

## Grands Réseaux d'Interaction

### TP n° 8 : Attachement préférentiel

Remarques importantes :

- Les règles générales en TP restent valides, voir feuille du TP 1. De plus il faut rendre un **rapport** au format PDF exclusivement, à joindre à l'archive (qui elle est au format tar non compressé exclusivement).
- Ce TP est à rendre sur Moodle pour le **30 novembre**

## I) le modèle Barabási–Albert

Le modèle de graphe aléatoire de Barabási–Albert<sup>1</sup> est le plus simple et le plus connu des générateurs aléatoires de graphes suivant l'attachement préférentiel. Il modélise l'adage *rich get richer*<sup>2</sup>. Il prend trois paramètres :  $d$ ,  $n_0$  et  $n$  avec  $d \leq n_0 \leq n$ . Initialement il y a  $n_0$  sommets tous connectés entre eux.  $n$  est le nombre de sommets final<sup>3</sup>. L'algorithme consiste à ajouter itérativement un sommet qui est connecté à  $d$  autres sommets existants.

La probabilité d'être connecté à un sommet dépend de son degré. En effet le sommet nouveau  $x$  se lie au sommet existant  $y$  avec une probabilité

$$\frac{\deg(y)}{\sum_{z \text{ éligible}} \deg(z)}$$

Par exemple si on ajoute le sommet  $x$  alors qu'il y a déjà  $m = 42$  arêtes dans le graphe et que le sommet  $y$  est de degré 7, alors une arête  $xy$  existera avec probabilité  $\frac{7}{84}$  puisque la somme des degrés est alors  $\sum \deg(z) = 2m = 2 \times 42 = 84$ .

Notez que l'on ne doit pas créer d'arêtes parallèles (il ne peut pas y avoir deux arêtes entre  $x$  et  $y$ ) ce qui veut dire qu'une fois que le nouveau sommet  $x$  est lié à  $y$ ,  $y$  devient inéligible pour être lié à  $x$ . Pour avoir une probabilité (quelque chose dont la somme est 1) on ne compte donc que le degré des sommets éligibles (par encore liés à  $x$ ). Ainsi si après avoir créé l'arête  $xy$  il y a un autre sommet  $y'$  de degré 7 aussi,  $x$  lui sera lié avec probabilité  $\frac{7}{77}$  car  $y$  ne compte plus dans  $\sum_{z \text{ éligible}} \deg(z)$ .

Enfin il y a deux variantes de ce modèle :

- une qui génère des graphes **non-orientés**, plus adaptée pour modéliser des situations symétriques comme le fait d'être ami sur Facebook<sup>tm</sup>.
- une qui génère des graphes **orientés**, plus adaptée pour modéliser des citations d'articles. On crée un arc depuis le nouveau sommet  $x$  vers un ancien sommet  $y$ , qui modélise la relation *l'article  $x$  cite l'article  $y$* . Dans ce cas  $\deg(y)$  est le **degré total (entrant plus sortant)** de  $y$ . Initialement on crée tous les  $n_0(n_0 - 1)$  arcs possibles entre les  $n_0$  sommets de départ. Notez que seul le degré entrant d'un sommet peut changer après qu'il ait été ajouté. Le degré sortant est toujours  $n_0 - 1$  pour les sommets du début et  $d$  pour les autres.

### Exercice 1 :

Faire un programme qui génère des graphes selon le modèle de Barabási–Albert de paramètres  $d$ ,  $n_0$  et  $n$ , orienté ou non (selon une option en ligne de commande) et les sauve au format `dot` (il n'y a aucun affichage à faire, sauf en cas d'erreur).

1. parfois aussi appelé à tort Albert-Barabási, car le prénom de Monsieur Barabási est Albert, mais il s'est associé à une Madame Albert pour créer ce modèle, d'où une certaine confusion...

2. cette expression typiquement anglo-saxonne ne peut être traduite en français, où l'on parle plutôt d'*accroissement des inégalités*, ce qui veut dire la même chose mais formulé négativement

3. attention, le paramètre appelé ici  $d$  est noté  $n$  dans le poly de Michel Habib et  $m$  sur Wikipedia

## II) Analyse des graphes Barabási–Albert

La suite du TP consiste à étudier les graphes ainsi générés. Vous devez rendre un rapport (au format PDF exclusivement) qui donne la réponse aux questions posées ainsi que les outils que vous avez utilisés pour y répondre. Ces outils sont normalement vos propres programmes, qui doivent être joints à l'archive. Toutefois pour la distribution des degrés il est conseillé d'utiliser Gnuplot ou un logiciel équivalent. Les quatre exercices sont classés par difficulté croissante. Il n'est pas nécessaire de les faire tous pour avoir une bonne note, car cela demande beaucoup de travail.

### Exercice 2 : k-cœur

Cet exercice concerne la version orientée comme la version non-orientée du modèle. Selon les valeurs de  $k$  et  $d$ , quelle est la taille (en fonction de  $n$ ) d'un  $k$ -cœur d'un graphe de Barabási–Albert ? Donner des statistiques et inférez une loi.

### Exercice 3 : diamètre et distance moyenne

On rappelle que le diamètre d'un graphe est la distance maximale entre deux sommets. Cet exercice consiste à analyser la variante non-orientée du modèle de Barabási–Albert. Vous devez comparer les graphes de Barabási–Albert avec les graphes d'Erdős–Rényi du TP 5 afin de valider, ou d'infirmer (et pour quelles valeurs de  $d$ , de  $n$  et de  $p$  ?) l'assertion communément admise que la moyenne des distances dans un graphe de Barabási–Albert est en

$$O\left(\frac{\log n}{\log \log n}\right)$$

et donc plus petite que celle d'un graphe d'Erdős–Rényi qui est en

$$O(\log n)$$

pour «les bonnes» valeurs de  $n$  et  $p$ . Vous pouvez aussi donner la répartition des distances : combien de sommets sont à distance 1, 2, 3 etc les uns des autres.

### Exercice 4 : distribution des degrés

On trouve aussi l'assertion que les degrés d'un graphe de Barabási–Albert suivent une *power law* de paramètre  $\gamma = -3$ . Est-ce vrai ? Pour quelles valeurs de  $d$  et  $n$  ? Vous pouvez aussi tracer la distribution des degrés avec Gnuplot. Qu'est-ce qui change entre les versions orientée et non-orientée du modèle ?

### Exercice 5 : coefficient de clustering

Cet exercice concerne seulement la version non-orientée du modèle. Le **coefficient de clustering** est la statistique disant si deux sommets ayant un voisin commun sont reliés<sup>4</sup>. Il modélise donc l'adage *les amis de mes amis sont mes amis* : si  $x$  et  $y$  ont un ami  $z$  en commun, seront-ils reliés ? Parmi toutes les paires de sommets d'un graphe  $G$  ayant un voisin en commun, la proportion de ceux qui sont reliés est notée  $clu(G)$ . Cela va de 1 pour un graphe complet à 0 pour un graphe sans triangle (par exemple, un graphe biparti). Formellement on a

$$clu(G) = \frac{3 * tri(G)}{nv(G)}$$

où

- un triangle dans un graphe est un ensemble de trois sommets tous reliés entre eux
- $tri(G)$  est le nombre de triangles de  $G$
- un  $\mathbb{V}$  dans un graphe est formé de deux arêtes incidentes, c'est-à-dire ayant un sommet en commun. Un  $\mathbb{V} xyz$  correspond donc à une arête  $xy$  et une arête  $yz$ , peu importe que l'arête  $yz$  existe ou pas. En particulier, un triangle correspond à trois  $\mathbb{V}$  différents.
- $nv(G)$  est le nombre de  $\mathbb{V}$  de  $G$ . On peut remarquer que

$$nv(G) = \sum_{\substack{x \text{ sommet de } G \\ \text{avec } deg(x) > 0}} \frac{deg(x)(deg(x) - 1)}{2}$$

4. attention il existe une autre définition de ce paramètre, nommée *clustering local moyen*. Celle donnée ici est le *clustering global*.