? et son degré ?*

Cette technique a d'abord été utilisée pour modéliser des réseaux d'interactions complexes qui présentent initialement un caractère biparti : on cite souvent l'exemple du graphe reliant les acteurs aux films dans lesquels ils jouent, mais aussi les graphes d'occurrence des mots dans les phrases d'un livre, ou encore le graphe des auteurs de publications scientifiques. Partant du graphe projeté, on cherche alors à reconstruire des graphes bipartis qui ont les mêmes caractéristiques que les graphes initiaux, entre autres :

– la distribution des degrés de $\bot$ suit une loi de puissance ;
– la distribution des degrés de $\top$ suit une loi de puissance ou une loi de poisson (en particulier il existe des sommets de degrés élevés) ;

– le nombre de sommets dans $\top$ et dans $\perp$ sont du même ordre de grandeur.

Pour obtenir un ensemble $\top$ relativement petit, on pourrait essayer de minimiser le nombre de clique couvrant le graphe. Malheureusement, c'est un problème NP-complet (minimal clique covering problem). Cependant on peut concevoir des heuristiques efficaces, par exemple :

Question 2.3. *Soit $G = (V, E)$ un graphe, et (u, v) une arête de E. En considérant $N(u, v)$, le sous-graphe induit par les sommets voisins de u et de v, proposez une méthode récursive pour trouver la clique maximale contenant l'arête (u, v). Quels hypothèse doit-on faire sur $N(u, v)$ pour que cet algorithme soit efficace ? Quel graphe est le pire cas pour cette méthode ?*

Pour ce qui est du graphe des pages Web (comme pour le graphe l'Internet d'ailleurs), on cherche à générer des graphes bipartis dont la projection aurait les mêmes caractéristiques que le graphe du Web.

3 Tirage aléatoire de graphe biparti

Question 3.1. *Étant donnée une distribution de degré pour $\top$ et $\perp$, proposez un algorithme permettant de générer de manière uniforme un graphe biparti qui suit ces distributions.*

Question 3.2. *Pour le cas du graphe du Web, on s'intéresse uniquement à la projection sur $\perp$ du graphe biparti. Adaptez votre algorithme pour générer directement ce graphe.*

On s'intéresse aux propriétés des graphes ainsi générés, afin de montrer qu'ils sont proches du graphe du Web.

4 Distribution des degrés dans le graphe projeté

On considère un graphe biparti généré comme à la partie précédente. On note $d(u)$ le degré d'un nœud u de $\perp$ dans le graphe biparti. On s'intéresse au degré des nœuds dans le graphe projeté.

Question 4.1. *Montrer que la probabilité pour que deux nœuds u et v ne soient pas voisins dans le graphe projeté est donnée par :*

$$\frac{\binom{|\top| - d(u)}{d(v)}}{\binom{|\top|}{d(v)}}$$

Question 4.2. *Montrer que*

$$\frac{\binom{|\top| - d(u)}{d(v)}}{\binom{|\top|}{d(v)}} \sim \frac{(|\top| - d(v))^{d(u)}}{|\top|^{d(u)}} \sim 1 - \frac{d(v)d(u)}{|\top|} + O\left(\left(\frac{d(v)d(u)}{|\top|}\right)^2\right)$$

Question 4.3. *En déduire le degré de u dans le graphe projeté (en fonction des degrés des nœuds dans le graphe biparti).*

Question 4.4. *Montrer que si dans le graphe biparti, la distribution des degrés de $\perp$ suit une loi de puissance d'exposant β, alors la distribution des degrés dans la projection sur $\perp$ suit une loi de puissance de même exposant.*

5 Distance moyenne dans le graphe projeté

Pour étudier la distance moyenne dans le graphe projeté, on utilise le résultat suivant de Lu [2] :

Théorème 1. *Soit $G = (V, E)$ un graphe à n sommets dont les sommets sont pondérés par des poids $w_1, \ldots, w_n$, tels que chaque lien $\{i, j\}$ apparaît avec une probabilité $w_i \times w_j \times p$. Si les poids sont tels que les degrés des sommets dans V suivent une loi de puissance d'exposant $\beta > 2$, alors le diamètre de g est en $O(log(n))$ avec forte probabilité.*

Question 5.1. *Montrer que dans si $G = (\top, \perp, E)$ est un graphe biparti dont la distribution de degrés de $\perp$ suit une loi de puissance d'exposant strictement supérieur à 2, alors la distance moyenne dans la projection sur $\perp$ de G est en $O(log(n))$ avec forte probabilité.*

On peut également montrer que le coefficient de clustering possède une borne inférieure indépendante en la taille du graphe, qui garantit que ce coefficient ne tend pas vers 0 quand la taille du graphe augmente.

Sources et références

[1] Jean-Loup Guillaume. *Analyse statistique et modélisation des grands réseaux d'interactions*. PhD thesis, Université Paris 7 - Denis Diderot, 2004.

[2] Linyuan Lu. The diameter of random massive graphs. In *SODA*, pages 912–921, 2001.