Introduction au domaine de recherche: Les graphes aléatoires

Table des matières

1 Modèle d'Erdős-Rényi **1**

2 Erdős-Rényi Mixture graph **2**
2.1 Construction du modèle, et notations utilisées 2
2.2 Premiers résultats : le cas simple 4
2.2.1 Présence de motifs 4
2.2.2 Taille de la plus grande composante connexe 5
2.3 Le cas général 6
2.3.1 Recherche de motif 6
2.3.2 Plus grande composante connexe 7

1 Modèle d'Erdős-Rényi

Ce modèle a été introduit en 1960 dans un article de P. Erdős et A. Rényi [3]. Il s'agit d'un modèle de graphe n'ayant pas de propriétés géométriques particulières. Plus précisément, Erdős et Rényi considèrent le graphe $G(n, N)$ défini comme le graphe aléatoire choisi uniformément parmi les graphes à n sommets et N arêtes. On notera $\mathbb{P}_{n,N}$ la loi de probabilité associée. Un modèle proche, quoique légèrement différent, consiste à considérer le graphe $G(n, p)$ défini comme le graphe aléatoire à n sommets et dont chacune des $\binom{n}{2}$ arêtes possibles est présente indépendamment des autres avec probabilité p. On notera $\mathbb{P}_{n,p}$ la loi de probabilité associée. Les deux modèles sont proches, pour n grand (qui est le cas qui nous intéresse), notamment d'après la loi des grands nombres.
Les problèmes étudiés sur ce modèle sont de la forme suivante. Si l'on a une propriété Q possible du graphe, à partir de quel $p = p(n)$ (ou de quel $N = N(n)$) est-elle vraie pour $G(n, p)$ (ou $G(n, N)$), pour $n \to \infty$?
Pour répondre à ces questions avec précisions, nous avons besoin de quelques définitions :

Définition 1. *Soit Q une propriété. On dira que Q est croissante si, à n fixé, elle est croissante pour l'ensemble des arêtes.*

Cela signifie que si Q est vraie pour G, alors Q est vraie pour G auquel on a ajouté une arête.

Définition 2. *Soit Q une propriété possible de G.*
- *On dit que Q est asymptotiquement presque sûrement (noté a.p.s.) vraie si* $\lim_{n\to\infty} P_{n,p(n)}(Q) = 1$
- *On dit que Q est asymptotiquement presque sûrement (noté a.p.s.) fausse si* $\lim_{n\to\infty} P_{n,p(n)}(Q) = 0$

Définition 3. *Soit Q une propriété croissante. On dit que $f : \mathbb{N} \to \mathbb{R}_+$ est une fonction seuil pour la propriété Q si l'on a :*
- $\lim_{n\to\infty} P_{n,p(n)}(Q) = 0$ *si* $p(n) = o(f(n))$
- $\lim_{n\to\infty} P_{n,p(n)}(Q) = 1$ *si* $f(n) = o(p(n))$

Remarque. *Une fonction seuil n'est définie qu'à une constante près : si f est une fonction seuil, λf est une fonction seuil, pour $\lambda \in \mathbb{R}_+^*$*

Erdős et Rényi ont montré l'existence de telles fonctions seuil pour un certain nombre de propriétés, comme la présence d'un motif donné pour une certaine classe de motifs (les graphes équilibrés) (résultat généralisé à tous les motifs par Bollobás en 1981[1]), la présence d'une composante connexe de taille βn avec $0 < \beta < 1$, la connexité. D'une manière générale, il existe également un grand nombre de résultats sur G, notamment sur son apparence, sur la présence de cycle, sur la taille de la plus grand composante connexe, etc. Néanmoins, ce modèle souffre d'un défaut, qui limite ses applications à des réseaux concrets : tous les sommets sont interchangeables, et ont la même connectivité.Récemment, des applications, notamment à la biologie ont motivé le besoin d'étudier un modèle plus général, appelé *Erdős-Rényi Mixture graph* (ERMG).

2 Erdős-Rényi Mixture graph

Ce modèle, introduit par Daudin et Al dans [2], cherche à tenir compte de l'existence de sommets de différents types, en faisant dépendre la probabilité d'existence d'une arête des types des sommets à ces extrémités. Cela permet de modéliser par exemple un modèle de réseau social avec plusieurs groupes, et beaucoup plus de liens intragroupes que intergroupes, ou un modèle avec quelques sommets très reliés, et les autres peu reliés. Le but de la suite de cette partie est de montrer quels résultats connus sur le modèle simple ont pu être portés sur l'ERMG, et de montrer quelques problèmes encore ouverts.

2.1 Construction du modèle, et notations utilisées

On cherche à construire un graphe ayant q types de sommets différents, que l'on qualifiera dans la suite de couleurs de sommets. D'une manière

générale, si G est un graphe, on notera par $E(G)$ l'ensemble de ses arêtes, $V(G)$ l'ensemble de ses sommets, $e(g) = |E(G)|$ son nombre d'arêtes de $v(G) = |V(G)|$ son nombre de sommets.
On se munit d'une suite de de fonctions $(\alpha_{i,n})_{i\in\{1,...,q\}n\in\mathbb{N}}$, telle que $\forall n, \sum \alpha_i = 1$, avec α_i la probablité qu'un sommet fixé soit de couleur i, ainsi que d'une suite de matrices de taille $q \times q$, symétriques et à coefficients réels et compris entre 0 et 1, notée A_n, dont le coefficient (i, j) est la probabilité que deux sommets donnés soient reliés sachant qu'ils sont de couleur respective i et j. D'une manière générale, lorsque cela ne prêtera pas à confusion, on oubliera le n en indice.
Une contruction possible du graphe aléatoire G_n à n points est la suivante : On se munit de n variables aléatoires X_k i.i.d., à valeurs dans $\{1, ..., q\}$ telles que $\mathbb{P}(X_1 = i) = \alpha_i$. Le sommet k sera de la couleur X_k.
On se munit de $\binom{n}{2}$ variables aléatoires $Y_{k,l}$ $(1 \leqslant k < l \leqslant n)$ i.i.d. et indépendantes des X_i et de loi uniforme sur $[0; 1]$. On dira alors que l'arête reliant les sommets k et l est présente si $Y_{k,l} \leqslant A_{X_k,X_l}$. On notera par $\mathbb{P}_{n,\alpha,A}$ la probabilité associée (le plus souvent on écrira $\mathbb{P}_n$ voire $\mathbb{P}$ lorsque le contexte le permettra)

Remarque. *Un autre modèle possible et proche, serait de fixer le nombre de sommets de chaque couleur (i.e. de se donner un q-uplets d'entiers $(n_1, ..., n_q)$ de somme n et de dire que les sommets de 1 à n_1 sont de couleur 1, ceux de $n_1 + 1$ à $n_1 + n_2$ de couleur 2, etc). Et ensuite d'utiliser les $Y_{k,l}$ de la même manière que précedemment. On voit que les deux modèles seront proches, à condition que $\forall i, n\alpha_{n,i} \xrightarrow[n\to\infty]{} \infty$.*

Remarque. *On voit que le modèle (simple) d'Erdős-Rényi est bien un cas particulier de l'ERMG, à condition de prendre $q = 1$, et $A = (p)$.*

On aura besoin d'une matrice auxiliaire, notée B_n définie par $B_{n,i,j} = n\alpha_j A_{i,j}$. Ce coefficient représente le nombre moyen de voisins de couleur j qu'aura un sommet de couleur i (plus précisement, pour un sommet donné le nombre moyen de voisins de couleur j sachant qu'il est de couleur i est $(n - 1)\alpha_j A_{i,j}$).

2.2 Premiers résultats : le cas simple

Dans ce cas, on suppose que les α_i sont constants (i.e. la proportion de sommets de chaque type ne dépend pas de n), et qu'il existe une fonction $f : \mathbb{N} \to \mathbb{R}$ et une matrice $\mathcal{A}$ de taille $q \times q$ à coefficients réels positifs tels que $A_n = f(n)\mathcal{A}$. Dans ce cas, on a que $B_n = nf(n)\mathcal{B}$, avec $\mathcal{B}_{i,j} = \alpha_j \mathcal{A}_{i,j}$. Pour simplifier, on suppose que les α_i et les $\mathcal{A}_{i,j}$ sont tous non nuls[1]. On peut étendre la notion de fonctions seuil de la manière suivante :

Définition 4. *Si Q est une propriété possible du graphe G, on dit que G est une fonction seuil pour Q si l'on a :*
- $\lim_{n\to\infty} P_n(Q) = 0$ *si* $f(n) = o(g(n))$
- $\lim_{n\to\infty} P_n(Q) = 1$ *si* $g(n) = o(f(n))$

Remarque. *En théorie la notion de fonction seuil dépend des paramètres α et $\mathcal{A}$, mais dans les cas qui nous intéressent, cela ne sera pas le cas.*
Par ailleurs, la notion de fonction seuil est encore définie à une constante multiplicative près.

On va dans un premier temps chercher des fonctions seuil pour différentes propriétés.

2.2.1 Présence de motifs

On va s'intéresser au problème suivant : soit H un graphe, on s'intéresse à la propriété Q_H suivante : « il existe un sous-graphe de G isomorphe à H ». Nous allons voir que cela dépend principalement du nombre d'arêtes et du nombre de sommets de H.

Définition 5. *Soit G un graphe. On rappelle que $v(G)$ représente le nombre de sommets et $e(G)$ le nombre d'arêtes.*

- *On appelle densité de G, notée $d(G)$ la quantité suivante :*

$$d(G) = \frac{e(G)}{v(G)}$$

- *On appelle densité intérieure de G, notée $\tilde{d}(G)$ la quantité suivante :*

$$\tilde{d}(G) = \max_{H \subset G} d(H)$$

[1] les conditions strictement nécessaires sont plus compliqués : il faut que pour le graphe H considéré, il existe une coloration de H, telle que la probabilité de chaque sommet (α_i, avec i la couleur de sommet) et de chaque arête ($\mathcal{A}_{k,l}$ avec k et l la couleur des deux extrémités) soit non nulle

On a alors le théorème suivant :

Théorème 1. *La fonction g définie par $g(n) = n^{-\frac{1}{\tilde{d}(H)}}$ est une fonction seuil pour Q_H.*

Remarque. *Ce résultat permet de retrouver le résultat dans le cas du modèle simple d'Erdős-Rényi, prouvé dans [3] dans le cas où $\tilde{d}(H) = d(h)$ et dans [1] dans le cas général.*

Arguments de la preuve Si l'on note Z_H le nombre de copie de H dans G, on a que

$$\mathbb{E}(Z_H) = \Theta(n^{v(H)} f(n)^{e(H)}) = \Theta((n^{\frac{1}{\tilde{d}(H)}} f(n))^{e(H)})$$

En particulier, si on note H_0 un sous-graphe de H tel que $d(H_0) = \tilde{d}(H)$, on a que $f(n) = o(g(n)) \Rightarrow \mathbb{E}(Z_{H_0}) \to 0$, donc que a.p.s, on n'a pas de copie de H_0 dans G, donc à plus forte raison de copies de H.
Pour montrer le cas inverse, il faut estimer la variance de Z_H puis montrer que a.p.s. $Z_H > 0$ en utilisant l'inégalité de Bienaymé-Chebyshev.

2.2.2 Taille de la plus grande composante connexe

On s'intéresse maintenant à la taille de la plus grande composante connexe de G, et en particulier à partir de quel paramètre pour f il existe une composante connexe de taille au moins βn, avec β une constante arbitraire entre 0 et 1. Soit Q_β cette propriété. On a alors :

Théorème 2. *Pour tout $\beta \in]0;1[$, la fonction seuil pour la propriété Q_β est $\frac{1}{n}$.*

On s'intéresse maintenant au cas plus précis où $f = \frac{1}{n}$, pour savoir dans quel cas on a une composante connexe de taille linéraire en n. On a maintenant que $A_n = \frac{1}{n}\mathcal{A}$ et $B_n = \mathcal{B}$. $\mathcal{B}$ est une matrice à coefficients positifs. D'après le théorème de Perron-Frobenius, la plus grande valeur propre réelle positive est aussi la plus grande valeur propre en module, et a un vecteur propre associé à gauche et à droite à coefficients positifs. Appelons $\rho(\mathcal{B})$ cette valeur propre. Selon la valeur de $\rho(\mathcal{B})$, des comportements différents arrivent.

Théorème 3.
- *Si $\rho(\mathcal{B}) < 1$, alors il existe un réel C (dépendant de α et $\mathcal{A}$) tel que a.p.s. la plus grande composante connexe est de taille au plus $Clnn$.)*
- *Si $\rho(\mathcal{B}) > 1$, a.p.s. il existe un réel $g(\mathcal{B}) \in]0;1[$ tel que la taille de la plus grande composante connexe de G est $g(\mathcal{B})n + o_p(n)$.*

Remarque. *Le cas $\rho(\mathcal{B}) = 1$ est encore, à ma connaissance, un problème ouvert.*

Idée de la preuve, lien avec les arbres de Galton-Watson multitypes Pour connaître la composante connexe à laquelle appartient un sommet donné v_0, on peut utiliser l'algorithme suivant :

- On va partager les sommets en trois sous ensembles disjoints, C l'ensemble des sommets déjà explorés, R l'ensemble des sommets déjà repérés, mais non encore explorés et S les autres sommets
- Initialement, on a $C = \emptyset$, $R = \{v_0\}$ et $S = \{1, n\} \setminus \{v_0\}$
- Tant que R n'est pas vide, on choisit un sommet v de R, on le déplace dans C, et on ajoute à R tous les sommets voisins de v encore dans S.
- Lorsque R est vide, C est la composante connexe de V_0

On voit que le résultat de cet algorithme ne dépend pas de l'ordre dans lequel on a considéré les sommets. Au moins pour un petit nombre d'étapes (i.e. petit devant n), on peut approximer cet algorithme par un arbre de Galton-Watson multitype, dont la matrice des premiers moments est $\mathcal{B}$. Pour que le sommet v soit dans une composante connexe « grande », il faut que l'algorithme dure assez longtemps, donc que l'arbre de Galton-Watson survive « assez » longtemps. La condition de survie d'un arbre de Galton-Watson multitype étant $\rho(\mathcal{B}) > 1$ (c'est l'analogue de $\mathbb{E}(n) > 1$ avec n le nombre de fils dans le cas de l'arbre de Galton-Watson standard (monotype)), on comprend la raison de cette condition. Dans le cas $\rho(\mathcal{B}) > 1$, on peut même exprimer que la proportion des sommets appartenant à la plus grande composante connexe (notée g dans le théorème) comme la probabilité de survie d'un arbre de Galton-Watson.

Remarque. *Si on se limite aux cas $\rho(\mathcal{B}) \neq 1$, on retrouve les résultats du modèle simple en prenant $q = 1$.*

2.3 Le cas général

Dans ce cas, on suppose que les α_i peuvent varier et que la matrice A est une fonction quelconque de n. La notion de fonction seuil est bien plus difficile à définir (il n'y a plusieurs paramètres qui varient indépendamment), mais on peut néanmoins avoir un résultat proche (quoique plus compliqué) sur la recherche de motif. On définit les fonctions seuil de la manière suivante.

Définition 6. *Soit Q une propriété possible de G. On dit qu'une fonction $F : [0;1]^q \times \mathcal{M}_{q,q}(\mathbb{R}) \to R$ est une fonction seuil si l'on a :*

- $F(\alpha, A) \xrightarrow[n\to\infty]{} \infty \Rightarrow P_n(Q) \xrightarrow[n\to\infty]{} 1$
- $F(\alpha, A) \xrightarrow[n\to\infty]{} 0 \Rightarrow P_n(Q) \xrightarrow[n\to\infty]{} 0$

2.3.1 Recherche de motif

On cherche le motif H dans G. On a pour cela besoin d'introduire l'ensemble C_H des colorations de H, i.e. l'ensemble des applications de $V(H)$ dans $\{1, ..., q\}$.

On définit alors la fonction F qui à un graphe H et une coloration r de H associe le réel $F(H,r)$ défini par :

$$F(H,r) = n^{v(H)} \prod_{i \in V(H)} \alpha_{r(i)} \prod_{e \in E(H)} A_{r(e_1),r(e_2)}$$

Remarque. *$F(H,r)$ est équivalent asymptotiquement à l'espérance du nombre de sous-graphes de G, isomorphes à (H,r) (i.e. les sous-graphes isomorphes à H et dont l'image de la coloration par l'isomorphisme est r).*

$$F(H) = \max_{r:V(H)\to\{1,\dots,q\}} \min_{H' \subset H} f(H', r_{|H'})$$

Théorème 4. *Alors, F est une fonction seuil pour Q_H*

2.3.2 Plus grande composante connexe

Le problème de la plus grande composante connexe est encore ouvert, en particulier lorsque l'un des α_i tend vers 0. Dans ce cas, il peut-ere judicieux de considérer séparément le modèle avec un nombre non aléatoire de sommets de chaque couleur, car celui-ci n'est pas forcément équivalent.

Conjectures.
- *Si $limsup \quad \rho(B_n) = c < 1$, alors il existe une constante (dépendant de c) telle que a.p.s. la plus grande composante connexe soit de taille inférieure à $clnn$.*
- *Si $liminf \quad \rho(B_n) = c > 1$, alors il existe une constante (dépendant de c) telle que a.p.s. la plus grande composante connexe soit de taille au moins cn.*

Références

[1] B. Bollobás. Random graphs. *Combinatorics*, 1981.

[2] J. Daudin, F.Picard, and S.Robin. A mixture model for random graph. *Technical report, INRIA*, 2006.

[3] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. 1960.