Probabilités sur le graphe complet : l'exemple des arbres couvrants uniforme et minimal

Grégory Miermont *

26 avril 2016

1 Introduction

1.1 Contexte

Cet exposé a pour but de donner quelques exemples d'analyse de grandes structures combinatoires aléatoires, que l'on peut définir naturellement en termes de modèles simples d'arbres couvrants sur le graphe complet.

Ce thème se situe à l'interface entre différents aspects des probabilités : physique statistique, combinatoire, optimisation combinatoire. Le thème de la physique statistique est de décrire l'état d'un « grand système » à l'aide d'une description microscopique au niveau particulaire : souvent, pour simplifier ou pour des raisons physiques, on considère que ces espaces de configurations sont définis sur les sites d'un graphe fini, par exemple un sous-graphe de $\mathbb{Z}^d$. Les modèles de référence sont la percolation, où chaque arête du graphe est conservée avec une probabilité donnée, indépendamment entre les arêtes, ou le modèle d'Ising du ferromagnétisme. La marche aléatoire auto-évitante uniforme, qui est un modèle naturel décrivant le repliement d'un polymère, traité dans les exposés de Vincent Beffara, en est un autre exemple emblématique. Nous allons nous intéresser à différents modèles d'arbres couvrants. Par ailleurs, en optimisation combinatoire, on cherche à minimiser une fonction aléatoire de l'espace des configurations (souvent interprétée comme une énergie), avec l'idée de décrire un système physique dans un état d'équilibre.

*ÉNS de Lyon & Institut universitaire de France

Nous allons considérer de tels systèmes définis sur le graphe complet, où chaque sommet est connecté à tous les autres. Cette « géométrie » est beaucoup plus simple à décrire que celle du réseau $\mathbb{Z}^d$, par exemple par un comptage entièrement explicite des configurations possibles (d'où le lien avec la combinatoire). mais il est instructif d'étudier les modèles ci-dessus dans ce cas simple : à la fois parce qu'ils peuvent donner une intuition précieuse sur les modèles sur réseau, mais aussi parce que leur simplicité fait surgir des objets fondamentaux que l'on retrouve en de nombreuses autres occasions en probabilités. Les deux objets-clés qui nous intéresseront dans ces notes ont été introduits par David Aldous [5, 6]. Ils sont deux modèles d'arbres « continus » de natures très différentes. Le premier (l'arbre continu aléatoire brownien) est également un acteur-clé de l'exposé d'Igor Kortchemski. Le second est une intrigante structure arborescente où tout sommet est à la fois de degré infini, et n'a pourtant qu'un nombre fini de voisins à distance bornée. Nous suivrons d'assez près le premier article, ainsi que l'article de revue d'Aldous et Steele [10].

1.2 Terminologie et notations

On emploiera les notations $\mathbb{N} = \{1, 2, \ldots\}$ et $\mathbb{Z}_+ = \mathbb{N} \cup \{0\}$. Pour $n \in \mathbb{N}$, on notera $[n] = \{1, 2, \ldots, n\}$.

Graphes. Rappelons qu'un graphe (simple) est un couple $G = (V, E)$, où $V = V(G)$ est un ensemble de *sommets*, et $E = E(G)$ est un ensemble de paires $\{x, y\}$ avec $x, y \in V$ et $x \neq y$. On notera souvent $|G| = \mathrm{Card}\,(G)$ le cardinal de l'ensemble des *arêtes* de G.

On dit que les sommets x, y sont adjacents si $e = \{x, y\} \in E$, ce que l'on note $x \sim y$ pour simplifier, et l'on dit que x, y sont les extrémités de l'arête e. Le graphe est orienté si l'on distingue un ordre $\vec{e} = (x, y)$ pour chaque arête $e = \{x, y\} \in E$. On dit alors que x est l'origine, et y la cible de $\vec{e}$.

Un chemin dans un graphe $G = (V, E)$ est une suite finie $(x_1, x_2, \ldots, x_k)$ de sommets avec $x_i \sim x_{i+1}$ pour tout $i \in \{1, 2, \ldots, k-1\}$, et on dit que $k-1$ est la longueur de ce chemin. On dit que c'est une chaîne si de plus tous les x_i sont deux-à-deux distincts, et un cycle si $x_1 = x_n$ et si $x_1, \ldots, x_{n-1}$ sont deux -à-deux distincts. Le graphe G est connexe si pour tout $x, y \in V$, il existe une chaîne commençant en x et finissant un y.

Un arbre est un graphe $G = (V, E)$ connexe, sans cycle. Si V est fini, cela équivaut au fait que G est connexe et a $\mathrm{Card}\,(V) - 1$ arêtes, ou encore,

que G est sans cycle et a $\mathrm{Card}\,(V)-1$ arêtes. Ainsi, $n-1$ est le nombre minimal (resp. maximal) d'arêtes permettant de connecter n sommets (resp. permettant de ne pas créer de cycle sur n sommets).

Un graphe enraciné est un couple (G,v) où $G=(V,E)$ est un graphe et $v\in V$. Un graphe marqué est un couple (G,f) où $f:E\to\Xi$ est une fonction positive définie sur les arêtes, à valeurs dans un ensemble de marques Ξ. Si $\Xi=(0,\infty)$, on parlera de graphe pondéré. Un sous-graphe *couvrant* de G est un graphe $G'=(V,E')$ avec $E'\subset E$: on peut également le voir comme le graphe marqué (V,E,f) avec $f(e)=\mathbf{1}_{\{e\in E'\}}$. Si G' est un sous-graphe couvrant qui est un arbre, on dit que G' est un arbre couvrant.

On notera $K_n=([n],E(K_n))$ le graphe complet à n sommets étiquetés $[n]=\{1,2,\ldots,n\}$, et avec $E(K_n)=\{\{i,j\}:i,j\in[n],i\neq j\}$.

Notions probabilistes. Les variables aléatoires que nous considérerons seront toujours supposées définies sur un espace de probabilités $(\Omega,\mathcal{F},\mathbb{P})$.

Soit $(X_n,n\geq 1),X$ des variables aléatoires à valeurs dans un espace métrique (M,d). On dit que X_n converge *en loi* vers X si pour toute fonction $f:M\to\mathbb{R}$ continue bornée, on a

$$\mathbb{E}[f(X_n)]\underset{n\to\infty}{\longrightarrow}\mathbb{E}[f(X)]\,.$$

Dans le cas où $M=\mathbb{R}^d$ normé, il suffit en fait de le vérifier pour les fonctions f continues à support compact. Voici quelques faits sur la convergence en loi qui nous serviront au cours du texte.

1. Par un théorème de Skorokhod, la convergence en loi de X_n vers X signifie que l'on peut trouver, sur un espace de probabilités éventuellement différent, des variables aléatoires $(X'_n,n\geq 1),X'$ telles que
   - pour tout $n\geq 1$, X'_n a même loi que X_n
   - X' a même loi que X
   - l'événement $\lim X'_n=X'$ lorsque $n\to\infty$ a probabilité 1 (il est presque sûr).

2. Si X_n est à valeurs dans $\mathbb{R}$, il y a équivalence entre la convergence en loi de X_n vers X et le faut que $\mathbb{P}(X_n>r)$ converge vers $\mathbb{P}(X>r)$, pour tout r tel que $\mathbb{P}(X=r)=0$.

3. Si X_n est à valeurs dans $\mathbb{Z}^d$ pour un entier $d\geq 1$, et si l'on a que pour une suite $a_n\to\infty$

$$\mathbb{P}(X_n=x)=\frac{1}{a_n^d}g\left(\frac{x}{a_n}\right)(1+o(1))\,,$$

ce uniformément pour tout $x = O(a_n)$, où $g : \mathbb{R}^d \to \mathbb{R}_+$ est une densité de probabilités (c'est-à-dire que $\int g = 1$), alors on a que X_n/a_n converge en loi vers X, variable aléatoire dont la loi est donnée par $g(x)\mathrm{d}x$ sur $\mathbb{R}^d$.

2 L'arbre couvrant uniforme

Soit $G = (V, E)$ un graphe fini (simple, non orienté) et connexe. Un arbre couvrant est un sous-graphe $T = (V, E')$, $E' \subset E$, qui est un arbre. On note $\mathbf{A}_G$ l'ensemble des arbres couvrants de G, qui est évidemment fini. On dit qu'une variable aléatoire A sur un espace de probabilités $(\omega, \mathcal{F}, \mathbb{P})$, et à valeurs dans $\mathbf{A}_G$, est un arbre couvrant uniforme de G si

$$\mathbb{P}(A = \mathbf{a}) = \frac{1}{\mathrm{Card}\,(\mathbf{A}_G)}\,, \qquad \mathbf{a} \in \mathbf{A}_G\,.$$

Dans le cas où $G = K_n$, l'ensemble $\mathbf{A}_n = \mathbf{A}_{K_n}$ est tout simplement l'ensemble des arbres sur $[n]$. On peut se demander pourquoi l'on insiste sur le fait que ces arbres "couvrent" K_n, la raison étant que l'ensemble des arbres couvrants d'un graphe est extrêmement étudié. Pour commencer, on peut simplement se demander combien il y en a.

Théorème 2.1 (Cayley)**.** *Le cardinal de* $\mathbf{A}_n$ *est* n^{n-2}*.*

Il existe de très nombreuses preuves de ce résultat, et nous en rencontrerons deux chemin faisant, une partielle et une complète.

Pour $v_0 \in V$ on notera également $\mathbf{A}_G(v_0)$ l'ensemble des arbres couvrants $(\mathbf{a}, v_0)$ de G enracinés en v_0, et dans ce cas, on conviendra que les arêtes $\{u, v\}$ de $\mathbf{a}$ sont systématiquement munies de l'orientation (u, v) telle que $d_{\mathbf{a}}(u, v_0) = d_{\mathbf{a}}(v, v_0) + 1$, où $d_{\mathbf{a}}(v, w)$ est la distance de graphe sur $\mathbf{a}$ entre v et w. Une feuille de $\mathbf{a}$ est par définition un sommet vers lequel aucune arête ne pointe, dans cette orientation canonique.

2.1 Dénombrement

On peut déterminer le cardinal de $\mathbf{A}_G$ en termes du déterminant du laplacien du graphe : soit M la matrice d'adjacence de $G = (V, E)$, qui est la matrice définie par

$$M(v, v') = \mathbf{1}_{\{v \sim v'\}}\,, \qquad v, v' \in V\,,$$

où l'on note $v \sim v'$ si $\{v, v'\} \in E$, et notons $D(v, v') = \deg(v)\mathbf{1}_{\{v=v'\}}$ où $\deg(v) = \mathrm{Card}\,\{v' \in V : v' \sim v\}$ est le degré de v. Enfin, on note $\Delta^G = D - M$.

Clairement, Δ^G n'est pas une matrice inversible puisque $\Delta^G \mathbf{1} = 0$, où $\mathbf{1}$ est le vecteur de $\mathbb{R}^V$ dont toutes les composantes sont égales à 1. En revanche, on peut montrer que si G est connexe, alors la matrice $\Delta^G_{v_0}$ obtenue en retirant la ligne et la colonne correspondant à $v_0 \in V$ est inversible pour tout v_0, et de plus

Théorème 2.2 (Kirchhoff)**.** *On a que* $|\det(\Delta^G_{v_0})| = \mathrm{Card}\,(\mathbf{A}_G)$ *pour tout* $v_0 \in V$.

Nous ne donnons pas la preuve de ce théorème ici. Une application directe au cas de $G = K_n$ et $v_0 = 1$ donne une première preuve du théorème de Cayley ci-dessus, par le calcul du déterminant de la matrice circulante $(n-1) \times (n-1)$:

$$\Delta_1^{K_n} = \begin{pmatrix} n-1 & -1 & \dots & -1 \\ -1 & n-1 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & n-1 \end{pmatrix}.$$

2.2 Simulation

Nous cherchons à présent à décrire la géométrie d'un arbre couvrant aléatoire de K_n. Par exemple, quel est le diamètre typique d'un tel arbre, en fonction de n, et comment ce diamètre évolue-t-il lorsque $n \to \infty$? Un problème naturel similaire consiste à étudier le comportement asymptotique d'une marche aléatoire

$$S_n = X_1 + X_2 + \ldots + X_n$$

où les variables aléatoires $X_1, X_2, \ldots$ sont indépendantes et de même loi, par exemple $\mathbb{P}(X_1 = 1) = 1 - P(X_1 = -1) = 1/2$. En effet, la loi forte des grands nombres stipule que $S_n/n \to 0$ presque surement lorsque $n \to \infty$, tandis que le théorème central limite donne

$$\mathbb{P}\left(a \leq \frac{S_n}{\sqrt{n}} \leq b\right) \underset{n\to\infty}{\longrightarrow} \int_a^b \sqrt{\frac{2}{\pi}} \exp(-2x^2)\,\mathrm{d}x$$

pour tout $a < b$. Ainsi, la marche aléatoire S_n est « typiquement » de l'ordre de $\sqrt{n}$ au sens où la probabilité que S_n n'est pas dans $[-K\sqrt{n}, K\sqrt{n}]$ tend vers 0 lorsque $K \to \infty$ uniformément en n.

Pour pouvoir répondre à ce genre de question pour un arbre uniforme, il ne suffit évidemment pas de connaître le nombre de possibilités, comme il serait insuffisant dans le cas de la marche aléatoire de savoir que S_n est déterminée par les 2^n valeurs que peut prendre $(X_1, X_2, \ldots, X_n)$. On est donc amené à essayer de décrire la structure d'un arbre couvrant uniforme, et pour cela on va avoir recours à un algorithme permettant d'en simuler un. Ce dernier a été proposé par Aldous [4] et Broder au début des années 1990.

Théorème 2.3. *Soit $v_0 \in V$ fixé. On se donne une marche aléatoire $X_0 = v_0, X_1, X_2, \ldots$ au plus proche voisin dans G, issue de v_0, c'est-à-dire que pour tout $x_0 = v_0, x_1, \ldots, x_i$ dans V,*

$$\mathbb{P}(X_{i+1} = y | X_0 = x_0, X_1 = x_1, \ldots, X_i = x_i) = \frac{1}{\deg(x_i)} \mathbf{1}_{\{y \sim x_i\}},$$

dès lors que l'événement par lequel on conditionne est de probabilité non nulle. Soit $T_v = \inf\{i \geq 0 : X_i = v\}$ le premier temps d'atteinte de $v \in V$ par cette marche. Alors le graphe aléatoire $A = (V, \{\{X_{T_v - 1}, X_{T_v}\} : v \in V \setminus \{v_0\}\})$, vu comme enraciné en v_0, est un arbre couvrant uniforme parmi l'ensemble $\mathbf{A}_G(v_0)$ des arbres couvrants de G enracinés en v_0.

On notera que les temps T_v sont tous finis presque surement, par un résultat classique sur les chaînes de Markov, dont nous n'aurons pas besoin explicitement dans ces notes et que nous omettons. Il faut prendre garde dans l'énoncé précédent qu'il n'est pas vrai en général que le graphe A « déraciné » est uniforme dans $\mathbf{A}_G$. Cela devient néanmoins vrai si l'on a aussi choisi le sommet v_0 aléatoirement, selon la loi $\pi(v) = \deg(v) / \sum_{w \in V} \deg(v)$ sur V.

Donnons un point de vue un peu différent sur ce résultat, qui met en évidence son caractère algorithmique.

Algorithme d'Aldous-Broder.

1. Initialiser $S = \{v_0\}$, $Z = \varnothing$, $I = 0$, $y = v_0$.
2. Tant que $S \neq V$, faire
   - $I \leftarrow I + 1$,
   - si $X_I \notin S$ alors $Z \leftarrow Z \cup \{\{y, X_I\}\}$,

– $S \leftarrow S \cup \{X_I\}$
– $y \leftarrow X_I$

3. Retourner Z.

Autrement dit, on construit un sous-graphe de G en traçant les arêtes visitées par la marche aléatoire les unes après les autres, sauf aux moments où l'on retourne sur un sommet déjà visité auparavant : en de tels instants, on se place en ce sommet, mais sans ajouter l'arête qui y mène. L'algorithme termine pour la même raison que précédemment, au premier instant C où $\{v_0, X_1, X_2, \ldots, X_C\} = V$ (ce temps C est appelé temps de couverture issu de v_0).

Notons $R_0 = 0$, et par récurrence, pour tout $i \geq 0$, soit

$$R_{i+1} = \inf\{j > i : X_j \in \{X_0, X_1, \ldots, X_{j-1}\}\},$$

c'est-à-dire que $R_1, R_2, \ldots$ sont les instants successifs où la suite $X_0, X_1, \ldots$ répète une valeur déjà prise par le passé. L'arbre A de l'algorithme peut alors s'interpréter comme le graphe dont les arêtes sont

$$\{\{X_{j-1}, X_j\} : j \geq 1, j \notin \{R_1, R_2, \ldots\}\}.$$

Nous allons donner une preuve de ce résultat *uniquement dans le cas où $G = K_n$ est le graphe complet*. Dans ce cas précis, la marche aléatoire $(X_1, X_2, \ldots)$ est particulièrement simple puisqu'il s'agit tout simplement d'une suite i.i.d. uniforme dans $[n]$, et ce, indépendamment du choix de $X_0 = v_0$. Pour fixer les idées, on prendra $X_0 = 1$ dans l'énoncé qui suit. Néanmoins, clairement, on peut oublier l'enracinement par symétrie.

Suivant Camarri-Pitman [14], on peut donner une description exacte du déroulé de l'algorithme, au sens suivant.

Lemme 2.4. *Soit $m \geq 1$ fixé, et $y_1, y_2, \ldots, y_m, x \in [n]$. On se donne un arbre $\mathbf{a}$ enraciné en 1 sur un sous-ensemble de $[n]$, dont les feuilles sont incluses dans $\{y_1, \ldots, y_m\}$. On note enfin $A[m]$ le sous-arbre de A obtenu en ne gardant que les arêtes $\{X_{j-1}, X_j\}$ avec $j \in \{1, 2, \ldots, R_m - 1\} \backslash \{R_1, R_2, \ldots, R_{m-1}\}$. Alors si $|\mathbf{a}|$ est le nombre d'arêtes de $\mathbf{a}$, et si x est un sommet de $\mathbf{a}$,*

$$\mathbb{P}(X_{R_i - 1} = y_i, 1 \leq i \leq m; X_{R_m} = x; A[m] = \mathbf{a}) = \frac{1}{n^{m+|\mathbf{a}|}}.$$

Démonstration. Notons que la contrainte que toute feuille de $\mathbf{a}$ soit de la forme y_i provient de la construction de l'algorithme : en effet, si i n'est pas

un instant de la forme $R_j - 1$ pour un $j \geq 1$, cela signifie que $i+1$ est le premier instant où le sommet X_{i+1} est visité, et l'arête $\{X_i, X_{i+1}\}$ est dans l'arbre construit par l'algorithme, avec l'orientation (X_{i+1}, X_i). Donc X_i n'est pas une feuille de A.

Notons que l'on a $R_m = m + |A[m]|$ puisque $A[m]$ est constitué de $R_m - m$ arêtes (une arête pour chaque itération de l'algorithme, sauf lorsque cela implique une répétition).

On montre alors le résultat par récurrence sur m. Pour $m = 1$, l'arbre A_1 est nécessairement une chaîne $1 = x_0, x_1, \ldots, x_k$, avec $k = |A_1|$, et $X_{R_1} \in \{x_0, \ldots, x_k\}$. De ce fait, pour $\mathbf{a}$ de cette forme, l'événement $\{X_{R_1-1} = y_1, X_{R_1} = x, A[m] = \mathbf{a}\}$ est l'événement $\{X_1 = x_1, \ldots, X_k = x_k, X_{k+1} = x\}$, et cela a probabilité $1/n^{k+1} = 1/n^{1+|\mathbf{a}|}$.

Supposons le résultat connu au rang m. Notons alors que $A[m+1]$ s'obtient en ajoutant la chaîne $\{X_{R_m}, X_{R_m+1}, \ldots, X_{R_{m+1}-1}\}$ au graphe $A[m]$. Soit alors $\mathbf{a}$ un arbre enraciné en 1 dont les feuilles sont incluses dans un ensemble $\{y_1, \ldots, y_{m+1}\}$, et soit $\mathbf{a}'$ le sous-arbre engendré par la racine et les sommets $y_1, \ldots, y_m$: il s'agit de la réunion des chaînes orientées de $y_1, \ldots, y_m$ vers la racine 1 dans $\mathbf{a}$. La chaîne orientée de y_{m+1} vers 1 rencontre $\mathbf{a}'$ pour la première fois en un sommet x', de sorte que $\mathbf{a}$ est la réunion de $\mathbf{a}'$ avec une chaîne, disons $\{x' = x_0, x_1, \ldots, x_k\}$, avec $k = |\mathbf{a}| - |\mathbf{a}'|$. L'événement $\{X_{R_i-1} = y_i, 1 \leq i \leq m+1; A[m+1] = \mathbf{a}, X_{R_{m+1}} = x\}$ peut alors se réécrire

$$\begin{aligned} &\{X_{R_i-1} = y_i, 1 \leq i \leq m; A_m = \mathbf{a}', X_{R_m} = x'\} \\ &\cap \{X_{R_m+1} = x_1, \ldots, X_{R_m+k} = x_k, X_{R_m+k+1} = x\}, \end{aligned}$$

mais sur le premier événement, on a que $R_m = m + |\mathbf{a}|$, et on voit que le premier événement ne dépend que des variables aléatoires $X_1, X_2, \ldots, X_{m+1+|\mathbf{a}|}$, tandis que le second dépend de $X_{m+2+|\mathbf{a}|}, X_{m+3+|\mathbf{a}|}, \ldots$. Par indépendance, on conclut que

$$\begin{aligned} &\mathbb{P}(X_{R_i-1} = y_i, 1 \leq i \leq m+1; A[m+1] = \mathbf{a}, X_{R_{m+1}} = x) \\ &= \frac{1}{n^{|\mathbf{a}|-|\mathbf{a}'|+1}} \mathbb{P}(X_{R_i-1} = y_i, 1 \leq i \leq m; A_m = \mathbf{a}', X_{R_m} = x') \\ &= \frac{1}{n^{|\mathbf{a}|-|\mathbf{a}'|+1}} \cdot \frac{1}{n^{m+|\mathbf{a}'|}} = \frac{1}{n^{m+1+|\mathbf{a}|}} \end{aligned}$$

où l'on a utilisé l'hypothèse de récurrence à l'avant-dernière étape. D'où le résultat. □

De ce résultat, on déduit le Théorème 2.3 dans le cas particulier $G = K_n$.

Corollaire 2.5. *La probabilité que l'algorithme d'Aldous-Broder sur le graphe complet donne un arbre couvrant* $\mathbf{a}$ *donné est*

$$\mathbb{P}(A = \mathbf{a}) = \frac{1}{n^{n-2}}\,.$$

En particulier, A *est uniforme dans* $\mathbf{A}_n$ *et* $\mathrm{Card}\,(\mathbf{A}_n) = n^{n-2}$.

De plus, la suite $X_{R_1-1}, X_{R_2-1}, \dots$ *est i.i.d. uniforme sur* $\{1, 2, \dots, n\}$*, et indépendante de* A

Démonstration. Soit $\mathbf{a}$ un arbre couvrant de K_n, et $m \geq n$. Si l'on somme la formule du lemme 2.4 sur toutes les familles $y_1, \dots, y_m$ telles que $\{y_1, \dots, y_m\} = [n]$ (on notera $\mathbf{y} = (y_1, \dots, y_m) \in C_{m,n}$), alors en notant que $A[m] = A$ sur l'événement où $(X_{R_i-1}, 1 \leq i \leq m) \in C_{m,n}$,

$$\begin{aligned}
&\mathbb{P}((X_{R_i-1}, 1 \leq i \leq m) \in C_{m,n}; A = \mathbf{a}) \\
&= \sum_{\mathbf{y} \in C_{m,n}, x \in [n]} \mathbb{P}(X_{R_i-1} = y_i, 1 \leq i \leq m; X_{R_m} = x; A[m] = \mathbf{a}) \\
&= \frac{n\mathrm{Card}\, C_{m,n}}{n^{m+1}} \cdot \frac{1}{n^{|\mathbf{a}|-1}} = \frac{\mathbb{P}((X_1, \dots, X_m) \in C_{m,n})}{n^{n-2}}\,.
\end{aligned}$$

Ici, on a utilisé le fait que $|\mathbf{a}| = n - 1$, qui vient du fait général qu'un arbre a toujours un sommet de plus que d'arêtes. Si l'on fait tendre m vers l'infini, chacun des événements $\{(X_{R_i-1}, 1 \leq i \leq m) \in C_{m,n}\}$ et $(X_i, 1 \leq i \leq m) \in C_{m,n}$ a une probabilité qui tend vers 1, n étant fixé. On conclut donc par un passage à la limite que l'on a la formule attendue.

La fin de l'énoncé est obtenu de façon similaire : cette fois on choisit $k \geq 1$ et des entiers $y_1, \dots, y_k \in [n]$ et on écrit, pour tout $m \geq n$,

$$\begin{aligned}
&\mathbb{P}((X_{R_i-1}, k < i \leq m + k) \in C_{m,n}; X_{R_i-1} = y_i, 1 \leq i \leq k; A = \mathbf{a}) \\
&= \sum_{\mathbf{y}' \in C_{m,n}} \mathbb{P}(X_{R_i-1} = y_i, 1 \leq i \leq k; X_{R_{k+i}-1} = y'_i, 1 \leq i \leq m; A = \mathbf{a}) \\
&= \frac{\mathbb{P}((X_1, \dots, X_m) \in C_{m,n})}{n^k \cdot n^{n-2}}\,,
\end{aligned}$$

enfin, on prend la limite $m \to \infty$ et on conclut. □

Remarque. Combien de temps faut-il à l'algorithme pour terminer ? Dans le cas du graphe complet, le temps de couverture est appelé le temps du « collectionneur de coupons », qui correspond au temps nécessaire pour rassembler

une collection complète si on tire chacun des n coupons uniformément au hasard . On montre que ce temps C_n a même loi qu'une somme de variables aléatoires indépendantes de lois géométriques, respectivement de paramètres $1, (n-1)/n, (n-2)/n, \ldots, 1/n$:

$$C_n = \sum_{i=1}^{n} \tau_i$$

où

$$\mathbb{P}(\tau_i > k) = \left(\frac{i-1}{n}\right)^k, \qquad i \in [n], k \geq 0\,.$$

et une application d'une inégalité de Bienaymé-Chebychev montre que l'on a que $C_n/n\ln n$ converge vers 1 en probabilité, c'est-à-dire que pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{C_n}{n\ln(n)} - 1\right| > \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 0\,.$$

On parle du problème du *collectionneur de coupons* : C_n est le temps que doit mettre un collectionneur pour obtenir une collection complète de n images, s'il se procure à chaque instant une image choisie uniformément au hasard dans la collection.

L'algorithme d'Aldous-Broder n'est pas le meilleur connu en termes de rapidité d'exécution : un autre algorithme également très simple, dit *algorithme de Wilson*, le surpasse de ce point de vue [21]. Néanmoins, pour notre propos, c'est l'algorithme d'Aldous-Broder qui donne le point de vue le plus simple sur l'arbre couvrant uniforme.

2.3 Une autre interprétation

Comme nous l'avons dit, l'arbre couvrant uniforme du graphe complet que nous avons décrit ci-dessus n'est rien d'autre qu'un arbre uniforme parmi les n^{n-2} arbres possibles avec sommets étiquetés par $1, 2, \ldots, n$. Il y a d'autres manières que l'algorithme d'Aldous-Broder d'envisager cet arbre d'un point de vue probabiliste, dont le recours aux arbres associés aux processus de branchement (Bienaymé-Galton-Watson). Rappelons que si μ est une mesure de probabilités sur $\mathbb{N}$, que l'on voit comme la loi de reproduction pour des individus d'une population asexuée, on peut lui associer un processus dans lequel chaque individu donne naissance indépendamment des autres à un

nombre d'individus aléatoire de loi μ. À ce processus est à son tour associé l'arbre généalogique correspondant.

On renvoie à l'exposé d'Igor Kortchemski pour la propriété suivante. Nous reviendrons brièvement sur cette interprétation au paragraphe 2.5.

Proposition 2.6. *Soit μ la loi de Poisson de paramètre $\lambda > 0$, et soit A l'arbre généalogique d'un processus de branchement de Bienaymé-Galton-Watson, de loi de reproduction μ, conditionné à avoir n sommets. On étiquette ses sommets uniformément au hasard. Le graphe sur $V = [n]$ obtenu est alors un arbre couvrant uniforme de K_n.*

2.4 Géométrie asymptotique : l'arbre continu aléatoire

Notons A_n une variable aléatoire qui est un arbre couvrant uniforme de K_n. On insiste sur la dépendance en n, ce dernier paramètre allant être envoyé vers l'infini. Que peut-on dire de la géométrie de A_n ?

2.4.1 Distance entre deux points marqués

L'algorithme d'Aldous-Broder permet de donner une réponse à cette question. Commençons par étudier la distance de graphe dans A_n entre deux sommets choisis uniformément au hasard (ou, ce qui revient au même, entre la racine 1 et un sommet choisi au hasard, pour des raisons claires de symétrie). Le corollaire 2.5 montre que cette distance de graphe a même loi que la première répétition $R_1 = R_1^{(n)}$ dans une suite $X_1, X_2, \ldots$. Il y a un lien clair avec un problème très connu en probabilités, appelé le « paradoxe des anniversaires », qui consiste à observer que la probabilité qu'au moins deux individus dans un groupe de m personnes aient la même date d'anniversaire subit une transition abrupte de 0 à 1 lorsque m passe, disons, de 20 à 26.

Ici, le problème est le même : n est, disons, le nombre de dates d'anniversaire possibles (365 dans le véritable problème des anniversaires), et l'on peut voir $R_1^{(n)}$ comme le premier instant où, en observant les individus d'un groupe un à un, l'on s'aperçoit de l'existence d'une coïncidence d'anniversaires.

Proposition 2.7. *Lorsque $n \to \infty$, la variable aléatoire $R_1^{(n)}/\sqrt{n}$ converge en loi vers une variable de loi de Rayleigh, au sens où*

$$\mathbb{P}(R_1^{(n)} > r\sqrt{n}) \underset{n\to\infty}{\longrightarrow} \exp(-r^2/2)\,,$$

où il convient d'interpréter $\exp(-r^2/2)$ *comme la queue de distribution de la loi à densité* $re^{-r^2/2}\mathrm{d}r\mathbf{1}_{\mathbb{R}_+}(r)$ *(loi de Rayleigh).*

Plus généralement, pour tout m *on a la convergence en loi des* m *premiers temps de répétitions* $n^{-1/2}(R_1^{(n)}, \ldots, R_m^{(n)})$ *vers un vecteur limite* $(L_1, \ldots, L_m)$*, dont la loi est à densité :*

$$\ell_1\ell_2\ldots\ell_m \exp\left(-\frac{\ell_m^2}{2}\right)\mathrm{d}\ell_1\mathrm{d}\ell_2\ldots\mathrm{d}\ell_m\mathbf{1}_{\{0<\ell_1<\ell_2<\ldots<\ell_m\}}$$

Démonstration. Remarquons que si a est un entier,

$$\mathbb{P}(R_1^{(n)} > a) = 1\cdot(1-1/n)\cdot(1-2/n)\ldots(1-a/n) = \exp(\sum_{i=1}^{a}\ln(1-i/n))\,.$$

Si a est maintenant autorisé à dépendre de n, mais avec $a = O(\sqrt{n})$, on obtient

$$\mathbb{P}(R_1^{(n)} > a) = \exp(-n^{-1}\sum_{i=1}^{a} i + o(1)) = \exp(-a^2/2n + o(1))\,,$$

d'où le résultat. En fait, on peut être plus précis et constater que

$$\mathbb{P}(R_1^{(n)} = a) = \frac{a-1}{n}\mathbb{P}(R_1^{(n)} > a) = \frac{a}{n}\exp(-a^2/2n + o(1))\,.$$

Cet argument se généralise si l'on considère les répétitions suivantes : par exemple, pour $a < b$, avec $b = O(\sqrt{n})$,

$$\begin{aligned}
&\mathbb{P}(R_1^{(n)} = a, R_2^{(n)} = b) = \mathbb{P}(R_2^{(n)} = b | R_1^{(n)} = a)\mathbb{P}(R_1^{(n)} = a)\\
&= (b/n)\cdot(1-(a-1)/n)\cdot(1-a/n)\ldots(1-(b-1)/n)\mathbb{P}(R_1^{(n)} = a)\\
&= (a/n)(b/n)\exp(-b^2/2n + o(1))\,.
\end{aligned}$$

On déduit le résultat pour $m = 2$ par les caractérisations de la convergence en loi rappelés dans l'introduction, et ceci se généralise sans difficulté à m quelconque. Il convient néanmoins de montrer que l'expression de $\ell_1, \ldots, \ell_m$ donnée dans l'énoncé est bien une densité pour tout m, ce que nous laissons en exercice. □

La loi jointe des variables aléatoires $L_1, L_2, \ldots$ apparaissant dans l'énoncé précédent peut se décrire très simplement en termes d'un processus de Poisson, c'est-à-dire une suite $\eta_1, \eta_2, \ldots$ de variables aléatoires telles que la suite

$(\eta_i - \eta_{i-1}, i \geq 1)$ (avec la convention $\eta_0 = 0$) est formée de variables aléatoires i.i.d. de loi exponentielle de paramètre 1, c'est-à-dire $\mathbb{P}(eta_1 \geq t) = \exp(-t)$. On a en effet que

$$(L_1, L_2, \ldots) \overset{\text{(loi)}}{=} (\sqrt{2\eta_1}, \sqrt{2\eta_2}, \ldots)\,.$$

Nous laissons cette preuve en exercice. On constate donc que les variables aléatoires $L_1, L_2, \ldots$ ont tendance à s'accumuler, au sens où les écarts $L_n - L_{n-1}$ convergent vers 0 presque surement, en contraste avec les η_n. Pour s'en convaincre, on peut utiliser la loi des grands nombres, qui montre que $\eta_n \sim n$ presque surement, et écrire, avec la représentation ci-dessus,

$$L_n - L_{n-1} = \sqrt{2\eta_{n-1}}(\sqrt{1 + \mathbf{e}_n/2\eta_{n-1}} - 1) = O(\ln(n)/n^{1/2})$$

où $\mathbf{e}_n = \eta_n - \eta_{n-1}$, et où l'on a utilisé le fait que

$$\limsup \frac{\mathbf{e}_n}{\ln(n)} \leq 1\,, \qquad \text{presque surement}\,.$$

Ce dernier fait est une conséquence aisée du lemme de Borel-Cantelli : on a en effet, comme $\mathbb{P}(\mathbf{e}_i > x) = e^{-x}$,

$$\mathbb{P}(\mathbf{e}_n > (1+\varepsilon)\ln(n)) = n^{-(1+\varepsilon)}\,,$$

qui est sommable, et donc presque surement $\mathbf{e}_n \leq (1+\varepsilon)\ln(n)$ pour tout n assez grand.

2.4.2 Construction de l'arbre continu

Comment interpréter ce résultat ? Rappelons que l'on peut voir A_n, dans la construction par l'algorithme d'Aldous-Broder, comme un processus de recollement de chaînes de longueurs $R_1^{(n)}, R_2^{(n)} - R_1^{(n)}, R_3^{(n)} - R_2^{(n)}, \ldots$, la chaîne de longueur $R_{m+1}^{(n)} - R_m^{(n)}$ étant branchée sur un sommet uniforme sur l'arbre réduit $A_n[m]$. Nous savons que les différences de longueurs, renormalisées par $\sqrt{n}$, convergent lorsque $n \to \infty$ vers la suite $L_1, L_2 - L_1, L_3 - L_2, \ldots$ de limite nulle que nous avons décrite ci-dessus. Ceci donne une intuition de ce que devrait être la structure limite de A_n, lorsque toutes les distances sont renormalisées par $\sqrt{n}$. La description suivante est appelée la *line-breaking construction* par Aldous [5]. Elle consiste à brancher récursivement des segments réels de longueur $L_i - L_{i-1}, i \geq 1$, en recollant une des extrémités en

un point uniformément choisi parmi la réunion des $i-1$ premiers segments recollés.

Soit $U_1, U_2, \ldots$ une suite de variables aléatoires i.i.d. uniformes sur $[0,1]$, indépendantes de $L_1, L_2, \ldots$. Considérons la distance D_∞ sur $\mathbb{R}_+$ obtenue en « coupant » $\mathbb{R}_+$ aux points L_i, c'est-à-dire que $D_\infty(x,y) = |x-y|$ s'il existe i tel que $x, y \in [L_{i-1}, L_i[$, et $D_\infty(x,y) = \infty$ sinon, où l'on note $L_0 = 0$ par convention. On considère la plus petite relation d'équivalence $\approx$ sur $\mathbb{R}_+$ telle que $L_i \approx U_i L_i$ pour tout $i \geq 1$. Enfin, on considère le quotient métrique de $\mathbb{R}_+$ par $\approx$. Il s'agit de l'espace métrique obtenu en recollant les (paires de) points identifiés par $\approx$. Plus précisément, on note

$$D(x,y) = \inf \left\{ \sum_{i=1}^{k} D_\infty(x_i, y_i) \right\}$$

où l'infimum porte sur toutes les suites finies de points $x_1, y_1, \ldots, x_k, y_k$ avec $x_1 = x, y_k = y$ et $y_i \approx x_{i+1}$ pour $1 \leq i \leq k-1$. Il n'est pas difficile de voir que cet infimum est en fait atteint, et que $D(x,y) = 0$ si et seulement si $x \approx y$. Le quotient métrique est l'espace quotient $\mathbb{R}_+/\approx$ muni de la métrique induite par la (pseudo) distance D, encore notée D.

De façon intuitive, nous avons donc collé l'extrémité gauche de l'intervalle $[L_i, L_{i+1}[$ au point $U_i L_i$, qui appartient à la réunion des i premiers intervalles $[L_{j-1}, L_j[, 1 \leq j \leq i$.

Définition 2.1. *L'arbre continu aléatoire brownien est l'espace complété de $(\mathbb{R}_+/\approx, D)$. On le note $\mathcal{T}$.*

Un autre nom communément utilisé est CRT, pour *Continuum Random Tree*. Pour énoncer quelques propriétés importantes de $\mathcal{T}$, donnons une définition.

Définition 2.2. *Un espace métrique (M,d) est un $\mathbb{R}$-arbre si*

- *pour tout $x, y \in M$, il existe un application isométrique $\phi : [0, d(x,y)] \to M$ avec $\phi(0) = x, \phi(d(x,y)) = y$ (on dit que ϕ est une géodésique de x à y), et*
- *il n'existe pas d'application continue injective $\mathbb{S}^1 \to M$.*

En particulier, si un espace (M,d) est un $\mathbb{R}$-arbre il est connexe (et même connexe par arcs) et sans cycle, au sens de la deuxième propriété caractéristique. On ajoute juste la propriété supplémentaire que la métrique

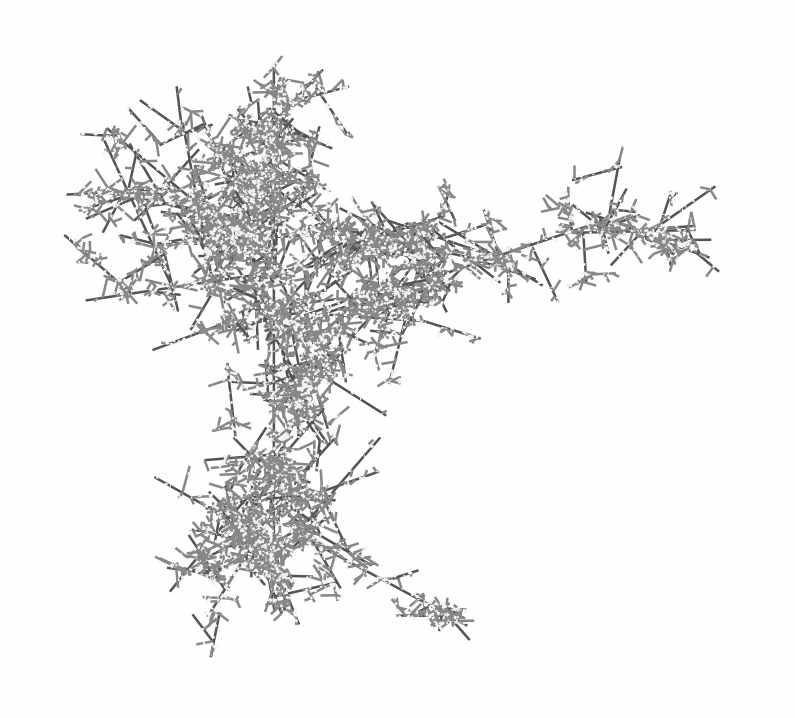

FIGURE 1 – Quelques milliers d'itérations de la *line-breaking construction*, plongée dans $\mathbb{R}^3$ avec une orientation choisie uniformément au hasard pour chaque segment

est une métrique de longueur, c'est-à-dire que la distance entre deux points s'interprète comme la longueur d'un (plus court) chemin continu entre ces points. On peut penser à un $\mathbb{R}$-arbre comme à un recollement de copies isométriques de segments réels, de façon à ne pas créer de cycle, comme le fait la *line-breaking construction*.

Théorème 2.8. *L'espace aléatoire $\mathcal{T}$ est un $\mathbb{R}$-arbre presque surement, et sa dimension de Hausdorff est égale à 2 presque surement.*

Aldous [5], montre la première partie de l'énoncé, ainsi qu'un résultat sur la dimension de Minkowski (le nombre de boules nécessaires pour recouvrir l'arbre), le résultat sur la dimension de Hausdorff apparaît dans Duquesne-Le Gall [17].

Nous ne redonnons pas ici la définition de dimension de Hausdorff, mais insistons sur le côté un peu surprenant que peut avoir ce résultat à première

vue : en effet, le CRT semble être un objet 1-dimensionnel, puisqu'il est obtenu à partir d'un recollement d'une quantité dénombrable de segments réels. Néanmoins, rappelons que $\mathcal{T}$ est en fait la *complétion* d'un tel recollement, et cette opération ajoute une quantité indénombrable de points (ces points sont nécessairement des *feuilles*, c'est-à-dire des points x tels que $\mathcal{T} \setminus \{x\}$ est connexe). Par analogie, on pourra penser aux ensembles de Cantor, que l'on obtient en retirant une collection dénombrable d'intervalles ouverts disjoints de $[0,1]$: les extrémités gauches de ces intervalles forment un sous-ensemble dénombrable dense dans l'ensemble de Cantor correspondant, mais ce dernier est lui-même non-dénombrable.

Enfin, nous avons indiqué dans quelle mesure on peut considérer que $\mathcal{T}$ est un objet qu'il est naturel de considérer comme la limite d'échelle de l'arbre couvrant uniforme A_n. On peut donner un sens précis à cela : on dit que la suite d'espaces métriques compacts (M_n, d_n) converge vers une limite (M, d) au sens de Gromov-Hausdorff s'il existe un espace métrique (Z, δ) et des isométries $\phi_n : M_n \to Z$, $\phi : M \to Z$ telles que

$$\delta_H(\phi_n(M_n), \phi(M)) \underset{n\to\infty}{\longrightarrow} 0\,,$$

où δ_H est la distance de Hausdorff associée à δ entre sous-ensembles fermés de Z, donnée par

$$\delta_H(A, B) = \max\left(\sup_{z\in A} d(z, B), \sup_{z\in B} d(z, A)\right).$$

On peut montrer [13] que cette notion de convergence correspond effectivement à une topologie, et même une métrique, sur l'ensemble des espaces métriques compacts vus à isométrie près. De ce fait, cela a un sens de parler de convergence en loi de variables aléatoires à valeurs dans cet ensemble, muni de la distance de Gromov-Hausdorff.

Théorème 2.9 (Aldous [5])**.** *On a que $(A_n, n^{-1/2} d_{A_n})$ converge vers $\mathcal{T}$ en loi au sens de Gromov-Hausdorff.*

Aldous n'a pas énoncé ce résultat en termes de la distance de Gromov-Hausdorff, mais a en fait fourni une représentation (isométrique) de $\mathcal{T}$ et des A_n renormalisés comme sous-ensembles compacts de $\ell^1(\mathbb{N})$, de sorte que la convergence a lieu en loi au sens de la distance de Hausdorff : cela implique immédiatement le résultat ci-dessus.

Les arguments que nous avons introduits impliquent assez facilement un résultat analogue, si nous remplaçons A_n par $A_n[m]$, et $\mathcal{T}$ par le recollement $\mathcal{T}[m]$ des m premiers segments de la « line-breaking construction », et ce pour tout m fixé. On voit que toute la difficulté pour passer d'un tel résultat à celui exposé ci-dessus est de montrer qu'on peut trouver m assez grand de sorte que $A_n[m]$ soit assez proche de A_n au sens de la distance de Hausdorff, et ce uniformément en n, au sens suivant : pour tout $\varepsilon > 0$,

$$\lim_{m\to\infty} \limsup_{n\to\infty} \mathbb{P}(\Delta(m,n) > \varepsilon\sqrt{n}) = 0$$

où $\Delta(m,n)$ est la distance maximale d'un sommet de A_n à un sommet de $A_n[m]$. Nous ne donnons pas ici la preuve de ce résultat technique, qui se démontre de façon analogue aux *inégalités maximales* sur les marches aléatoires.

2.5 Perspectives sur le CRT

L'arbre Brownien est un objet qui intervient un peu partout en probabilités, un peu au même titre que le mouvement brownien. Suivant l'approche originale d'Aldous, nous avons motivé sa définition par une approche algorithmique du problème de l'arbre couvrant uniforme du graphe complet, en insistant sur la *line-breaking construction* dont la description est peu technique — elle revient après tout à se donner les deux suites infinies $(L_1, L_2, \ldots)$ et $(U_1, U_2, \ldots)$, dont la description est élémentaire.

Néanmoins, une définition plus moderne de l'arbre brownien, qui explique son nom, se fait en termes d'un objet bien plus élaboré, que l'on peut définir en termes du mouvement brownien $(B_t, t \geq 0)$ en dimension 1, et qui consiste à isoler une de ses excursions, c'est-à-dire la restriction de sa trajectoire à une des composantes connexes de l'ouvert $\{t \geq 0 : B_t > 0\}$. Nous ne rentrerons pas dans les détails de cette construction, issue des idées de Neveu, Le Gall et Aldous, et qui est plus proche dans l'esprit de l'approche des arbres décrite dans l'exposé d'Igor Kortchemski.

La première raison qui explique le caractère naturel de l'arbre brownien est le fait qu'il peut se voir comme la limite universelle des processus de branchement, et non seulement la limite de l'arbre couvrant uniforme de K_n, qui, rappelons-le, peut se voir comme l'arbre généalogique d'un processus de branchement dont la loi de reproduction est une loi de Poisson (et conditionné à avoir n individus).

Théorème 2.10 (Aldous [7]). *Soit $\mu = (\mu(k), k \geq 0)$ une mesure de probabilités sur $\mathbb{Z}_+$, telle que*

$$\mu(1) < 1\,, \quad \sum_{k\geq 0} k\mu(k) = 1\,, \quad \text{et} \quad \sigma^2 = \sum_{k\geq 0} k^2\mu(k) - 1 < \infty\,.$$

Soit A_n l'arbre généalogique d'un processus de branchement de Bienaymé-Galton-Watson de loi de reproduction μ, conditionné à avoir n sommets (on suppose que n est restreint aux entiers tels que le conditionnement a un sens.)

Alors $(A_n, (\sigma/\sqrt{n})d_{A_n})$ converge dans le même sens que le Théorème 2.9 vers le CRT $\mathcal{T}$.

Par ailleurs, le CRT est présent, comme on l'a dit, dans l'étude asymptotique de très nombreux problèmes combinatoires. On pourra citer le problème de parking de Knuth [15], les arbres sur réseau [16], les processus de coalescence et de fragmentation [8, 12], les modèles d'attachement préférentiel, les partitions non-croisées, les laminations discrètes du cercle (voir les exposés d'Igor Kortchemski), et les cartes aléatoires [20]. C'est peut-être dans ce dernier domaine, qui consiste à étudier de grands graphes aléatoires plongés dans le plan, que l'intervention du CRT est la plus surprenante, puisque ce dernier semble émerger d'une géométrie de type « champ moyen » (dans le graphe complet, tous les sommets sont voisins les uns des autres), alors que les cartes planaires sont des objets intrinsèquement 2-dimensionnels.

3 L'arbre couvrant minimal

On s'intéresse à présent à un modèle d'arbre couvrant aléatoire de K_n complètement différent, et qui correspond à ce que l'on appelle un « problème d'optimisation combinatoire ». Comme tout problème d'optimisation, il s'agit de minimiser une certaine fonction (« l'énergie ») définie sur un espace de configurations. Dans un problème d'optimisation « combinatoire », on insiste sur le fait que l'espace de configurations est fini, mais, typiquement, beaucoup trop grand pour espérer résoudre le problème par une fouille exhaustive des configurations. Beaucoup de ces problèmes peuvent se poser en termes d'un graphe pondéré $G = (V, E, w)$, c'est-à-dire un graphe muni d'une fonction $w : E \to (0, \infty)$ de poids sur les arêtes du graphe. Parmi ceux-ci :

– trouver l'arbre couvrant dont le poids $w(\mathbf{a}) = \sum_{e\in E(\mathbf{a})} w(e)$ est le plus petit possible : c'est le problème de l'arbre couvrant minimal.

– si (V, E) est le graphe complet à n sommets, trouver le cycle Hamiltonien (visitant tous les sommets une fois et une seule), disons $C = (x_0, x_1, x_2, \ldots, x_n = x_0)$ dont le poids $w(C) = \sum_{i=0}^{n} w(\{x_{i-1}, x_i\})$ est le plus petit possible : c'est le problème du voyageur de commerce (*travelling salesman problem* en anglais).

Les deux problèmes présentés sont emblématiques et sont situés en quelque sorte aux deux extrémités du spectre en termes de difficulté : le problème du voyageur de commerce est *NP-hard*, c'est-à-dire « au moins aussi difficile que tout problème NP » (un problème NP est un problème dont il est « facile » de vérifier si une solution donnée est effectivement une solution).

À l'inverse, le problème de l'arbre couvrant minimal est des plus simples, au sens où il peut se résoudre par des algorithmes *gloutons*, c'est-à-dire faisant à chaque étape un choix d'optimum *local*, et fonctionnant de plus en temps polynomial. Néanmoins, le problème devient étonnamment riche lorsqu'on y ajoute une part d'aléa.

Reprenons le cas d'un graphe général fini connexe $G = (V, E)$ Soit donc $w_e, e \in E$ une famille de poids positifs indexés par les arêtes de G. Si $\mathbf{a}$ est un arbre couvrant de G, on note

$$w(\mathbf{a}) = \sum_{e \in E(\mathbf{a})} w_e \,,$$

appelé poids de $\mathbf{a}$. Par la suite, on supposera **toujours** que les poids sont deux-à-deux distincts.

Définition 3.1. *L'arbre couvrant minimal de G est l'arbre $\mathbf{a}$ qui réalise le minimum des poids $w(\mathbf{a})$.*

Il est clair que l'arbre couvrant minimal existe, puisqu'il n'y a qu'un nombre fini d'arbres couvrants. En revanche, on peut se demander s'il est bien défini de façon unique.

Proposition 3.1. *L'arbre couvrant d'un graphe pondéré dont les poids sont distincts deux-à-deux est unique.*

Démonstration. Supposons qu'il existe deux arbres couvrants minimaux distincts $\mathbf{a}, \mathbf{a}'$. Soit e l'arête de plus petit poids qui est dans l'un mais pas dans l'autre : on suppose sans perte de généralité que e est une arête de $\mathbf{a}$. Alors le graphe $\mathbf{a}' \cup \{e\}$ contient un cycle C. Par définition, toutes les arêtes de poids plus petit que $w(e)$ qui sont dans $\mathbf{a}'$ sont aussi dans $\mathbf{a}$, et par conséquent, il

existe une arête e' de C dont le poids est plus grand que celui de e (sinon toutes les arêtes de C seraient dans $\mathbf{a}$, mais ce dernier ne contient pas de cycle). Finalement, le graphe $(\mathbf{a}' \cup \{e\}) \setminus \{e'\}$ est un arbre couvrant, de poids $w(\mathbf{a}') - w(e') + w(e) < w(\mathbf{a}')$. Contradiction avec la minimalité de $w(\mathbf{a}')$. □

3.1 L'algorithme de Kruskal, et applications

L'algorithme de Kruskal (qui, semble-t-il, remonte en fait à Boruvka) est un algorithme « glouton ». Nous l'énonçons dans le cas d'un graphe fini connexe pondéré général $G = (V, E, w)$. L'algorithme consiste, partant du graphe vide, à ajouter une à une les arêtes du graphe par ordre croissant de poids, sauf lorsque l'ajout de l'arête en question crée un cycle. En pseudo-code, cela donne l'algorithme suivant.

Algorithme de Kruskal.

1. Numéroter les arêtes de $G = (V, E)$ en $e_1, e_2, \ldots$ de sorte que $w(e_1) < w(e_2) < \ldots$.
2. initialiser $S = Z = \varnothing$ et $I = 1$.
3. tant que $S \neq V$ faire
   - si $e_I = \{a, b\} \not\subset S$ alors $S \leftarrow S \cup e_I$, $Z \leftarrow Z \cup \{e_I\}$
   - $I \leftarrow I + 1$.
4. retourner Z.

Théorème 3.2. *L'algorithme de Kruskal produit l'unique arbre couvrant minimal de G pondéré par w.*

Démonstration. Soit $G = (V, E)$ le sous-graphe de K_n produit par l'algorithme. Par définition, G n'a aucun cycle, et donc définit une forêt couvrante du graphe complet, c'est-à-dire que chaque composante connexe est un arbre. Par ailleurs, soit V_1 l'ensemble des sommets de G connectés à 1, et supposons par l'absurde que $V_1 \neq [n]$. L'ensemble (non vide) des arêtes ayant une extrémité dans V_1 et l'autre dans $[n] \setminus V_1$ a un élément de plus petit poids, et clairement, cette arête aurait dû être ajoutée à l'étape de l'algorithme où elle est considérée. Donc G est un arbre couvrant de K_n.

Il reste à montrer que G est un arbre couvrant minimal. Pour cela, montrons par récurrence que pour tout $i \geq 1$, le graphe $G(i)$ construit au i-ème instant de l'algorithme où une arête est ajoutée peut être prolongé en un arbre couvrant minimal.

Pour $i = 0$ c'est évident en prenant n'importe quel arbre couvrant minimal. Si cela est vrai au rang $i < |V| - 1$, soit $G'(i)$ un arbre couvrant minimal contenant $G(i)$. Soit $e(i+1)$ la $i+1$-ème arête ajoutée dans l'algorithme de Kruskal. Si $e(i+1)$ est dans $G'(i)$ alors il n'y a rien à montrer, et sinon, l'ajout de $e(i+1)$ à $G'(i)$ crée un cycle C. Il existe forcément une arête e' de ce cycle qui n'est pas dans $G(i)$: en effet, dans le cas contraire, c'est l'apparition de e à l'étape i qui aurait créé le cycle C. Alors $G'(i+1) = (G'(i) \setminus \{e'\}) \cup \{e\}$ est un arbre couvrant qui prolonge $G(i+1)$, et son poids est $w(G'(i)) + w(e) - w(e') < w(G'(i))$, ce qui est absurde par minimalité. □

Nous en déduisons le critère suivant déterminant si une arête est dans l'arbre couvrant minimal, qui nous sera très utile par la suite.

Proposition 3.3. *Une arête $e = \{x, y\} \in E$ est dans $E(\mathbf{a}_m)$ si et seulement si pour tout chemin $x = x_1, x_2, \ldots, x_k = y$, il existe une arête $e' = \{x_i, x_{i+1}\}$ telle que $w(e) \geq w(e')$.*

Démonstration. Supposons que $e = e_j$ soit la j-ème arête dans l'ordre croissant de poids. Cette arête est ajoutée à la j-ème étape de l'algorithme de Kruskal si et seulement si les extrémités x et y ne sont pas connectées à l'étape $j-1$. Si tout chemin de x à y passe par une arête de poids $> w(e)$, alors, comme aucune de ces arêtes ne peut être ajoutée avant l'étape j de l'algorithme, on voit que x et y ne peuvent être connectés à cette étape, et donc $e \in E(\mathbf{a}_m)$.

Réciproquement, s'il existe un chemin d'arêtes de $\{e_1, \ldots, e_{j-1}\}$ connectant x et y, c'est-à-dire que x et y sont dans la même composante connexe C du graphe sur les arêtes $\{e_1, \ldots, e_{j-1}\}$, alors l'algorithme de Kruskal à l'étape $j-1$ produit un arbre couvrant de C, et connecte donc x et y. Donc e n'est pas dans $E(\mathbf{a}_m)$. □

Un autre algorithme, dû à Prim, consiste à ajouter l'arête de plus petit poids disponible, mais uniquement parmi celles incidentes à une « graine » fixée au départ. On fixe donc un sommet $v_0 \in V$.

Algorithme de Prim.

1. initialiser $S = \{v_0\}$ et $Z = \varnothing$ et $I = 1$.
2. tant que $S \neq V$ faire
   - trouver l'arête $e = \{x, y\}$ de poids minimal avec $x \in S$ et $y \in V \setminus S$

– $S \leftarrow S \cup \{y\}$ et $Z \leftarrow Z \cup \{e\}$

3. retourner Z.

Nous laissons au lecteur le soin de démontrer que l'algorithme produit effectivement l'arbre couvrant minimal.

3.2 Le « théorème $\zeta(3)$ » de Frieze

Supposons maintenant que les poids $w_e, e \in E(K_n)$ soient des variables aléatoires i.i.d. de loi exponentielle de paramètre 1 : $\mathbb{P}(w_e > r) = e^{-r}$ pour tout $r \geq 0$. La loi exponentielle étant à densité, il est effectivement vérifié presque surement que les poids sont deux-à-deux distincts, comme nous l'exigeons. Tacitement, on travaillera systématiquement en restriction à cet événement de probabilité 1.

On notera M_n l'arbre couvrant minimal associé à la suite des poids ci-dessus. On peut poser au moins deux questions naturelles au sujet de cet arbre :

– quelle est la géométrie asymptotique de M_n lorsque $n \to \infty$, au sens de la première partie (convergence de Gromov-Hausdorff)

– quel est le poids de M_n ?

On connaît essentiellement la réponse à ces deux questions, et nous allons évoquer la seconde avec quelque détail. Le résultat suivant est dû à Frieze en 1985 [19].

Théorème 3.4. *On a que* $\mathbb{E}[w(M_n)] \to \zeta(3)$ *lorsque* $n \to \infty$*, et de plus, pour tout* $\varepsilon > 0$*, on a que*

$$\mathbb{P}(|w(M_n) - \zeta(3)| > \varepsilon) \underset{n\to\infty}{\longrightarrow} 0\,.$$

Il peut paraître surprenant de voir apparaître la quantité $\zeta(3)$ dans ce contexte. Mais la première surprise vient certainement du fait qu'il n'y a aucune renormalisation dans la convergence, alors que les poids sont des variables aléatoires exponentielles, et que l'arbre minimal contient $n-1$ arêtes (comme tout arbre couvrant de K_n). Néanmoins, le mystère est dissipé par le résultat suivant, qui sera utile par la suite.

Proposition 3.5. *Soit* $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ *des variables aléatoires indépendantes exponentielles de paramètre* 1*. Notons* $\mathbf{e}_{(1)} < \mathbf{e}_{(2)} < \ldots < \mathbf{e}_{(n)}$ *le réordonnement*

croissant de $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$. *Alors* $(\mathbf{e}_{(1)}, \ldots, \mathbf{e}_{(n)})$ *a même loi que la suite des sommes partielles renormalisées*

$$\sum_{i=1}^{j} \frac{\mathbf{e}_i}{n-i+1}\,, \qquad 1 \leq j \leq n$$

Démonstration. Soit $f : \mathbb{R}^n \to \mathbb{R}_+$ une fonction mesurable. Alors

$$\mathbb{E}[f(\mathbf{e}_{(1)}, \ldots, \mathbf{e}_{(n)})] = n! \int_{0 \leq x_1 \leq \ldots \leq x_n} \mathrm{d}x_1 \ldots \mathrm{d}x_n e^{-(x_1 + \ldots + x_n)} f(x_1, \ldots, x_n)\,.$$

On fait le changement de variables $y_i = x_i - x_{i-1}, 1 \leq i \leq n$ (avec la convention $x_0 = 0$), ce qui donne

$$\begin{aligned}&\mathbb{E}[f(\mathbf{e}_{(1)}, \ldots, \mathbf{e}_{(n)})] \\ &= n! \int_{R_+^n} \mathrm{d}y_1 \ldots \mathrm{d}y_n e^{-\sum_{i=1}^n (n-i+1) y_i} f(y_1, y_1 + y_2, \ldots, y_1 + \ldots + y_n)\,,\end{aligned}$$

et l'on conclut en utilisant le fait que $\mathbf{e}_1/k$ a une loi à densité $ke^{-kx}\mathrm{d}x\mathbf{1}_{\{x \geq 0\}}$. □

On voit en particulier que tout sommet du graphe complet, qui est incident à $n-1$ arêtes, est typiquement incident à au moins une arête de poids $O(1/n)$, et ce sont certainement ces très petites arêtes qui seront préférées par l'algorithme de Kruskal. Ceci explique l'absence de renormalisation dans le théorème de Frieze.

Nous allons donner l'idée d'une approche introduite par Aldous et Steele [9, 10] pour calculer la limite de $\mathbb{E}[w(M_n)]$. Cette méthode, appelée la *méthode objective*, consiste à construire une limite (une sorte de version $n = \infty$) des objets étudiés. L'approche est longue à mettre en place, mais elle introduit en chemin des objets fondamentaux qui permettent d'obtenir une meilleure intuition du résultat et des objets avec lesquels on travaille. Nous insisterons davantage sur ces objets en eux-mêmes que sur les détails de la preuve.

Pour mettre en action la méthode objective, on est amené à étudier les propriétés *locales* de l'arbre couvrant du graphe complet, c'est-à-dire des propriétés que l'on peut décrire en étudiant seulement un voisinage d'une racine choisie au hasard.

Pourquoi une telle approche a-t-elle une chance de marcher ? Constatons que

$$\begin{aligned}\mathbb{E}[w(M_n)] &= \mathbb{E}\left[\sum_{e\in E(M_n)} w(e)\right] \\ &= \frac{1}{2}\mathbb{E}\left[\sum_{x\in[n]}\sum_{y\in[n]} w(\{x,y\})\mathbf{1}_{\{\{x,y\}\in E(M_n)\}}\right] \\ &= \frac{1}{2n}\sum_{x\in[n]}\mathbb{E}\left[\sum_{y:\{x,y\}\in E(M_n)} nw(\{x,y\})\right] \\ &= \frac{1}{2}\mathbb{E}\left[\sum_{y:\{x_*,y\}\in E(M_n)} nw(\{x_*,y\})\right],\end{aligned}$$

où x_* est une variable aléatoire uniforme dans $[n]$, indépendante de $w(e), e \in E(K_n)$. Nous avons donc exprimé l'espérance de la variable « globale » $w(M_n)$ comme celle de la variable « locale » (autour d'un sommet x_* choisi au hasard) $\sum_{y\sim x_*} nw(\{x_*,y\})$. L'idée est alors de trouver une forme de limite lorsque $n\to\infty$ de cette quantité. En fait, pour des raisons de symétrie, on peut remplacer x_* par 1, et nous sommes amenés à étudier la convergence en loi de

$$\frac{1}{2}\sum_{y:\{1,y\}\in E(M_n)} nw(\{1,y\})\,. \tag{1}$$

Remarquons que la seule convergence en loi d'une suite de variables aléatoires X_n positives ne permet pas de montrer la convergence de l'espérance : il faut pour cela montrer un résultat d'uniforme intégrabilité, c'est-à-dire que $\sup_n \mathbb{E}[X_n\mathbf{1}_{\{X_n>a\}}]\to 0$ lorsque $a\to\infty$. Nous admettrons que ceci est vérifié pour les variables définies en (1).

3.3 La convergence locale

La méthode objective utilise comme principal outil la notion de *convergence locale* d'une suite de graphe, que l'on peut exprimer ainsi : on dit que la suite de graphes G_n à n sommets converge localement vers un graphe enraciné infini aléatoire (G,o) si, en choisissant o_n uniforme parmi les sommets de G,

on a que pour tout r, la boule de rayon r centrée en o_n dans G_n converge en loi vers la boule de centre o et de rayon r dans G (pour la distance de graphe, et on voit $B_r(G,o)$ comme un graphe en y adjoignant toutes les arêtes de G reliant deux sommets à distance au plus r de o), c'est-à-dire que pour tout graphe enraciné (g,ρ) on a

$$\mathbb{P}(B_r(G_n,o_n) \approx (g,\rho)) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(B_r(G,o) \approx (g,\rho))\,,$$

où la notation $\approx$ signifie que les graphes (enracinés) de part et d'autre sont isomorphes.

On peut se demander ce que signifie un « graphe infini aléatoire », puisque cela nécessite de préciser une notion de tribu : en fait, il est facile de voir que la convergence ci-dessus correspond à la convergence en loi pour une métrique sur les graphes connexes enracinés ayant une infinité dénombrable de sommets, par exemple

$$d_{\mathrm{loc}}((G,o),(G',o')) = (1+\sup\{r\geq 0 : B_r(G,o)\approx B_r(G',o')\}+1)^{-1}\,.$$

On considère la tribu borelienne associée.

Si maintenant $G=(V,E,w)$ est pondéré, on peut modifier la définition précédente, en remplaçant les boules combinatoires de rayon r par les boules de rayon r dans la métrique induite par w : $d_w(x,y)=\inf\{\sum_{e\in\gamma:x\to y} w(e)\}$ où l'infimum est pris sur tous les chemins de x à y. On notera $B_r(G,o,w)$ la boule de rayon r pour cette distance, ce que l'on voit comme un graphe (non pondéré, mais que l'on peut munir naturellement de la restriction de w).

On dit que la suite de graphes pondérés enracinés (G_n,o_n,w_n) converge vers le graphe infini enraciné $(G=(V,E),o,w)$ si pour tout $r\notin\{d_w(o,v) : v\in V\}$, et pour tout n assez grand, il existe un isomorphisme de graphes $\phi_{n,r} : B_r(G,o)\to B_r(G_n,o_n)$, tel que $w_n(\phi_{n,r}(e))$ converge vers $w(e)$ pour tout e arête de $B_r(G,o,w)$. La raison pour laquelle on exclut le cas où $r=d_w(o,v)$ pour un $v\in V$ est que les boules de rayon s proche de r peuvent inclure ou exclure v selon que $s<r$ ou $s>r$: on ne peut donc pas exiger d'une suite de graphes approchant G que ses boules de rayon r soient toutes isomorphes.

Ici encore on peut munir l'ensemble des graphes enracinés pondérés infinis dénombrables d'une métrique correspondant à cette notion de convergence,

par exemple

$$d_{\mathrm{loc}}((G,o,w),(G',o',w'))$$
$$= \int_0^\infty \mathrm{d}r e^{-r} \left((\inf \left\{ \sup_{e \in E(B_r(G,o,w))} |w(e) - w'(\phi(e))|, \phi \in \mathrm{Isom}_r \right\} \wedge 1 \right) ,$$

où Isom_r est l'ensemble des isomorphismes de graphes entre $B_r(G,o,w)$ et $B_r(G',o',w')$, ce dernier pouvant être vide.

Définition 3.2. *On dit alors que la suite de graphes pondérés (éventuellement aléatoires) $(G_n, w_n), n \geq 1$, tel que G_n a n sommets, converge localement vers un graphe infini pondéré enraciné aléatoire (G,o,w) si, o_n étant un sommet choisi uniformément dans $V(G_n)$ conditionnellement à G_n, on a que (G_n, o_n, w_n) converge en loi vers (G,o,w) pour la distance d_{loc}.*

La notion de convergence locale a été formalisée par Benjamini et Schramm [11], même si elle apparaît en germe dans des travaux antérieurs. Son importance dans l'étude de grands graphes aléatoires est cruciale.

3.4 Le graphe complet pondéré vu comme un arbre

Revenons à notre but de décrire un espace limite au graphe complet pondéré. L'objet limite se décrit de la façon suivante. Notons $\mathcal{U}$ l'ensemble des mots d'entiers

$$\mathcal{U} = \bigcup_{n \geq 0} \mathbb{N}^n ,$$

où $\mathbb{N}^0 = \{\varnothing\}$. On notera $\mathcal{U}^* = \mathcal{U} \backslash \{\varnothing\}$. On peut interpréter $\mathcal{U}$ comme un arbre infini enraciné en $\varnothing$, et où le sommet correspondant au mot $(u_1, u_2, \ldots, u_n)$, souvent noté $u_1 u_2 \ldots u_n$, est vu comme le u_n-ième enfant du mot $(u_1, \ldots, u_{n-1})$. En tant que graphe, $\mathcal{U}$ est donc muni des arêtes $\{u, uk\}$ avec $u \in \mathcal{U}$ et $k \geq 1$. De façon équivalente, l'ensemble des arêtes s'identifie canoniquement à $\mathcal{U}^*$.

Donnons-nous à présent une famille de variables aléatoires $(\xi_u, u \in \mathcal{U}^*)$, de telle sorte que pour tout $u \in \mathcal{U}$, la famille $(\xi_{uk}, k \geq 1)$ ait même loi qu'un processus de Poisson $(\eta_1, \eta_2, \ldots)$ tel qu'introduit au paragraphe 2.7. On demande de plus que ces familles de variables aléatoires soient indépendantes entre elles.

On interprète $(\xi_u, u \in \mathcal{U})$ comme des longueurs d'arêtes, au sens où ξ_{uk} est la longueur de l'arête $e = \{u, uk\}$, et l'on notera indifféremment $\xi(e)$ cette

quantité. Le graphe pondéré enraciné $\mathfrak{T} = (\mathcal{U}, \varnothing, \xi)$ a été baptisé *Poisson weighted infinite tree* par Aldous, ce qu'on pourrait traduire par *arbre infini avec pondération poissonnienne*. L'acronyme, PWIT, est semble-t-il une facétie de l'auteur.

Théorème 3.6. *La limite locale du graphe complet (K_n, nw) pondéré par des variables exponentielles indépendantes de paramètre* 1*, renormalisées par n, est le PWIT.*

Une chose remarquable est que ce théorème reste valable pour des poids i.i.d. qui suivent une loi admettant une densité f continue dans un voisinage à droite de 0, avec $f(0) = 1$ (ou $f(0) > 0$, mais cela nécessite une renormalisation par une constante multiplicative dans l'énoncé).

Nous ne donnerons pas une preuve complète de ce théorème, mais nous contenterons de quelques remarques élémentaires. Tout d'abord, pour des raisons évidentes de symétrie, le choix de la racine o_n n'importe pas, et peut être aussi bien choisi comme étant le sommet 1 ; ce que nous supposerons.

Le PWIT est bien un graphe localement fini. Nous utiliserons ici et par la suite le fait suivant.

Lemme 3.7. *Si $\mathbf{e}_1, \ldots, \mathbf{e}_k$ sont des variables aléatoires exponentielles de paramètre* 1 *indépendantes, alors leur somme $\mathbf{e}_1 + \ldots + \mathbf{e}_k$ a pour loi la loi Gamma(k) dont la densité par rapport à la mesure de Lebesgue*

$$\frac{x^{k-1}e^{-x}}{k!}\,\mathrm{d}x\,\mathbf{1}_{\{x\geq 0\}}\,.$$

La preuve est une application simple de la formule de changement de variables, et est omise.

Soit $u = u_1u_2\ldots u_n \in \mathcal{U}$ un mot. La loi de $d_\xi(u, \varnothing)$ celle de la somme des variables ξ_v où v décrit l'ensemble des préfixes de u, et ces variables aléatoires sont indépendantes, respectivement de lois Gamma$(u_i), 1 \leq i \leq n$. Leur somme est donc de loi Gamma$(|u|_1)$ par une application du lemme 3.7. Or si X suit une loi Gamma(k) on a que $P(X < r) = \frac{1}{k!}\int_0^r x^{k-1}e^{-x}\mathrm{d}x \leq r^k/k!$. Par conséquent, l'espérance du nombre de sommets à d_ξ-distance inférieure à r de l'originie (à l'exception de celle-ci) est

$$\mathbb{E}\left[\sum_{u\in\mathcal{U}^*}\mathbf{1}_{\{d_\xi(u,\varnothing)\leq r\}}\right] \leq \sum_{u\in\mathcal{U}^*}\frac{r^{|u|_1}}{|u|_1!} = \sum_{k\geq 1}\sum_{l\geq 0}\frac{r^{k+l}}{(k+l)!} \leq \sum_{k,l\geq 0}\frac{r^{k+l}}{k!l!} \leq e^{2r}\,,$$

puisque $(k+l)! \geq k!l!$. D'où le résultat.

La première génération du PWIT. Nous avons donné précédemment la loi du réarrangement croissant $(\mathbf{e}^n_{(1)}, \ldots, \mathbf{e}^n_{(n-1)})$ de $n-1$ variables exponentielles indépendantes : on peut les représenter sous la forme

$$\mathbf{e}^n_{(i)} = \sum_{j=1}^{i} \frac{\mathbf{e}'_j}{n-j}\,, \qquad 1 \leq i \leq n-1\,,$$

où $\mathbf{e}'_i, i \geq 1$ sont également des variables exponentielles indépendantes. Par conséquent, on voit immédiatement que dans cette représentation,

$$n(\mathbf{e}^n_{(i)}, i \geq 1) \underset{n\to\infty}{\longrightarrow} \left(\sum_{j=1}^{i} \mathbf{e}'_i, i \geq 1 \right)\,,$$

(presque surement !) et on reconnaît à droite un processus de Poisson. Si l'on interprète les variables $(\mathbf{e}^n_{(i)}, 1 \leq i \leq n)$ comme les distances aux voisins de la racine dans le graphe complet, on voit donc qu'elles correspondent à la limite aux distances de la racine à ses voisins dans le PWIT. On notera que cette limite décrit en réalité les voisins de la racine qui en sont très proches : la plupart des voisins se trouvent à une distance renormalisée de l'ordre de n, et « disparaissent » à la limite.

La preuve du théorème 3.6 consiste essentiellement à réitérer l'argument ci-dessus génération par génération, en justifiant le fait que les distances des B plus proches voisins de la racine de K_n à leurs B plus proches voisins distincts de la racine peuvent être décrits à la limite par les B premiers atomes d'un processus de Poisson indépendants, et aussi que tous les sommets ainsi visités sont tous distincts. Nous omettons les détails techniques de la preuve. Ici, B est un grand entier fixé lorsque $n \to \infty$.

3.5 L'arbre minimal vu comme une forêt couvrante du PWIT

Rappelons que nous voulons décrire non seulement une limite locale du graphe complet, mais aussi de l'arbre couvrant minimal M_n. On peut se demander à quel objet cela peut correspondre sur le PWIT, puisque ce dernier est *déjà un arbre* ! Mais un moment de réflexion montre que l'on ne saurait voir toutes les arêtes du PWIT comme correspondant à la limite de

l'arbre couvrant minimal, pour la raison suivante. Rappelons-nous la Proposition 3.3 disant qu'une arête $e = \{x, y\}$ est dans l'arbre couvrant minimal si et seulement si tout chemin de x à y emprunte au moins une arête e' de poids $w(e') > w(e)$. Ou autrement dit, si les extrémités de e deviennent déconnectées si l'on enlève toutes les arêtes de poids $\geq w(e)$.

Cette remarque permet de définir l'analogue de l'arbre couvrant minimal dans le PWIT : on voudrait dire que l'arête $e = \{u, uk\}$ est dans la *forêt couvrante minimale* $F(\mathfrak{T})$ si u et uk ne sont pas connectés dans le sous-graphe du PWIT obtenu en retirant toutes les arêtes de poids $\geq \xi(uk)$. Evidemment, cela est encore vide de sens, puisque le PWIT est un arbre, et donc deux sommets ne peuvent être connectés que par un seul chemin. Néanmoins, on peut rattraper cela en disant que les sommets u et uk ne sont pas connectés (hors de l'arête e) si l'une au moins des deux composantes connexes de ces deux sommets, formée des arêtes du PWIT de poids $< \xi(e)$ est finie. À l'inverse, on dira que l'arête e n'est pas dans $F(\mathfrak{T})$ s'il existe deux chemins infinis issus de u et uk et n'empruntant que des arêtes de poids $< \xi(e)$, l'idée étant que dans ce cas, u et uk sont connectés « à l'infini » par un chemin de tels arêtes. Nous allons expliquer pourquoi cette idée a une chance de fonctionner, mais tout d'abord nous donnons le résultat.

Définition 3.3. *Soit $\mathfrak{T} = (\mathcal{U}, \varnothing, \xi)$ le PWIT. La forêt minimale couvrante $F(\mathfrak{T})$ de $\mathfrak{T}$ est le sous-graphe de $\mathfrak{T}$ défini par les arêtes $e = \{u, v\} \in E(\mathfrak{T})$ telles que l'une au moins des deux composantes connexes du sous graphe $(V(\mathfrak{T}), \{e' \in E(\mathfrak{T}) : \xi(e') < \xi(e)\})$ contenant u et v est finie.*

On a alors le résultat suivant.

Théorème 3.8. *On a la convergence locale suivante*

$$(K_n, M_n, nw) \underset{n\to\infty}{\longrightarrow} (\mathfrak{T}, F(\mathfrak{T}), \xi)$$

Il faut expliquer ce qu'on entend dans le théorème précédent par la convergence jointe de K_n et sont arbre couvrant minimal M_n. Ce dernier est un sous-graphe de K_n, il est donc immédiat d'étendre la notion de convergence locale à (K_n, M_n, w), si l'on voit par exemple M_n comme une fonction de marques $m : E(K_n) \to \{0, 1\}$, les pondérations w servant toujours à définir les distances : dans la définition ci-dessus de la distance locale, il faut se restreindre aux isomorphismes $\phi : B_r(G, o, w) \to B_r(G', o', w')$ qui préserve les marques : $m(e) = m'(\phi(e))$ pour tout $e \in E(B_r(G, o, w))$.

Le théorème 3.8 sera admis. On peut penser qu'il est un peu miraculeux que deux sommets voisins incidents à l'arête e du PWIT, parce qu'ils sont reliés à l'infini par des chemins n'empruntant pas e faits d'arêtes de poids $< \xi(e)$, peuvent effectivement être interprétés comme étant reliés ensemble. La raison est que, s'il existe un tel chemin, il en existe en fait énormément, et ces chemins envahissent littéralement le graphe. On a en effet le résultat suivant.

Lemme 3.9. *La composante connexe $\mathfrak{T}_\lambda$ du sous-graphe $(V(\mathfrak{T}), \{e \in E(\mathfrak{T}) : \xi(e) < \lambda\})$ contenant $\varnothing$ est l'arbre généalogique d'un processus de branchement de loi de reproduction Poisson de moyenne λ.*

Démonstration. La propriété fondamentale du processus de Poisson, encore noté $(\eta_1, \eta_2, \ldots)$ est que $N_\lambda = \sup\{i \geq 0 : \eta_i < \lambda\}$ (avec la convention $\eta_0 = 0$) est une variable aléatoire de Poisson de paramètre λ. En effet, pour $n \geq 0$, on a

$$\begin{aligned}
\mathbb{P}(N_\lambda = n) &= \mathbb{P}(\eta_n \leq \lambda < \eta_{n+1}) \\
&= \int_{x_1+\ldots+x_n \leq \lambda < x_1+\ldots+x_n+x_{n+1}} \mathrm{d}x_1 \ldots \mathrm{d}x_n \mathrm{d}x_{n+1}\, e^{-(x_1+\ldots+x_n+x_{n+1})} \\
&= \int_{y>\lambda} \mathrm{d}y e^{-y} \int_{y_1<\ldots<y_n\leq\lambda} \mathrm{d}y_1 \ldots \mathrm{d}y_n \\
&= \frac{\lambda^n\, e^{-\lambda}}{n!}\,,
\end{aligned}$$

comme voulu. Par construction du PWIT, on voit donc que le sous graphe de $\mathfrak{T}$ formé des arêtes de poids $< \lambda$ est constitué des N_λ^u premiers voisins de chaque sommet $u \in V(\mathfrak{T})$, où $N_\lambda^u, u \in V(\mathfrak{T})$ est une famille i.i.d. de loi de Poisson de paramètre λ. Il est immédiat de vérifier la composante connexe de $\varnothing$ a la loi voulue. □

Les résultats standards sur le processus de branchement (rappelons que $\mathbb{E}[N_\lambda] = \lambda$) montrent que $\mathfrak{T}_\lambda$ est fini presque sûrement si $\lambda \in [0, 1]$, et que $\mathbb{P}(|\mathfrak{T}_\lambda| = \infty)$ est l'unique racine $q = q(\lambda)$ dans $]0, 1[$ de l'équation

$$e^{-\lambda q} = 1 - q$$

(voir par exemple l'exposé d'Igor Kortchemski). En fait, il est également connu que sur l'événement où $\mathfrak{T}_\lambda$ est infini, la génération n contient de l'ordre

de $W\lambda^n$ individus, où $W > 0$ est une variable aléatoire. On voit donc qu'il existe une quantité de chemins vers l'infini qui croît de façon exponentielle avec la génération : en fait, si l'approximation de K_n par le PWIT est bonne jusqu'à une hauteur de l'ordre de $c\ln(n)$ pour une constante $c > 0$, on voit qu'une proportion positive des sommets de K_n est reliée aux sommets correspondant à u et uk par des arêtes de poids $< \xi(e)$. Il devient alors plausible que l'on puisse en fait relier ces sommets dans K_n par un tel chemin.

Du théorème 3.8 on en déduit la convergence en loi

$$\frac{1}{2}\sum_{i:\{1,i\}\in E(M_n)} nw(\{1,i\}) \underset{n\to\infty}{\longrightarrow} \frac{1}{2}\sum_{i:\{\varnothing,i\}\in F(\mathfrak{T})} \xi(i)\,,$$

où la quantité à gauche est celle qui apparaît en (1). Comme nous avons admis que nous avions aussi la convergence des espérances, il nous reste à calculer

$$\mathbb{E}\left[\sum_{i:\{\varnothing,i\}\in F(\mathfrak{T})} \xi(i)\right] = \mathbb{E}\left[\sum_{i\geq 1} \xi(i)\mathbf{1}_{\{|\mathfrak{T}_{\xi(i)}|<\infty \text{ ou } |\mathfrak{T}^i_{\xi(i)}|<\infty\}}\right],$$

où $\mathfrak{T}_{\xi(i)}$ est $\mathfrak{T}_\lambda$ pour la valeur $\lambda = \xi(i)$, et $\mathfrak{T}^i_{\xi(i)}$ est la composante connexe contenant le sommet $i \in \mathcal{U}$ de $\{e \in E(\mathfrak{T}) : \xi(e) < \xi(i)\}$. Notons que les événements $\{|\mathfrak{T}^i_{\xi(i)}| = \infty\}$ et $\{\mathfrak{T}_{\xi(i)} = \infty\}$ sont indépendants conditionnellement à $\{\xi(u) : u \notin i\mathcal{U}^*\}$ où u décrit tous les sommets du PWIT, à l'exception des descendants stricts de i, car ils font intervenir des poids d'arêtes sur des générations disjointes du PWIT (à l'exception du poids $\xi(i)$ de l'arête $\{\varnothing, i\}$), et de plus la probabilité conditionnelle du premier est $\mathbb{P}(|\mathfrak{T}^i_{\xi(i)}| = \infty \,|\, \xi(u), u \notin i\mathcal{U}^*) = q(\xi(i))$, où $q(\lambda)$, défini au-dessus, est la probabilité de non-extinction de $\mathfrak{T}_\lambda$. On a donc

$$\mathbb{E}\left[\sum_{i:\{\varnothing,i\}\in F(\mathfrak{T})} \xi(i)\right] = \mathbb{E}\left[\sum_{i\geq 1} \xi(i)\left(1 - q(\xi(i))\mathbf{1}_{\{|\mathfrak{T}_{\xi(i)}|=\infty\}}\right)\right]. \qquad (2)$$

Nous utilisons une dernière propriété des processus de Poisson : qui est que

$$\mathbb{E}\left[\sum_{i\geq 1} H(\eta_i, (\eta_j, j\geq 1, j\neq i))\right] = \int_0^\infty \mathrm{d}\lambda\, \mathbb{E}[H(\lambda, (\eta_j, j\geq 1))]\,,$$

où H est une fonction mesurable positive (pouvant aussi être fonction de variables aléatoires indépendantes de $\eta_1, \eta_2, \ldots$). On laisse au lecteur le soin de montrer cette propriété dans le cas où $H(\lambda, (t_j, j \geq 1))$ est de la forme $f(\lambda)g(t_1, \ldots, t_k)$, le résultat s'en déduisant par un argument de classe monotone.

L'application de cette formule à (2) donne

$$\begin{aligned}\frac{1}{2}\mathbb{E}\left[\sum_{i:\{\varnothing,i\}\in F(\mathfrak{T})} \xi(i)\right] &= \int_0^\infty \lambda(1-\mathbb{P}(|\mathfrak{T}_\lambda| = \infty)q(\lambda)) \\ &= \frac{1}{2}\int_0^\infty \lambda(1-q(\lambda)^2) \\ &= \frac{1}{2}\int_0^\infty \lambda(1-q(\lambda)^2) \\ &= \frac{1}{2}\int_1^\infty \lambda^2 q(\lambda)q'(\lambda)\mathrm{d}\lambda\end{aligned}$$

où l'on a utilisé une intégration par parties, puis le fait que q est constante égale à 1 sur $[0,1]$. Comme on a que $\exp(-\lambda q(\lambda)) = (1-q(\lambda))$, soit $\lambda = -\ln(1-q(\lambda))/q(\lambda)$, on peut faire le changement de variables $q = q(\lambda)$, ce qui donne, avec un dernier changement de variables $u = -\log(1-q)$,

$$\begin{aligned}\frac{1}{2}\int_1^\infty \lambda^2 q(\lambda)q'(\lambda)\mathrm{d}\lambda = \int_0^1 \frac{\log^2(1-q)}{2q}\mathrm{d}q &= \int_0^1 \frac{u^2}{2}\frac{e^{-u}}{1-e^{-u}}\mathrm{d}u \\ = \int_0^1 \frac{u^2}{2}\sum_{k\geq 1} e^{-ku}\mathrm{d}u &= \sum_{k\geq 1}\frac{1}{k^3} = \zeta(3)\,.\end{aligned}$$

C'est ce qu'on voulait !

3.6 Géométrie asymptotique

On peut légitimement se poser la question de la convergence au sens de Gromov-Hausdorff de l'arbre minimal M_n, après renormalisation idoine. C'est l'objet d'un résultat récent [3].

Théorème 3.10. *La suite $(M_n, n^{-1/3}d_{M_n})$ converge en loi au sens de Gromov-Hausdorff vers une limite $\mathcal{M}$, qui est un $\mathbb{R}$-arbre, dont la dimension de Minkowski égale 3 presque surement.*

On voit en particulier que la géométrie asymptotique de M_n est très différente de celle du CRT, dont la dimension (de Hausdorff, ou de Minkowski) est 2. En fait, on peut décrire la limite $\mathcal{M}$ en termes d'une opération un peu sophistiquée sur le CRT, mais nous n'entrerons pas dans ces considérations.

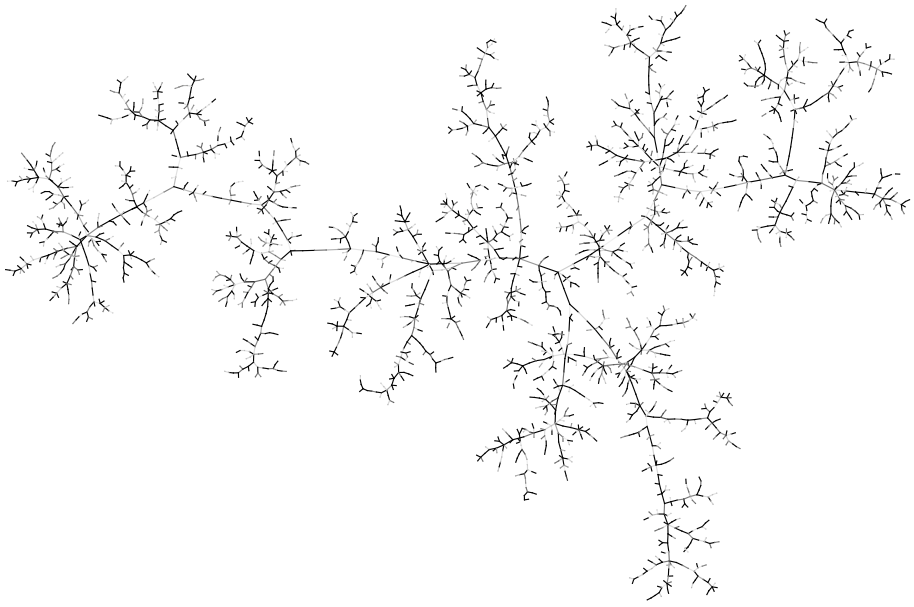

FIGURE 2 – Une simulation de l'arbre couvrant minimal M_n avec $n = 3000$, par Louigi Addario-Berry

Ce résultat fait suite aux importants travaux [2, 1] dont l'objet est de décrire la limite d'échelle du *graphe d'Erdős-Rényi* [18] au paramètre critique, qui consiste à garder chaque arête du graphe complet avec probabilité p, et à enlever les autres. Les composantes connexes de ce graphe, lorsque p est de l'ordre de $1/n$, sont étroitement liées au CRT, tandis que l'émergence de ces composantes alors que p augmente est lié au déroulement de l'algorithme de Kruskal.

Références

[1] L. Addario-Berry, N. Broutin, and C. Goldschmidt. Critical random graphs : limiting constructions and distributional properties. *Electron. J. Probab.*, 15 :no. 25, 741–775, 2010.

[2] L. Addario-Berry, N. Broutin, and C. Goldschmidt. The continuum limit of critical random graphs. *Probab. Theory Related Fields*, 152(3-4) :367–406, 2012.

[3] L. Addario-Berry, N. Broutin, C. Goldschmidt, and G. Miermont. The scaling limit of the minimum spanning tree of the complete graph. *arXiv :1301.1664 [math]*, Jan. 2013. arXiv : 1301.1664.

[4] D. Aldous. The Random Walk Construction of Uniform Spanning Trees and Uniform Labelled Trees. *SIAM J. Discrete Math.*, 3(4) :450–465, Nov. 1990.

[5] D. Aldous. The continuum random tree. I. *Ann. Probab.*, 19(1) :1–28, 1991.

[6] D. Aldous. Asymptotics in the random assignment problem. *Probab. Theory Related Fields*, 93(4) :507–534, 1992.

[7] D. Aldous. The continuum random tree. III. *Ann. Probab.*, 21(1) :248–289, 1993.

[8] D. Aldous and J. Pitman. The standard additive coalescent. *Ann. Probab.*, 26(4) :1703–1726, 1998.

[9] D. Aldous and J. M. Steele. Asymptotics for Euclidean minimal spanning trees on random points. *Probab. Theory Related Fields*, 92(2) :247–258, 1992.

[10] D. Aldous and J. M. Steele. The objective method : probabilistic combinatorial optimization and local weak convergence. In *Probability on discrete structures*, volume 110 of *Encyclopaedia Math. Sci.*, pages 1–72. Springer, Berlin, 2004.

[11] I. Benjamini and O. Schramm. Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab.*, 6 :no. 23, 13 pp. (electronic), 2001.

[12] J. Bertoin. *Random fragmentation and coagulation processes*, volume 102 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2006.

[13] D. Burago, Y. Burago, and S. Ivanov. *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001.

[14] M. Camarri and J. Pitman. Limit distributions and random trees derived from the birthday problem with unequal probabilities. *Electron. J. Probab.*, 5 :no. 2, 18 pp. (electronic), 2000.

[15] P. Chassaing and G. Louchard. Phase transition for parking blocks, Brownian excursion and coalescence. *Random Structures Algorithms*, 21(1) :76–119, 2002.

[16] E. Derbez and G. Slade. The scaling limit of lattice trees in high dimensions. *Comm. Math. Phys.*, 193(1) :69–104, 1998.

[17] T. Duquesne and J.-F. Le Gall. Probabilistic and fractal aspects of Lévy trees. *Probab. Theory Related Fields*, 131(4) :553–603, 2005.

[18] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5 :17–61, 1960.

[19] A. M. Frieze. On the value of a random minimum spanning tree problem. *Discrete Appl. Math.*, 10(1) :47–56, 1985.

[20] J.-F. Le Gall and G. Miermont. Scaling limits of random trees and planar maps. In *Probability and Statistical Physics in two and more Dimensions*, Clay Mathematics Proceedings. American Mathematical Society, 2012. To appear.

[21] D. B. Wilson. Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996)*, pages 296–303. ACM, New York, 1996.