TD 06 – Graphes aléatoires et partiel de l'an dernier (corrigé)

Exercice 1. *K4*

Soit G un graphe aléatoire de loi $G_{n,p}$. L'objectif de cet exercice est de montrer qu'il y a un seuil $p_0 := n^{-2/3}$ tel que pour $p = o(p_0)$, le graphe G n'a pas de clique de taille 4 avec bonne probabilité, et que pour $p = \omega(p_0)$, le graphe G a au moins une clique de taille 4 avec bonne probabilité.

Rappels / définitions :

— Un graphe aléatoire G suit la loi $G_{n,p}$ s'il a n sommets et que chaque arrête est présente dans G avec probabilité p ;

— une clique de taille 4 est un ensemble de 4 sommets tous reliés deux à deux par des arêtes ;

— $p = o(p_0)$ signifie $\frac{p}{p_0} \to 0$ quand $n \to +\infty$;

— $p = \omega(p_0)$ signifie $\frac{p_0}{p} \to 0$ quand $n \to +\infty$.

1. Pour p quelconque, calculer $\mathbf{E}[X]$, où X est le nombre de cliques du graphe G.

☞ On a $\binom{n}{4}$ ensembles possibles de 4 sommets. Pour chacun de ces ensembles, on définit X_i qui vaut 1 si c'est une clique et zero sinon. On a $X = \sum_i X_i$. Et $\mathbf{E}[X_i] = \mathbf{P}\{X_i = 1\} = p^6$. D'où

$$\mathbf{E}[X] = \binom{n}{4} p^6.$$

2. Soit $p = o(p_0)$, montrer que $\Pr(X \neq 0) \to 0$ quand n tend vers l'infini.

☞ On a $\mathbf{E}[X] \le n^4 p^6 = o(n^{4-6\times 2/3}) = o(1)$. Donc $\mathbf{E}[X]$ tend vers zero quand n tend vers l'infini. Comme X est a valeur entières, positives ou nulle, on conclut que $\mathbf{P}\{X \neq 0\} = \Pr(X \ge 1) \le \mathbf{E}[X]$ tend aussi vers 0.

On suppose maintenant $p = \omega(p_0)$, et on veut montrer que $\mathbf{P}\{X = 0\} \to 0$ quand n tend vers l'infini.

3. Montrer que $\mathbf{P}\{X = 0\} \le \frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2}$. Il suffira donc de montrer que $\frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2} \to 0$.

☞ On utilise l'inégalité de Chebychev

$$\Pr(X = 0) \le \mathbf{P}\{|X - \mathbf{E}[X]| \ge \mathbf{E}[X]\} \le \frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2}.$$

4. Soit X_i des variables aléatoires à valeur dans $0, 1$ (et non indépendantes). Montrer que

$$\mathbf{Var}\left[\sum_i X_i\right] \le \mathbf{E}\left[\sum_i X_i\right] + \sum_{i \neq j} \mathbf{E}\left[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])\right].$$

☞ On a

$$\begin{aligned}\mathbf{Var}\left[\sum_i X_i\right] &= \mathbf{E}\left[\left(\sum_i (X_i - \mathbf{E}[X_i])\right)^2\right] \\ &= \mathbf{E}\left[\sum_{i,j} (X_i - \mathbf{E}[X_i]) \cdot (X_j - \mathbf{E}[X_j])\right] \\ &= \sum_i \mathbf{E}\left[(X_i - \mathbf{E}[X_i])^2\right] + \sum_{i \neq j} \mathbf{E}\left[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])\right].\end{aligned}$$

On observe ensuite que $\mathbf{E}\left[(X_i - \mathbf{E}[X_i])^2\right] = \mathbf{E}\left[X_i^2\right] - \mathbf{E}[X_i]^2 \le \mathbf{E}\left[X_i^2\right]$. Mais comme X_i est à valeur dans $0, 1$, on a $\mathbf{E}\left[X_i^2\right] = \mathbf{E}[X_i]$. D'où l'inégalité.

5. En déduire que $\mathbf{Var}[X] = o(\mathbf{E}[X]^2)$ et conclure.

☞ On note C_i les ensembles de 4 sommets du graphe G, et on défini X_i la variables aléatoire qui vaut 1 si C_i forme une clique et 0 sinon. On a $X = \sum_i X_i$ et d'après la question précédente $\mathbf{Var}[X] \le \mathbf{E}[X] + \sum_{i \neq j} \mathbf{E}\left[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])\right]$. Fixons $i \neq j$ et considérons $\mathbf{E}\left[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])\right] = \mathbf{E}[X_i X_j] - \mathbf{E}[X_i]\,\mathbf{E}[X_j]$. On a $\mathbf{E}[X_i X_j] = \mathbf{P}\{C_i \text{ et } C_j \text{ sont des cliques}\} = p^k$, où k est le nombre d'arêtes nécessaires pour que C_i et C_j soient des cliques. Ce nombre d'arêtes va dépendre nu nombre de sommets communs entre C_i et C_j. On distingue donc les cas suivants

— Si $|C_i \cap C_j| = 0$ ou $|C_i \cap C_j| = 1$, alors $k = 12$ et $\mathbf{E}\left[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])\right] = 0$.

— Si $|C_i \cap C_j| = 2$, alors $k = 11$ et $\mathbf{E}\left[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])\right] = p^{11}(1-p)$. Il y a $\binom{n}{4} \cdot \binom{4}{2} \cdot \binom{n-4}{2}$ tels couples (C_i, C_j).
— Si $|C_i \cap C_j| = 3$, alors $k = 9$ et $\mathbf{E}\left[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])\right] = p^{9}(1-p^3)$. Il y a $\binom{n}{4} \cdot \binom{4}{3} \cdot \binom{n-4}{1}$ tels couples (C_i, C_j).
— Le cas $|C_i \cap C_j| = 4$ est impossible car $C_i \neq C_j$.

On a donc $\mathbf{Var}[X] \leq \mathbf{E}[X] + \binom{n}{4} \cdot \binom{4}{2} \cdot \binom{n-4}{2} p^{11}(1-p) + \binom{n}{4} \cdot \binom{4}{3} \cdot \binom{n-4}{1} p^9(1-p^3)$. Chacun des trois termes de cette somme est un $o(\mathbf{E}[X]^2)$ (car $p = \omega(p_0)$), d'où la réponse à la question. On conclut grâce aux questions précédentes que $\Pr(X \neq 0) \to 0$ quand n tend vers l'infini.

Exercice 2. *Graphe Aléatoire Bipartite*

Soit $0 < p < 1$ et $n \in \mathbb{N}^*$. On définit un graphe aléatoire non orienté $H_{2n,p}$ de la manière suivante. On se donne une famille $\{X_{i,j} \,:\, 1 \leq i \leq n,\ n+1 \leq j \leq 2n\}$ de v.a. i.i.d. de loi de Bernoulli de paramètre p. On pose alors $H_{2n,p} = (V, E)$, avec $V = \{1, \dots, 2n\}$ et

$$E = \{(i,j) \,:\, X_{i,j} = 1\} \subset \{1, \dots, n\} \times \{n+1, \dots, 2n\}.$$

1. Quelle est la loi du nombre d'arêtes de $H_{2n,p}$?

☞ Le nombre d'arêtes de $H_{2n,p}$ suit la loi $B(n^2, p)$.

2. Quelle est l'espérance du nombre de sommets isolés de $H_{2n,p}$?

☞ Soit N le nombre de sommets isolés. Si A_i est l'événement « le sommet i est isolé », on a par linéarité de l'espérance, on a $\mathbb{E}[N] = \sum \mathbb{P}(A_i) = 2n(1-p)^n$.

3. Dans cette question on pose $p = c\log(n)/n$ pour un nombre réel $c > 0$.
   1. Montrer que si $c > 1$, alors
$$\lim_{n\to\infty} \mathbb{P}(H_{2n,p} \text{ a un sommet isolé}) = 0.$$
   2. Montrer que si $c < 1$, alors
$$\lim_{n\to\infty} \mathbb{P}(H_{2n,p} \text{ a un sommet isolé}) = 1.$$

☞
1. Si $c > 1$, on a $\mathbb{E}[N] = 2n\exp(n\log(1 - \frac{c\log n}{n})) \to 0$ et donc $\mathbb{P}(N \geq 1) \leq \mathbb{E}[N] \to 0$.
2. Si $c < 1$, on calcule
$$\mathbb{E}[N^2] = \sum_{i,j=1}^{2n} \mathbb{P}(A_i \cap A_j) = 2n(1-p)^n + 2n(n-1)(1-p)^{2n} + 2n^2(1-p)^{2n-1}$$
d'où il vient que $\mathbb{E}[N^2]/\mathbb{E}[N]^2$ tend vers 1. On utilise l'inégalité de Tchebychev pour conclure que
$$\mathbb{P}(N=0) = \mathbb{P}(\mathbb{E}[N] - N \geq \mathbb{E}[N]) \leq \mathbb{P}(|N - \mathbb{E}[N]| \geq \mathbb{E}[N]) \leq \frac{\mathbf{Var}[N]}{\mathbb{E}[N]^2} = \frac{\mathbb{E}[N^2]}{\mathbb{E}[N]^2} - 1 \to 0.$$

4. Dans cette question on pose $p = 1/2$. Montrer qu'il existe une constante $C > 0$ telle que
$$\lim_{n\to\infty} \mathbb{P}\left(\text{tous les sommets de } H_{2n,p} \text{ ont un degré inférieur à } \frac{n}{2} + C\sqrt{n\log n}\right) = 1.$$

☞ Le degré d_i du sommet i suit la loi $B(n, 1/2)$. Par l'inégalité de Chernoff I, on a donc
$$\mathbb{P}(d_i \geq \frac{n}{2} + a) \leq \exp(-2a^2/n).$$
Ainsi, par la borne de l'union,
$$\mathbb{P}(\max_i d_i \geq \frac{n}{2} + a) \leq 2n\exp(-2a^2/n).$$
Cette quantité tend vers 0 si $a = C\sqrt{n\log n}$ avec $2C^2 > 1$.

Exercice 3. *Clique Number*

Le clique number $\omega(G)$ d'un graphe G est l'entier k maximal tel que G contient une clique de taille k (c'est-à-dire un ensemble de k sommets tous reliés deux à deux par des arêtes). Soit G un graphe aléatoire de loi $G_{n,1/2}$ (i.e. G a n sommets et chaque arête de G est présente avec probabilité $1/2$). L'objectif de cet exercice est de montrer que pour tout $\varepsilon > 0$,

$$\lim_{n\to+\infty} \mathbf{P}\{(2-\varepsilon)\log n \leq \omega(G) \leq (2+\varepsilon)\log n\} = 1.$$

Asymptotiquement, $\omega(G)$ est donc de l'ordre de $2\log n$.
Remarque : La méthode utilisée pour montrer ce résultat ressemble beaucoup à celle utilisée dans l'exercice K4.

1. Pour un entier k quelconque, on définit X_k la variable aléatoire comptant le nombre de cliques de taille k dans G. Calculer $\mathbf{E}[X_k]$

   ☞ Une solution pour l'exercice peut être trouvée ici : `https://www.win.tue.nl/~nikhil/courses/2015/2W008/probabilistic_method-2.pdf`, section 3.2.

2. Soit $\varepsilon > 0$. Montrer que $\lim_{n\to+\infty} \mathbf{P}\{\omega(G) \geq (2+\varepsilon)\log n\} = 0$ (pour simplifier, on pourra supposer que $k = (2+\varepsilon)\log n$ est un entier).

3. On considère maintenant $k = (2-\varepsilon)\log n$ (encore une fois, on le suppose entier). Montrer que $\lim_{n\to+\infty} \mathbf{P}\{\omega(G) \leq (2-\varepsilon)\log n\} = 0$. *Indice : on pourra utiliser les questions 3 et 4 de l'exercice K4.*

Exercice 4. *Estimer l'intersection avec un rectangle*

Let $P \subset \mathbb{Z}^2$ of size n. Our objective is to be able to answer quickly queries of the form what is the fraction of points in P that are in the rectangle $r = [a_1, b_1] \times [a_2, b_2]$? We write $r[P] = \frac{|P\cap r|}{n}$ for this fraction. We consider a simple data structure to approximate $r[P]$ efficiently for any query r. The data structure is just a random subset $S \subset P$ of size m. On query r, the estimate for $r[P]$ we output is $\frac{|S\cap r|}{m}$. The structure S defines an ε-approximation if for all queries r, we have $|r[P] - \frac{|S\cap r|}{m}| \leq \varepsilon$.

1. What m should we take to obtain an ε-approximation with probability $1-\delta$?

   ☞ On va prendre un sample S dont *l'espérance de la taille* est m (plutôt que taille exactement m).
   Pour $p \in P$, soit X_p une variable aléatoire de Bernouilli de paramètre m/n, et l'on définit S par la relation suivante : si $X_p = 1$ alors $p \in S$, et si $X_p = 0$ alors $p \notin S$. Fixons un rectangle r et soit $X(r) = \sum_{p\in R} X_p = |S \cap R|$ de telle sorte que $X(r)/m$ soit notre estimateur. Alors $\mathbf{E}[X(r)] = \sum_{p\in R} \mathbf{P}\{p \in S\} = \sum_{p\in R} \mathbf{P}\{m/n\} = mr[P]$. On peut donc appliquer Chernoff à $X(r)$ car :

$$\mathbf{P}\{|X(r)/m - r[P]| \geq \varepsilon\} = \mathbf{P}\{|X(r) - mr[P]| \geq \varepsilon m\} = \mathbf{P}\{|X(r) - \mathbf{E}[X(r)]| \geq \varepsilon/r[P]\cdot\mathbf{E}[X(r)]\} \leq 2e^{-\frac{\varepsilon'^2}{2+\varepsilon'}\mathbf{E}[X(r)]} .$$

   avec $\varepsilon' = \varepsilon/r[P]$, ce qui donne (en utilisant $r[P] \leq 1$ pour la dernière inégalité) :

$$2e^{-\frac{\varepsilon'^2}{2+\varepsilon'}\mathbf{E}[X(r)]} \leq 2e^{-\frac{\varepsilon^2}{2r[P]+\varepsilon}m} \leq 2e^{-\frac{\varepsilon^2}{2+\varepsilon}m}.$$

   Cette inégalité est vraie pour un rectangle r fixé, mais nous avons besoin d'une Union-Bound sur tous les rectangles. Or, il y a une infinité de rectangles possibles dans $\mathbb{Z}^2$, donc nous devons être un peu plus malin. Il faut remarquer que si r et r' sont des rectangles pour lesquels $P \cap r = P \cap r'$, alors $r[P] = r'[P]$ et l'estimation sera la même, donc l'erreur sur l'un sera exactement la même que l'erreur sur l'autre. En d'autre termes, on veut trouver un certain nombre de rectangle $r_1, r_2, \ldots, r_k$ tels que pour tout rectangle r de $\mathbb{Z}^2$, il existe i tel que $P \cap r = P \cap r_i$. Ainsi,

$$\mathbf{P}\{\exists r \text{ s.t. } |r[P] - X(r)| \geq \varepsilon\} \leq \sum_{i=1}^{k} \mathbf{P}\{|r_i[P] - X(r_i)| \geq \varepsilon\} \leq k2e^{-\frac{\varepsilon^2}{2+\varepsilon}m} .$$

   Montrons maintenant qu'on peut obtenir $k = n^4$: pour chaque 4-uplet des points de P $((x_1, y_1)(x_2, y_2), (x_3, y_3), (x_4, y_4))$, on définit un rectangle $r_i = [x_1, x_2] \times [y_3 y_4]$. Cet ensemble de n^4 rectangles a bien la propriété demandée car si r est un rectangle quelconque, on peut "pousser" sa limite verticale gauche le plus à droite possible jusqu'à rencontrer un point de P, auquel cas on s'arrête de "pousser". On fait de même pour les quatre côtés du rectangle (on "pousse" vers l'intérieur jusqu'à rencontrer un point de P), et on tombe sur un r_i pour lequel $r \cap P = r_i \cap P$.
   En résumé, nous voulons m tel que

$$n^4 2e^{-\frac{\varepsilon^2}{2+\varepsilon}m} \leq \delta ,$$

   ce qui est possible pour

$$m \geq \frac{2+\varepsilon}{\varepsilon^2}(4\ln n + \ln 2 - \ln\delta) = \Omega(\ln n) .$$

Exercice 5. *Algorithme probabiliste pour calculer la médiane*

On étudie un algorithme probabiliste [1] pour déterminer la médiane d'un ensemble $E = \{x_1, \ldots, x_n\}$ de n nombres réels en temps $O(n)$. On rappelle que m est une médiane de E si au moins $\lceil n/2 \rceil$ des élements de E sont inférieurs ou égaux à m, et au moins $\lceil n/2 \rceil$ des élements de E sont supérieurs ou égaux à m. Pour simplifier on suppose n impair (ce qui fait que la médiane est unique) et on suppose aussi que les éléments de E sont tous distincts.

Voici comment fonctionne l'algorithme

1. Remarque : il existe un algorithme déterministe de même performance

(a) Soit $(Y_i)_{1\leq i\leq n}$ une suite de v.a. i.i.d. de loi de Bernoulli de paramètre $n^{-1/4}$. On considère le sous-ensemble aléatoire de E défini par $F = \{x_i : Y_i = 1\}$. Si card $F \leq \frac{2}{3}n^{3/4}$ ou card $F \geq 2n^{3/4}$ on répond «`ERREUR 1`».

(b) On trie F et on appelle d le $\lfloor \frac{1}{2}n^{3/4} - \sqrt{n} \rfloor$ème plus petit élément de F, et u le $\lfloor \frac{1}{2}n^{3/4} - \sqrt{n} \rfloor$ème plus grand élément de F.

(c) On détermine le rang de d et de u dans E (l'élément minimal a rang 1, l'élément maximal a rang n), que l'on note respectivement r_d et r_u. Si $r_d > n/2$ ou $r_u < n/2$ on répond «`ERREUR 2`».

(d) On note $G = \{x_i \in E : d < x_i < u\}$. Si card $G \geq 4n^{3/4}$ on répond «`ERREUR 3`».

(e) On trie G et on renvoie le $(\lceil n/2 \rceil - r_d)$ème élement de G.

1. Justifier pourquoi l'algorithme retourne la médiane en temps $O(n)$ lorsqu'il ne répond pas de message d'erreur.

☞ Si aucun message d'erreur n'est renvoyé, l'algorithme s'exécute en temps $O(n)$; en effet la génération des (Y_i) prend un temps $O(n)$, le tri de F et G prend un temps $O(m \log m)$ pour $m = O(n^{3/4})$, et la détermination de r_d, de r_u et de G nécessite $O(n)$ comparaisons. De plus, l'absence de message d'erreur numéro 2 garantit que la médiane est dans l'intervalle $[d, u]$, donc dans G.

2. Montrer que pour $i \in \{1, 2, 3\}$, on a

$$\lim_{n\to\infty} \Pr\left(\text{l'algorithme retourne «}\texttt{ERREUR } i\text{»}\right) = 0.$$

Pour simplifier l'analyse et éviter d'écrire des symobles $\lfloor\cdot\rfloor$ ou $\lceil\cdot\rceil$, on pourra supposer implicitement que des nombres tels que $\sqrt{n}$, $\frac{1}{2}n^{3/4}$, … sont des entiers

☞

1. Pour l'erreur 1 : comme card $F = Y_1 + \cdots + Y_n$ a la loi $B(n, n^{-1/4})$, on a par l'inégalité de Chernoff II

$$\mathbb{P}(\text{card } F \geq 2n^{3/4}) \leq \exp(-n^{3/4}/3), \quad \mathbb{P}(\text{card } F \leq \frac{2}{3}n^{3/4}) \leq \exp(-n^{3/4}/18).$$

2. Pour l'erreur 2 : on note E^- l'ensemble des éléments de E inférieurs où égaux à la médiane, et on remarque que $r_d > n/2$ équivaut à card $(F \cap E^-) < \frac{1}{2}n^{3/4} - \sqrt{n}$. La v.a. card $(F \cap E^-)$ suit la loi $B(\lceil n/2 \rceil, n^{-1/4})$ (notons μ sa moyenne) donc par l'inégalité de Chernoff II

$$\mathbb{P}(\text{card } (F \cap E^-) < \frac{1}{2}n^{3/4} - \sqrt{n}) \leq \mathbb{P}(\text{card } (F \cap E^-) \leq (1 - 2n^{-1/4})\mu) \leq \exp(-\mu\sqrt{n}) \to 0$$

Un argument symétrique traite le cas de $r_u > n/2$ et considérant E^+ l'ensemble des éléments de E supérieurs où égaux à la médiane

3. Pour l'erreur 3 : si card $G \geq 4n^{3/4}$, alors ou bien card $(G \cap E^-) \geq 2n^{3/4}$ ou bien card $(G \cap E^+) \geq 2n^{3/4}$; ces deux événements ayant même probabilité, il suffit de montrer que $\mathbb{P}(\text{card } (G \cap E^-) \geq 2n^{3/4}) \to 0$. On remarque que si card $(G \cap E^-) \geq 2n^{3/4}$, alors $r_d \leq \frac{n}{2} - 2n^{3/4}$ et donc l'ensemble F contient au moins $\frac{1}{2}n^{3/4} - \sqrt{n}$ parmi les $\frac{n}{2} - 2n^{3/4}$ plus petits éléments de E. La probabilité de ce dernier événement est $\mathbb{P}(X \geq (1+\varepsilon)\varepsilon X)$, où $X \sim B(\frac{n}{2} - 2n^{3/4}, n^{-1/4})$ et $\varepsilon = \frac{\sqrt{n}}{n^{3/4}/2 - 2\sqrt{n}} = O(n^{-1/4})$. Une dernière application de l'inégalité de Chernoff II permet de conclure que la probabilité considérée tend vers 0.