Graphes aléatoires

Vincent Picard [1]

Résumé

Dans ce court rapport, on présente l'utilisation des probabilités en théorie des graphes. Les graphes aléatoires ont été introduits par Erdös et Rényi en 1959 pour prouver certains résultats combinatoires sur les graphes, tel qu'une borne inférieure sur les nombres de Ramsey par exemple. Les graphes aléatoires peuvent être également utilisés pour évaluer la complexité en moyenne d'algorithmes ou encore pour modéliser de vrais réseaux de la vie de tous les jours (réseaux sociaux, internet, etc). Plusieurs modèles de graphes alétoires existent. Le choix du modèle doit être fait avec soin car les propriétés des graphes peuvent être très différentes selon le modèle choisi. On montre ainsi, dans ce rapport que le modèle de Erdös-Rényi fournit pratiquement toujours des graphes connexes et que la distribution des degrés est nécessairement Poissonienne. Le modèle à distribution des degrés fixés permet d'obtenir des graphes qualitativement différents et leurs propriétés peuvent être calculés grâce aux fonctions génératrices.

Table des matières

1 Introduction **2**

2 Préliminaires mathématiques **3**
2.1 Probabilités 3
2.2 Fonction génératrice 4

3 Présentation rapide des deux modèles **6**
3.1 Modèle de Erdös-Rényi 6
3.2 Modèle à distribution des degrés fixée 6

4 Taille des composantes connexes **7**
4.1 Exploration du voisinage d'un sommet 7
4.2 Quelques calculs de G_0 8
4.2.1 Loi de Poisson 8
4.2.2 Loi exponentielle 8
4.2.3 Loi des degrés choisis 9
4.3 Fonction génératrice de la taille des composantes connexes 9
4.4 Espérance de la taille des composantes connexes 10

1. vincent.picard@ens-cachan.fr

5 Conclusion **10**

1 Introduction

Les graphes aléatoires ont été introduits pour la première en 1959 par Erdös et Rényi [1]. Brièvement, un graphe aléatoire de taille n est un graphe à n sommets dont on a choisi aléatoirement les arêtes.

La méthode probabiliste Les graphes aléatoires ont été utilisés pour la première dois pour démontrer des résultats combinatoires sur les graphes en utilisant ce qu'on appelle maintenant *la méthode probabiliste*. Dans sa forme la plus simple, elle consiste à prouver l'existence d'un graphe possédant une certaine propriété. Pour cela, on considère un ensemble de graphes muni d'une loi de probabilité et dont l'apparition de l'événement souhaité constitue un événement aléatoire. Si on arrive à prouver que la probabilité de cet événement est strictement positive alors on a l'existence de l'objet voulu. La méthode probabiliste permet donc d'obtenir des résultats déterministes en utilisant des méthodes probabilistes. Elle a permi d'obtenir des démonstrations plus simples de certains résultats combinatoires de théorie des graphes.

Évaluation des algorithmes Les graphes aléatoires jouent également un rôle important dans l'évaluation des performances et de la complexité des algorithmes. Certains algorithmes fréquemment utilisés ont une mauvaise complexité dans le pire des cas, mais ont en moyenne une complexité raisonnable. On peut par exemple penser à l'algorithme du simplexe pour la résolution des problèmes de programmation linéaire qui s'éxécute en moyenne en temps polynomial. Toutefois l'utilisation de polytopes particuliers tel que le cube de Goldfarb a permis de donner une borne inférieure exponentielle dans le pire cas. Pour évaluer les peformances d'un algorithme travaillant sur un graphe en pratique, on peut donc mesurer ses performances pour une distribution de graphes aléatoires.

Modélisation En modélisation, les graphes occupent une place importante. Ils servent par exemple à modéliser les réseaux de communications, les liens entre pages internet, les réseaux d'intéractions sociales, les réseaux d'intéractions professionnelles (publications scientifiques), ou encore les réseaux d'intéractions biologiques (réseaux métaboliques, réseaux de régulation génétique, etc). L'utilisation de graphes aléatoires permet donc de produire des graphes de modélisation "typiques" des objets étudiés. Par exemple, Stuart Kauffman a introduit la notion de réseau booléen aléatoire pour modéliser les réseaux de régulations génétiques.

Choix du modèle La question fondamentale qui se pose lorsqu'on veut utiliser des graphes aléatoires pour l'évaluation d'algorithmes ou en modélisation est quelle distribution de probabilité choisir pour notre ensemble de graphes ? La méthode naïve consiste à choisir la distribution de probabilité qui correspond aux données réelles. Cela est difficile en pratique car le nombre de graphes à n sommets donnés est exponentiel en n ce qui rend cette approche impossible (enregistrement de la loi de probabilité, génération des graphes aléatoires). De plus, on ne connait pas forcément la distribution de probabilité voulue. C'est souvent le cas en biologie, certains réseaux de régulation génétique sont mal connus. Enfin, il semble difficile d'obtenir des résultats théoriques si la distribution de probabilité ne possède pas une certaine structure. L'idée est alors d'utiliser un modèle de graphes aléatoires qui va générer une certaine loi de probabilité sur les graphes. Par exemple, le modèle d'Erdös-Rényi consiste à considérer que l'existence de chaque arc est indépendante de celle des autres et que chaque arc a une probabilité p d'exister. On montre que, dans ce cas, les degrés des sommets du graphe suivent une loi de Poisson. Or, il a été montré que de nombreux réseaux de la vie de tous les jours ont une distribution de degrés très éloignée d'une loi de poisson (internet, réseaux biologiques). L'article de Newman *et al.* [2] présente un modèle de graphes aléatoires à distribution de degrés fixés et montre qu'on peut calculer des quantités d'intérêt sur ces graphes.

Ce rapport est avant tout une introduction aux graphes aléatoires qui constitue un sujet trop vaste pour être traité ici dans son intégralité. On se limitera au cas des graphes non-orientés. La section 2 donne des rappels rapides sur la théorie des probabilités et l'utilisation des fonctions génératrices.

2 Préliminaires mathématiques

2.1 Probabilités

Dans cette partie, on rappelle les définitions et les propriétés de la théorie des probabilités utiles pour la compréhension du rapport. On se limitera aux ensembles probabilisés discrets qui sont suffisants à cette étude.

Espace probabilisé et variables aléatoires

Définition 1. *Un* espace probabilisé *discret est un ensembe* Ω *au plus dénombrable muni d'une application* $\mathbb{P} : \mathfrak{P}(\Omega) \to [0,1]$ *appelée* mesure de probabilité *vérifiant les axiomes*

1. $\mathbb{P}[\Omega] = 1$;
2. $\mathbb{P}\left[\bigcup_{n=0}^{+\infty} A_n\right] = \sum_{n=0}^{+\infty} A_n$ *pour toute famille dénombrable de parties* $(A_n)_{n\geq 0}$ *disjointes deux à deux.*

Il est facile de vérifier que dans un espace discret, il suffit de se donner la probabilité des singletons $p_x = \mathbb{P}[\{x\}]$ pour définir la mesure de probabilité. Un *événement* est une partie de Ω. Une *variable aléatoire* est une application $X : \Omega \to \mathbb{R}$. On notera $X = a$ l'événement X prend la valeur $a \in \mathbb{R}$ et $\mathbb{P}[X = a] = \mathbb{P}\left[X^{-1}(\{a\})\right]$ sa probabilité. Deux événements A et B sont indépendants si $\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]$. Deux variables aléatoires sont indépendantes si les évenements relatifs à chacune d'entre elles sont indépendants. La donnée de la famille $(p_X(k) = \mathbb{P}[X = k])_{k\in\mathbb{N}}$ consitue la *loi de probabilité* ou *distribution* de la variable aléatoire X à valeurs discrètes. Enfin, on se servira de la proposition suivante qui permet de prouver l'existence d'un objet satisfaisant une propriété voulue.

Proposition 1. *Soit X une variable aléatoire sur un espace probabilisé discrèt Ω. Si $\mathbb{P}[X = a] > 0$ alors il existe $\omega \in \Omega$ tel que $X(\omega) = a$.*

Éspérance mathématique L'espérance est la valeur moyenne que prend une variable aléatoire.

Définition 2. *L'*espérance *d'une variable aléatoire X prenant ses valeurs dans $\mathbb{N}$ est définie par*

$$\mathbb{E}[X] = \sum_{k=0}^{+\infty} k\mathbb{P}[X = k]. \tag{1}$$

Proposition 2. *Pour toute variable aléatoire discrète d'espérance finie, il existe $\omega \in \Omega$ tel que $X(\omega) \geq \mathbb{E}[X]$.*

Démonstration. Par l'absurde, on aurait $\forall\omega \in \Omega, X(\omega) < \mathbb{E}[X]$ et donc $\mathbb{E}[X] = \sum_{k=0}^{\mathbb{E}[X]-1} x\mathbb{P}[X = k] < \mathbb{E}[X]\sum_{k=0}^{\mathbb{E}[X]-1}\mathbb{P}[X = k] \leq \mathbb{E}[X]$ et donc $\mathbb{E}[X] < \mathbb{E}[X]$. □

2.2 Fonction génératrice

À toute suite de réels $(a_n)_{n\in\mathbb{N}}$ on peut associer une unique série formelle $\sum_{n=0}^{+\infty} a_n X^n$ qui caractérise cette suite. On peut ainsi considérer la série associée aux $(p_X(k))_k$ d'une variable aléatoire X à valeurs discrètes.

Définition 3. *La* fonction génératrice *d'une variable aléatoire discrète est la série entière*

$$G_X(x) = \sum_{k=0}^{+\infty} \mathbb{P}[X = k]\, x^k \tag{2}$$

qui converge absolument pour tout $x \in [-1, 1]$.

Dérivées de la fonction génératrice La fonction génératrice d'une variable aléatoire discrète caractérise entièrement la loi de probabilité de cette dernière grâce à la proposition suivante.

Proposition 3. *Si $G_X^{(n)}$ désigne la n-ième dérivée de G_X, on a*

$$k!\, p_k = G_X^{(k)}(0). \tag{3}$$

Moments d'ordre n La fonction génératrice permet également de calculer rapidement les moments d'ordre n en utilisant la proposition suivante.

Proposition 4. *Le moment d'ordre n d'une variable aléatoire $\mu_n = \mathbb{E}[X^n]$ est donné par*

$$\mu_3 = \mathbb{E}[X^n] = \left[\left(x\frac{\mathrm{d}}{\mathrm{d}x}\right)^{(n)} G_X(x)\right]_{x=1}, \tag{4}$$

où $\left(x\frac{\mathrm{d}}{\mathrm{d}x}\right)^{(n)}$ correspond à l'opération dériver par rapport à x et multiplier par x que l'on itère n fois.

On a en particulier $\mu_1 = \mathbb{E}[X] = G'_X(1)$.

Somme de variables aléatoires indépendantes La proposition donne la fonction génératrice de la somme de deux variables aléatoires indépendantes.

Proposition 5. *Si X et Y sont deux variables aléatoires discrètes indépendantes alors*

$$G_{X+Y} = G_X G_Y. \tag{5}$$

Démonstration.

$$\begin{aligned}
G_{X+Y}(z) &= \sum_{k=0}^{+\infty} \mathbb{P}[X+Y=k]\, z^k \\
&= \sum_{k=0}^{+\infty} \sum_{x,y} \mathbb{P}[\{x+y=k\} \cap \{X=x\} \cap \{Y=y\}]\, z^k \\
&= \sum_{k=0}^{+\infty} \sum_{x,y} \mathbb{P}[x+y=k]\, \mathbb{P}[X=x]\, \mathbb{P}[Y=y]\, z^k \text{ par indépendance} \\
&= \sum_{k=0}^{+\infty} \sum_{x+y=k} \mathbb{P}[X=x]\, z^x \mathbb{P}[Y=y]\, z^y \\
&= \left(\sum_{x=0}^{+\infty} \mathbb{P}[X=x]\, z^x\right)\left(\sum_{y=0}^{+\infty} \mathbb{P}[Y=y]\, z^y\right) \text{ par produit de Cauchy} \\
&= G_X(z) G_Y(z).
\end{aligned}$$

□

3 Présentation rapide des deux modèles

Dans cette section, on présente rapidement les modèles de Erdös-Rényi et à distribution de degrés fixés.

3.1 Modèle de Erdös-Rényi

Le premier modèle de graphes aléatoires a avoir été utilisé est celui de Erdös et Rényi en 1959 [1]. Il est l'un des modèles les plus simples qu'on puisse imaginer. On considère N sommets, et on relie chaque couple de sommets distincts par un arc avec une probabilité p indépendamment les uns des autres.

On voit immédiatement qu'un tel modèle est trop simple pour pouvoir modéliser des réseaux de la vie réelle. Par exemple si on modélise la relation d'amitié par un graphe, on imagine facilement que les relations d'amitié ne sont pas indépentes les une des autres. Les réseaux de régulation génétique sont un autre exemple. Uri Alon [3] a montré l'existence de petites structures topologiques appelée motifs (par exemple des triangles de régulation) qui apparaissent plus fréquemment dans les réseaux biologiques que dans des réseaux aléatoires. A fortiori, on se rend compte que ce modèle est trop pauvre pour modéliser des structures topologiques de cette espèce.

L'avantage du modèle de Erdös-Rényi est qu'il est relativement simple à comprendre et à étudier. Par exemple on peut facilement s'intéresser à la loi de probabilité des degrés des sommets. Un sommet est de degré k s'il est relié à exactement k autres des $N-1$ sommets du graphe. Cela arrive donc avec une probabilité

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} \underset{N\to+\infty}{\sim} \frac{z^k e^{-z}}{k!}, \tag{6}$$

où $z = p(N-1) \sim_{N\to+\infty} pN$ est le degré moyen de chaque sommet. On a ainsi la proposition suivante

Proposition 6. *Dans la limite $N \to +\infty$, la distribution des degrés suit une loi de Poisson de paramètre $z = pN$.*

Cette proposition montre que ce modèle est trop simple pour modéliser beaucoup de réseaux de la vie réelle, puisqu'on a montré que beaucoup d'entre eux ont des distributions des degrés très différentes d'une loi de Poisson.

3.2 Modèle à distribution des degrés fixée

Le modèle de graphes aléatoires à distribution des degrés fixée est présenté dans l'article de Newman. Il s'agit de se donner une loi de probabilité pour les degrés et de considérer que les graphes sont par ailleurs totalement aléatoires.

Formellement, cela signifie qu'on s'intéresse à un espace probabilisé de graphes $\mathcal{G}$ de taille N vérifiant

$$\forall G = (V, E) \in \mathcal{G}, \forall x \in V, \mathbb{P}\left[\deg(x) = k\right] = p_k, \tag{7}$$

où $(p_k)_{k\in\mathbb{N}}$ est donné.

On peut construire un tel ensemble en se donnant une suite de degrés pour chaque sommet et en considérant une loi uniforme sur tous les graphes possédant exactement les degrés voulus à chaque sommet. Les dégrés étant choisis pour approximer au mieux la distribution des degrés voulue.

Le second modèle a l'avantage de permettre de modéliser des situations plus réalistes. Par exemple, une problématique de recherche est la métrologie du réseau internet où l'on a obtenu des résultats montrant que les degrés suivent des lois de puissance (scale-free networks).

On remarquera que le modèle de Erdös correspond au modèle à loi de probabilité des degré fixée lorsqu'on choisit une loi de Poisson de paramètre $z = pN$.

4 Taille des composantes connexes

Dans cette section nous montrons comment les fonctions génératrices permettent de calculer des quantités d'intérêt sur les graphes aléatoires. On se place dans le modèle plus général de distribution des degrés fixée, et on note p_k la probabilité qu'un sommet soit de degré k. Enfin, on note G_0 la série génératrice associée aux p_k.

4.1 Exploration du voisinage d'un sommet

Si on choisit un arc aléatoirement dans le graphe et qu'on le suit jusqu'à un soommet, ce sommet a un certain nombre de voisins. On peut déterminer la fonction génératrice correspondant à la distribution du nombre de voisins. Un arc tiré aléatoirement arrive sur un certain sommet avec une probabilité proportionnelle au degré du sommet et cette probabilité et elle-même proportionnelle à kp_k. Si on normalise la distribution pour que sa somme vaille 1, on obtient la série génératrice souhaitée

$$\frac{\sum_k kp_k x^k}{\sum_k kp_k} = x\frac{G_0'(x)}{G_0'(1)}. \tag{8}$$

On considère maintenant un sommet choisi au hasard. Si l'on suit ses arcs sortants pour explorer le voisinage immédiat, la distribution des degrés de chaque

voisin immédiat est donnée à l'équation précédente. On va toute fois la diviser par x pour ne pas compter le sommet depuis lequel on est arrivé, ce qui donne

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)} = \frac{G_0'(x)}{z}, \tag{9}$$

où z est l'espérance de la distribution des degrés.

La probabilité qu'un des arcs sortants du voisinage immédiat revienne sur le sommet initial est en $\frac{1}{N}$, on la négligera donc pour des N grands, de telle sorte que le graphe soit localement un arbre.

Puisque $G_1(x)^k$ est la fonction génératrice de la somme du nombre de noeuds sortant de k voisins, on obtient la série génératrice du nombre de 2-voisins (voisins à distance 2) par

$$\sum_k p_k G_1(x)^k = G_0(G_1(x)). \tag{10}$$

En itérant le même principe on peut obtenir le nombre de 3-voisins par $G_0(G_1(G_1(x)))$.

Ainsi on peut calculer l'espérance du nombre de voisins à distance 2 par

$$z_2 = \left[\frac{\mathrm{d}}{\mathrm{d}x} G_0(G_1(x))\right]_{x=1} = G_0'(1)G_1'(1) = G_0''(1). \tag{11}$$

4.2 Quelques calculs de G_0

On donne ici la forme de G_0 pour quelques loi de probabilités des degrés simples.

4.2.1 Loi de Poisson

On a montré dans la présentation du modèle de Erdös-Rényi que $p_k \sim \binom{N}{k} p^k (1-p)^{N-k} x^k$ ainsi,

$$G_0(x) = \sum_{k=0}^{+\infty} \binom{N}{k} p^k (1-p)^{N-k} x^k \tag{12}$$

$$= (1-p+px)^N \underset{N\to+\infty}{=} e^{z(x-1)}. \tag{13}$$

On retrouve alors facilement que l'espérance du degré est z et en développant en série entière le dernier terme de l'égalité, on retrouve la loi de poisson $p_k = \frac{z^k e^{-z}}{k!}$.

4.2.2 Loi exponentielle

Une distribution de degrés exponentielle a pour forme

$$p_k = (1 - e^{-\frac{1}{\kappa}}) e^{-\frac{k}{\kappa}}. \tag{14}$$

On obtient alors

$$G_0(x) = (1 - e^{-\frac{1}{\kappa}}) \sum_k e^{-\frac{k}{\kappa}} x^k = \frac{1 - e^{-\frac{1}{\kappa}}}{1 - xe^{-\frac{1}{\kappa}}}. \tag{15}$$

4.2.3 Loi des degrés choisis

Dans le cadre où l'on souhaite modéliser des données réelles, et qu'on possède un échantillon statistiquement significatif, on peut se fixer les p_k en fonction de nos connaissances. Par exemple, si les données contiennent respectivement $100, 200, 300, 400$ sommets de degrés $0, 1, 2, 3$, on peut poser

$$G_0(x) = \frac{100 + 200x + 300x^2 + 400x^3}{1000}. \tag{16}$$

4.3 Fonction génératrice de la taille des composantes connexes

Dans toute cette section, on considère que le graphe ne possède pas de composante géante, c'est-à-dire, qu'il y a suffisamment peu d'arcs pour considérer que le graphe est formée de plusieurs petites composantes connexes ("état liquide"). De même comme précédemment on considère qu'il n'y a pas de cycle puisque leur présence disparait en $\frac{1}{N}$.

Comme on l'a fait précédemment avec le voisinage, on commence par calculer H_1 la fonction génératrice de la taille de la composante connexe qu'on obtient après avoir suivi un arc choisi au hasard. Losrqu'on suit cet arc on obtient un sommet qui peut avoir k nouveaux arcs sortants, menant à de nouvelles composantes connnexes suivant la même distribution. Si on considère tous les k possibles on a donc

$$H_1(x) = xq_0 + xq_1 H_1(x) + xq_2 H_1(x)^2 + \cdots \tag{17}$$

La multiplication par x correspond à la presence du premier sommet qu'on rencontre. Maintenant on remarque que les q_k correspondent aux coefficients de G_1 ce qui permet d'écrire

$$H_1(x) = xG_1(H_1(x)). \tag{18}$$

Toujours dans la même idée de démarche que pour le calcul des voisins, on calcule la fonction génératrice de la taille des composants en considèrant maintenant qu'on a un sommet de départ desquels partent des arcs pour lesquels la distribution de la taille des composantes est connue grâce à H_1, ce qui donne

$$H_0(x) = xG_0(H_1(x)). \tag{19}$$

On peut donc obtenir une solution de H_1 en résolvant l'équation implicite puis obtenir H_0 en utilisant la dernière équation. Toutefois, cela s'avère compliqué en théorie de résoudre l'équation implicite.

4.4 Espérance de la taille des composantes connexes

Même s'il est difficile d'obtenir une expression pour H_0, on peut toutefois estimer la taille moyenne des composantes. Si on note C l'espérance de la taille des composantes connexes, on a

$$C = H_0'(1) = 1 + G_0'(1)H_1'(1), \tag{20}$$

l'équation implicite sur H_1 donne

$$H'(1) = 1 + G_1'(1)H_1'(1), \tag{21}$$

ce qui permet d'obtenir

$$C = 1 + \frac{G_0'(1)}{1 - G'(1)} = 1 + \frac{z_1^2}{z_1 - z_2} \tag{22}$$

où $z_1 = z$ est le espérance des degré du graphe et z_2 le nombre moyen de 2-voisins.

On remarque que cette expression diverge quand $G_1'(1) = 1$ ce qui montre une transition de phase : on obtient une seule composante géante au moment de la transition de phase.

Dans le cas du modèle de Erdös-Rényi, on a

$$G_1'(x) = \frac{G_0''(x)}{z} = \frac{(e^{z(x-1)})''}{z} = ze^{z(x-1)}, \tag{23}$$

et donc la transition de phase a lieu quand $z = 1$ c'est-à-dire quand $p = \frac{1}{N}$. On obtient donc le théorème suivant

Théorème 1. *Pour une probabilité p fixée, le modèle de Erdös-Rényi fournit presque surement un graphe connexe, c'est-à-dire, que la probabilité que le graphe soit connexe tend vers 1 quand $N \to +\infty$.*

Dans le cas général, on peut avoir selon les paramètres de la loi, des phases solides et liquides pour le graphe même quand N tend vers l'infini. L'article de Newman, détaille par exemple le cas de la loi exponentielle quand $\kappa \to +\infty$. Ainsi, selon les lois de probabilités choisies, les graphes aléatoires résultants peuvent être totalement différents. Les calculs utilisant les fonctions génératrices permettent de déterminer les caractéristiques des graphes aléatoires.

5 Conclusion

Dans ce rapport, on a présenté le modèle de graphes alétoires de Erdös-Rényi et les modèles à loi de probabilité des degrés fixée. Le premier modèle peut être vu comme un cas particulier du second où la loi des degrés choisie est une loi de Poisson. En plus des résultats combinatoires qu'ils permettent d'obtenir,

les graphes aléatoires peuvent servir à évaluer la complexité en moyenne d'algorithmes ou à modéliser des réseaux de la vie réelle. Il faudra toutefois être prudent dans le choix du modèle. On a montré que le modèle de Erdös-Rényi fournit nécessairement une distribution des degrés Poissonienne, ce qui est très éloigné de nombreux réseaux réels. De plus, on a vu qu'un graphe aléatoire du modèle de Erdös-Renyi est presque surement connexe, ce qui est une propriété assez forte alors que l'idée de départ est d'avoir des "graphes au hasard". Enfin, on a vu que l'utilisation des fonctions génératrices dans le cas du second modèle permette d'obtenir des quantités d'intérêt sur les graphes étudiés.

Références

[1] Stuart A. West. *Introduction to Graph Theory.* Prentice Hall, USA, 2nd edition, 2001.

[2] M. E. J. Newman, Strogatz S. H., and Watts D. J. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 2001.

[3] Uri Alon. Network motifs : theory and experimental approaches. *Nature Reviews Genetics*, 8 :450–461, 2007.