MU5IN075 - Networks Analysis and Mining

Sorbonne Université
Master d'Informatique spécialité Réseaux

Examen

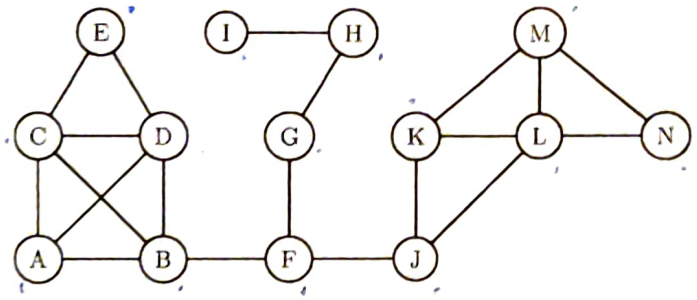
8 Février 2022

Esteban Bautista Ruiz et Lionel Tabourier

L'examen dure 2h. Tous les documents sont autorisés. Le barème est indicatif. Bon courage !

Exercice 1 — ***Mesures sur un petit graphe (6pts)***

Considérez le graphe suivant:



Q1. Tracez la distribution de degré cumulative (DDC) de ce graphe.
Rappel: la DDC d'un graphe représente sur l'axe Y le nombre de nœuds de degré inférieur ou égal à la valeur sur l'axe X.

Q2. Calculez les cœfficients de clustering des nœuds K, F et C.

Q3. Faites un parcours en largeur (*BFS*) depuis le nœud G. Donnez le résultat de ce parcours en largeur sous la forme d'un arbre.

Q4. Déduire de la Q3 la distribution des distances du nœud source G à tous les autres nœuds du graphe.

Q5. Déduire également de la Q3 la centralité de proximité (*closeness centrality*) du nœud G.

Exercice 2 — *Compréhension du cours (5pts)*

Dans cette exercice, on demande de répondre à ces questions **en justifiant votre réponse** par une ou deux phrases simples.

Q1. On rappelle que l'expérience de Milgram du petit-monde consistait à demander à plusieurs personnes situées dans le Nebraska de faire parvenir un dossier à une cible située dans le Michigan en utilisant exclusivement le réseau de connaissances. Dans cette expérience, on constate qu'une grande fraction (25%) des dossiers ayant effectivement atteint la cible ont transité par un homme appelé M.Jacobs.

Supposant qu'on connaisse exactement le réseau des connaissances, quelle mesure réalisée sur ce réseau rendrait compte du rôle particulier de M.Jacobs?

Q2. Nous avons affirmé dans le cours que la distance moyenne entre deux nœuds dans un réseau réel est en général faible et que cette propriété n'est pas étonnante.

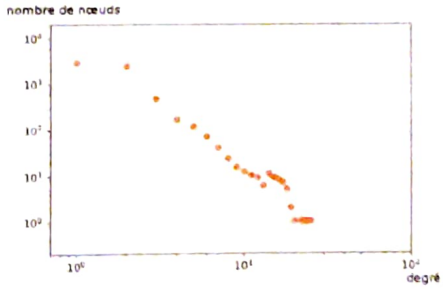
Quel modèle de réseau aléatoire permet de faire cette affirmation et pourquoi?

Q3. Quelle caractéristique de la structure des réseaux réels (qui n'existe pas dans un modèle d'Erdős-Rényi par exemple), explique qu'une diffusion épidémique en voie de s'éteindre peut soudainement redémarrer dans un réseau réel?

Q4. Est-il vrai qu'un système de recommandation basé sur le filtrage collaboratif a tendance à toujours proposer le même type de produits à un utilisateur?

Q5. On a mesuré la distribution de degré d'une sous-partie d'Internet (au niveau IP) en répétant des mesures de traceroute d'une source S vers 3000 destinations différentes. La superposition des chemins obtenus par traceroute est le *réseau observé*. Le réseau réel sous-jacent des IP est simplement nommé *réseau réel*.

On mesure que le *réseau observé* a la distribution de degré suivante:

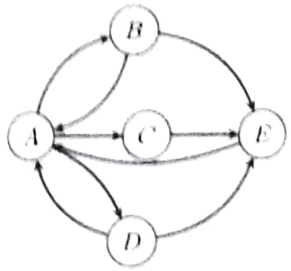


Est-ce que les affirmations suivantes vous semblent justes? Si non, expliquez pourquoi.

- Il y a environ 3000 nœuds de degré 1 dans le *réseau observé*. ✓
- La distribution de degré du *réseau réel* suit une loi de puissance. ×
- Le degré maximum du *réseau réel* est inférieur ou égal à 25. ✗
- Le degré maximum du *réseau réel* est supérieur ou égal à 25. ✓

Exercice 3 — ***Comparaison entre PageRank et HITS (5pts)***

On considère le graphe orienté suivant G_d:



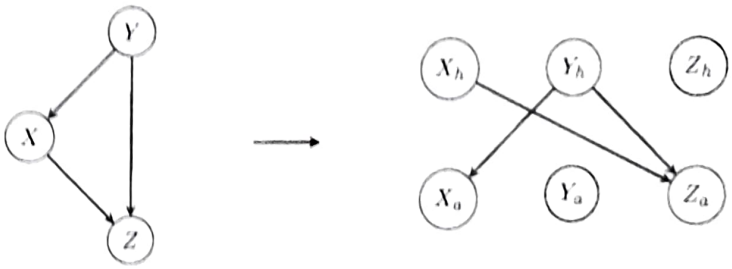
Q1. Calculez l'état stationnaire de l'algorithme de PageRank (développé dans le Cours 11 sur la Recherche d'Information) sur ce graphe.

Q2. Que se passerait-il si l'arête (C, E) n'existait pas? Quelle phase de l'algorithme de PageRank est indispensable pour éviter ce problème ?

Nous proposons maintenant de comparer l'algorithme de PageRank à celui de HITS. Pour cela, nous commençons par transformer le graphe initial en graphe biparti de la manière suivante:

- chaque nœud N est transformé en 2 nœuds N_h et N_a (h pour hub, a pour autorité)
- s'il existe un arc (Y, Z) dans le graphe initial, celui-ci est remplacé par l'arc (Y_h, Z_a).

À titre d'illustration on donne la transformation de graphe suivant:



Q3. Dessinez le graphe transformé du graphe initial G_d.

Q4. Appliquer 2 itérations de l'algorithme HITS sur ce graphe. Pour vous aider, on rappelle ci-dessous l'algorithme HITS.

Q5. Sans itérer l'algorithme davantage, pouvez-vous faire une hypothèse sur le(s) nœud(s) qui obtiendra(ont) le score d'autorité le plus élevé après un grand nombre d'itérations? Sur le(s) nœud(s) qui obtiendra(ont) le score de hub le plus élevé après un grand nombre d'itérations?

Rappel sur l'algorithme HITS:

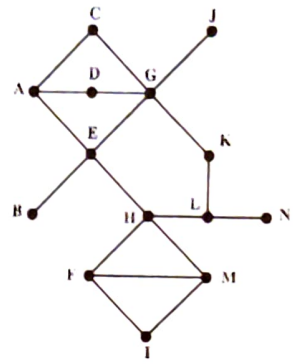
- *initialisation: on initialise les scores de hub de tous les nœuds hub à 1 et tous les scores d'autorité de tous les nœuds autorité à 0*

puis on itère jusqu'à convergence des scores la boucle suivante:

1. *mise à jour des autorités: pour chaque nœud autorité, son score d'autorité est remplacé par la somme des scores de hub des nœuds hubs qui pointent vers lui,*
2. *mise à jour des hubs: pour chaque nœud hub, son score de hub est remplacé par la somme des scores d'autorité des nœuds vers lesquels il pointe,*
3. *normalisation des autorités: chaque score d'autorité est remplacé par lui-même divisé par la somme des scores de toutes les autorités,*
4. *normalisation des hubs: chaque score de hub est remplacé par lui-même divisé par la somme des scores de tous les hubs.*

Exercice 4 — ***Prédiction de liens par la méthode de Borda (8pts)***

Le réseau ci-dessous représente un réseau social l'année A. On cherche à prédire les liens apparaissant l'année $A+1$ connaissant le réseau l'année A.



Comme dans le cours, on note l'ensemble des voisins du nœud i: $\mathcal{N}(i)$.

Pour simplifier les calculs, nous considérons pour la suite que les seuls liens susceptibles d'apparaître l'année $A+1$ sont à choisir parmi les 20 paires de nœuds suivantes:

(A,B), (A,F), (A,G), (A,H), (A,J), (B,F), (C,D), (C,E), (C,J), (D,E),
(E,F), (E,K), (G,H), (G,L), (H,K), (I,H), (J,K), (L,E), (L,M), (M,N).

On veut prédire l'apparition de 5 nouveaux liens suivant la méthode des classements.

Q1. On utilise le nombre de voisins communs $CN(i,j) = |\mathcal{N}(i) \cap \mathcal{N}(j)|$ pour ordonner les paires de nœuds. Donnez les 5 paires de nœuds candidates les plus probablement connectées d'après ce classement.

Q2. On utilise l'attachement préférentiel $PA(i,j) = |\mathcal{N}(i)|.|\mathcal{N}(j)|$ pour ordonner les paires de nœuds. Donnez les 5 paires de nœuds candidates les plus probablement connectées d'après ce classement.

À chaque nœud est aussi associé une caractéristique s représentant le niveau d'étude de la personne (de 1 à 9):

nœud	A	B	C	D	E	F	G	H	I	J	K	L	M	N
s	6	2	2	5	4	2	8	4	1	9	8	8	3	4

Q3. Par ailleurs, on sait que dans ce réseau social, deux individus ayant un niveau d'étude similaire ont de plus grandes chances d'être liés. En utilisant la mesure de similarité suivante pour ordonner les paires de nœuds,

$$sim(i,j) = 9 - |s(j) - s(i)|$$

donnez les 5 paires de nœuds candidates les plus similaires au sens de cette mesure.

On observe expérimentalement que 5 liens sont effectivement apparus l'année $A+1$, ce sont les liens

(A,G), (B,F), (G,L), (J,C), (J,K).

Q4. Donnez le nombre de vrais positifs prédits et le nombre de faux positifs prédits en utilisant la méthode des classements pour chacune des trois scores proposés (CN, PA et sim).

Q5. En déduire la précision et le rappel des prédictions pour chacun de ces trois scores.

On propose maintenant d'améliorer nos prédictions en utilisant la méthode de Borda, dont le principe est décrit ici:

- *Si une paire apparaît première dans un classement, on lui attribue 5 points, si elle apparaît seconde 4 points et ainsi de suite jusqu'à la cinquième place (1 point).*
- *À noter que si plusieurs paires sont en même position, elles reçoivent toutes le même nombre de points. Par exemple si trois paires sont à égalité pour la deuxième place parce qu'elles ont le même score, elles reçoivent toutes 4 points et la peire suivante sera classée cinquième et recevra 1 point.*
- *Une paire qui n'apparaît pas dans les 5 premières ne reçoit pas de point pour ce classement.*
- *On somme les points reçus par chaque paire candidate sur tous les classements, puis on crée un nouveau classement basé sur ce nombre de points.*

Q6. Faites le classement obtenu par la méthode de Borda depuis les 3 classements proposés précédemment.

Q7. Calculez le nombre de vrais positifs et de faux positifs si l'on prédit que les 5 paires les mieux classées dans ce nouveau classement sont effectivement liées.

Q8. En déduire la précision et le rappel pour cette nouvelle prédiction.

Exercice 5 — *Détection de communautés (6pts)*

Rappels sur la notion de conductance

Considérez un graphe non-dirigé $G(V, E)$. Pour une paire de nœuds quelconque $u, v \in V$, on note l'existence d'un lien de la manière suivante:

$$a_{uv} = \begin{cases} 1 & u \text{ est connecté à } v \\ 0 & \text{sinon} \end{cases}$$

Soit $S \subseteq V$ un sous-ensemble de nœuds de G. Une coupe S, notée $cut(S, S^c)$, est le nombre de liens entre l'ensemble S et son complémentaire S^c. Donc si S représente les nœuds d'une communauté, une coupe S est le nombre de liens qui connectent la communauté S aux autres communautés du graphe. Plus précisément, la coupe S est définie par:

$$cut(S, S^c) = \sum_{u \in S} \sum_{v \in S^c} a_{uv}.$$

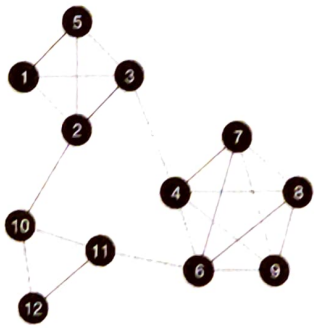
Le volume de S, noté $vol(S)$, est la somme des degrés des nœuds de S. Donc si S représente une communauté, le volume de S donne le nombre de liens qui ont (au moins) une extrémité dans la communauté S. Il est donc défini par:

$$vol(S) = \sum_{u \in S} d(u).$$

La conductance de l'ensemble S, notée h_S, est une métrique qui permet d'évaluer si on ensemble S peut être considéré comme une bonne communauté ou non. Elle est définie de la manière suivante:

$$h_S = \frac{cut(S, S^c)}{min(vol(S), vol(S^c))}.$$

On considère maintenant le graphe suivant G:



Q1. (question ouverte) Expliquez pourquoi h_S peut être utilisé pour dire si S est une bonne communauté ou non. Votre réponse doit aborder les trois points suivants:

(a) la relation entre h_S et la définition des communautés,

(b) l'intervalle de valeurs que h_S peut prendre,

(c) comment on peut utiliser h_S pour évaluer des communautés.

Q2. Pour le graphe G, quelle est la conductance des groupes suivants:

(a) $S = \{1, 2, 3, 5\}$

(b) $S = \{4, 6, 7, 8, 9\}$

(c) $S = \{10, 11, 12\}$

(d) $S = \{7, 8, 9\}$

(e) $S = \{2, 12, 7\}$

Q3. Donnez l'expression du problème d'optimisation basé sur la conductance que l'on doit résoudre pour trouver une partition du graphe entier entre des communautés disjointes.

Q4. Donnez la complexité temporelle du problème de trouver la partition communautaire avec une conductance optimale.

Q5. (question ouverte) La modularité est une autre métrique qui permet de vérifier si un ensemble S est une bonne communauté ou non. Expliquez en quoi la modularité diffère de la conductance. Il n'est pas nécessaire d'écrire des équations mais votre réponse doit aborder les trois points suivants:

(a) en quoi la modularité vérifie que S satisfait le critère d'une communauté,

(b) l'intervalle de valeurs que la modularité peut prendre,

(c) quel est le problème d'optimisation basé sur la modularité.