Revisiter l'Attachement Préférentiel, et ses applications aux Réseaux Sociaux

Guillaume Ducoffe, Frédéric Giroire, Stéphane Pérennes, Thibaud Trolliet

Revisiter l'Attachement Préférentiel, et ses applications aux Réseaux Sociaux

Ducoffe Guillaume[1], Giroire Frédéric[2], Pérennes Stéphane[2], Trolliet Thibaud[3]

[1] *Université de Bucharest & Institut National de Recherche et Développement en Informatique, Roumanie*
[2] *Université Côte d'Azur/CNRS, France*
[3] *INRIA Sophia-Antipolis, France*

Calculer la distribution des degrés de graphes aléatoires, exactement ou même par approximation, est une question de grande importance. Motivés par l'analyse des systèmes complexes, en particulier des réseaux sociaux, nous étudions cette question dans le cas des modèles d'attachement préférentiel. Plus précisément, nous montrons de nouvelles connexions entre le calcul asymptotique de la distribution des degrés dans ces modèles, et la distribution stationnaire de certains processus de Markov associés. En particulier, il nous permet d'analyser des modèles d'attachements préférentiels complexes, tels que ceux avec des distributions de degrés corrélés, et ceux avec suppressions d'arêtes.

Nous appliquons notre approche à un nouveau modèle, créé afin de mieux prendre en compte certaines propriétés importantes du graphe de Twitter. Un autre résultat de l'application de notre méthode est la détermination de la distribution bivariée des degrés entrant et sortant pour le modèle proposé par Bollobas et al. Notre approche est générique, et simplifie et unifie les calculs. Nous nous attendons donc à d'autres applications dans le domaine des graphes aléatoires.

Mots-clefs : Graphes aléatoires, Systèmes complexes, Attachement préférentiel, Distribution des degrés, Chaines de Markov, Réseaux sociaux, Twitter.

1 Introduction

Twitter est l'un des réseaux sociaux les plus utilisés au monde. L'étude d'une capture du graphe de Twitter, obtenue en 2012 par Gabielkov et al. [GRL14], nous a permis de mettre en évidence certaines propriétés encore inconnues de ce graphe, en particulier la grande présence de liens bidirectionnels, et la forte corrélation entre les degrés sortant et bidirectionnel. Ces nouvelles propriétés ne sont, à notre connaissance, reproduites dans aucun modèle de graphe aléatoire. Nous proposons donc un nouveau modèle pour le graphe de Twitter, basé sur le modèle d'attachement préférentiel (noté AP dans la suite), permettant une représentation plus fidèle de celui-ci que ne le sont les modèles dirigés classiques. On définit un modèle d'attachement préférentiel comme un modèle dans lequel un graphe aléatoire évolue tel que les noeuds de celui-ci ont une probabilité proportionnelle à leurs degrés de former un nouveau lien.

Cependant, l'etude des propriétés de ce modèle, en particulier de la distribution des degrés, est complexe. Plus généralement, l'étude rigoureuse de la distribution des degrés n'a jusqu'alors été faite que pour les modèles d'AP les plus simples, les équations des modèles plus complexes devenant difficiles à résoudre. En particulier pour les modèles de graphes dirigés, l'intrication des degrés entrants et sortants complexifient l'étude de leurs distributions. Dans cet article, nous proposons une nouvelle méthode de calcul de la distribution des degrés dans les graphes d'AP, en reliant celle-ci à une chaine de Markov continue. Cette nouvele méthode permet de calculer de façon exacte la distribution des degrés de modèles complexes, en particulier ceux avec retraits d'arêtes au cours du temps, ainsi que les modèles avec des degrés corrélés - en particulier les graphe dirigés. Nous appliquons cette méthode au modèle classique proposé par Barabási et Albert [BA99], ainsi qu'au modèle dirigé de Bollobás et al. [BBCR03], pour lequel la distribution des degrés joints (in,out) n'a jusqu'alors jamais été donnée. Nous l'appliquons ensuite au modèle de Twitter afin de déterminer la distribution jointe des degrés entrant, sortant et bidirectionnel.

Nous commençons par introduire notre nouvelle méthode de calcul de la distribution des degrés en section 2, et l'appliquons à deux cas classiques en section 3. Puis nous présentons l'étude du graphe de Twitter

en section 4.1. Enfin, nous présentons en section 4.2 le modèle de Twitter, ainsi que le calcul de sa distribution des degrés en section 4.3.

2 Méthode de calcul de la distribution des degrés dans les graphes d'attachement préférentiel

Considérons la distribution d'état normalisée $\bar{x_i}(t) = x_i(t)/s(t)$ où $x_i(t)$ est le nombre d'individus dans l'état i à l'instant $t \geq 0$, et $s(t)$ est le nombre total d'individus, que l'on suppose concentré autour de sa moyenne. En pratique, l'état i représentera les degrés, $\bar{x_i}(t)$ la distribution des degrés.

On définit un processus de Markov $(\mathcal{S}, m, \mu)$, où $\mathcal{S}$ est un espace dénombrable, $(m_{i,j})_{i,j\in\mathcal{S},i\neq j}$ une séquence fixe, et $\mu : \mathcal{S} \to (0;1)$ la distribution selon laquelle est choisie l'état initial d'un nouvel individu. Le nombre d'états initiaux possibles est supposé fini, et le nombre d'individus pouvant changer d'état entre deux étapes de temps est supposé borné. Soit Q la matrice de taux de transition de ce processus. D'après les travaux sur les processus de Markov avec retour à l'origine [APZ13], ce proccessus de Markov admet une mesure invariante unique π. Le théorème qui suit énonce que, sous certaines hypothèses techniques (toujours satisfaites pour les modèles d'AP), la distribution $\bar{x_i}$ tend vers la mesure invariante d'un processus de Markov associé, c'est-à-dire que $\lim_{t\to\infty} \bar{x_i}(t) = \pi_i$ pour chaque état i.

Theorem 1. *Si les conditions suivantes sont remplies, il y a convergence en loi de $\bar{x}(t)$ vers l'unique mesure invariante pour Q :*

1. *Les transitions sont locales, c'est-à-dire qu'il existe une séquence de sous-ensembles finis croissants $(B_n)_{n\in\mathbb{N}^*}$ telle que $\bigcup_{n\geq 1} B_n = \mathcal{S}$ et, $\forall n \geq 1, \forall i \in B_n$, on a $m_{i,j} \neq 0 \Longrightarrow j \in B_{n+1} \setminus B_{n-2}$.*
2. *Il existe $n_0 \geq 1$ tel que, pour tout $n \geq n_0$ et $i \in B_n$, $\sum_{j\neq i} m_{i,j} \leq n-1$;*
3. *Pour tout $n \geq 1$ et $i \in B_n \setminus B_{n-1}$, $\sum_{j\notin B_n} m_{j,i} \leq \sum_{j\in B_{n-1}} m_{j,i}$;*
4. *Pour un certain choix de $\lambda > 0$ nous avons, pour tout $i \in \mathcal{S}$: $\sum_{j\neq i} m_{j,i} \leq \left(\sum_{j\neq i} m_{i,j}\right) - 2\lambda$.*

Notons que toutes ces conditions ont une interprétation très intuitive pour les modèles d'AP. La condition 1 résulte du fait qu'à chaque instant t, on ne peut ajouter et/ou supprimer qu'un nombre constant d'arêtes. Un choix naturel pour B_n serait donc l'ensemble des sommets de degré total au plus n. La condition 2 nous dit qu'un individu change son état avec une vitesse légèrement sous-linéaire à son degré. Grossièrement, la condition 3 garantit qu'il est plus probable pour un individu d'augmenter son degré plutôt que de le diminuer. La condition 4 implique que nous quittons tout état i à un taux plus élevé que celui auquel nous entrons dans i — en d'autres termes, nous privilégions les sommets avec de grands degrés.

Nous présentons brièvement ici une idée de la facon dont nous avons prouvé ce théorème. Tout d'abord, nous utilisons des inégalités de concentration classiques afin de réduire l'étude de $\bar{x}(n)$ à sa moyenne, $\mathbb{E}[\bar{x}(n)]$. Ensuite, nous comparons $\mathbb{E}[\bar{x}(n)]$ avec $p_Q(T_n)$, la distribution d'état du processus définie par Q à un certain instant $T_n = \Theta(\log n)$. Pour cela, en utilisant la condition 2, nous commençons par prouver $p_Q(B_n, \log(n+1)) \geq 1 - \frac{2e^{-1}}{n+1}$, c'est-à-dire que presque toute la masse de $p_Q(T_n)$ est contenue dans B_n. Étant donné que le support de $\bar{x}(n)$ est contenu dans B_n, il nous reste à comparer $\mathbb{E}[\bar{x}(n)]$ avec la distribution d'état *limitée à* B_n. Ensuite, nous utilisons une approximation d'Euler de $p_Q(T_n)$ afin de limiter l'écart entre les deux processus. Afin de rendre l'approche continue rigoureuse, nous avons besoin d'une propriété de contraction locale de l'opérateur $Id + \frac{1}{n+1} \cdot Q$ afin de contrôler la croissance de certains termes d'erreur dans nos inéquations. C'est pourquoi nous avons également besoin des conditions 3 et 4. Ce faisant, nous pouvons déduire que la distribution d'état converge exponentiellement vite vers la mesure invariante de Q.

3 Application à quelques modèles classiques

3.1 Modèle de Barabási-Albert [BA99]

Dans le cas du modèle de Barabási et Albert généralisé, la transition est de la forme $m_{i,i+1} \stackrel{def}{=} m_i = ai + b$ (le modèle originel est simplement $m_i = \frac{i}{2}$). Dans ce cas simple, le théorème 1 permet de dériver une formule exacte pour la distribution stationnaire $\pi = (S_i)_{i\in\mathbb{N}}$:

Theorem 2. *Pour $m_i = ai + b, a \in]0, +\infty[$, on a $\forall i \geq i_0$:*

$$S_i = S_{i_0} \cdot \frac{\Gamma\left(i + \frac{b}{a}\right)\Gamma\left(i_0 + 1 + \frac{b+1}{a}\right)}{\Gamma\left(i_0 + \frac{b}{a}\right)\Gamma\left(i + 1 + \frac{b+1}{a}\right)} \underset{i\to\infty}{\sim} \left[S_{i_0} \cdot \frac{\Gamma\left(i_0 + 1 + \frac{b+1}{a}\right)}{\Gamma\left(i_0 + \frac{b}{a}\right)}\right] \cdot i^{-\left(1+\frac{1}{a}\right)}.$$

De plus, le processus de taux m_i et de taux de retour à l'origine 1 à i_0 est stable et admet S_i comme distribution stationnaire.

Notons que $S_i \sim_{i\to\infty} c(a,b) i^{-(1+\frac{1}{a})}$, avec $c(a,b)$ une constante. Dans le cas du modèle proposé dans [BA99] pour lequel $(a,b) = (\frac{1}{2}, 0)$, on retrouve bien la dépendance connue $S_i \sim_{i\to\infty} Ci^{-3}$, où C est une constante.

3.2 Modèle de Bollobás et al. [BBCR03]

En utilisant le théorème 1, nous pouvons aussi calculer la distribution des degrés joints (d_{in}, d_{out}) dans le modèle dirigé proposé par Bollobás et al. [BBCR03]. Dans ce modèle, les degrés entrants et sortants des noeuds sont corrélés, rendant difficile l'analyse de la distribution des degrés joints. Le théorème nous a malgré tout permis de calculer celle-ci :

Theorem 3. *Dans le modèle de Bollobás et al. [BBCR03], la distribution des degrés joints est :*

$$S_{i,j} = \Theta\big(min(i^{-\frac{1+a_0+a_1+b_1}{a_0}} j^{\frac{b_1}{a_1}}, i^{\frac{b_0}{a_0}} j^{-\frac{1+a_0+a_1+b_0}{a_1}})\big)$$

avec $a_0 = \frac{(\alpha+\beta)}{(1-\beta)(1+(1-\beta)\cdot\delta_{in})}$, $b_0 = a_0\delta_{in}$, $a_1 = \frac{(\gamma+\beta)}{(1-\beta)(1+(1-\beta)\cdot\delta_{out})}$, $b_1 = a_1\delta_{out}$, et $(\alpha, \beta, \delta_{in}, \delta_{out})$ les paramètres du modèle définis dans [BBCR03].

4 Étude du graphe de Twitter

4.1 Propriétés

Comme discuté en introduction, cette nouvelle méthode de calcul a initialement été introduite afin de calculer la distribution des degrés corrélés d'un nouveau modèle de Twitter. Celui-ci est basé sur des propriétés mises en évidence lors de l'étude d'une capture du graphe de Twitter récupérée par Gabielkov et al. [GRL14], comportant 505 millions de noeuds et 23 milliards d'arêtes.

Nous avons constaté en particulier que 35 % des arêtes étaient impliquées dans un lien bidirectionnel. De plus, 71% des noeuds possèdent au moins un lien bidirectionnel. Ces liens bidirectionnels ont une signification sociale importante ; un modèle adéquat de Twitter devrait donc prendre en compte ceux-ci.

Nous avons aussi constaté une forte corrélation entre les degrés sortants et les degrés bidirectionnels, avec un coefficient de Pearson de 0.95. A l'inverse, les degrés entrants n'étaient corrélés ni avec les degrés sortants, ni avec les bidirectionnels - les deux coefficients de Pearson ayant une valeur de 0.15.

4.2 Modèle du graphe de Twitter

Le modèle presenté ici étend celui de Bollobás et al., en introduisant trois modifications : (i) certains évènements créent des liens bidirecionnels, (ii) nous choisissons les deux extrémités d'un lien bidirectionnel en fonction des degré sortants de noeuds, (iii) nous ne considérons pas les liens bidirectionnels lorsque l'on s'intéresse au degré entrant d'un noeud.

Le modèle commence par un graphe orienté G_0 qui grandit à chaque étape de temps avec l'ajout d'arêtes simples ou doubles. À chaque étape, on a une probabilité α d'ajouter un sommet en plus d'un lien. Lorsqu'un lien apparaît, il s'agit d'un arc double avec probabilité γ, et d'un arc simple avec probabilité $1-\gamma$.

Soit $t_0 \geq 1$ l'instant initial. On note $G(t)$ le graphe à l'instant $t \geq t_0$, $e(t)$ son nombre (aléatoire) d'arcs et $n(t)$ son nombre (aléatoire) de sommets. Nous définissons $\overline{d}_{in}(v)$ le nombre de liens entrants du noeud v qui ne sont pas impliqués dans un lien bidirectionnel (lorsque $d_{in}(v)$ les inclut).

Soit $(\alpha, \gamma, \delta_{in}, \delta_{out}) \in \mathbb{R}^{+4}$, $\alpha, \gamma \in [0,1]$. Le graphe $G(t)$ évolue en $G(t+1)$ selon les règles suivantes :

— (A) Avec probabilité $\alpha(1-\gamma)$, on ajoute un nouveau sommet v avec un lien allant de v à un sommet existant w, où w est choisi proportionnellement à $\frac{\overline{d}_{in}(w)+\delta_{in}}{e(t)+\delta_{in}n(t)}$;

— (B) Avec probabilité $\alpha\gamma$, on ajoute un nouveau sommet v avec un lien allant de v à un sommet existant w et un lien dans la direction inverse, où w est choisi proportionnellement à $\frac{d_{out}(w)+\delta_{out}}{e(t)+\delta_{out}n(t)}$;

— (C) Avec probabilité $(1-\alpha)(1-\gamma)$, on ajoute un lien d'un sommet existant v à un sommet existant w, où v est choisi proportionnellement à $\frac{d_{out}(v)+\delta_{out}}{e(t)+\delta_{out}n(t)}$ et w proportionnellement à $\frac{\overline{d}_{in}(w)+\delta_{in}}{e(t)+\delta_{in}n(t)}$;

— (D) Avec probabilité $(1-\alpha)\gamma$, on ajoute deux liens entre les sommets existants v et w, où v est choisi proportionnellement à $\frac{d_{out}(v)+\delta_{out}}{e(t)+\delta_{out}n(t)}$ et w proportionnellement à $\frac{d_{out}(w)+\delta_{out}}{e(t)+\delta_{out}n(t)}$.

4.3 Calcul de la distribution des degrés joints du nouveau modèle

Pour plus de lisibilité, on pose $i := \overline{d}_{in}$ le nombre de liens entrants non impliqués dans un lien bidirectionnel, $j := \overline{d}_{out}$, et $k := d_{out} - \overline{d}_{out} = d_{in} - \overline{d}_{in}$ le nombre de liens bidirectionnels associé au noeud.

Dans le modèle présenté, la distribution stationnaire satisfait l'équation de récurrence suivante :

$$\begin{aligned} &c_1(i-1+\delta_{in})(S_{i,j,k}-S_{i-1,j,k})+c_2(j+k-1+\delta_{out})(S_{i,j,k}-S_{i,j-1,k}) \\ &+c_3(j+k-1+\delta_{out})(S_{i,j,k}-S_{i,j,k-1})+(c_1+c_2+c_3)S_{i,j,k} \;=\; 0 \end{aligned}$$

avec $c_1 = \frac{1-\gamma}{1+\gamma+\delta_{in}\alpha}$, $c_2 = \frac{(1-\gamma)(1-\alpha)}{1+\gamma+\delta_{out}\alpha}$, et $c_3 = \frac{\gamma(2-\alpha)}{1+\gamma+\delta_{out}\alpha}$.

On remarque que, en ne considérant que les deux dimensions i et $j+k$, nous nous ramenons au cas bidimensionnel présenté dans la partie 3.2, avec $a_0 = c_1,\ b_0 = c_1 \cdot \delta_{in},\ a_1 = c_2 + c_3,\ b_1 = (c_2+c_3)\cdot\delta_{out}$. Nous connaissons donc la distribution correspondante $\hat{S}_{i,j+k}$. On peut alors exprimer la distribution jointe des degrés $S_{i,j,k}$ en fonction de cette distribution :

$$S_{i,j,k} = \hat{S}_{i,j+k} \cdot \mathbb{P}r[j \mid j+k].$$

Ce dernier terme correctif correspond à la probabilité d'avoir j événements réussis parmi $j+k$ épreuves de Bernouilli, avec une probabilité de succès de $c_2/(c_2+c_3)$. On a donc par conséquent :

$$S_{i,j,k} = \binom{j+k}{j} \cdot \frac{c_2^j c_3^k}{(c_2+c_3)^{j+k}} \cdot \hat{S}_{i,j+k}.$$

5 Conclusion

Dans cet article, nous avons étudié certaines propriétés du graphe des interactions sociales de Twitter, l'un des plus grands graphes dirigés de réseaux sociaux disponibles à ce jour. Nous avons remarqué que celui-ci a un nombre élevé de liens bidirectionnels (environ 35 % des arêtes) et que, si le degré entrant est peu corrélé avec les degrés sortant et bidirectionnel, ces derniers sont fortement corrélés entre eux.

Cela nous a conduit à proposer un nouveau modèle d'attachement préférentiel pour prendre en compte ces propriétés. Cependant, les méthodes classiques d'analyse n'étaient pas suffisantes pour calculer la distribution des degrés de notre modèle. Nous avons donc proposé un nouveau cadre théorique pour calculer les distributions de degrés d'un vaste ensemble de modèles d'attachement préférentiel. L'idée clé est de réduire les calculs à l'analyse de la distribution stationnaire d'un processus de Markov continu. À titre d'exemple, nous avons utilisé ce cadre pour calculer, pour la première fois dans la littérature à notre connaissance, la distribution jointe des degrés du modèle de Bollobás et al. [BBCR03], ainsi que celle de notre modèle avec trois dimensions corrélées. Nous pensons que ce cadre peut être utilisé sur un ensemble encore plus large de modèles d'attachement préférentiel. En particulier, nous pensons qu'il serait possible de pousser l'étude des solutions analytiques à des états de dimensions $\mathbb{N}^k$, avec des calculs similaires à ceux présentés ici.

Références

[APZ13] Konstantin Avrachenkov, Alexey Piunovskiy, and Yi Zhang. Markov processes with restart. *Journal of Applied Probability*, 50(4) :960–968, 2013.

[BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439) :509–512, 1999.

[BBCR03] Béla Bollobás, Christian Borgs, Jennifer Chayes, and Oliver Riordan. Directed scale-free graphs. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 132–139. Society for Industrial and Applied Mathematics, 2003.

[GRL14] Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. Studying social networks at scale : macroscopic anatomy of the twitter social graph. In *ACM SIGMETRICS Performance Evaluation Review*, volume 42, pages 277–288. ACM, 2014.