Graphes d'Erdös-Rényi et grandes déviations
Introduction au domaine de recherche

Paul Thevenin

7 octobre 2016

1 Résumé

Les graphes d'Erdös-Rényi sont un outil fondamental de la théorie des graphes aléatoires. Il s'agit d'un modèle à la fois simple et très riche, dont l'étude permet d'obtenir de nombreux résultats portant sur des graphes de grande taille.
Dans ce modèle, on considère un ensemble de sommets. Un couple de sommets est relié par une arête avec une probabilité p fixée, et ce indépendamment des autres couples.

Ce modèle est souvent étudié d'un point de vue asymptotique. La question principale est : à quoi ressemble un graphe d'Erdös-Rényi lorsque le nombre de sommets n est grand ? Quelles propriétés possède-t-il ?

Nous introduisons d'abord le modèle d'Erdös et Rényi et certaines de ses propriétés les plus intéressantes. Puis nous exposons deux modèles différents permettant d'obtenir des résultats de grandes déviations sur ces graphes.

Table des matières

2 Cadre et premières définitions

Les graphes d'Erdös-Rényi sont des objets probabilistes introduits à la fin des années 1950 par les deux mathématiciens hongrois Paul Erdös et Alfréd Rényi dans [6]. Il s'agit du premier modèle de graphes aléatoires, c'est-à-dire de graphes générés par un processus aléatoire.

2.1 Graphes d'Erdös-Rényi

Ici, nous définissons rigoureusement les graphes d'Erdös-Rényi et la notion de seuil critique pour une propriété de graphe. Puis nous donnons des résultats sur certaines propriétés de ces graphes.

Définition 2.1. *Soit $p \in [0,1]$ et n un entier. Soit $(X_{i,j})_{1 \leq i < j \leq n}$ une famille de variables de Bernoulli indépendantes de paramètre p*
$(\mathbb{P}(X=1) = 1 - \mathbb{P}(X=0) = p)$.
On appelle graphe d'Erdös-Rényi, noté $G(n,p)$, la variable aléatoire à valeurs dans l'ensemble des graphes à n sommets, dont les sommets sont $1, 2, ..., n$ et dont l'ensemble des arêtes est $\{(i,j) | i < j, X_{i,j} = 1\}$.

En réalité, le modèle introduit par Erdös et Rényi dans leur article de 1959 ([6]) est légèrement différent. Leur idée consistait à étudier un graphe choisi uniformément parmi tous les graphe à n sommets et M arêtes. Le modèle décrit plus haut est dû à Gilbert, en 1959 également ([4]). Cependant, les deux modèles sont très similaires.

2.2 Seuil critique pour une propriété

Une des façons d'étudier les graphes d'Erdös-Rényi est la suivante : on se donne une propriété de graphe, par exemple la connexité, dont la probabilité croît avec le paramètre p. On se donne une suite de paramètres $p(n)$, et on regarde asymptotiquement, quand $n \to \infty$, avec quelle probabilité la propriété qui nous intéresse est vérifiée.

Pour certaines propriétés, on dispose même d'un seuil critique en dessous duquel la probabilité d'observer cette propriété tend vers 0, et au dessus duquel elle tend vers 1.

Définition 2.2. *On dit qu'une propriété $\mathcal{P}$ admet la suite $s(n)$ comme seuil critique si pour tout $\epsilon > 0$:*
- *Pour $p(n) \leq (1-\epsilon)s(n)$, on a $\mathbb{P}(G(n,p(n))$ vérifie $\mathcal{P}$)$= o(1)$*
- *Pour $p(n) \geq (1+\epsilon)s(n)$, on a $\mathbb{P}(G(n,p(n))$ vérifie $\mathcal{P}$)$= 1 - o(1)$*

Obtenir des seuils critiques pour des propriétés telles que le nombre de clique ou la connexité permet de comprendre en détail la structure d'un graphe d'Erdös-Rényi ayant un très grand nombre de sommets.

2.3 Propriétés intéressantes

La propriété la plus naturelle à étudier dans un graphe est la connexité : étant donnés deux sommets du graphe, existe-t-il toujours un chemin les reliant ?

Cette propriété a même été la première étudiée par Erdös et Rényi, qui ont pu déterminer le seuil critique associé.

Théorème 2.1 (Connexité). $s(n) = \frac{\log n}{n}$ *est seuil critique pour la connexité du graphe d'Erdös-Rényi.*
Plus précisément, pour tout $\epsilon > 0$*, on a :*

- $p(n) \geq (1+\epsilon)\frac{\log n}{n} \Rightarrow \lim_{n\to\infty} \mathbb{P}(G(n, p(n))$ *est connexe* $) = 1.$
- $p(n) \leq (1-\epsilon)\frac{\log n}{n} \Rightarrow \lim_{n\to\infty} \mathbb{P}(G(n, p(n))$ *est connexe* $) = 0.$

Plus tard, Matula ([5], 1972) et Bollobas ([3], 1988) ont pu déterminer les comportements typiques de deux quantités inhérentes au graphe, respectivement le nombre de clique (c'est-à-dire la taille du plus grand sous-graphe complet de notre graphe) et le nombre chromatique (c'est-à-dire le nombre minimal de couleurs nécessaire pour colorier les sommets du graphe, de sorte que deux sommets reliés par une arête soient toujours de couleurs différentes).

L'une des propriétés cruciales pour comprendre la structure d'un graphe d'Erdös-Rényi est la présence d'une composante dite "géante", contenant une fraction constante des sommets du graphe (cela n'a de sens qu'asymptotiquement). On connait beaucoup de choses sur cette composante géante, et notamment le seuil critique associé à sa présence dans le graphe (par exemple, voir Alon et Spencer, [2]).

Lemme 2.1. *Composante géante*
Posons $p(n) = \frac{c}{n}$*, où* $c > 0$.

- *Pour* $c < 1$*, avec probabilité tendant vers* 1*, la plus grande composante connexe de* $G(n, p(n))$ *est de taille* $O(\log n)$ *quand* $n \to \infty$.

- *Pour* $c > 1$*, il existe avec probabilité tendant vers* 1 *une unique composante dite géante, de taille* $Kn(1+o(1))$*, où* K *est une constante dépendant de* c*. Les autres composantes connexes sont de taille* $O(\log n)$*.* K *est en outre l'unique solution non nulle de l'équation* $x = 1 - e^{-cx}$*.*

Ainsi, il y a une transition dans le comportement du graphe. Si p est assez faibles, toutes les composantes sont très petites. Mais lorsqu'on dépasse le seuil $p = \frac{1}{n}$, il existe une unique composante géante, toutes les autres étant de nouveau très petites. Puis, à partir de $p = \frac{\log n}{n}$, le graphe est même connexe (avec probabilité tendant vers 1).

3 Générer des graphes d'Erdös-Rényi

Une façon de générer un graphe d'Erdös-Rényi est de procéder comme suit : On fixe $d \in \mathbb{N}^*$ et on se donne pour chaque arête $(i,j)_{1 \leq i < j \leq n}$ une variable gaussienne standard $Y_{i,j}$ à valeurs dans $\mathbb{R}^d$. Puis on choisit un vecteur $a \in S^{d-1}$. Enfin, on regarde le graphe dont les sommets sont $1, ..., n$ et les arêtes sont les $\{(i,j) | \langle Y_{i,j}, a \rangle \geq 0\}$.
Il s'agit simplement de garder les arêtes situées d'un côté d'un hyperplan, et d'enlever les autres.
On obtient alors un graphe distribué comme un $G(n, \frac{1}{2})$.
Cette méthode, introduite par Addario-Berry et al. [1], présente deux avantages :

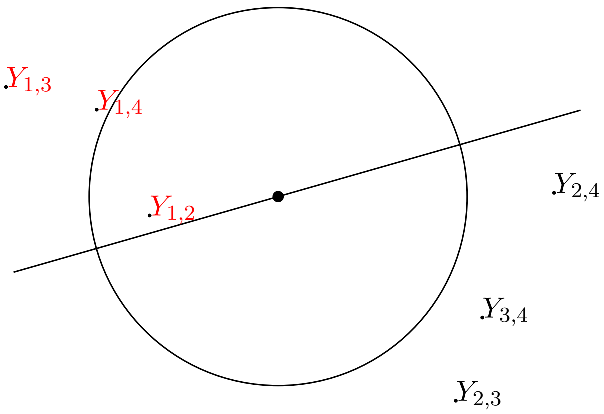
— En regardant $\{(i,j) | \langle Y_{i,j}, a \rangle \geq r\}$, où r est un réel quelconque, on peut aisément générer des graphes $G(n,p)$ pour tout $p \in]0,1[$. On doit alors garder les arêtes situées d'un côté d'un certain hyperplan affine, cette fois.

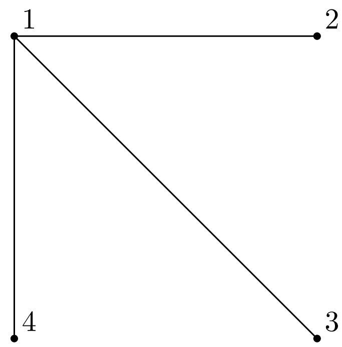
— Pour chaque vecteur a, on obtient un graphe distribué comme un graphe d'Erdös-Rényi de même paramètre. Cependant, au lieu de les étudier séparément, on peut étudier la famille des graphes obtenus lorsque a décrit S^{n-1}.

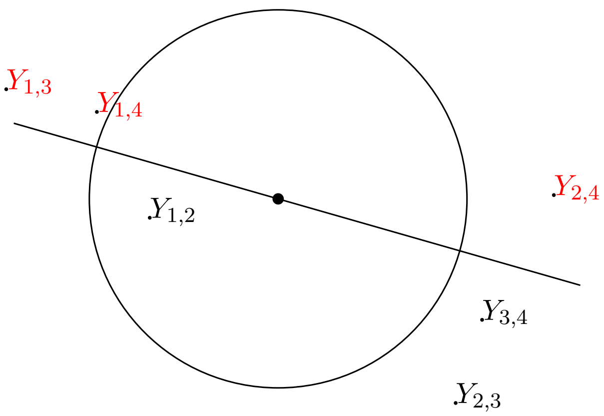
Remarque 3.1. *On fixe d'abord les variables $Y_{i,j}$, puis on fait varier l'hyperplan.*

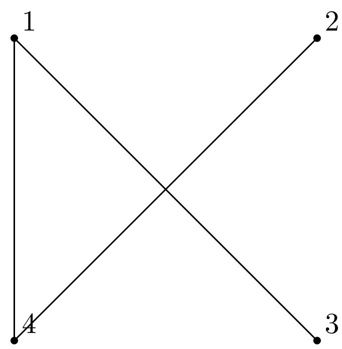
Ce faisant, on obtient une famille de graphes d'Erdös-Rényi dépendant les uns des autres, appelée famille sphérique. Selon les valeurs de d, cette famille aura des propriétés différentes.

Voici un exemple, pour $d = 2$ et $n = 4$:

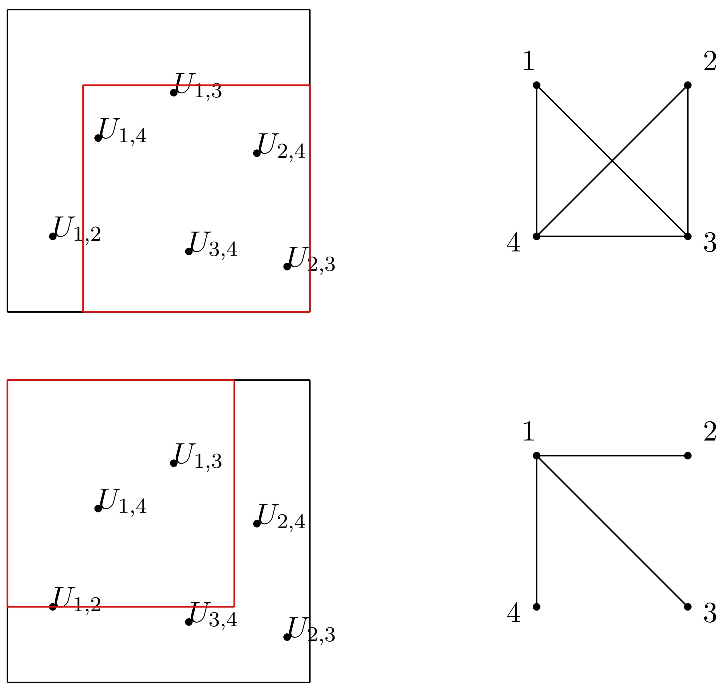

$Y_{1,3}$
$Y_{1,4}$
$Y_{1,2}$
$Y_{2,4}$
$Y_{3,4}$
$Y_{2,3}$


1
2
4
3


$Y_{1,3}$
$Y_{1,4}$
$Y_{1,2}$
$Y_{2,4}$
$Y_{3,4}$
$Y_{2,3}$


1
2
4
3

Une seconde façon similaire de générer ce type de graphes consiste à prendre, pour modéliser les arêtes du graphe, non pas des variables gaussiennes dans $\mathbb{R}^d$, mais des variables $(U_{i,j})_{i<j}$ uniformes sur $\mathbb{T}^d = (S^1)^d$, le tore de dimension d. Puis on se donne un point $a \in \mathbb{T}^d$, et on regarde $\{(i,j), ||U_{i,j} - a||_\infty \leq r\}$, où r est un réel de $[0,1]$.
De même que dans le modèle précédent, on peut faire varier r et ainsi générer des graphes $G(n,p)$ pour tout p.
De plus, en faisant varier a, on obtient une autre famille de graphes, appelée famille torique.



4 Grandes déviations

Les premières idées de la théorie des grandes déviations ont été introduites par Laplace et Cramer, puis développées par Chernoff et Varadhan notamment. Il s'agit d'étudier la queue des lois de probabilité : peut-on décrire le comportement asymptotique d'une suite de variables aléatoires donnée, conditionnellement à ce qu'elle soit différente de son comportement typique (par exemple, celui prédit par la loi des grands nombres) ?

Cette théorie s'applique particulièrement bien aux graphes d'Erdös-Rényi. Par exemple, si p est très petit devant $\frac{\log n}{n}$, quelle est la probabilité que le graphe d'Erdös-Rényi $G(n,p)$ soit tout de même connexe ?

Une façon plus amusante de voir le problème serait la suivante : si on prend un grand nombre de graphes $G(n,p)$, assez indépendants les uns des autres (dans un sens qui serait à préciser), quelle est la probabilité que certains soient connexes ?

Cette façon de voir les choses est à la base des travaux présentés ici.

5 Famille sphérique de graphes d'Erdös-Rényi

5.1 Objectif

L'étude de la famille sphérique a été entamée dans un article d'Addario-Berry et al. [1] datant de 2015.

L'idée est simple : Soit $\mathcal{P}$ une propriété possédant un seuil critique $s(n)$. Un graphe d'Erdös-Rényi $G(n,p)$ aura asymptotiquement un comportement typique, selon que $p \geq (1+\epsilon)s(n)$ ou $p \leq (1-\epsilon)s(n)$, pour un $\epsilon > 0$.

Cependant, puisqu'on étudie ici un grand nombre de graphes, il est possible que certains aient un comportement atypique. L'objectif est de déterminer à partir de quelle dimension d (dépendant de n) de tels graphes apparaissent au sein de notre famille.

Ainsi, certaines propriétés typiques seront robustes, dans le sens où il faut de grands d pour qu'un graphe de notre modèle ne la vérifie plus. Au contraire, pour d'autres propriétés, on trouvera même pour de petits d des graphes atypiques.

Ce sont ces considérations qui ont amené à définir la notion de dimension critique.

Définition 5.1. *(Dimension critique) Soit $\mathcal{P}$ une propriété de graphe. On dit que d_c (dépendant de n) est dimension critique pour $\mathcal{P}$ si pour tout $\epsilon > 0$:*

— *pour $d \leq (1-\epsilon)d_c$, avec probabilité tendant vers 1, tous les graphes du modèle de dimension d ont un comportement typique.*
— *pour $d \geq (1+\epsilon)d_c$, avec probabilité tendant vers 1, il existe un graphe dans le modèle qui a un comportement non-typique.*

La dimension critique détermine la "richesse" de notre modèle. Si d est trop faible, le modèle est trop pauvre et tous les graphes se comportent comme ils le devraient. Par contre, plus d augmente et plus le modèle devient riche. Certains graphes peuvent alors dévier de leur comportement typique. Nous en verrons plusieurs exemples par la suite.

Remarque 5.1. *Pour chaque propriété, il existe potentiellement deux dimensions critiques : une dans le cas où on est au dessus du seuil critique, et donc où les graphes du modèle devraient vérifier la propriété ; et une dans le cas où on est sous le seuil critique, et où les graphes ne devraient pas la vérifier. On ne s'occupera pas de ce qui se passe lorsque l'on est exactement au seuil critique.*

Remarque 5.2. *L'existence de ces dimensions critiques n'est pas garantie.*

5.2 Résultats

Dans leur article, Addario-Berry et al. démontrent de nombreux résultats sur ce modèle et les dimensions critiques associées. Pour cela, il s'agit d'allier résultats sur les graphes d'Erdös-Rényi (comme ceux présentés plus haut) et lemmes sur la structure de la sphère S^{d-1}. Dans cette optique, un des lemmes fondamentaux utilisés est le suivant :

Lemme 5.1. *Pour $0 < \eta < 1$, il existe un ensemble C_η de cardinal au plus $(\frac{4}{\eta})^d$ tel que pour tout $s \in \mathbb{S}^{d-1}$, il existe $a \in C_\eta$ avec $||s-a||_2 \leq \eta$.*

Ce simple lemme permet d'obtenir des bornes inférieures sur la dimension critique de certains phénomènes. En voici un exemple :

Théorème 5.1. *Posons $p = c\frac{\log n}{n}$.*
Si $c < 1$ alors pour tout $\epsilon > 0$, si $d = O(n^{1-c-\epsilon})$, alors avec probabilité tendant vers 1 tous les graphes du modèle sont non-connexes.

Démonstration. La preuve mêle le lemme précédent et un lemme concernant les graphes d'Erdös-Rényi.
Pour $s \in S^{d-1}$, posons $G(s)$ le graphe associé à s dans notre modèle.
Posons également $\eta \in]0,1[$ et C_η comme dans le lemme précédent et $s_0 = (1,0,...,0) \in S^{d-1}$.

Alors, en utilisant les symétries de notre modèle, on a

$$\mathbb{P}(\exists s \in S^{d-1}, G(s) \text{ est connexe }) \leq |C_\eta| \mathbb{P}(\exists s \in S^{d-1}, ||s - s_0|| \leq \eta, G(s) \text{ est connexe }).$$

On remarque alors que, si l'un des $G(s)$, où s est proche de s_0, est connexe, alors l'union de tous ces graphes (c'est-à-dire le graphe dont les arêtes sont toutes celles qui apparaissent dans au moins un de ces graphes) est connexe. Or, $\cup_{||s-s_0||\leq\eta} G(s)$ se comporte comme un graphe d'Erdös-Rényi, de paramètre plus grand.
Il suffit alors d'avoir une estimation de la probabilité qu'un unique graphe d'Erdös-Rényi soit connexe. Ces estimées existent (mais sont très techniques, et nous ne les aborderons pas).
On peut majorer C_η par le lemme précédent, puis optimiser sur $\eta \in]0, 1[$. On obtient alors le résultat voulu. $\square$

Il est conjecturé que la dimension critique pour la connexité, dans le cas où $c < 1$, existe et est de l'ordre de n^{1-c}. Cependant, on n'a pour l'instant qu'une borne supérieure en $d = O(n\sqrt{\log n})$.
Dans le cas où $c > 1$, les résultats obtenus sont bien plus précis, et nous avons l'encadrement :
$(c-1)\frac{\log n}{\log \log n} \leq d_{critique} \leq 2(c-1)\frac{\log n}{\log \log n}$.

Cela indique une asymétrie entre les cas $c > 1$ et $c < 1$: il est beaucoup plus difficile d'obtenir un graphe connexe pour un paramètre petit que d'obtenir un graphe non-connexe pour un paramètre grand.

6 Famille torique de graphes d'Erdös-Rényi

Le modèle torique a été introduit plus récemment, pour tenter d'obtenir des résultats plus précis. En effet, déterminer les dimensions critiques correspondant au modèle sphérique est difficile, de par la structure même de S^{n-1}. On rencontre moins de difficultés dans l'étude du modèle torique. Intuitivement, le fait qu'on ne puisse pas partitionner la sphère en disques empêche de mener des calculs exacts, alors qu'il est simple de partitionner le tore en petits cubes.

Rien n'assure que les dimensions critiques pour ces deux modèles soient égales (ni même qu'elles existent, dans le modèle torique). Il existe déjà des phénomènes pour lesquels il est même certain qu'elles sont différentes. Cependant, pour la plupart des propriétés étudiées (connexité, apparition d'une composante géante, nombre chromatique, nombre de clique, ...), les résultats obtenus dans le modèle torique sont simplement des raffinements de ceux obtenus dans le modèle sphérique, les preuves étant quasiment identiques.

Déterminer l'influence de la structure du modèle sur les dimensions critiques, ou découvrir de nouveaux modèles similaires, sont des problèmes encore ouverts.

Les résultats les plus intéressants dans ce modèle concernent la présence d'une composante géante.
On sait déjà que le seuil critique pour ce phénomène se situe à $p = \frac{1}{n}$.

Conjecture 6.1. *Mettons-nous alors dans le cas où $p = \frac{c}{n}$, avec $c > 1$.*
Nous savons alors que le graphe d'Erdös-Rényi $G(n,p)$ a, avec probabilité tendant vers 1, une composante géante.
Dans ce cas, il est conjecturé, dans les deux modèles, que la dimension critique associée, à partir de laquelle on observe des graphes sans composante géante, vaut :

$$d_{critique} = f(c)\frac{n}{\log n}$$

où $f(c) = \frac{1}{2}(c - 1 - \log c)$ est la constante donnée par le principe de grande déviation de Cramer.

Ceci n'est prouvé dans aucun des deux modèles. Cependant, on a les résultats suivants :

— Dans le modèle sphérique, on a

$$K\frac{n}{\log n} \leq d_{critique} \leq O(n \log^2 n)$$

Ici, K est une constante, fonction de c uniquement, mais n'est pas exprimable facilement. Cependant, elle est inférieure à $f(c)$.

— Dans le modèle torique, on a

$$\frac{1}{2}(c - \frac{1}{c} - 2\log c)\frac{n}{\log n} \leq d_{critique} \leq f(c)\frac{n}{\log n}.$$

Ainsi, les résultats obtenus dans le cadre du modèle torique sont bien meilleurs. On sait déjà, notamment, que la dimension critique est de l'ordre de $\frac{n}{\log n}$, ce qui n'est pas encore prouvé dans le modèle sphérique.
Une piste pour prouver ce résultat dans le modèle sphérique serait d'affiner les calculs d'Addario-Berry et al. ([1]) pour pouvoir transposer la preuve du modèle torique vers celui-ci.

Références

[1] L. Addario-Berry et al. Exceptional rotations of random graphs : A vc theory. *Journal of Machine Learning Research*, 16 :1893–1922, 2015.

[2] N. Alon and J. Spencer. *The Probabilistic Method*, chapter 11, page 185. Wiley, 2008.

[3] B. Bollobas. The chromatic number of random graphs. *Combinatorica*, 8 :49–55, 1988.

[4] E. Gilbert. Random graphs. *Annals of Mathematical statistics*, 30 :1141–1144, 1959.

[5] D. Matula. Employee party problem. *Notices of the American Mathematical Society*, 19 :A382–A382, 1972.

[6] P.Erdös and A.Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5 :17–61, 1960.