Les modèle de Barabasi et Albert, et de Price en liaison avec les lois puissance[1]

Barabasi et Albert ont proposé en 1999 un modèle de croissance d'un graphe complexe du type de celui d'Internet, en faisant l'hypothèse d'une apparition permanente de nouveaux nœuds qui sont connectés aux anciens sur la base d'un *attachement préférentiel :* les connexions de ces nouveaux nœuds se font préférentiellement avec ceux qui ont déjà beaucoup de liens. Cependant leur modèle est une variante d'un autre élaboré en 1975 par Price concernant le réseau des citations scientifiques, où les nœuds sont les articles et les liens sont les citations qu'un article fait des autres articles. Le graphe ainsi constitué est orienté : un article i « pointe », en le citant, vers l'article k. Dans le modèle de Barabasi et Albert par contre, le graphe n'est pas orienté. Une autre différence entre les deux modèles est que dans celui de Price un article quelconque fait deux types de citations, vers des articles beaucoup cités (en probabilité, proportionnellement aux citations dont il bénéficie déjà, suivant le principe d'attachement préférentiel) mais aussi de façon aléatoire. Ainsi un nouveau nœud fait c citations en moyenne, dont une partie n est « préférentielle » et une autre partie a, aléatoire. Barabasi et Albert ne considèrent que l'attachement préférentiel. Nous présentons le modèle de Price et donnons à la fin les formules obtenues par Barabasi et Albert.

Le réseau des publications est en croissance permanente. Un nœud j déjà présent (un article déjà publié) est cité k_j fois. Un nouvel article i apparait lorsque la taille du réseau est déjà de N nœuds, et ce nouvel arrivant fait, en moyenne, c citations. La probabilité P_i^j qu'il cite l'article j est proportionnelle à $k_j + a$, où a est une constante : tout article présent a au moins une probabilité uniforme (la même pour tous les articles) d'être cité par la nouvelle publication, à laquelle s'ajoute la probabilité dépendant du nombre des citations qu'il a déjà reçues (attachement préférentiel).. La condition de normalisation sur les probabilités donne[2]

$$P_i^j = \frac{k_j + a}{\sum_l (k_l + a)} = \frac{k_j + a}{N\bar{k} + Na} = \frac{k_j + a}{N(n + a)}$$

Puisque le nœud i doit distribuer ses n citations non aléatoires sur tous les nœuds ($n = N.\bar{k}$). Si l'on s'intéresse à la distribution des citations lorsque le réseau a une taille N, on considère combien de nœuds ont déjà k citations. On définit cette valeur comme $N.p_k(N)$, où $p_k(N)$ est la proportion des nœuds avec k citations. Ainsi, le nombre total de nouvelles citations en direction d'articles en ayant déjà k, est[3] :

$$N \times p_k(N) \times c \times \frac{k + a}{N(n + a)} = \frac{n(k + a)}{n + a} \times p_k(N)$$

Ceci nous permet de calculer l'équation de « transition ». A l'apparition d'un nouvel article, la cohorte de ceux qui possèdent déjà k citations va s'enrichir du nombre de ceux de la cohorte possédant (k-1) citations avant l'arrivée du nouvel article et qui sont cités par ce nouvel arrivé, et va perdre ceux qui avaient k citations et qui sont aussi cités (et passent donc dans la cohorte de ceux qui ont k+1 citations).

La probabilité, pour un article cité k-1 fois d'obtenir une citation nouvelle et de passer dans la cohorte des articles cités k fois est :

$$n \frac{(k - 1 + a)}{n + a} p_{k-1}(N)$$

On a une expression similaire pour ceux cités k fois qui passent à k+1 citations. Du coup la taille de la cohorte à k citations lorsque le réseau a incorporé le nouvel arrivant et a une taille N+1, est $(N+1)p_k(N+1)$, soit :

[1] Nous suivons ici au plus près la présentation de l'excellent livre de Newman (2009) pages 486 et suivantes.
[2] Newman, cité, p 490
[3] Ibid.p 491

$$(N+1)p_k(N+1) = Np_k(N) + \frac{n(k-1+a)}{n+a}p_{k-1}(N) - \frac{n(k+a)}{(n+a)}p_k(N) \qquad [1]$$

Le deuxième terme du membre de droite représente les nouveaux venus de la cohorte (k-1) ayant obtenu une citation, et le troisième ceux de la cohorte k ayant également obtenu 1 citation. Par ailleurs le nouvel article vient enrichir la cohorte des articles non encore cités, donc

$$(N+1)p_0(N+1) = N.p_0(N) + 1 - \frac{n \times a}{n+a}p_0(N) \qquad [1']$$

On s'intéresse à l'évolution des équations [1] et [1'] lorsque le réseau devient très grand, c'est-à-dire lorsque N-> ∞. A la limite $p_k(N)$, $p_k(N+1)$ et $p_k(N-1)$) se confondent en une valeur limite p_k et l'équation [1] se transforme en :

$$p_k = \frac{n}{n+a}\left[(k-1+a)p_{k-1} - (k+a)p_k\right] \quad <=>$$

$$p_k = \frac{k+a-1}{k+a+1+a/n}p_{k-1} \qquad [2]$$

Et l'équation [1'] devient :

$$p_0 = \frac{1+a/n}{a+1+a/n} \qquad [2']$$

L'équation [2] définit une suite itérative et [2'] fournit sa valeur initiale. Un peu de manipulation algébrique permet de trouver la formule donnant p_k en fonction de p_0. On obtient

$$p_q = \frac{(k+a-1)(k+a-2)...a}{(k+a+1+a/n)....(a+2+a/n)} x \frac{1+a/n}{(a+1+a/n)} \qquad [3]$$

La forme de cette relation invite à penser à utiliser la fonction Γ (une généralisation de la factorielle k!) qui possède la propriété suivante

$$\Gamma(x+1) = x\Gamma(x)$$

Ce qui permet d'écrire :

$$\frac{\Gamma(x+n)}{\Gamma(x)} = (x+n-1)(x+n-2)...(x+1)x$$

Et en reportant dans [3]

$$p_k = (1+a/n)\frac{\Gamma(k+a)\Gamma(k+1+k/n)}{\Gamma(k)\Gamma(k+a+2+a/n)}$$

On peut simplifier encore plus en utilisant la fonction d'Euler B définie par

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \qquad [4]$$

Ce qui nous donne :

$$p_k = \frac{B(k+a,2+a/n)}{B(a,1+a/n)} \qquad [5]$$

On s'intéresse aux grandes valeurs de k, c'est-à-dire celles où les papiers sont cités de nombreuses fois car le nombre de papiers publiés est devenu très grand. Si k est grand par rapport à a, on peut utiliser la formule d'approximation de Stirling qui précise que

$$\Gamma(x) \approx \sqrt{2\pi} e^{-x} x^{x-\frac{1}{2}}$$

Quand x -> ∞. Avec cette formule on aboutit alors à l'approximation :

$$B(x,y) \approx x^{-y}\Gamma(y)$$

En reportant cette approximation dans l'expression [5] ci-dessus, cela donne une valeur approchée de p_k :

$$p_k \approx k^{-\alpha}$$

Où

$$\alpha = 2 + \frac{a}{n}$$

Ainsi, la « traîne » de la distribution (ie les grandes valeurs de k, ou les papiers avec beaucoup de citations) suit une ***loi puissance***.

Dans le cas du modèle de Barabasi et Albert où, rappelons-le, a =0 (pas de connexion au hasard) et le graphe n'est pas orienté, on arrive également, en suivant une démarche analogue, à une loi puissance, mais avec une valeur de a nécessairement égale à 3. Ceci représente une limite de ce modèle par rapport à celui de Price, et c'est pour cela que nous avons choisi de présenter celui-ci. Pour démontrer le résultat de Barabasi et Albert, nous donnons une autre méthode très utilisée dans la littérature qui consiste à prendre une approximation continue du mécanisme d'apparition de nouveaux noeuds. Cela revient à supposer que chaque intervalle de temps est infinitésimal et qu'il arrive un seul nœud nouveau durant l'intervalle dt, qui crée n liens avec les nœuds existants[4], en suivant une probabilité qui dépend du nombre de liens dont dispose déjà le nœud en question. Par conséquent la probabilité de recevoir un lien du nouvel arrivant est :

$$p(i\Re j) = n\frac{d_j(i)}{\sum_{k=1}^{i} d_k(i)}$$

Chaque lien permet à deux nœuds (les extrémités du lien) d'augmenter son degré d'une unité, puisque le graphe n'est pas orienté. Par conséquent, puisqu'il y a i*n liens non orientés dans le réseau, on a deux fois plus de degrés :

$$\sum_{k=1}^{i} d_k(i) = 2.i.n$$

Car tout lien entre deux nœuds augmente d'une unité le degré de chacun de ces nœuds. Dans l'approximation continue, la probabilité p(i$\Re$j) est une probabilité instantanée pour le nœud j d'obtenir un lien du nouvel arrivant i. Par ailleurs on utilise également « l'approximation de champ moyen » qui consiste à faire disparaître le caractère aléatoire du phénomène des attachements de liens et à remplacer en quelque sorte les événements aléatoires par leur moyenne. L'idée est que si la loi des grands nombres s'applique, c'est bien ce qui va se passer à terme.

On peut donc écrire que l'accroissement du degré du nœud j entre l'instant i et i+dt (période d'arrivée du noeud i) est donnée par l'équation différentielle suivante (où l'on a reporté la valeur 2.i.n à la place de la somme du dénominateur de l'expression p(i$\Re$j)) :

[4] Pour démarrer le processus on a supposé que m liens étaient étaient arrivés en même temps initialement et se trouvaient tous connectés entre eux par m liens.

$$\frac{d(d_j(i))}{dt} = \frac{d_j(i)}{2.dt}$$

Dans cette équation dt représente l'écoulement du temps et i l'arrivée du noeud supplémentaire à l'instant t=i. Cette équation s'intègre aisément et compte tenu de la condition initiale $d_i(j)=n$, on obtient :

$$d_j(t) = n.\left(\frac{t}{j}\right)^{1/2}$$

A l'instant i, la proportion des nœuds qui ont un degré inférieur ou égal à k, correspond à tous les nœuds arrivés après la date τ, donnée par l'équation :

$$k = n.\left(\frac{i}{\tau}\right)^{1/2}$$

Soit :
$$\tau = i.\left(\frac{n}{k}\right)^{2}$$

En effet, le nœud τ par définition arrivé à la date τ, a exactement à l'instant i (i>τ), un degré égal à k. Il y a donc (i-τ)/i ou 1-τ/i nœuds de degré inférieur à k à l'instant i. Du coup la fonction de répartition des nœuds en fonction de leur degré est :

$$F_\tau(k) = 1 - \left(\frac{n}{k}\right)^{2}$$

Et par conséquent la fonction de densité est :

$$f_\tau(k) = 2.n^2 k^{-3}$$

Il s'agit bien d'une loi puissance, mais dont le coefficient vaut –3. Or il n'y a aucune raison pour que dans les réseaux réels, une loi puissance ait un coefficient égal à une valeur précise. Pour contourner cette limitation, Jackson propose la modélisation suivante : on suppose qu'il n'arrive pas un mais plusieurs nouveaux nœuds dans l'instant infinitésimal entre t et t+dt, mais qui établissent une proportion α de leur lien avec le réseau existant et le reste (1-α) entre eux. Dans ce cas là, la probabilité, pour un nœud τ existant, d'obtenir un lien à l'instant i est $\alpha.n.d_\tau/2n.i$. Du coup, après un calcul similaire à celui effectué ci-dessus, la distribution s'écrit (en posant $\gamma = 2/\alpha$):

$$f_\tau(k) = \gamma.n^\gamma.k^{-\gamma-1}$$

Ainsi le coefficient de la loi puissance n'est pas égal à -3 mais à –γ-1, et peut prendre n'importe quelle valeur entre -1 et -∞.