GRAPHES ALÉATOIRES D'ERDÖS-RÉNYI

Un grand réseau de communication (internet, Facebook®, *etc.*) est modélisé par un *graphe* $G = (V, E)$, dont les *sommets* $v \in V$ sont les agents de ce réseau, et dont les *arêtes* $e = \{u, v\}$ sont les paires de sommets distincts u et v tels qu'existe une connection entre les agents u et v. Par exemple, le graphe suivant représente une partie des connections du réseau internet :

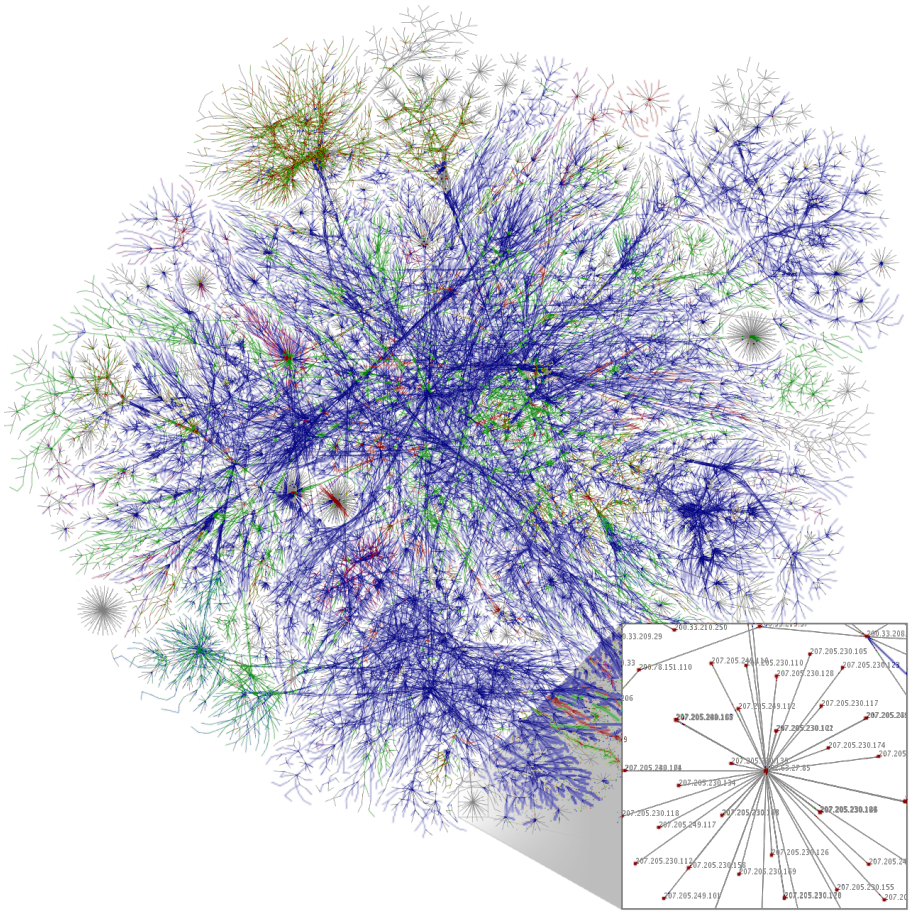

FIGURE 1. Connections entre une partie des sites du réseau internet (d'après `opte.org`).

Pour un tel graphe, les questions suivantes se posent :

(1) Étant données deux personnes u et v dans le réseau, existe-t-il toujours une suite de connections $\{u_0, u_1\}, \{u_1, u_2\}, \ldots, \{u_{d-1}, u_d\}$ avec $u_0 = u$ et $u_d = v$? Autrement dit, le graphe est-il *connexe* ?

(2) Si le graphe n'est pas connexe, quelle est la *taille typique de ses composantes connexes* ? Ont-elles toutes à peu près la même taille, ou, au contraire, existe-t-il des composantes "géantes" entourées d'îlots de taille beaucoup plus petite ?

(3) Si l'on fixe une configuration de voisins (par exemple, quatre voisins a, b, c, d avec a connecté à b, et b, c, d mutuellement connectés), combien de configurations de ce type peut-on identifier à l'intérieur du réseau ?

Dans ce qui suit, les connections du réseau n'étant pas forcément connues, on le remplacera par un *graphe aléatoire* $G = G_n$, dont on va étudier le comportement lorsque le nombre n de sommets tend vers l'infini.

1. Modèles d'Erdös-Rényi

On fixe un entier $n \geq 1$, et un paramètre $p = p_n \in [0, 1]$. Le *graphe aléatoire d'Erdös-Rényi* $G = G(n, p)$ est le graphe dont les sommets sont les entiers $i \in [\![1, n]\!]$, et tel que les variables aléatoires

$$X_{ij} = \begin{cases} 1 & \text{si } \{i, j\} \text{ est une arête de } G, \\ 0 & \text{sinon} \end{cases}$$

avec $1 \leq i < j \leq n$ sont des variables de Bernoulli indépendantes de même paramètre p : $\mathbb{P}[X_{ij} = 1] = 1 - \mathbb{P}[X_{ij} = 0] = p$. Autrement dit, chaque arête possible entre les sommets de G apparaît avec probabilité p, indépendamment des autres arêtes.

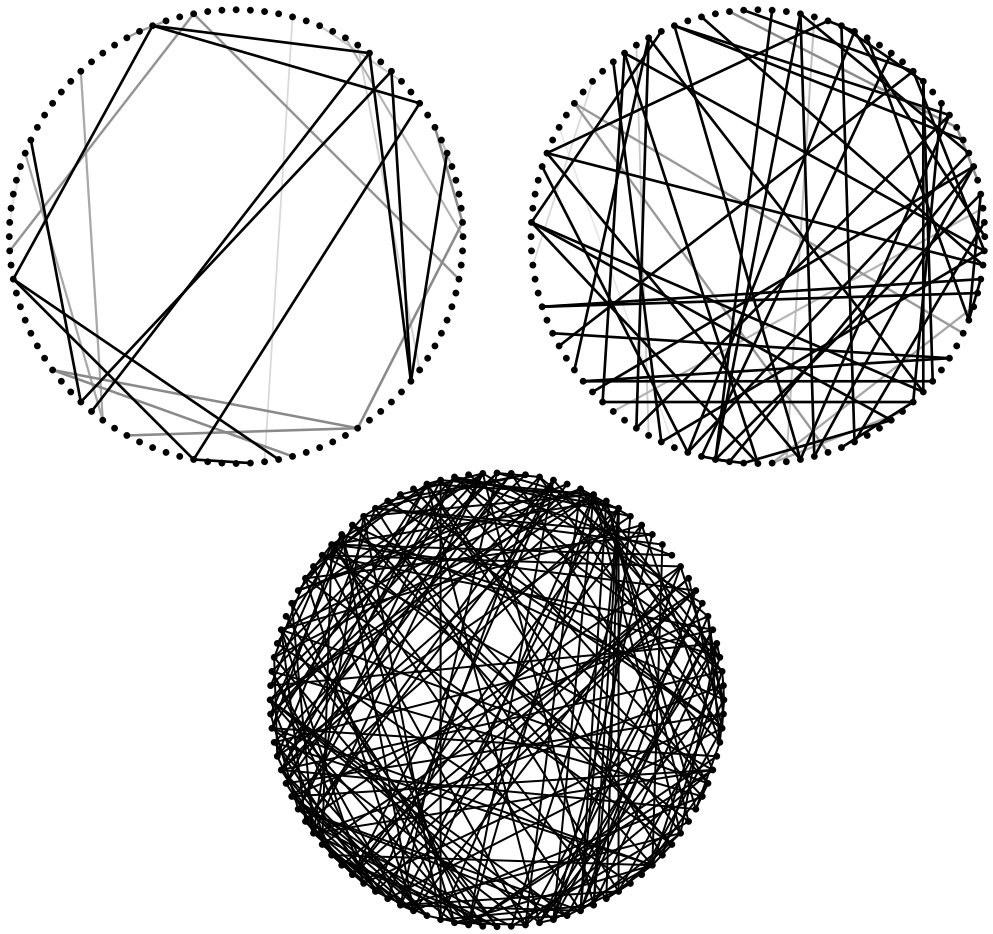

Figure 2. Trois graphes aléatoires d'Erdös-Rényi de taille $n = 100$ et paramètres $p = 0.007$, 0.015 et 0.05. Lorsqu'il y a plusieurs composantes connexes, la plus grande est en noir et les autres en gris.

Selon la valeur du paramètre p, on observe trois comportements très différents. Si p est suffisamment petit, alors le graphe a de nombreuses composantes connexes, toutes de "petite" taille (par rapport à n). Puis, à partir d'une certaine valeur p_1, le graphe aléatoire a une unique composante "géante", de taille de l'ordre de n, et ses autres composantes

connexes sont très petites. Ainsi, pour $n = 100$ et $p = 0.015$, sur l'exemple précédent, on a une composante connexe géante de taille 56 :

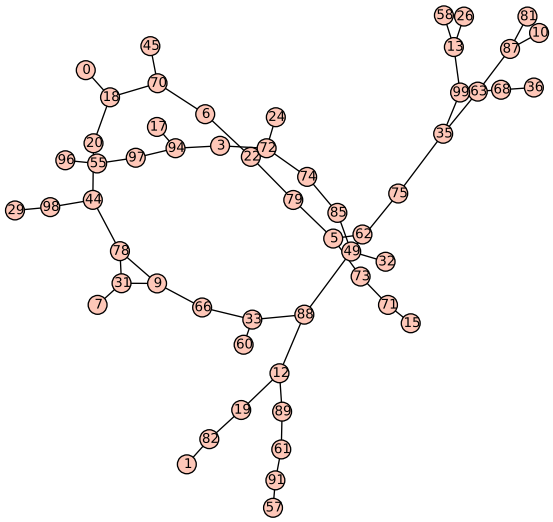

FIGURE 3. La composante connexe géante du graphe d'Erdös-Rényi de paramètres $n = 100$ et $p = 0.015$.

et toutes les autres sont de taille plus petite que 3. Enfin, si l'on augmente encore la valeur de p, alors à partir d'une certaine valeur p_2, le graphe est connexe avec très grande probabilité.

2. Limite d'échelle gaussienne

Il existe plusieurs limites d'échelle pertinentes pour les graphes aléatoires d'Erdös-Rényi. Dans un premier temps, on peut supposer $p_n = p$ indépendant de n, et décrire le comportement asymptotique de G_n en utilisant les *nombres de sous-graphes*, définis comme suit. Si $G = (V_G, E_G)$ et $H = (V_H, E_H)$ sont deux graphes, on dit que H est un *sous-graphe* de G s'il existe une application injective $i : V_H \to V_G$ telle que, si $i(\{v, w\}) = \{i(v), i(w)\}$ pour $e = \{v, w\} \in E_H$, alors

$$\forall e \in E_H, \;\; i(e) \in E_G$$

Ainsi, on peut plonger H dans G. On note $I(H, G)$ le nombre d'injections i vérifiant les hypothèses précédentes, et

$$N(H, G) = \frac{I(H, G)}{I(H, H)}$$

le nombre de sous-graphes de type H dans G. Par exemple, si

$H_2 =$ $H_3 =$ 

alors $I(H_2, H_2) = 2$ et $I(H_3, H_3) = 6$ sont les nombres d'automorphismes des graphes H_2 et H_3 (permutations des sommets qui sont compatibles avec les arêtes). Pour un autre graphe G, $N(H_2, G)$ et $N(H_3, G)$ sont respectivement le nombre d'arêtes $\{i, j\}$ de G ; et le nombre de triangles $\{i, j, k\}$ tels que $\{i, j\}$, $\{j, k\}$ et $\{i, k\}$ sont des arêtes de G.

Le nombre d'arêtes de $G_n = G(n,p)$ suit une loi binomiale $\mathcal{B}(\binom{n}{2}, p)$, donc satisfait le théorème central limite :

$$\frac{N(H_2, G_n) - \binom{n}{2} p}{n \sqrt{\frac{p(1-p)}{2}}} \longrightarrow^{\text{loi}} \mathcal{N}(0,1), \tag{1}$$

où $\mathcal{N}(0,1)$ désigne une loi normale centrée de variance 1. Pour le nombre de triangles, notons $n^{\downarrow k}$ la factorielle décroissante $n(n-1)(n-2)\cdots(n-k+1)$; c'est le nombre d'injections d'un ensemble à k éléments dans un ensemble à n éléments. On peut calculer

$$\mathbb{E}[N(H_3, G_n)] = \frac{n^{\downarrow 3}}{6} p^3; \tag{2}$$

$$\mathbb{E}[(N(H_3, G_n))^2] = \frac{1}{36} \left(n^{\downarrow 6} p^6 + 9 n^{\downarrow 5} p^6 + 18 n^{\downarrow 4} p^5 + 6 n^{\downarrow 3} p^3 \right), \tag{3}$$

où pour le second moment on a compté les paires de triangles $(\{i \neq j \neq k\}, \{i' \neq j' \neq k'\})$ selon le cardinal de leur intersection. Par suite,

$$\operatorname{var}(N(H_3, G_n)) = n^4 \frac{p^5(1-p)}{2} + O(n^3)$$

et l'on peut alors raisonnablement conjecturer le théorème central limite :

$$\frac{N(H_3, G_n) - \binom{n}{6} p^3}{n^2 \sqrt{\frac{p^5(1-p)}{2}}} \longrightarrow^{\text{loi}} \mathcal{N}(0,1). \tag{4}$$

Cette asymptotique est vérifiée, et plus généralement :

Théorème 1. *Si H est un graphe avec k sommets et h arêtes, et si $G_n = G(n,p)$, alors*

$$\frac{I(H, G_n) - n^{\downarrow k} p^h}{h\, n^{k-1} \sqrt{2p^{2h-1}(1-p)}} \longrightarrow^{\text{loi}} \mathcal{N}(0,1).$$

3. Limite d'échelle poissonienne

Une autre limite d'échelle intéressante est la limite poissonienne, où $p_n = \frac{\lambda}{n}$ avec $\lambda > 0$ fixé. Dans ce cas, le nombre d'arêtes issues d'un sommet (*degré*) suit la loi binomiale $\mathfrak{B}(n-1, \frac{\lambda}{n})$, donc, dans l'approximation poissonienne,

$$\deg(\text{un sommet fixé de } G_n) \longrightarrow^{\text{loi}} \mathcal{P}(\lambda),$$

où $\mathcal{P}(\lambda)$ désigne la loi de Poisson $\mathbb{P}[X = k] = \mathrm{e}^{-\lambda} \frac{\lambda^k}{k!}$. En particulier, chaque agent du réseau a un nombre de voisins/connections directes qui n'explose pas avec la taille n du réseau, ce qui est pertinent pour modéliser des réseaux réels.

Le comportement asymptotique du nombre $N(H, G_n)$ de sous-graphes de type H dans $G(n, \frac{\lambda}{n})$ dépend maintenant de la forme de H. Supposons H connexe, et notons $\operatorname{ex}(H)$ l'*excès* de H, qui est défini par

$$\operatorname{ex}(H) = h - k = \operatorname{card} E_H - \operatorname{card} V_H \geq -1.$$

Si H est connexe, alors l'excès de H est toujours plus grand que -1, et il est égal à -1 si et seulement si H est un arbre (graphe connexe sans cycle).

Proposition 2. *Si H est un graphe avec h arêtes et k sommets, et si $G_n = G(n, \frac{\lambda}{n})$, alors*

$$\mathbb{E}[N(H, G_n)] = \frac{1}{I(H,H)}\, n^{\downarrow k} \left(\frac{\lambda}{n}\right)^h \simeq \frac{1}{I(H,H)}\, \lambda^h\, n^{-\mathrm{ex}(H)}.$$

Par conséquent, si $\mathrm{ex}(H) \geq 1$*, alors* $N(H, G_n) \longrightarrow^{probabilité} 0$.

Autrement dit, les sous-configurations observables de G_n dans la limite poissonienne sont seulement celles d'excès 0 et -1. Lorsque H a excès -1 et est un arbre, l'étude du cas où $H = H_2$ permet de conjecturer que

$$\frac{N(H, G_n) - \mathbb{E}[N(H, G_n)]}{\sqrt{\mathrm{var}(N(H, G_n))}} \longrightarrow^{\mathrm{loi}} \mathcal{N}(0, 1).$$

Supposons finalement que H a excès 0. C'est par exemple le cas lorsque $H = H_3$, ou plus généralement lorsque $H = H_k$ est un cycle de longueur $k \geq 3$; on regardera seulement ce cas, remarquant qu'alors $I(H_k, H_k) = 2k$. Il existe une indexation des sommets de $H = H_k$ par les entiers de $[\![1, k]\!]$, de sorte que les arêtes de H_k sont les $\{j, j+1\}$, $j \in \mathbb{Z}/k\mathbb{Z}$. D'autre part, on notera $((v_1, \ldots, v_k))$ un ensemble de sommets distincts dans $[\![1, n]\!]$, modulo permutation cyclique et retournement du cycle. Ainsi, $((v_1, \ldots, v_k)) = ((w_1, \ldots, w_k))$ si et seulement si les ensembles

$$\big\{\{v_j, v_{j+1}\},\ j \in \mathbb{Z}/k\mathbb{Z}\big\} = \big\{\{w_j, w_{j+1}\},\ j \in \mathbb{Z}/k\mathbb{Z}\big\}$$

sont les mêmes. On a pour tout $r \geq 1$

$$\begin{aligned}(N(H, G_n))^{\downarrow r} &= \text{nombre de } r\text{-uplets de } k\text{-cycles distincts dans } G_n \\ &= \sum_{\substack{V_1 \neq V_2 \neq \cdots \neq V_r \\ V_i = ((v_{i1}, v_{i2}, \ldots, v_{ik}))}} \prod_{i=1}^{r} X((v_{i1}, \ldots, v_{ik}))\end{aligned}$$

où $X((v_1, \ldots, v_k))$ est la variable aléatoire qui vaut 1 si l'application $j \in V_H \to v_j \in [\![1, n]\!]$ est un plongement du cycle H dans G_n (*i.e.*, $\{v_j, v_{j+1}\} \in E_{G_n}$ pour tout $j \in \mathbb{Z}/k\mathbb{Z}$), et 0 sinon. Autrement dit,

$$X((v_1, \ldots, v_k)) = \prod_{j=1}^{k} 1_{\{v_j, v_{j+1}\} \in E_{G_n}}.$$

Or, étant donnés r ensembles cycliques $V_1, \ldots, V_r$, le nombre de paires $\{v_{ij}, v_{i(j+1)}\}$ distinctes est toujours plus grand que le nombre de sommets distincts v_{ij} :

$$\mathrm{card}\,\big\{\{v_{ij}, v_{i(j+1)}\},\ i \in [\![1, r]\!],\ j \in [\![1, k]\!]\big\} \geq \mathrm{card}\,\{v_{ij},\ i \in [\![1, r]\!],\ j \in [\![1, k]\!]\}. \tag{5}$$

Notons m le cardinal à droite dans l'inégalité (5), et supposons dans un premier temps l'inégalité stricte. Alors, la variable aléatoire $\prod_{i=1}^{r} X((v_{i1}, \ldots, v_{ik}))$ est une variable de Bernoulli d'espérance au plus égale à $(p_n)^{m+1} = \left(\frac{\lambda}{n}\right)^{m+1}$, et d'autre part, il y a $\binom{n}{m} \leq n^m$ choix possibles pour un ensemble de sommets $\{v_{ij}\}$ de taille m. On en déduit que dans l'égalité

$$\mathbb{E}[(N(H, G_n))^{\downarrow r}] = \sum_{\substack{V_1 \neq V_2 \neq \cdots \neq V_r \\ V_i = ((v_{i1}, v_{i2}, \ldots, v_{ik}))}} \mathbb{E}\left[\prod_{i=1}^{r} X((v_{i1}, \ldots, v_{ik}))\right] \tag{6}$$

les termes correspondants à des ensembles de sommets $V_1, V_2, \ldots, V_r$ avec l'inégalité (5) strictement vérifiée et $\mathrm{card}\,(V_1 \cup V_2 \cup \cdots \cup V_r) = m$ donnent une contribution d'ordre inférieur à

$$O\left(\left(\frac{\lambda}{n}\right)^{m+1} n^m\right) = O\left(\frac{1}{n}\right).$$

Ainsi, pour calculer l'asymptotique de (6), il suffit de regarder les termes tels que (5) soit une égalité. Cette égalité implique pour un ensemble de cycles $V_1, \dots, V_r$ que si $i \neq i'$, alors les cycles V_i et V_i' sont disjoints, ou sont identiques. Comme la somme porte sur les cycles différents, on conclut que

$$\mathbb{E}[(N(H,G_n))^{\downarrow r}] \simeq \sum_{\substack{V_1,\dots,V_r \text{ cycles disjoints} \\ V_i=((v_{i1},v_{i2},\dots,v_{ik}))}} (p_n)^{kr} = \frac{n^{\downarrow kr}}{(2k)^r} \left(\frac{\lambda}{n}\right)^{kr} \simeq \left(\frac{\lambda^k}{2k}\right)^r.$$

Comme l'unique distribution discrète dont les moments factoriels $\mathbb{E}[X^{\downarrow r}]$ sont les puissances μ^r est la distribution de Poisson $X = \mathcal{P}(\mu)$, on conclut :

Théorème 3. *Si H est un cycle de longueur k, et si $G_n = G(n, \frac{\lambda}{n})$, alors*

$$N(H,G_n) \longrightarrow^{\text{loi}} \mathcal{P}\left(\frac{\lambda^k}{2k}\right).$$

Questions

Pour la rédaction des programmes, on pourra soit écrire le code en un langage de programmation (n'importe lequel), soit donner une description détaillée de l'algorithme (pseudo-code). Les questions difficiles sont signalées d'une étoile.

• On regarde dans un premier temps des graphes d'Erdös-Rényi arbitraires $G(n,p_n)$, avec un paramètre $p = p_n$ qui peut varier avec n.

1.1 Écrire un algorithme `RandomGraph` qui construit un graphe aléatoire d'Erdös-Rényi de paramètres n et p arbitraires. Le résultat sera une liste aléatoire de paires $(i < j)$, les sommets de $G(n,p)$.

1.2 À partir de `RandomGraph`, écrire un programme `DrawGraph` qui dessine le graphe aléatoire, et un autre programme `Components` qui renvoie la liste des tailles des composantes connexes du graphe aléatoire $G(n,p)$. Dessiner un graphe $G(50, 1/25)$. Y-a-t'il une composante connexe géante ?

1.3 On admet que le paramètre p_1, tel que pour $p > p_1$ il y ait avec très grande probabilité une composante connexe géante, est de la forme $p_1 = \frac{\kappa_1}{n}$. À l'aide des programmes précédemment écrits, conjecturer la valeur de κ_1 (on prendra n suffisamment grand, entre 100 et 500).

1.4 (*) On note $I_n = I(G_n)$ le nombre de sommets isolés de $G(n,p)$, c'est-à-dire,

$$I_n = \sum_{i=1}^{n} 1_{\deg(i)=0}.$$

On note $p = \frac{\lambda_n}{n}$. Montrer que si $1 \leq \lambda_n \leq \sqrt{n}$, alors

$$\mathbb{E}[I_n] = n\,\mathrm{e}^{-\lambda_n}\left(1 + O\left(\frac{(\lambda_n)^2}{n}\right)\right);$$

$$\mathrm{var}(I_n) \leq \mathbb{E}[I_n] + \frac{\lambda_n}{n-\lambda_n}(\mathbb{E}[I_n])^2.$$

En utilisant l'inégalité de Bienaymé-Chebyshev, en déduire que si $\lambda_n - \log n \to -\infty$, alors $G(n, \frac{\lambda_n}{n})$ a avec très grande probabilité au moins un sommet isolé, donc en particulier n'est pas connexe. Ainsi, $p_2 \geq \frac{\log n}{n}$.

1.5 On admet que le paramètre p_2, tel que pour $p > p_2$ le graphe $G(n,p)$ soit avec très grande probabilité connexe, est de la forme $p_2 = \kappa_2 \frac{\log n}{n}$ (d'après la question précédente, $\kappa_2 \geq 1$). Écrire un programme qui permette de conjecturer la valeur de la constante κ_2.

• Dans les questions suivantes, p est fixé entre 0 et 1 et ne dépend pas de n (approximation gaussienne).

2.1 Démontrer les formules (1), (2) et (3).

2.2 (*) Montrer plus généralement que si H est un graphe à k sommets et h arêtes, alors

$$\mathbb{E}[I(H, G_n)] = n^{\downarrow k}\, p^h;$$
$$\mathrm{var}(I(H, G_n)) = 2h^2 n^{2k-2}\, p^{2h-1}(1-p) + O(n^{2k-3}).$$

On pourra introduire $X(H; i_1, \ldots, i_k)$, qui vaut 1 si l'injection $j \in [\![1,k]\!] \mapsto i_j \in [\![1,n]\!]$ est un plongement de H dans G_n, et 0 sinon ; et écrire $I(H, G_n)$ comme somme de ces variables.

2.3 Écrire un algorithme `NumberTriangles` qui compte le nombre de triangles dans un graphe aléatoire $G(n,p)$. Vérifier par l'expérience la loi limite (4). On pourra par exemple dessiner un histogramme évaluant la distribution de $N(H_3, G_n)$ avec n grand, et construire des estimateurs de $\mathbb{E}[N(H_3, G_n)]$ et de $\mathrm{var}(N(H_3, G_n))$.

• Finalement, on suppose que $p = \frac{\lambda}{n}$ avec $\lambda > 0$ fixé et qui ne dépend pas de n (approximation poissonienne).

3.1 Rappeler pourquoi on a la convergence en loi $\mathcal{B}(n, \frac{\lambda}{n}) \longrightarrow^{\text{loi}} \mathcal{P}(\lambda)$.

3.2 Démontrer complétement la Proposition 2. Donner des exemples de graphes H avec excès plus grand que 1.

3.3 Décrire un algorithme qui permette de construire tous les graphes connexes avec k arêtes et qui ont excès -1 ou 0.

3.4 Si $X \sim \mathcal{B}(n,p)$, calculer les moments factoriels $\mathbb{E}[X^{\downarrow r}]$ pour $r \geq 1$. En déduire les moments factoriels d'une distribution de Poisson $X \sim \mathcal{P}(\mu)$.

3.5 Utiliser l'algorithme `NumberTriangles` pour vérifier par l'expérience la loi limite donnée par le Théorème 3, dans le cas $H = H_3$ (on dessinera de nouveau un histogramme de $N(H_3, G_n)$, et on construira un estimateur du paramètre de la loi limite).