Random trees, graphs and recursive partitions

Nicolas Broutin

Université Pierre et Marie Curie – Paris 6
Ecole Doctorale ED 386

Habilitation à Diriger les Recherches

Spécialité – Mathématiques

Random trees, graphs and recursive partitions

Nicolas Broutin

Rapporteurs:

M. David Aldous	*UC Berkeley*
M. Philippe Chassaing	*Université de Nancy*
M. Remco van der Hofstad	*TU Eindhoven*

Soutenue le 5 juillet 2013 devant le jury composé de:

M. Jean Bertoin	*Université de Zurich*
M. Philippe Chassaing	*Université de Nancy*
M. Thomas Duquesne	*Université Paris 6*
M. Remco van der Hofstad	*TU Eindhoven*
Mme. Claire Mathieu	*Ecole Normale Supérieure*
M. Gilles Schaeffer	*Ecole Polytechnique*

Acknowledgement

Les travaux de David Aldous, Philippe Chassaing, et Remco van der Hofstad ont beaucoup inspiré mes recherches et je suis très honoré qu'ils aient accepté d'être les rapporteurs de ce mémoire.

Je remercie aussi très chaleureusement Jean Bertoin, Philippe Chassaing, Thomas Duquesne, Remco van der Hofstad, Claire Mathieu, et Gilles Schaeffer d'avoir accepté de faire partie de mon jury. Doodle n'a plus de secret pour aucun d'entre eux.

Un grand merci à mes collaborateurs Bruce, Cecilia, Christina, Erin, Gábor, Grégory, Henning, Jean-François, Louigi, Luc, Mikael, Minmin, Nicolas, Olivier, Omar, Philippe, Ralph, Ross H., Ross K., Stefan et Stéfane. Ça a toujours été un plaisir de travailler avec vous, et j'espère de tout coeur que nous continuerons à prouver de jolis théorèmes.

Je n'oublie pas non plus le grand ouest Parisien à l'Inria, ou plutôt, chez Inria je devrais dire... Merci en particulier à tous ceux qui ont contribué à l'ambiance du bâtiment 9: feu l'équipe Algo(righms) – Alin, Bruno, Frédéric, Philippe D. – qui se retrouve maintenant dispersée aux quatre vents. Merci aussi à l'équipe RAP qui m'a gentiment accueilli depuis septembre dernier: Christine, Emanuele, Jim, Nada, Philippe, et Virginie. Bien sûr, il m'est impossible d'oublier Philippe Flajolet, qui tenait une place si importante (dans tous les sens du terme) et qui m'a aidé comme jamais je ne l'aurais espéré.

Finalement, et c'est le plus important, je pense à Patricia qui me supporte et qui éclaire mes journées.

Résumé

Je présente dans ce mémoire mes travaux sur les limites d'échelle de grandes structures aléatoires. Il s'agit de décrire les structures combinatoires dans la limite des grandes tailles en prenant un point de vue *objectif* dans le sens où on cherche des limites des *objets*, et non pas seulement de paramètres caractéristiques (même si ce n'est pas toujours le cas dans les résultats que je présente).

Le cadre général est celui des *structures critiques* pour lesquelles on a typiquement des distances caractéristiques polynomiales en la taille, et non concentrées. Sauf exception, ces structures ne sont en général pas adaptées aux applications informatiques. Elles sont cependant essentielles de part l'universalité de leurs propriétés asymptotiques, prouvées ou attendues.

Je parle en particulier d'arbres uniformément choisis, de graphes aléatoires, d'arbres couvrant minimaux et de partitions récursives de domaines du plan:

- CHAPITRE 2 – ARBRES ALÉATOIRES UNIFORMES. Il s'agit ici de mieux comprendre un objet limite essentiel, l'arbre continu brownien (CRT). Je présente quelques résultats de convergence pour des modèles combinatoires "non-branchants" tels que des arbres sujets aux symétries [P20] et les arbres à distribution de degrés fixée [P22]. Je décris enfin une nouvelle décomposition du CRT basée sur une destruction partielle [P6].

- CHAPITRE 3 – GRAPHES ALÉATOIRES. J'y décris la construction *algorithmique* de la limite d'échelle des graphes aléatoires du modèle d'Erdős–Rényi dans la zone critique [P4], et je fais le lien avec le CRT et donne des constructions *structurelles* de l'espace métrique limite [P3].

- CHAPITRE 4 – ARBRES COUVRANT MINIMAUX. J'y montre qu'une connection avec les graphes aléatoires permet de quantifier les distances dans un arbre convrant aléatoire. On obtient non seulement l'ordre de grandeur de l'espérance du diamètre [P10], mais aussi la limite d'échelle en tant qu'espace métrique mesuré [P5].

- CHAPITRE 5 – PARTITIONS RÉCURSIVES. Sur deux exemples, les arbres cadrant [P23] et les laminations du disque [P25], je montre que des idées basées sur des théorèmes de point fixe conduisent à des convergences de processus, où les limites sont inhabituelles, et caractérisées par des décompositions résursives.

Contents

Chapter 1

Introduction

1.1 General context and overview

This document presents in a synthetic and (hopefully) gentle way the results about scaling limits of large combinatorial structures I have contributed to: we will talk about *random trees*, *random graphs*, but also about *randomized data structures*. The subset of results I will present form, in my opinion, a very coherent collection that has a common story, which I will do my best to tell properly. I also like my other results very much, which are more in line with the research directions which I started pursuing during my PhD (on random geometric graphs, branching random walk, combinatorial testing). However, they now certainly occupy only a smaller proportion of my time, and my research perspectives and long term projects do not seem to modify this trend.

Most of my research revolve around the estimation of distances in combinatorial structures. From the point of view of applications, distances in data structures give a handle on their performances, and distances in networks allow to estimate their navigability. From a theoretical point of view, studying distances is a way a natural way get access to the object itself by considering a graph as a metric space (endowed with the graph distance). It is this point of view that will occupy us from now; when it is possible to rescale the graph distances in such a way that it resembles in the limit a non-trivial (usually compact) random metric space, we call that limit object the *scaling limit*. Most applications to computer science require logarithmic distances for algorithms to be efficient, and unfortunately, the concentration of pairwise distances forbid the existence of any non-trivial and compact, or even interesting, scaling limit. This is why we will focus on some *critical* combinatorial structures, in which distances are polynomial and not concentrated; although it makes them inefficient from a concrete perspective (unless nothing better can be done), they are nevertheless essential since they keep popping up everywhere.

The results we present in the first three chapters are all tightly connected. I always find it tricky to decide how to tell a story: should I take the path that let the audience discover the connections along the way, or tell in advance some of the important guiding ideas, even if that may spoil some of the surprises? Here the story actually starts with Chapter 3, and the question of the metric structure of the minimum spanning tree of a randomly weighted complete graph. This question appears as Research Problem 23 in [53], where it is asked whether the diameter scales like $n^{1/2}$, relaying an earlier idea of Aldous [4]. I learned this question almost ten years ago from Claire Mathieu during my Master research. A few months ago only, I was happy to tell her I knew the answer, and much more: it took some time trying to push a lower bound of $n^{1/3}$ up to $n^{1/2}$ until we realized that we were trying to tighten the wrong end of the range... The first three chapters are in some sense the written version of this tale: understanding distances in the minimum spanning tree was done using a connection with random graphs, and an unexpectedly precise result about the scaling limit of such graphs made it possible to go back to the fine structure of the minimum spanning tree and to construct its scaling limit. Chapter 1 essentially tells about some of the results about (uniformly) random trees which were gathered along the way. I have also included the results

about the distribution of extreme distances in non-plane binary trees, which originally did not make it into the script, but end up fitting nicely in the picture.

The results of Chapter 5 are in some sense fundamentally different: they originate in the analysis of algorithms and the chase for finer estimation of the performance of data structures, and provide a bridge to the topics of my PhD. The main question was initially to estimate the variance of the cost of a search query in a random k-d or quad tree (a multidimensional tree-like data structure). The problem of the expected value had been solved by Flajolet et al. [49, 51], but every tentative to estimate higher moments had resulted in failure due some pernicious mistake... The result of our investigations in [P23] prove the existence of a continuous limit cost process (where the variable is the location of the query). But arguably, although I was not part of it, the most important outcome of this project is the general theory of Banach-space valued random variables using the contraction method by Neininger and Sulzbach [90] on which our result in [P23] heavily relies. I am very happy that Philippe has seen this question he liked very much solved before he left. The results discussed in Section 5.3 are also a by-product of the initial project on the data structure. There, the processes of interest actually encode trees, and yield yet another natural real tree that does not come from an excursion of a Lévy process.

After some short initial preliminaries in Section 1.2, I will present the context and motivations for the results described in this document in Sections 1.3 to Section 1.6. The presentation in the introduction does not aim at giving a comprehensive overview of the background, but merely give enough information to allow the reader to understand the motivation underlying the results.

A list of publications together with links to download the files is available on page 46; the ones which have not been submitted by the time of the redaction of this document should be put on arXiv soon, and are available upon request.

1.2 Preliminaries

1.2.1 Convergence of metric spaces

GROMOV–HAUSDORFF DISTANCE. The metric space approach of the scaling limit requires to compare metric spaces, sometimes measured, and we introduce the relevant distance here. Comparing two subsets of a single metric space is done using the *Hausdorff distance*. To compare two metric spaces, Gromov's idea was to embed them into a single one in the best possible way (so as to minimize the Hausdorff distance). More precisely, given two compact metric spaces (X, d) and (X', d'), define the Gromov–Hausdorff distance $\mathrm{d_{GH}}(X, X')$ between X and X' by

$$\mathrm{d_{GH}}(X, X') := \inf\{\mathrm{d_H^Z}(\phi(X), \phi'(X'))\},$$

where the infimum ranges over the choice of compact metric spaces (Z, d_Z), and isometries $\phi : X \to Z$ and $\phi' : X' \to Z$, and $\mathrm{d_H^Z}$ denotes the Hausdorff distance in Z. The distance $\mathrm{d_{GH}}$ is a pseudo-metric between compact metric spaces, and induces a metric on the quotient space which identifies two compact metric spaces if they are isometric [see, e.g., 42, 56, 74].

GROMOV–HAUSDORFF–PROKHOROV DISTANCE. The analysis of the minimum spanning tree is made easier if we also control the *amount of vertices* at different locations. This requires to consider the graphs as *measured metric spaces*. The "measure" part is controlled using Prokhorov's distance in the same way that the "metric" part is controlled using Hausdorff distance. Let $\mathcal{M}$ be the set of measured isometry-equivalence classes of compact measured metric spaces, and let $\mathrm{d_{GHP}}$ denote the Gromov–Hausdorff–Prokhorov distance on $\mathcal{M}$; the pair $(\mathcal{M}, \mathrm{d_{GHP}})$ forms a Polish space.

SEQUENCES OF MEASURED METRIC SPACES. The random graphs we will study are *not* connected, and to look at them as a whole one actually needs metrics on a space of sequences of metric spaces, or measured metric spaces. The product topology is actually too weak for us, and we introduce ℓ^p-like spaces built on

the Gromov–Hausdorff and Gromov–Hausdorff–Prokhorov distances. For finite sequences, we append an infinite number of metric (or measured metric) spaces consisting of a single point (with no mass).

For two sequences of compact metric spaces $\mathbf{A} = (A_i, i \geq 1)$ and $\mathbf{B} = (B_i, i \geq 1)$ we write

$$\mathrm{d}_{\mathrm{GH}}^{p}(\mathbf{A}, \mathbf{B}) = \left(\sum_{i \geq 1} \mathrm{d}_{\mathrm{GH}}(A_i, B_i)^p \right)^{1/p}.$$

Similarly, if $\mathbf{A} = (A_i, i \geq 1)$ and $\mathbf{B} = (B_i, i \geq 1)$ are now two sequences of compact measured metric spaces, we write

$$\mathrm{d}_{\mathrm{GHP}}^{p}(\mathbf{A}, \mathbf{B}) = \left(\sum_{i \geq 1} \mathrm{d}_{\mathrm{GHP}}(A_i, B_i)^p \right)^{1/p}.$$

1.2.2 Trees, exploration, and encodings

For convenience we write $\mathbb{N} = \{1, 2, \dots\}$ for the set of positive natural numbers. First recall some definitions related to standard rooted plane trees. Let $\mathcal{U} = \bigcup_{n \geq 0} \mathbb{N}^n$ be the set of finite words on the alphabet $\mathbb{N}$, where $\mathbb{N}^0 = \{\varnothing\}$, and $\varnothing$ denotes the empty word. Denote by uv the concatenation of u and v; by convention $\varnothing u = u\varnothing = u$.

A subset T of $\mathcal{U}$ is a *plane tree* if it satisfies the following properties:

- it contains $\varnothing$ (called the root),
- it is stable by prefix (if $uv \in T$ for u and v in $\mathcal{U}$, then $u \in T$), and
- if ($uk \in T$ for some $k > 1$ and $u \in \mathcal{U}$) then $uj \in T$ for j in $\{1, \dots, k\}$.

This last condition appears necessary to get a unique tree with a given genealogical structure. The set of plane trees will be denoted by $\mathbf{T}$.

Notice that the lexicographical order $<$ on $\mathcal{U}$, also named the depth-first order, induces a total order on any tree t; this is of prime importance for the encodings of t we will present. For $t \in \mathbf{T}$, and $u \in t$, let $c_t(u) = \max\{i \ : ui \in t\}$ be the number of children of u in t. The depth of u in t, its number of letters as a word in $\mathcal{U}$, is denoted $|u|$. The notation $|t|$ refers to the cardinality of t, its number of nodes including the root $\varnothing$.

With a tree $t \in \mathbf{T}$, one can associate its degree sequence $\mathbf{s}(t) = (n_i(t), i \geq 0)$, where $n_i(t) = \#\{u \in t : c_t(u) = i\}$ is the number of nodes with degree i in t.

EXPLORATION AND ENCODINGS. We will use the usual encodings: *height process* H and *depth-first walk* S (or Łukasiewicz path). These encodings are defined by first fixing their values at the integral points, and then linear interpolation in between (See Figure 1.1). For a tree $t \in \mathbf{T}$, let $\tilde{u}_1 = \varnothing < \tilde{u}_2 < \cdots < \tilde{u}_{|t|}$ denote the nodes of t sorted according to the lexicographic order. Then we define $H = H_t$ by $H(i) = |\tilde{u}_{i+1}|$, $S = S_t$ by $S_t(i) = \sum_{j=1}^{i} (c_t(\tilde{u}_j) - 1)$; the process H_t is defined on $[0, |t| - 1]$ and S_t on $[0, |t|]$.

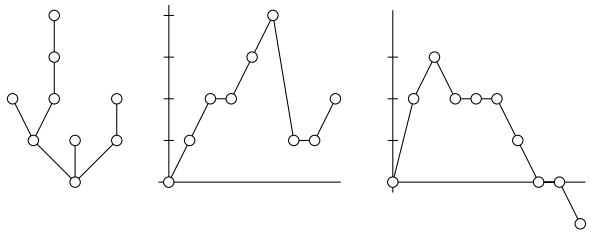

Figure 1.1: *A plane tree $t \in \mathbf{T}$, its height process H_t, and its depth-first walk (Łukasiewicz walk) S_t.*

GALTON–WATSON TREES, SIMPLY GENERATED TREES AND BROWNIAN ASYMPTOTICS. The most classical models of random trees come in two (more or less equivalent) versions the probabilistic point of view of Galton–Watson trees (the family tree of a Galton–Watson process) and the (more) combinatorial version in the *simply generated trees*.

To define the Galton–Watson tree with progeny distribution ξ, one first consider a family of independent copies of ξ, $\{\xi(u), u \in \mathcal{U}\}$. A node $u = i_1 i_2 \dots i_k \in \mathcal{U}$ is in the tree if one has $u = \varnothing$ or $k \geq 1$ and $i_1 \leq \xi(\varnothing)$ and for all $2 \leq j \leq k$ one has $i_j \leq \xi(i_1 i_2 \dots i_{j-1})$.

Simply generated trees are the combinatorial counterpart. In this model trees are sampled with probability proportional to some weight which has a product form in functions of the degrees of all the nodes. One is given a collection of non-negative weights $(a_i, i \geq 0)$. Then, a tree t is given weight

$$w(t) := \prod_{u \in t} a_{c_t(u)} = \prod_{i \geq 0} a_i^{n_i(t)}.$$

Then, a fixed tree t of size n is chosen with probability $w(t)/\sum_{t' \in \mathbf{T}_n} w(t)$. This is easily seen to be similar to the Galton–Watson model.

Some important trees which are not plane, such as uniform labelled trees or Cayley trees, are easily obtained (at least their shape, the labels are irrelevant but for the definition of the distribution) using the previous model. Some others, where the progenies are indistinguishable (e.g., non-plane and unlabelled binary trees) *do not* fit in the model.

1.3 Large random trees

ASYMPTOTICS AND SCALING LIMITS FOR LARGE RANDOM TREES. Investigations of the asymptotics for distances in such random trees started with Rényi and Szekeres [99] who proved in particular that the average height of *labelled* non-plane trees of size n is asymptotic to $2\sqrt{\pi n}$. De Bruijn, Knuth, and Rice [35] gave a similar result for general plane trees. Rényi and Szekeres actually also derived the limit law for the height. The first hints of *universality* are due to Flajolet and Odlyzko [47] who gave the limit law and the convergence of all moments in a paper whose title could not be more humble. They treated the case of Galton–Watson trees with a progeny distribution having small exponential moments.

Theorem 1.1 (Flajolet and Odlyzko [47]). *The height H_n of a random tree taken uniformly from $\mathcal{Y}_n$ admits a* limiting theta distribution*: for any fixed $x > 0$, there holds*

$$\lim_{n \to \infty} \mathbf{P}\left(H_n \geq \frac{2}{\sigma} x \sqrt{n}\right) = \theta(x) := 2 \sum_{k \geq 1} (4k^2 x^2 - 1) e^{-2k^2 x^2}.$$

For those who could read them, this together with the results of Meir and Moon [84] about the altitude of a random node (which says that the distribution of a random node follows a Rayleigh distribution with density $xe^{-x^2/2}$) provides the first signs of *universal Brownian asymptotics*: for a standard Brownian excursion $\mathbf{e}$ and U, an independent random variable uniform on $[0, 1]$, the distribution of $\mathbf{e}(U)$ is Rayleigh while $\sup_{0 \leq s \leq 1} \mathbf{e}(s)$ is distributed according to $\theta(x)$ [68]. Aldous was the first to make the connection explicit for general Galton–Watson trees (see also [73]).

Proposition 1.1 (Aldous [4]). *Let $\mu = (\mu_i, i \geq 0)$ be a distribution with mean one and variance $\sigma^2 \in (0, +\infty)$, and let P_μ be the distribution of a Galton–Watson tree with offspring distribution μ. Along the subsequence $\{n : P_\mu(|\mathbf{t}| = n) > 0\}$, under $P_\mu(\ \cdot\ \big| |\mathbf{t}| = n)$*

$$\left(\frac{H_{\mathbf{t}}(nx)}{\sqrt{n}}\right)_{x \in [0,1]} \xrightarrow[n \to \infty]{} \left(\frac{2\mathbf{e}(x)}{\sigma_\mu}\right)_{x \in [0,1]}$$

in distribution, where $\mathbf{e}$ denotes a standard Brownian excursion, the convergence holding in the space $\mathcal{C}[0, 1]$ equipped with the topology of uniform convergence.

REAL TREES AND THE BROWNIAN CONTINUUM RANDOM TREE. It is possible to re-interpret the previous result about the convergence of the encodings as a convergence of the trees themselves, as metric spaces. The natural scaling limits for large trees are *real trees*. A compact metric space (X, d) is called a real tree if it is geodesic and acyclic:

- for every $x, y \in X$ there exists a unique isometry $\phi_{x,y} : [0, d(x, y)] \to X$ such that $\phi_{xy}(0) = x$ and $\phi_{xy}(d(x, y)) = y$, and
- if q is a continuous injective map from $[0, 1]$ to X such that $q(0) = x$ and $q(1) = y$ then $q([0, 1]) = \phi_{x,y}([0, d(x, y)])$.

Continuous excursions may be seen as encoding real trees [7, 42, 74]: Consider a continuous function $f : [0, 1] \to [0, \infty)$ such that $f(0) = f(1) = 0$ and $f(s) \geq 0$ for all $s \in (0, 1)$. Define $d_f := [0, 1]^2 \to [0, \infty)$ by

$$d_f(x, y) = f(x) + f(y) - 2 \inf\{f(s) : x \wedge y \leq s \leq x \vee y\}.$$

One easily verifies that d_f is a pseudo-metric on $[0, 1]$. Let $x \sim y$ if $d_f(x, y) = 0$. Write $\mathcal{T}_f$ for the quotient $[0, 1]/\sim$; then $(\mathcal{T}_f, d_f)$ is a real tree.

The Brownian continuum random tree is defined as $\mathcal{T}_{2\mathbf{e}}$, the real tree encoded by twice a Brownian excursion [4, 6, 7]. The following version of Proposition 1.1 is due to Le Gall [74].

Theorem 1.2. *Let T_n be a Galton–Watson tree with progeny distribution ξ, conditioned to have size n. Let d_n be the graph distance in T_n. Then,*

$$(T_n, n^{-1/2} d_n) \xrightarrow{d} (\mathcal{T}_{2\mathbf{e}}, \sigma^{-1} d_{2\mathbf{e}})$$

in distribution in the Gromov–Hausdorff sense.

Aldous [4] conjectured that many other models should also rescale as the Brownian CRT. Models of trees which are more combinatorial in nature are harder to represent probabilistically; they are especially interesting since their treatment must put new ideas on the table.

In Sections 2.1 and 2.2, we present our contributions to the analysis of two such models: random non-plane unlabeled binary trees (Otter trees) [P19, P20] and trees with a fixed degree sequence [P22].

CUTTING DOWN TREES AND A NEW DECOMPOSITION OF THE BROWNIAN CRT. Random trees also exhibit some surprising asymptotics. The subject of cutting down trees was introduced by Meir and Moon [82, 83]. One is given a rooted tree T which is pruned by random removal of edges. At each step, only the portion containing the root is retained (we refer to the portions not containing the root as the *pruned* portions) and the process continues until eventually the root has been isolated. The main parameter of interest is the random number of cuts necessary to isolate the root. (The dual problem of isolating a leaf or a node with a specific label has been considered by Kuba and Panholzer [71, 72].)

For conditioned trees emerging from a progeny distribution with variance $\sigma^2 \in (0, \infty)$, once divided by $\sigma\sqrt{n}$, the number of cuts required to isolate the root of a conditioned tree of size n converges in distribution to a Rayleigh random variable with density $xe^{-x^2/2}$ on $[0, \infty)$. (In this form, under only a second moment assumption, this was proved by Janson [64]; below we discuss earlier, partial results in this direction.) The fact that the Rayleigh distribution appears here with a $\sqrt{n}$ scaling in a setting involving conditioned trees struck us as deserving of explanation. Indeed, as we already mentioned, the Rayleigh distribution also arises as the limiting distribution of the length of a path between two uniformly random nodes in a conditioned tree, after appropriate rescaling.

In Section 2.3 we show that the existence of a Rayleigh limit in both cases is not fortuitous. We prove using a coupling method that the number of cuts and the distance between two random vertices are asymptotically equal in distribution (modulo a constant factor). Our proof also yields a novel reversible decomposition of the Brownian continuum random tree.

1.4 Erdős–Rényi random graphs

Since its introduction by Erdős and Rényi [41], the model $G(n,p)$ of random graphs has received an enormous amount of attention [28, 66]. In this model, a graph on n labeled vertices $\{1,2,\dots,n\}$ is chosen randomly by joining any two vertices by an edge with probability p, independently for different pairs of vertices. A simple construction of the entire process $(G(n,p), p \in [0,1])$ consists in assigning an independent $[0,1]$-uniform weight to every pair of vertices, and declaring that two nodes are bound by an edge in $G(n,p)$ if the corresponding weight is at most p. From now on, we will assume that the graphs are coupled in this way.

THE PHASE TRANSITION. This model exhibits a radical change in structure (or *phase transition*) for large n when the average degree approaches one, that is for $p = p(n) \sim 1/n$. For $p \sim c/n$ with $c < 1$, the largest connected component has size (number of vertices) $O(\log n)$. On the other hand, when $c > 1$, there is a connected component containing a positive proportion of the vertices (the *giant component*). The cases $c < 1$ and $c > 1$ are called *subcritical* and *supercritical*, respectively. This phase transition was discovered by Erdős and Rényi in their seminal paper [41]; they further observed that in the *critical* case, when $p = 1/n$, the largest components of $G(n,p)$ have sizes of order $n^{2/3}$. For this reason, the phase transition in random graphs is sometimes dubbed the *double jump*.

THE CRITICAL WINDOW. The apparent double jump is actually only an artefact of the parametrization which is much too crude, and understanding the critical random graph (when $p = p(n) \sim 1/n$) requires a different and finer scaling: the natural parametrization turns out to be of the form $p = p(n) = 1/n + \lambda n^{-4/3}$, for $\lambda = o(n^{1/3})$ [27, 77, 78]. We will restrict our attention to $\lambda \in \mathbb{R}$; this parameter range is then usually called the *critical window*. One of the most significant results about random graphs in the critical regime was proved by Aldous [10]. He observed that one could encode various aspects of the structure of the random graph (specifically, the sizes and number of edges of the components) using stochastic processes. His insight was that standard limit theory for such processes could then be used to get at the relevant limiting quantities which could, moreover, be analyzed using powerful stochastic-process tools. Fix $\lambda \in \mathbb{R}$, set $p = 1/n + \lambda n^{-4/3}$ and write Z_i^n and S_i^n for the size and surplus (that is, the number of edges which would need to be removed in order to obtain a tree) of $\mathcal{C}_i^n$, the i-th largest connected component of $G(n,p)$. Set $\mathbf{Z}^n = (Z_1^n, Z_2^n, \dots)$ and $\mathbf{S}^n = (S_1^n, S_2^n, \dots)$.

Theorem 1.3 (Aldous [10])**.** *As $n \to \infty$.*

$$(n^{-2/3}\mathbf{Z}^n, \mathbf{S}^n) \xrightarrow{d} (\mathbf{Z}, \mathbf{S}).$$

Here, the convergence of the first co-ordinate takes place in $\ell^2_{\searrow}$, the set of infinite sequences $(x_1, x_2, \dots)$ with $x_1 \geq x_2 \geq \cdots \geq 0$ and $\sum_{i\geq 1} x_i^2 < \infty$. (See also [65, 78].) The limit $(\mathbf{Z}, \mathbf{S})$ is described in terms of a Brownian motion with parabolic drift, $(W^\lambda(t), t \geq 0)$, where

$$W^\lambda(t) := W(t) + t\lambda - \frac{t^2}{2}$$

and $(W(t), t \geq 0)$ is a standard Brownian motion. The limit $\mathbf{Z}$ has the distribution of the ordered sequence of lengths of excursions away from zero of the reflected process $W^\lambda(t) - \min_{0\leq s\leq t} W^\lambda(s)$, while $\mathbf{S}$ is the sequence of numbers of points of a Poisson point process with rate one in $\mathbb{R}^+ \times \mathbb{R}^+$ lying under the corresponding excursions. Aldous' limiting picture has since been extended to many other models, with for instance to random graphs with "immigration" [14], hypergraphs [55], percolation on random regular graphs with fixed degree [87], random graphs with a prescribed degree sequence [67, 100], and rank-1 random graphs [26, 62, 103].

In Chapter 3, we explain why the representation underlying Theorem 1.3 actually carries much more information than it seems. In particular, we show that the Brownian process with parabolic drift, and the location of the Poisson points allow to recover the *metric structure of the graph*.

1.5 The minimum spanning tree

Given a connected graph together with edge weights, a *minimum weight spanning tree* or *minimum spanning tree* of G is a connected subgraph of G that minimizes the sum of the weights of its edges. We will assume that all weights are distinct, which ensures that the minimum spanning tree (MST) is unique. Minimum spanning trees have been studied under various models of randomness, in particular the Euclidean model (distances in the graph arise from distances between points in random points $\mathbb{R}^d$) and the mean-field model (distances are independent and identically distributed). *It is the latter mean-field setting we are interested in: We consider a complete graph whose edges are weighted by i.i.d. random variables uniform in* $[0, 1]$.

Although *local* parameters of have received a lot of attention, the *global* structure of the minimum spanning tree remained until recently mostly untouched. The study is partly motivated by Research Problem 23 of [53] who relays a question/conjecture of Aldous [4] that the Brownian CRT might actually be the scaling limit of the minimum spanning tree of the complete graph with random weights (described as IMST in [4], for which the local structure matches).

KRUSKAL'S ALGORITHM. Kruskal's algorithm provides a strong connection between the minimum spanning tree and Erdős–Rényi random graphs. One can build a *forest* $F(n, p)$ as follows: taking the edges in increasing order of weight (which we assume all distinct), add the current edge to the forest unless it creates a cycle; stop before the first edge with weight greater than p. If the weights are the independent and uniform used to construct $G(n, p)$, then the collection of vertex sets of the connected component of $F(n, p)$ and $G(n, p)$ are the same (we only removed edges which were already binding two vertices of the same connected component). In other words, the sizes of the connected components of $F(n, p)$ exhibit the same phase transition as those of $G(n, p)$ when $p \sim 1/n$. Note that distances in $F(n, p)$ never change once they become finite.

In Chapter 4, we explain how this connection between the minimum spanning tree and the random graph allows to locate the range of values of p where the metric structure of the MST is built. This yields estimates of the expected diameter [P9, P10], and a construction of the scaling limit, using knowledge of the scaling limit of the random graphs [P5].

1.6 Random recursive partitions

QUAD TREES AND MULTIDIMENSIONAL SEARCH. The quadtree [46] allows to manage multidimensional data by extending the divide-and-conquer approach of the binary search tree. Consider the point sequence $p_1, p_2, \ldots, p_n \in [0, 1]^2$. As we build the tree, regions of the unit square are associated to the nodes where the points are stored. Initially, the root is associated with the region $[0, 1]^2$ and the data structure is empty. The first point p_1 is stored at the root, and divides the unit square into four regions $Q_1, \ldots, Q_4$, each assigned to a child of the root. More generally, when i points have already been inserted, we have a set of $1 + 3i$ disjoint (lower-level) regions that cover the unit square. The point p_{i+1} is stored in the node (say u) that corresponds to the region it falls in, divides it into four new regions that are assigned to the children of u. See Figure 1.2.

ANALYSIS OF PARTIAL MATCH RETRIEVAL. For the analysis, we will focus on the model of *random quadtrees*, where the data points are independent and uniformly distributed in the unit square. In the present case, the data are just points, and the problem of partial match retrieval consists in reporting all the data with one of the coordinates (say the first) being $s \in [0, 1]$. It is a simple observation that the number of nodes of the tree visited when performing the search is precisely $C_n(s)$, the number of regions in the quadtree that insersect a vertical line at s. The first analysis of partial match in quadtrees is due to Flajolet et al. [51] (after the pioneering work of Flajolet and Puech [49] in the case of k-d trees). They studied the singularities of a differential system for the generating functions of partial match cost to prove that, for a

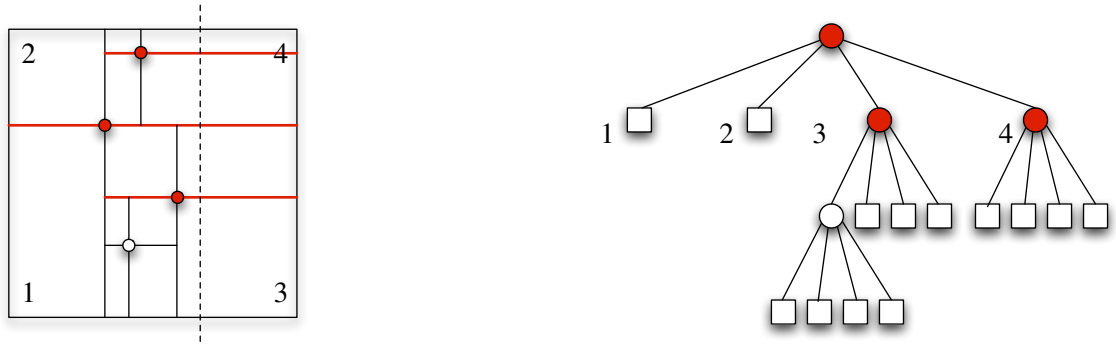

Figure 1.2: *An example of a (point) quadtree: on the left the partition of the unit square induced by the tree data structure on the right (the children are ordered according to the numbering of the regions on the left). Answering the partial match query materialized by the dashed line on the left requires to visit the points/nodes coloured in red. Note that each one of the visited nodes correspond to a horizontal line that is crossed by the query.*

random query ξ, being independent of the tree and uniformly distributed on $[0,1]$,

$$\mathbf{E}[C_n(\xi)] \sim \kappa\, n^{\beta}, \tag{1.1}$$

for some explicit constants κ and $\beta \in (1/2, 1)$.

This analytic approach only provides estimates for the expected value, and furthermore only when the query is itself *uniformly random*. It is a long standing open problem to estimate the variance or any kind of tail bounds that would guarantee that the expected value is indeed a legitimate estimate of the cost; the few documents that were claiming to have found the variance or the limit distribution were wrong for they assumed that subtrees where independent conditionally on the query line, which is false.

In Chapter 5, we present how ideas pertaining to fixed-point theorems can be used to obtain limit processes in the setting of recursive partitions of the plane. In particular, in Section 5.2 we give one of the first non-trivial applications of the results of Neininger and Sulzbach [90] for weak convergence of càdlàg processes.

RECURSIVE LAMINATION OF THE DISK. Our work on quadtrees led us to a related partitioning scheme in which random chords are sequentially added to the unit disk, unless they intersect an already inserted chord. Motivated by earlier work of Aldous [8, 9] on *uniform triangulations*, Curien and Le Gall [34] introduced the model of *random recursive triangulations* of the disk. The construction goes as follows: At $n = 1$, two points are sampled independently with uniform distribution on the circle. They are connected by a chord which splits the disk into two fragments. Later on, at each step, two independent points are sampled uniformly at random on the circle and are connected by a chord if the latter does not intersect any of the previously inserted chords; in other words the two points are connected by a chord if they both fall in the same fragment. At time n this gives rise to a lamination L_n of the unit disk which consists of the union of the chords inserted up to time n. As an increasing closed subset of the disk, L_n converges, and it is proved in [34] that

$$L_\infty = \overline{\bigcup_{n\geq 1} L_n}$$

is a triangulation of the disk in the sense that any face of the complement is an open triangle whose vertices lie on the circumference of the circle (see [8]). Curien and Le Gall also show that the limit lamination L_∞ is encoded by a continuous process $\mathscr{M}$ in the sense that will be made precise later [8, 9, 34].

The approach in [34] consists in estimating distances in the *planar dual* of the recursive lamination as the chords are inserted. Unfortunately, the fragmentation-based arguments only provide asymptotics for distances between two points which is not enough to obtain the scaling limit of the *dual tree* itself. Note in particular that the dual tree is a finer object, since it is actually *not* characterized by the limit lamination.

In Section 5.3 we present results proving convergence of the dual tree of the *self-similar* lamination converges to a limit compact real tree which is encoded by the process $\mathscr{M}$ [P25]. We also give the limit dual tree of the related *homogeneous* lamination, thus providing two very different trees which both encode the same infinite lamination.

Chapter 2

Around the Brownian CRT

In this chapter we present some results related to the Brownian continuum random tree. Sections 2.1 and 2.2 are about the question of its universality and we follow the presentation in [P19, P20] and [P22], written in collaboration with Philippe Flajolet, Jean-François Marckert, respectively. In Section 2.3, we follow [P6], that is joint work with Louigi Addario-Berry and Cecilia Holmgren, and present a novel random reversible decomposition of the CRT.

2.1 Extreme distances in non-plane binary trees

2.1.1 Non-plane binary trees

The case of trees (as are considered here) with *indistinguishable neighbourhoods* is essentially different from the framework of Galton–Watson tree presented in the introduction. Such trees are not easily amenable to direct random walk approach, due to the inherent presence of symmetries.

The analysis of unlabelled non-plane trees finds its origins in the works of Pólya [96] and Otter [91]. However, these authors mostly focused on enumeration—the problem of characterizing typical parameters of these random trees remained largely untouched. Recently, in an independent study, Drmota and Gittenberger [36] have examined the profile of *"general" trees* (where all degrees are allowed) and shown that the joint distribution of the number of nodes at a finite number of levels converges weakly to the finite dimensional distribution of Brownian excursion local times. They further extended the result to a convergence of the entire profile to the Brownian excursion local time. (See also [54].)

We consider trees that are *binary, non-plane, unlabelled, and rooted*; that is, a tree is taken in the graph-theoretic sense and it has nodes of (out)degree two or zero only; a special node is distinguished, the root, which has degree two. In this model, *the nodes are indistinguishable*, and no order is assumed between the neighbours of a node. Let $\mathcal{Y}$ denote the class of such trees, and let $\mathcal{Y}_n$ be the subset consisting of trees with n external nodes (i.e., nodes of degree zero). Our aim is to study the (random) *height* H_n of a tree sampled uniformly from $\mathcal{Y}_n$ (largest number of edges of a simple path to the root). Uniform bounds on the rescaled height are crucial in proving tightness in the proof of convergence of these trees to the Brownian CRT by Marckert and Miermont [79].

Our approach is entirely based on *generating functions*. The class $\mathcal{Y}$ of (non-plane, unlabelled, rooted) binary trees is defined to include the tree with a single external node. A tree has *size* n if it has n external nodes, hence $n-1$ internal nodes. The cardinality of the subclass $\mathcal{Y}_n$ of trees of size n is denoted by y_n and the generating function (GF) of $\mathcal{Y}$ is

$$y(z) := \sum_{n\geq 1} y_n z^n = z + z^2 + z^3 + 2z^4 + 3z^5 + 6z^6 + 11z^7 + 23z^8 + \cdots ,$$

the coefficients corresponding to the entry A001190 of Sloane's *On-line Encyclopedia of Integer Sequences*. The trees of $\mathcal{Y}$ with size at most 6 are shown in Figure 2.1.

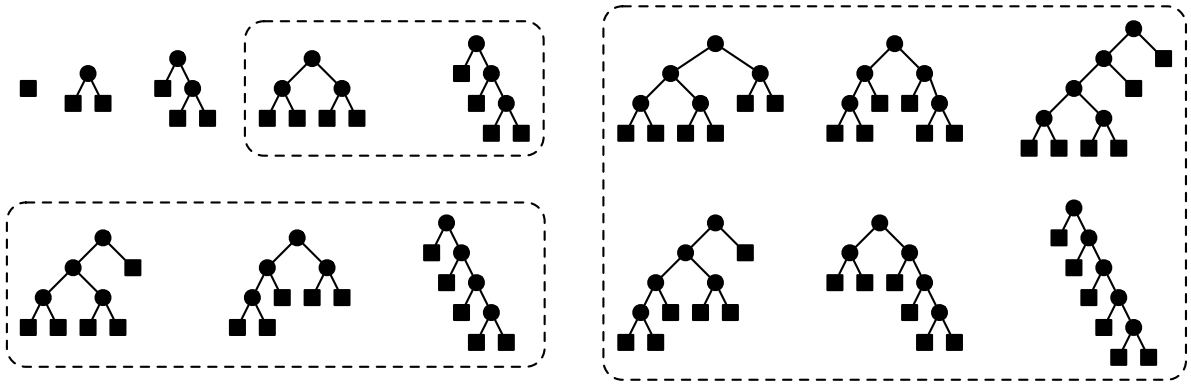

Figure 2.1: *The binary unlabelled trees of size less than six.*

A binary tree is either an external node or a root appended to an unordered pair of two (not necessarily distinct) binary trees. In the language of analytic combinatorics [50], this corresponds to the (recursive) specification

$$\mathcal{Y} = \mathcal{Z} + \mathrm{MSet}_2(\mathcal{Y}),$$

where $\mathcal{Z}$ represents a generic atom (of size 1) and MSet_2 forms multisets of two elements. The basic functional equation

$$y(z) = z + \frac{1}{2}y(z)^2 + \frac{1}{2}y(z^2), \tag{2.1}$$

closely related to the early works of Pólya [96, 97], and first studied by Otter [91], follows from fundamental principles of combinatorial enumeration [50, 59]. The term $\frac{1}{2}y(z^2)$ accounts for potential symmetries—hereafter, we refer to such terms involving functions of $z^2, z^3, \dots$, as *Pólya terms*. According to the general theory of analytic combinatorics, we shall operate in an essential manner with properties of generating functions in the *complex plane*. The Pólya terms, although modifying the nature of the generating function y, do not change the nature of the dominant singularity. In particular, y has a dominant singularity of the square-root type, as the generating functions for simply generated trees.

Lemma 2.1 (Otter [91])**.** *Let ρ be the radius of convergence of $y(z)$. Then, one has $1/4 \le \rho < 1/2$, and ρ is determined implicitly by $\rho + \frac{1}{2}y(\rho^2) = \frac{1}{2}$. As $z \to \rho^-$, the generating function $y(z)$ satisfies*

$$y(z) = 1 - \lambda\sqrt{1 - z/\rho} + O\left(1 - z/\rho\right), \qquad \lambda = \sqrt{2\rho + 2\rho^2 y'(\rho^2)}. \tag{2.2}$$

Furthermore, the number y_n of trees of size n satisfies asymptotically

$$y_n = \frac{\lambda}{2\sqrt{\pi}} \cdot n^{-3/2}\rho^{-n}\left(1 + O(1/n)\right), \tag{2.3}$$

Lemma 2.1 also suggests that, although there is no clear *exact* reduction of unlabeled non-plane trees to random walks, such trees should largely behave like simply generated families of ordered trees. In particular, it suggests that the rescaled height $H_n/\sqrt{n}$ is likely to admit a limit distribution of the theta-function type [39, 47, 68, 99]. The purpose of [P19, P20] is to prove formally that this is indeed the case.

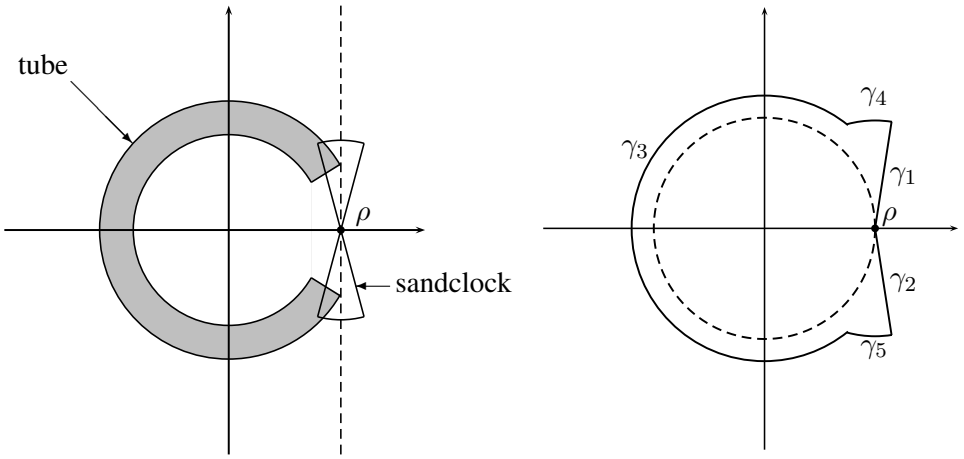

Figure 2.2: Left*: the "tube" and "sandclock" regions.* Right*: the Hankel contour used to estimate* $e_{h,n}$.

2.1.2 Approximations for the generating functions

Let $y_{h,n}$ be the number of trees of size n and height *at most* h and let $y_h(z) = \sum_{n\geq 1} y_{h,n} z^n$ be the corresponding generating function. Since trees of height at most $h+1$ are either a leaf, or made of trees of height at most h, the arguments leading to (2.1) yield the fundamental recurrence

$$y_{h+1}(z) = z + \frac{1}{2} y_h(z)^2 + \frac{1}{2} y_h(z^2), \qquad h \geq 0, \tag{2.4}$$

with initial value $y_0(z) = z$. A central rôle in what follows is played by the generating function of trees with height *exceeding* h:

$$e_h(z) \equiv \sum_{n\geq 1} e_{h,n} z^n := y(z) - y_h(z),$$

Then, a trite calculation shows that the $e_h(z)$ satisfy the main recurrence

$$e_{h+1}(z) = y(z) e_h(z) \left(1 - \frac{e_h(z)}{2y(z)}\right) + \frac{e_h(z^2)}{2}, \qquad e_0(z) = y(z) - z, \tag{2.5}$$

on which our treatment of height is entirely based.

THE GENERAL ANALYTIC APPROACH. The distribution of height is accessible by

$$\mathbf{P}\left(H_n > h\right) = \frac{y_n - y_{n,h}}{y_n} = \frac{e_{h,n}}{y_n}, \tag{2.6}$$

where $e_{h,n} = [z^n] e_h(z)$ (and in general we note $[z^n] \sum_{i\geq 0} a_i z^i := a_n$). Lemma 2.1 provides an estimate for y_n, and we shall get a handle on the asymptotic properties of $e_{h,n}$ by means of Cauchy's coefficient formula,

$$e_{n,h} = \frac{1}{2i\pi} \int_\gamma e_h(z) \frac{dz}{z^{n+1}}, \tag{2.7}$$

upon choosing a suitable integration contour γ in (2.7), of the form commonly used in singularity analysis theory [50]; see Figure 2.2 below. This task necessitates first developping suitable estimates of $e_h(z)$, for values of z both *inside* and *outside* of the disc of convergence $|z| < \rho$. Precisely, we shall need estimates valid in a *"tube"* around an arc of the circle $|z| = \rho$, as well as inside a *"sandclock"* anchored at ρ (we shall not give a formal definition for these regions, see Figure 2.2).

ESTIMATES FOR e_h ALONG THE CONTOUR. Estimates of the sequence of generating functions $(e_h(z))$ within the disc of convergence and a tube, where z stays away from the singularity ρ, are comparatively easy and follow essentially from a continuity argument. To deal with the portion of the contour lying inside the tube $(\gamma_3 \cup \gamma_4 \cup \gamma_5)$, convergence of $e_h \to 0$ thus boundedness is sufficient: taking a contour that lies sufficiently outside the disk of convergence ($(\log^2 n)/n$ away suffices) ensures that the Cauchy kernel in (2.7) is small enough to yield a contribution of order $O(\rho^{-n}\exp(-\log^2 n)) = o(y_n)$.

The bulk of the technical work is relative to the sandclock. One first needs the existence of a suitable sandclock for convergence of $e_h(z) \to 0$, uniformly in z, as $h \to \infty$. Furthermore, in the sandclock, the contour $(\gamma_1 \cup \gamma_2)$ needs to come too close to the disk of convergence to remain negligible compared to y_n: one needs to obtain refined information about the asymptotics for e_h near the singularity. But once convergence is guaranteed, it is possible to bootstrap it to develop an approximation of the form:

$$e_h(z) \equiv\ y(z) - y_h(z) \approx 2\frac{1-y}{1-y^h}y^h. \tag{2.8}$$

The form of the approximation in (2.8) is similar to that in the original paper by Flajolet and Odlyzko [47] where trees are ordered. Its justification closely follows the general strategy in [47]; however, non-trivial adaptations are needed, due to the presence of Pólya terms, so that the problem is no longer of a "pure" iteration type (in the ordered case, the recurrence only involves the generating functions at a single point z).

2.1.3 Asymptotics for the height

A quantified version of the approximation in (2.8) is all that is needed to reap the crop; from there, the work is classical and relies on a Tauberian transfer theorem from singularity analysis [48, 50]. There, we use (2.6), the approximation in (2.8) and the square root singularity of y at ρ to prove the following theorem relative to the distribution of height H_n:

Theorem 2.1 (Limit law of height)**.** *The height H_n of a random tree taken uniformly from $\mathcal{Y}_n$ admits a* limiting theta distribution*: for any fixed $x > 0$, there holds*

$$\lim_{n\to\infty} \mathbf{P}\left(H_n \geq \lambda^{-1}x\sqrt{n}\right) = \Theta(x), \qquad \lambda := \sqrt{2\rho + 2\rho^2 y'(\rho^2)},$$

where $$\Theta(x) := \sum_{k\geq 1}(k^2x^2 - 2)e^{-k^2x^2/4}.$$

The approximation we have is also strong enough to obtain a local limit theorem as well as convergence of all moments.

Theorem 2.2 (Local limit law of height)**.** *The distribution of the height H_n of a random tree taken uniformly from $\mathcal{Y}_n$ admits a local limit: for x in a compact set of $\mathbb{R}_{>0}$ and $h = \lambda^{-1}x\sqrt{n}$ an integer, there holds uniformly*

$$\mathbf{P}\left(H_n = h\right) \sim \frac{\lambda}{\sqrt{n}}\vartheta(x),$$

where $$\vartheta(x) = -\Theta'(x) = (2x)^{-1}\sum_{k\geq 1}(k^4x^4 - 6k^2x^2)e^{-k^2x^2/4}.$$

Theorem 2.3 (Moments of height)**.** *Let $r \geq 1$. The r*th *moment of height H_n satisfies*

$$\mathbf{E}\left[H_n\right] \sim \frac{2}{\lambda}\sqrt{\pi n} \qquad \textit{and} \qquad \mathbf{E}\left[H_n^r\right] \sim r(r-1)\zeta(r)\Gamma(r/2)\left(\frac{2}{\lambda}\right)^r n^{r/2}, \quad r \geq 2. \tag{2.9}$$

The asymptotics in Theorem 2.1 have been used by Marckert and Miermont [79] as a tighness argument to wrap up a proof that the scaling limit of random non-plane binary trees is Aldous' Brownian CRT, hence confirming its universal character. Finally, Haas and Miermont [58] devised a general approach based on the way the mass is fragmented when moving away from the root, which generalized the results in [79].

The general theorem in [58] requires the branching property: given the sizes the subtrees should behave independently. This suggest investigating the robustness of the Brownian CRT limit for models which are not Markov branching in the sense of Haas and Miermont [58], or which do not have the branching property, even conditional on the size of the subtrees; in the following section, we discuss a model which lacks the branching property, yet rescales to the Brownian CRT.

2.2 Trees with a prescribed degree sequence

2.2.1 Model and notations

Let t be a rooted tree and $n_i(t)$ the number of nodes in t having i children. The sequence $(n_i(t), i \geq 0)$ is called the degree sequence of t, and satisfies $\sum_{i\geq 0} n_i(t) = 1 + \sum_{i\geq 0} i n_i(t) = |t|$, the number of nodes in t.

Our aim in this section is to discuss trees chosen under $\mathbf{P}_{\mathbf{s}}$, the uniform distribution on the set of rooted plane trees with specified degree sequence $\mathbf{s} = (n_i, i \geq 0)$, and size $|\mathbf{s}| := \sum_{i\geq 0} n_i$. More precisely, a sequence of degree sequences $(\mathbf{s}(\kappa), \kappa \geq 0)$ with $\mathbf{s}(\kappa) = (n_i(\kappa), i \geq 0)$, corresponding to trees with size $n_\kappa := |\mathbf{s}(\kappa)| \to +\infty$ is given, and the investigations concern the limiting behaviour of tree under $\mathbf{P}_{\mathbf{s}(\kappa)}$.

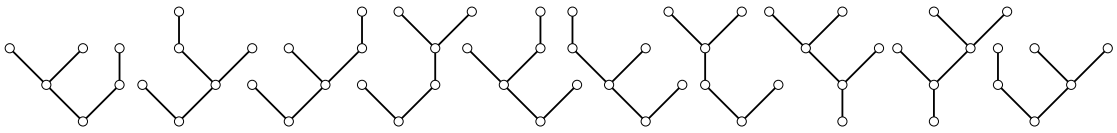

Figure 2.3: *The ten trees of* $\mathbb{T}_{\mathbf{s}}$ *for the degree sequence* $\mathbf{s} = (3, 1, 2, 0, 0, \dots)$.

We denote by $\mathbf{p}(\kappa) = (p_i(\kappa), i \geq 0)$ the degree distribution under $\mathbf{P}_{\mathbf{s}(\kappa)}$:

$$p_i(\kappa) = \frac{n_i(\kappa)}{n_\kappa} \qquad \text{and} \qquad \sigma_\kappa^2 := \sum_{i\geq 1} \frac{n_i(\kappa)}{n_\kappa - 1} i^2 - 1; \tag{2.10}$$

σ_κ^2 is "almost" the associated variance, this choice of definition yields shorter formulae. The maximum degree of any tree with degree sequence $\mathbf{s}(\kappa)$ is $\Delta_\kappa = \max\{i : n_i(\kappa) > 0\}$.

2.2.2 Motivations and discussions

A GENERALIZATION OF GALTON–WATSON TREES. The model $\mathbf{P}_{\mathbf{s}}$ is related to Galton–Watson trees [15, 60], simply generated trees in the combinatorial literature, by a simple conditioning: the distribution $\mathbf{P}_{\mathbf{s}}$ coincides with the distribution of the family tree $\mathbf{t}$ of a Galton–Watson process with offspring distribution $(\nu_i, i \geq 0)$ (which satisfies $\nu_i > 0$ if $n_i > 0$) conditioned on $\{n_i(\mathbf{t}) = n_i, i \geq 0\}$. Indeed, $\mathbf{P}_{\mathbf{s}}$ assigns the same probability to all trees with the same degree sequence. In this sense, the distribution ν plays a role of secondary importance, and $\mathbf{P}_{\mathbf{s}}$ appears to be a model of combinatorial nature, far from the world of Galton–Watson processes. Nevertheless, we will see that our convergence theorem implies Aldous' result that Galton–Watson trees rescale to the Brownian continuum random tree. The argument morally relies on the fact that under $\mathbf{P}_\mu(\,.\,||\mathbf{t}| = n)$, the empirical degree sequence satisfies the hypotheses of Theorem 2.4 (stated later) with probability going to one.

A CONSTRAINED COALESCENT PROCESS. In the same way that uniform random trees or forests may be seen as the results of coagulation/fragmentation processes involving particles [94, 95], trees with a prescribed degree sequence appear naturally in similar aggregation processes. The model where particles have constrained valence may appear more "physically" grounded. The relevant underlying coalescing procedure is the additive coalescent [12, 18], a Markov process whose dynamics are such that particles merge at a rate proportional to the sum of their masses/sizes. The additive coalescent is the aggregation process appearing in Knuth's modification of Rényi's parking problem [61, 98] or the hashing with linear probing [21, 31]. The reader may find more information about coagulation/fragmentation processes in the monograph by Bertoin [19] or the recent survey by Berestycki [17].

TOWARDS GRAPHS WITH A PRESCRIBED DEGREE SEQUENCE. We are also motivated by the metric structure of graphs with a prescribed degree sequence. Introduced by Bender and Canfield [16] and by Bollobás [29] in the form of the configuration model, these graphs have received a lot of attention since the first tight analysis of the size of connected components by Molloy and Reed [85, 86]. This is mainly because the model allows for a lot of flexibility in the degree sequence. In particular, the model provides a construction of random graphs with degree sequences that may match the observations in large real-world networks.

Of course, random graphs with a prescribed degree sequence are much more complex than trees with a prescribed degree sequence, but there is no doubt that the analysis of trees is a first step towards the identification of the metric structure of the corresponding graphs. Indeed, recent results of Joseph [67] show that under some moment condition, the sizes of the connected components of random graphs with a prescribed *critical* degree sequence are similar to those of Erdős–Rényi $G(n,p)$ random graphs [28, 41, 66]: they may be asymptotically described in terms of the lengths of the excursions of a Brownian motion with parabolic drift above its current minimum, as demonstrated by Aldous [10] and discussed in Section 1.4. (See also [100], where it is supposed that the maximum degree is bounded.) On the other hand, as we will see in Chapter 3, the metric structure of $G(n,p)$ inside the critical window may be identified in terms of modifications of Brownian CRT [P3, P4]. In other words, the present analysis is one more building block towards an invariance principle for scaling limits of random graphs, i.e., that critical random graphs with a prescribed degree sequence have (under a suitable moment condition on the degree distribution) the same scaling limit (as sequence of compact metric spaces) as classical random graphs [P4]. This is at least what is suggested by the results of Bhamidi, van der Hofstad, and van Leeuwaarden [25, 26], Hofstad [62], Joseph [67] and Riordan [100].

2.2.3 Convergence

The best conditions one could hope for which are sufficient to ensure convergence to the Brownian continuum random tree are weak convergence of the degree distribution with convergence of the second moment and $\Delta_k = o(\sqrt{n_\kappa})$. The following theorem confirms that these conditions are sufficient.

Let $\mathbf{p}$ be a probability distribution with mean one and variance σ^2.

Theorem 2.4. *Let $(\mathbf{s}(\kappa), \kappa \geq 0)$ be a sequence of degree sequences such that $n_\kappa \to +\infty$, $\Delta_\kappa = o(n_\kappa^{1/2})$, $\mathbf{p}(\kappa) \to \mathbf{p}$ weakly with $\sigma_\kappa^2 \to \sigma_{\mathbf{p}}^2$. Let $\mathbf{t}$ be a plane tree chosen under $\mathbf{P}_{\mathbf{s}(\kappa)}$ and let $d_{\mathbf{t}}$ be the graph distance in $\mathbf{t}$. Under $\mathbf{P}_{\mathbf{s}(\kappa)}$,*

$$(\mathbf{t}, \sigma_\kappa n_\kappa^{-1/2} d_{\mathbf{t}}) \xrightarrow[\kappa\to\infty]{} (\mathcal{T}_{2\mathbf{e}}, d_{2\mathbf{e}})$$

in distribution in the Gromov–Hausdorff sense.

Our approach uses a phenomenon observed in Marckert & Mokkadem [80] in the case of critical Galton–Watson tree (having a variance) that (under some mild assumptions) the Łukasiewicz path $S_{\mathbf{t}}$ and the height process $H_{\mathbf{t}}$ are asymptotically proportional, that is, up to a scalar normalisation, the difference between these processes converge to the zero function. It turns out that a similar phenomenon occurs when the degree sequence is prescribed, and this is the basis of our proof. In order to prove Theorem 2.4

we proceed in two steps: the first one consists in showing that the depth-first walk $S_{\mathbf{t}}$ associated to a tree sampled under $\mathbf{P}_{\mathbf{s}(\kappa)}$ converges to a Brownian excursion.

CONVERGENCE OF THE DEPTH-FIRST WALK. The process $S_{\mathbf{t}}$ is much easier to deal with than $H_{\mathbf{t}}$, since $S_{\mathbf{t}}$ is *essentially* a random walk conditioned to stay non-negative, and forced to end up at the origin (precisely at -1). An urn process containing n_i copies of $i-1$, $i \geq 0$ produces a random permutation of the degrees (-1); from there, there is a unique way to cyclically shift the sequence so that its partial sums form the depth-first walk of a tree. The tree obtained is distributed as a tree under $\mathbf{P}_{\mathbf{s}}$. So, modulo a very moderate amount of verifications, the convergence towards the Brownian excursion follows from standard theorems on urn sampling.

DISCREPANCY ŁUKASIEWICZ PATH / HEIGHT PROCESS. The core of the work lies in the second step, which consists in proving that rescaled versions of $S_{\mathbf{t}}$ and $H_{\mathbf{t}}$ are indeed close, uniformly on $[0,1]$; from the previous paragraph, this would imply that a rescaled $H_{\mathbf{t}}$ indeed converges to a Brownian excursion, hence proving Theorem 2.4. The value $S_{\mathbf{t}}(i)$ of the depth-first walk at position i, may be expressed as a sum of contributions for the nodes on the path between the root and $\tilde{u}_i$, the i-th node in the lexicographic order. These contributions can be proved to be *essentially independent* so that one has concentration, hence proportionality (at this given location i). Using this argument at a uniformly random location and then proving tightness using general Gaussian tail bounds from Addario-Berry [2] suffices to complete the proof.

RECENT RELATED WORK. Recent results of Rizzolo [101] and Kortchemski [70] have a flavor similar to our Theorem 2.4 (although neither implies the other): they proved scaling limits for Galton–Watson trees conditioned on the number of nodes having their degrees in a subset A of the support of the measure μ, the number of nodes with other out-degrees being left free. For instance, they consider trees conditioned on the number of leaves. The proofs in Rizzolo [101] rely ultimately on the approach based on Markov branching trees developed by Haas and Miermont [58].

2.3 Cutting down and typical distance

2.3.1 Motivation and approach

The problem of estimating the number of cuts to isolate the root of a random tree is not new, and our results in this direction are not new. However, apart from the results in [32], all the other proofs [45, 63, 64, 92, 93] rely on resolutions of recurrence relations or calculation of moments and do not yield any intuition as of the reason why the Rayleigh limit law appears. The main motivation for the work we present in this section is to *explain* why the Rayleigh distribution appears both in the number of cuts required to isolate the root of a random tree, and the distance between two uniformly chosen random nodes.

The first step consists in studying a *canonical* instance of Galton–Watson tree which would behave best with respect to the phenomenon at hand. This special instance is the Cayley tree (uniform labeled tree): there is actually a bijection which turns a tree and its cutting sequence into a tree with an independent distinguished node, in such a way that the number of cuts is turned into the length of the path to the distinguished node. This approach also yields very simple constructive proofs of the results concerning the distribution of the number of cuts obtained in [45, 63, 64, 92]. This exact discrete correspondence may also be lifted to the level of real trees: there is a simple way to recombine the pruned subtrees of a logged Brownian CRT that yields an other Brownian CRT with a distinguished node. This correspondence in the continuous setting may be interpreted as a new random reversible transformation between a Brownian excursion and a Brownian bridge.

The construction also generalizes to processes where one must isolate more than one node. This has been considered by Bertoin [23] who proved that the limit number of cuts required to isolate k independent nodes in a Cayley tree converges to the length of the subtree spanning $k+1$ independent nodes. This has recently been extended to the case of Galton–Watson trees with a finite variance by Bertoin and Miermont

[24]. The limit picture which explains constructively the connection between the number of cuts and the height of a uniformly random node has also recently been generalized to general Lévy trees by Abraham and Delmas [1].

2.3.2 A bijection for labeled Cayley

At the heart of our approach is a coupling which yields the *exact* distribution of the number of cuts for every fixed n, for the special case of uniform Cayley trees (uniformly random labeled rooted trees).

The transformation is very easily explained as follows. Consider an initial tree T. This tree is then pruned by iterative removal of vertices, uniformly chosen in the connected component which contains the root; it is this number of vertices we should keep track of. Each time, at least one vertex is removed and the procedures terminates at time $\kappa = \kappa(T) \leq |T|$. Write $v_1, v_2, \ldots, v_\kappa$ for the vertices chosen and $T_1, T_2, \ldots, T_\kappa$ for the trees pruned at the stages $1, 2, \ldots, \kappa$. In particular, v_i is the root of T_i and v_κ is the root of the initial tree T.

Theorem 2.5. *If T is uniformly random in $\mathbb{T}_n$, then the ordered labeled forest $(T_1, T_2, \ldots, T_\kappa)$ is uniformly random.*

Since the number of cuts is turned into the number of vertices on the path between two uniformly random nodes of a uniformly random labeled tree, it is then clear that it converges in distribution to a Rayleigh random variable:

Theorem 2.6. *Let $\kappa(T)$ the number of cuts required to isolate the root of a tree T. If T is uniform in $\mathbb{T}_n$, then for every $x \geq 0$,*

$$\mathbf{P}\left(\kappa(T) \leq x\sqrt{n}\right) \xrightarrow[n\to\infty]{} e^{-x^2/2}.$$

From the forest floor picture of [22], one sees that the random forest $(T_1, T_2, \ldots, T_\kappa)$ can be re-arranged into a random tree by connecting their roots into a path of length $\kappa(T)$ and making it rooted at v_1, the root of T_1. This tree is a labeled tree and has a distinguished path, or a distinguished node (the root of T). We call the new re-arranged tree $\hat{T}$.

Theorem 2.7. *If T is uniformly random in $\mathbb{T}_n$, then the re-arranged tree $\hat{T}$ rooted at v_1 is also uniformly random in $\mathbb{T}_n$, and v_κ is uniform in $\{1, 2, \ldots, n\}$ and independent of $\hat{T}$.*

Aldous [5] studied the subtree rooted at a uniformly random node in a critical, finite variance Galton–Watson tree conditioned to have size n. In particular, he showed that such a subtree converges in distribution to an *unconditioned* critical Galton–Watson tree. It is then straightforward that, for fixed $k \geq 1$, the first k trees that are cut converge in distribution to a forest of k critical Galton–Watson trees. On the other hand, a critical Galton–Watson tree conditioned to be large converges locally (in the sense of local weak convergence of [13], i.e., inside balls of arbitrary fixed radius k around the root) to the following infinite tree:

- there is an infinite backbone of nodes having a size-biased number of children (exactly one of which is again on the infinite path), and
- the children of the nodes of the backbone which are not themselves on the backbone are the root of an unconditioned critical Galton–Watson tree.

This is the incipient infinite cluster for critical, finite variance Galton–Watson trees [69]. When the progeny distribution is Poisson(1), one can equivalently consider the backbone as being $\mathbb{N}$, and every node $u \in \mathbb{N}$ is the root of a critical Galton–Watson tree with Poisson(1) offspring distribution. Theorem 2.7 then appears as a strengthening of this latter picture (valid only for Poisson Galton–Watson trees) in which k is allowed to grow with n.

This construction can be extended to the isolation of more than one node; explaining it reasonably well would require more space than I can give here. The details may be found in [P6]. The consequence for the Brownian continuum random tree is much more interesting and we move on to the real tree version of the discrete correspondence we have just described.

2.3.3 Lifting the transformation to the continuum random tree

It turns out that our coupling approach allows us to prove results about a natural "continuum version" of the random cutting procedure which takes place on the Brownian continuum random tree (CRT), here denoted by $(\mathcal{T}, d)$. Although we work principally in the language of $\mathbb{R}$-trees, the correspondence we will discuss can be viewed as a new, invertible random transformation between the Brownian excursion and a reflecting Brownian bridge. Though the precise statement requires a fair amount of set-up, if this set-up is taken for granted the result can be easily described.

CHOOSING "UNIFORMLY" RANDOM POINTS IN THE CRT. The first step in setting the continuous analog consists in choosing the random points correctly. The continuum random tree comes equipped with a natural probability measure μ, the push-forward of Lebesgue measure on $[0, 1]$ into the canonical projection from $[0, 1]$ onto $\mathcal{T}$. Unfortunately, this measure is concentrated on the leaves of $\mathcal{T}$, so that with probability one, for a point $x \in \mathcal{T}$ sampled according to μ, $\mathcal{T} \setminus \{x\}$ has a single connected component, and in particular the connected component of $\mathcal{T} \setminus \{x\}$ containing the root still has mass one. In other words, if we wish to log the continuum random tree, we ought to sample the points using a different measure.

There is actually an other "uniform" measure on $\mathcal{T}$, called the *length measure*. It is the only sigma-finite measure ℓ on $\mathrm{skel}(\mathcal{T})$ such that for any $a, b \in \mathrm{skel}(\mathcal{T})$, $\ell([\![a, b]\!]) = d(a, b)$. The length measure is the measure one wants to use to log the continuum random tree.

LOGGING THE BROWNIAN CONTINUUM RANDOM TREE. Let $(\mathcal{T}, d)$ be a Brownian CRT with root ρ and mass measure μ, write $\mathrm{skel}(\mathcal{T})$ for its skeleton, and let $\mathcal{P}$ be a homogeneous Poisson point process on $\mathrm{skel}(\mathcal{T}) \times [0, \infty)$ with intensity measure $\ell \otimes dt$, where ℓ is the length measure on the skeleton. We think of the second coordinate as a time parameter. View each point (p, τ) of $\mathcal{P}$ as a *potential* cut, but only make a cut at p if no previous cut has fallen on the path from the root ρ to p. At each time $0 \leq t < \infty$, this yields a forest of countably many rooted $\mathbb{R}$-trees; we write $\mathcal{T}_t$ for the component of this forest containing ρ. Run to time *infinity*, this process again yields a countable collection of rooted $\mathbb{R}$-trees, later called $(f_i, i \in I_\infty)$. Furthermore, each element f_i of the collection comes equipped with a time index τ_i (the time at which it was cut). This logging process is a *filtered* version of the Aldous–Pitman [11] fragmentation that only keeps those cuts which fall inside the connected component containing ρ.

PUTTING THE PIECES BACK TOGETHER. As in the discrete case, one wants to re-arrange this ordered forest into a single real tree. (Note that we are aware that the space in which this so-called forest lives is not clear, but this object is only used for pedagogical reasons.) Although we did not insist on this point in the discrete setting, putting back the trees together requires to put some *missing length* back in the game; this is how the backbone is created. Similarly, in the continuous version, we have to build a path on which to glue the pruned subtrees. The length of that path should correspond to the length we have removed in the process. In the continuous setting, the missing length is expressed in terms of a *local time*, but to avoid the annoying justification of the existence of this local time we define the length as follows.

For $0 \leq t < \infty$, let $L(t) = \int_0^t \mu(\mathcal{T}_s)ds$, and let $L(\infty) = \lim_{t\to\infty} L(t)$. It turns out that $L(\infty)$ is almost surely finite. Next, create a single compact $\mathbb{R}$-tree $(\mathcal{T}', d')$ from the collection $(f_i, i \in I_\infty)$ and the closed interval $[0, L(\infty)]$ by identifying the root of f_i with the point $L(\tau_i) \in [0, L(\infty)]$, for each $i \in I_\infty$, then taking the completion of the resulting object. (This completion only adds countably many points.) Let μ' be the push-forward of μ under the transformation described above.

Theorem 2.8. *The triples $(\mathcal{T}', d', \mu')$ and $(\mathcal{T}, d, \mu)$ have the same distribution. Furthermore, $0 \in \mathcal{T}'$ and $L(\infty) \in \mathcal{T}'$ are independent and both have law μ'.*

Using the standard encoding of the CRT by a Brownian excursion, we may take the triple $(\mathcal{T}, d, \mu)$, together with the point ρ, to be encoded by a Brownian excursion. Similarly, it is possible to view the triple $(\mathcal{T}', d', \mu')$, together with the points 0 and $L(\infty)$, as encoded by a reflecting Brownian bridge; see Section 10 of [11] (this is also closely related to the "forest floor" picture of [22]). From this perspective, the transformation from $(\mathcal{T}, \rho)$ to $(\mathcal{T}', 0, L_\infty)$ becomes a new, random transformation from Brownian excursion to reflecting Brownian bridge.

As an immediate consequence of the above development, we reprove the following well-known result. Let $\mu(t)$ be the mass of the tagged fragment in the Aldous–Pitman [11] fragmentation (dual to the standard additive coalescent) at time t, that is the fragment containing a random point (for instance the root).

Corollary 2.1. *The random variable $\int_0^\infty \mu(t)dt$ has the standard Rayleigh distribution.*

A REVERSE TRANSFORMATION. We are also able to explicitly describe the inverse of the transformation which takes a real tree together with a distinguished node, and reshuffles the subtrees rooted at the branch-points of the distinguished path.

Let $(\mathcal{T}, d, \mu)$ be a measured CRT, and let ρ, ρ' be independent random points in $\mathcal{T}$ with law μ. Let B be the set of branch points of $\mathcal{T}$ on the path from ρ to ρ'. For each $b \in B$ let $\mathcal{T}_b$ the the set of points $x \in \mathcal{T}$ for which the path from b to ρ contains a point $b' \in B$ with $d(\rho, b') > d(\rho, b)$. In words, $\mathcal{T}'$ is the set of points in subtrees that "branch off the path from ρ to ρ' after b." Then, independently for each point $b \in B$, let y_b be a random element of $\mathcal{T}_b$, with law $\mu/\mu(\mathcal{T}_b)$. Delete all non-branch points on the path between ρ and ρ'; then, for each $b \in B$, identify the points b and y_b. Write $(\mathcal{T}', d')$ for the resulting tree, and μ' for the push-forward of μ to $\mathcal{T}'$.

Theorem 2.9. *The triples $(\mathcal{T}, d, \mu)$ and $(\mathcal{T}', d', \mu')$ have the same distribution. Furthermore, the point $\rho' \in \mathcal{T}'$ has law μ'.*

Chapter 3

The scaling limit of critical random graphs

In this chapter, we describe the scaling limit of the Erdős–Rényi random graphs $G(n,p)$ when $p = 1/n + \lambda n^{-4/3}$ for some fixed real number λ. The presentation is based on the results in [P3] and [P4] which are in collaboration with Louigi Addario-Berry and Christina Goldschmidt.

3.1 Intuition and overview

Theorem 1.3 of Aldous [10] gives a very precise description of the sizes and surplus of the largest components in the random graph $G(n,p)$ when $p = 1/n + \lambda n^{-4/3}$, for $\lambda \in \mathbb{R}$. Furthermore, the independence of the edges ensures that given a connected component's vertex set and number of edges, the induced connected graph is *uniformly* random among all connected graphs with this size and number of edges, and independent of the rest of the graph. In other words, given the sizes and surpluses of the connected components, the entire graph may be recovered (more precisely, an identically distributed copy) by sampling independent connected components with the required sizes and number of edges.

Information about the graph $G(n,p)$ is gathered thanks to an *exploration process* (we will define the one we are interested in precisely in Section 3.2). For now, we only need to know that the exploration is encoded by a random walk in such a way that excursions away from zero correspond to distinct connected components; the length of the excursion encodes the size of the connected component. The exploration process of the graph $G(n,p)$, for $p = 1/n + \lambda n^{-4/3}$, converges to the process B^λ defined by

$$B^\lambda(t) := W^\lambda(t) - \inf_{s\in[0,t]} W^\lambda(s) \qquad \text{and} \qquad \mathrm{W}^\lambda(t) := t\lambda - t^2/2 + W(t),$$

where $(W(t), t \geq 0)$ is a standard Brownian motion. The fact that connected components have the same distribution given their sizes should somewhat appear in the formulas, but this fact is not immediately clear from the limit picture given in Theorem 1.3. However, an excursion theory calculation (see [10, P4]) shows that, conditional on their lengths, the distributions of the excursions of B^λ above zero do not depend on their starting points. Write $\tilde{\mathbf{e}}^{(\sigma)}$ for such an excursion conditioned to have length σ; in the case $\sigma = 1$, we will simply write $\tilde{\mathbf{e}}$. The distribution of $\tilde{\mathbf{e}}^{(\sigma)}$ is most easily described via a change of measure with respect to the distribution of a Brownian excursion $\mathbf{e}^{(\sigma)}$ conditioned to have length σ: for any test function f,

$$\mathbf{E}[f(\tilde{\mathbf{e}}^{(\sigma)})] = \frac{\mathbf{E}\left[f(\mathbf{e}^{(\sigma)}) \exp\left(\int_0^\sigma \mathbf{e}^{(\sigma)}(x)dx\right)\right]}{\mathbf{E}\left[\exp\left(\int_0^\sigma \mathbf{e}^{(\sigma)}(x)dx\right)\right]}. \tag{3.1}$$

We refer to $\tilde{\mathbf{e}}^{(\sigma)}$ as a *tilted excursion*. The surplus of a connected component in Aldous' view using the Brownian motion with parabolic drift is built as the number of points of a Poisson point process with unit rate in $[0,\infty)\times[0,\infty)$ which fall under a given excursion of B^λ. In other words, given $\tilde{\mathbf{e}}^{(\sigma)}$, the surplus of the corresponding connected component is distributed as Poisson$(\int_0^\sigma \tilde{\mathbf{e}}^{(\sigma)}(u)du)$.

Both tilted excursions and the Poisson point process in the quarter plane are crucial in our construction. We will make them appear in such a way that explains the connection with the *metric structure* of the connected components. It should be noted that the tilted excursions $\tilde{\mathbf{e}}$ are present in the work of Aldous [10], however, the breadth-first exploration procedure he uses makes it difficult to recover the metric structure of the graph.

Although it does not give a direct access to the distances, Aldous' result on the sizes and surplus of components gives a pretty good intuition about what should happen for the metric structure. Roughly, any single of the largest connected components of the critical graph has size of order $n^{2/3}$, and a number of surplus edges which is $O(1)$. This suggest that, although these *uniform* connected components are not *uniform trees*, they should remain fairly close to such trees and after rescaling by $\sqrt{n^{2/3}} = n^{1/3}$, one should obtain a non-trivial compact object. In this chapter, we explain more precisely why it is the case, and the distribution of the limit *continuum graph* is described in terms of the Brownian continuum random tree. We also give three constructions of the limit metric space which shed complementary light on the object and its properties.

THREE CONSTRUCTIONS. Our first construction is the one which is closest to the Brownian with parabolic drift of [10] and also relies on a careful algorithmic exploration of the graph. In a second construction, we use a more structural point of view that exhibits the Brownian continuum random trees hidden in the first construction. Finally, as is well known, the Brownian continuum random tree can be obtained by a stick breaking construction which is a continuous analog of the Aldous–Broder algorithm in [3, 30]. It is then clear from the second construction that the limit continuum connected component may be constructed using a stick-breaking procedure, or more precisely, with stick-breaking *procedureS*, one for each CRT. The third construction shows that it is actually possible to build the limit connected component with a *single* such procedure.

Before explaining the constructions for a single connected component, we give some properties of the *entire* graph which may be derived using a strengthening of the type of convergence which allows us to describe the diameter of the whole graph. We will actually give this description now, since the strengthening of the convergence of a single connected components to the following convergence only requires looking carefully at the tails, and I do not intend to give any detail about this.

Theorem 3.1. *Suppose that $p = 1/n + \lambda n^{-4/3}$, for $\lambda \in \mathbb{R}$. Let $C_1^n, C_2^n, \dots$ be the connected components of $G(n,p)$ in decreasing order of their sizes (ties being broken using the labels). Then, there exists a sequence of non-trivial random compact metric spaces $(\mathscr{C}_i, i \geq 1)$ such that, as $n \to \infty$,*

$$(n^{-1/3}C_i^n, i \geq 1) \to (\mathscr{C}_i, i \geq 1)$$

in distribution for $\mathrm{d}_{\mathrm{GH}}^4$.

Theorem 3.1 implies results about the actual diameter of $G(n,p)$, defined as the maximum diameter of one of its connected components. For a metric space (M,d), we write $\mathrm{diam}(M) = \sup\{d(x,y) : x,y \in M, d(x,y) < \infty\}$ for its diameter.

Theorem 3.2. *Suppose that $p = 1/n + \lambda n^{-4/3}$ for $\lambda \in \mathbb{R}$. Then*

$$(n^{-1/3}\mathrm{diam}(C_i^n), i \geq 1) \xrightarrow[n\to\infty]{d} (\mathrm{diam}(\mathscr{C}_i), i \geq 1).$$

Furthermore, $\mathscr{D} := \sup_{i\geq 1} \mathrm{diam}(\mathscr{C}_i)$ has an absolutely continuous distribution, $\mathbf{E}[\mathscr{D}] < \infty$ and we have $n^{-1/3}\mathrm{diam}(G(n,p)) \to \mathscr{D}$ in distribution, as $n \to \infty$.

3.2 Exploring and generating connected graphs

We now only consider connected graphs with a fixed surplus s. Our aim is to find a way to study the distances in random connected graphs with a fixed surplus s. For this the approach is the following: For a given connected graph, we first define a *canonical spanning tree*. This tree has only s edges less than the initial graph, and to recover the initial graph from the spanning tree one only needs to specify which pairs of vertices should be linked by these s edges.

THE DEPTH-FIRST SEARCH TREE. We extract a tree from a connected labeled graph $G = ([n], E)$ using the depth-first search procedure. The version we use is not completely standard, and we present it in detail. We proceed using a *stack*: items first inserted are the last one to get out (*last in, first out*). The contents of the stack at step i is kept in $\mathcal{S}_i = \mathcal{S}_i(G)$. We will also maintain the edge set of the tree we build; the current set of edges at time i is called $E_i = E_i(G)$. Of course, at every time step, we have $E_i \subset E$, and the tree we build is $T[G] = ([n], E_n)$. The sets $\mathcal{J}_i = \mathcal{J}_i(G)$ will keep track of the vertices which have already been explored.

- Initially, the stack contains the vertex 1: $\mathcal{S}_0 = \{1\}$, $\mathcal{J}_0 = \varnothing$ and $E_0 = \varnothing$;
- at the i-th step, we take out one item of the stack (in the order we defined above), say u_i, for every node v such that $\{u_i, v\} \in E$ that is not yet in the stack ($v \notin \mathcal{S}_{i-1}$), we add the corresponding edge to our tree, then put all the nodes in the stack in such a way that the nodes with smaller labels are pulled out first. In other words we set $E_i = E_{i-1} \cup \{\{u_i, v\} \in E : v \notin \mathcal{S}_{i-1} \text{ and } v \neq u_j, j \leq i\}$, $\mathcal{J}_i = \mathcal{J}_{i-1} \cup \{u_i\}$ and $\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{v : \{u_i, v\} \in E \text{ and } v \neq u_j, j \leq i\}$.
- the process stops when all n nodes have been explored, that is just after step n: at this point there is no node left that could have been added to the stack, which is then empty.

The order in which the nodes $u_1, u_2, \ldots, u_n$ are explored is called the depth-first order, and the tree $T[G]$ is called the *depth-first tree*.

THE SURPLUS EDGES. The depth-first procedures extracts a tree $T[G]$ (a *labeled* tree on $[n]$) with $n-1$ edges, from the graph G with $n-1+s$ edges. Rather than trying to figure out exactly where these edges were, we want to understand *where they could have been*. The underlying idea is the following: if the graph G is random (say uniform among all graphs with surplus s, for instance) it is enough to understand the distribution of the location of the edges rather than their exact location. The first step consists in understanding which edges there could have been in G, whose presence or absence from G does not affect the depth-first tree.

In other words, we want to see the set $\mathbb{C}_n^L$ of connected labeled graphs on $[n]$ as the following disjoint union:

$$\mathbb{C}_n^L = \bigsqcup_{t \in \mathbb{T}_n^L} \{G \in \mathbb{C}_n^L : T[G] = t\},$$

and then use this partition to provide an alternate two-step procedure to generate random connected graphs: the procedure would first choose in which set the random graph lies (which amounts to generating $T[G]$), and then pick a graph from this set (which amounts to adding some extra edges to $T[G]$).

Consider the depth-first search procedure on the labeled tree $t = ([n], E)$. Then of course $T[t] = t$. Say that an edge $\{u, v\} \notin E$ is *allowed* by the depth-first search procedure if at some stage u and v have both been discovered, but neither has been explored, that is, for some i, u and v are both in $\mathcal{S}_i$. Let $\mathcal{A}(t)$ be the set of edges which are allowed by the depth-first search procedure on t, and write $a(t) = \#\mathcal{A}(t)$.

Observe that, for a tree t, $(\#\mathcal{S}_i(t), i = 0, \ldots, n)$ is precisely $S_t + 1$, where S_t is the depth-first process/Lukasiewicz path associated to the tree t (with the canonical ordering). It is reasonably easy to verify that $a(t)$ actually corresponds to the *discrete area* of the depth-first process S_t:

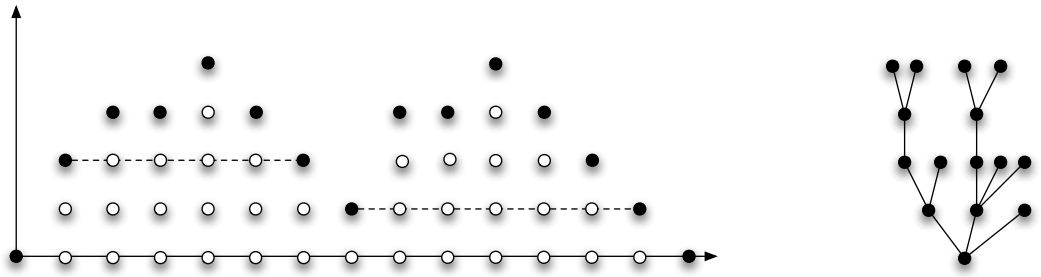

Figure 3.1: *A depth-first walk and the corresponding plane tree. The edges number of edges allowed by the depth-first is precisely the (integral) area under the depth-first walk (Lemma 3.1).*

Lemma 3.1. *For any tree t, one has*

$$a(t) = \#\mathcal{A}(t) = \sum_{i=0}^{n-1} S_t(i).$$

The first step towards a sampling procedure consists in counting the graphs in $\{G \ : \ T[G] = t\}$, for $t \in \mathbb{T}_n$. The following simple characterization is crucial:

Lemma 3.2. *Let G be a connected labeled graph on $[n]$. Then $T[G] = t$ if and only if G can be obtained from t by adding a subset of the allowed edges in $\mathcal{A}(T)$.*

The following is then clear from Lemma 3.2.

Corollary 3.1. *For any tree $t \in \mathbb{T}_n^L$, we have $\#\{G \in \mathbb{C}_n^L : T(G) = t\} = 2^{a(t)}$.*

One can now design a few "new" sampling procedures for random various kinds of connected graphs. Let C_n be a labeled graph generated as follows: pick a tree t on $[n]$ with probability proportional to $2^{a(t)}$. Then, add a uniformly random subset of $\mathcal{A}(t)$.

Corollary 3.2. *The graph C_n is a uniformly random connected graph on $[n]$.*

As we have seen, the surplus of large connected components of $G(n,p)$, for $p = 1/n + \lambda n^{-4/3}$ is $O(1)$, which is certainly not the case for C_n. So, for an integer $s \geq 0$, let C_n^s be a labeled graph generated by first picking a tree t on $[n]$ with probability proportional to $\binom{a(t)}{s}$, and then adding a uniformly random s-subset of the allowed edges $\mathcal{A}(t)$.

Corollary 3.3. *The graph C_n^s is a uniformly random among the set of connected graph on $[n]$ with surplus s.*

Finally, connected components of $G(n,p)$ may be obtained by *mixing* the previous sampling procedure. Let $\tilde{C}_m^p$ be constructed by first choosing a random tree $\tilde{T}_m^p$ in such a way that $\mathbf{P}(\tilde{T}_m^p = t) \propto (1-p)^{-a(t)}$. Then each edge of $\mathcal{A}(t)$ is added with probability p, independently of the others.

Corollary 3.4. *The graph $\tilde{C}_{m,p}$ is distributed as a connected graph of $G(n,p)$, $n \geq m$, conditioned to have size m.*

Putting together the results above and the remark about the labels, it is easy to see that one may actually sample the graphs (with their labels removed) using the depth-first process (for the canonical tree) and some points under the depth-first process (for the surplus edges). Rather than any bijection, we will need the following specific one which permits to follow the influence of the surplus edges on the metric structure. Consider a fixed tree t, and its depth-first walk S_t. Let $u_1, u_2, \ldots, u_n$ be the nodes of

t enumerated in the depth-first order. At time i, the value $S_t(i)$ of the depth-first walk is precisely the number of nodes u_j, $j > i$, such that u_j is a child of an ancestor of u_i. Write v^i_j, $j = 1, \dots, S_t(i)$ for these nodes, in the reverse depth-first order.

Definition 3.1. *For a depth-first walk S, and a pointset $P \subset \{(x, y) : 0 \le x < n, S(x) > y\}$ let $G(S, P)$ be the graph obtained by*
a. first taking the only plane tree t such that $S_t = S$;
b. then adding the edges $\{u_x, v^x_y\}$, for every $(x, y) \in P$;

Note that the graph $G(S, P)$ has its nodes labeled in the depth-first order, however the actual labels are unimportant for the metric structure; one could of course relabel the graph with the *correct* distribution by enforcing that the first node be labeled one, and distributing uniformly groups of k labels to the nodes of (out-)degree k, $k \ge 0$ which should then be assigned to the nodes so that at every node, the labels of the children are increasing in depth-first order.

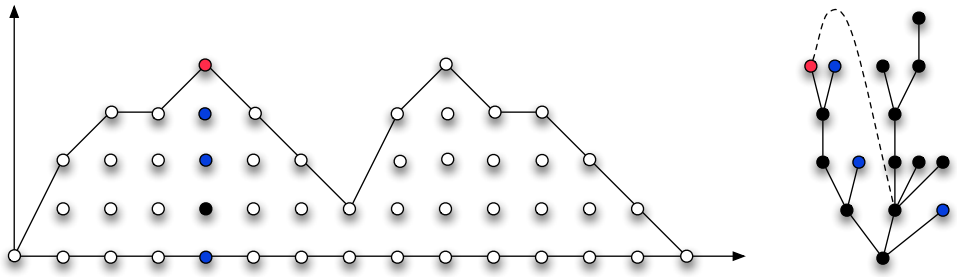

Figure 3.2: *An illustration of the bijection in Definition 3.1: the red point correspond to the node u_i currently explored; the blue ones are the ones in the stack at time i, the black one is the one to which we add an edge.*

3.3 Asymptotics for connected graphs

3.3.1 Convergence

In the previous section, we have seen a sampling procedure for connected graphs distributed like the connected components of $G(n, p)$. We now describe how this may help in getting information about the asymptotic metric structure of such graphs. Most of the metric information is contained in the canonical spanning tree, and we first focus on scaling limit of this tree; the effect of the surplus edges can be recovered by a modification of the scaling limit of the spanning tree.

LIMIT OF THE TILTED TREE. We know that for a sequence of *uniformly* random labeled trees T_n, $n \ge 1$, the rescaled depth-first search process $(n^{-1/2} S_{T_n}(nx))_{x\in[0,1]}$ converges in distribution to a standard Brownian excursion $\mathbf{e} = (\mathbf{e}(x))_{x\in[0,1]}$. We want to use this piece of information to find the limit in distribution of the rescaled depth-first process of the depth-first search tree of the connected graphs we have introduced in Section 3.2. Let C^p_m be distributed like a connected component of $G(n, p)$, conditioned on its size being $m \le n$. To simplify the exposition, we suppose that $m \sim n^{2/3}$, so that here $\sigma = 1$; the general case where $m \sim \sigma n^{2/3}$ is easily recovered using Brownian scaling. By Corollary 3.4, we have $\mathbf{P}(T[C^p_m] = t) \propto (1 - p)^{-a(t)}$.

If this change of measure is well-behaved one can expect that for $\tilde{T}^p_m = T[\tilde{C}^p_m]$, the rescaled depth-first walk $(n^{-1/2} S_{\tilde{T}^p_m}(nx))_{x\in[0,1]}$ converges in distribution to a continuous excursion $\tilde{\mathbf{e}}$ whose distribution is given by, for any Borel set $\mathcal{B}$

$$\mathbf{P}\left(\tilde{\mathbf{e}} \in \mathcal{B}\right) = \frac{\mathbf{E}[\mathbf{1}_{\{\mathbf{e}\in\mathcal{B}\}} \cdot \exp(\int_0^1 \mathbf{e}(x)dx)]}{\mathbf{E}[\exp(\int_0^1 \mathbf{e}(x)dx)]}, \tag{3.2}$$

where $\mathbf{e}$ denotes a standard Brownian excursion. The expression in (3.2) is the very equivalent in the continuous setting of the bias by $(1-p)^{-a(t)}$, since as $n \to \infty$, one expects that $a(t)$ should be of the order of $m \times \sqrt{m} \sim n$, so that $(1-p)^{-a(t)} \approx \exp(a(t)/n)$ should remain bounded and converge. Going back to the case where σ is general, we write $\tilde{\mathbf{e}}^{(\sigma)}(\cdot) = \sqrt{\sigma} \cdot \tilde{\mathbf{e}}(\cdot/\sigma)$. Writing $\tilde{H}^m$ for the height process of $\tilde{T}_m^p$ One then deduces that

Theorem 3.3. *Suppose that $p = p(m)$ is such that $mp^{2/3} \to \sigma$ as $m \to \infty$. Then, as $m \to \infty$,*

$$((m/\sigma)^{-1/2}\tilde{H}^m(\lfloor (m/\sigma)t \rfloor), 0 \le t \le \sigma) \to (2\tilde{\mathbf{e}}^{(\sigma)}(t), 0 \le t \le \sigma)$$

in distribution in the sense of $\mathbb{D}([0,\sigma], \mathbb{R}^+)$.

LIMIT SURPLUS EDGES. Using the bijection of the previous section, the extra edges are represented by a random subset of the integral points under the discrete excursion defined by the Łukasiewicz walk of $S_{\tilde{T}_n^s}$, where each point is present with probability p independently of the others. As $n \to \infty$, this point process should converge to a Poisson process of points under the limit excursion $\tilde{\mathbf{e}}^{(\sigma)}$.

Lemma 3.3. *Let $p = p(m)$ be such that $mp^{2/3} \to \sigma$ as $m \to \infty$. Pick a labeled tree $\tilde{T}_m^p$ on $[m]$ in such a way that $\mathbf{P}(\tilde{T}_m^p = T) \propto (1-p)^{-a(T)}$ and let $\tilde{X}^m$ be the associated depth-first walk. Let $\mathcal{Q}^p \subset \mathbb{Z}^+ \times \mathbb{Z}^+$ be a Binomial pointset of intensity p. Let $\mathcal{P}_m = \{((m/\sigma)^{-1}i, (m/\sigma)^{-1/2}j) : (i,j) \in \mathcal{Q}^p\}$. Then*

$$\left((m/\sigma)^{-1/2}\tilde{X}^m(\lfloor (m/\sigma)\cdot \rfloor), \mathcal{P}_m \cap ((m/\sigma)^{-1/2}\tilde{X}^m(\lfloor (m/\sigma)\cdot \rfloor))\right) \xrightarrow{d} \left(\tilde{\mathbf{e}}^{(\sigma)}, \mathcal{P} \cap \tilde{\mathbf{e}}^{(\sigma)}\right)$$

as $n \to \infty$, where $\mathcal{P}$ is a homogeneous Poisson point process with intensity measure the Lebesgue measure $\mathscr{L}$ on $\mathbb{R}^+ \times \mathbb{R}^+$, and $\mathcal{P}$ is independent of $\tilde{\mathbf{e}}^{(\sigma)}$. Convergence in the first co-ordinate is in $\mathbb{D}([0,\sigma], \mathbb{R}^+)$, and in the second co-ordinate is in the sense of the Hausdorff distance.

Lemma 3.3 gives a nice description of the limit of the *bijective* encoding. Unfortunately, it is not immediately clear that the location of the edges should behave nicely with respect to the *metric*. One way to see that it *must* uses the observation by Marckert and Mokkadem [80] that the height process, which does encode the metric, and the depth-first walk whose limit we know, are actually proportional: in other words, the *bijective* picture of the pair (depth-first walk;point process) asymptotically corresponds exactly to the pair (height process; point process) and the location of the edges are then easily seen to behave nicely. One can deduce that the connected component should converge to a metric space formed by a spanning tree in which one identifies some points as follows:

- the canonical tree converges to the tree encoded by (twice) the limit depth-first process (since this is also the height process!), and
- the surplus edges encoded by the discrete point process should converge to *point identifications* given by the limit point process; each identification occurs between a point, and one of its ancestors.

3.3.2 The limit of connected graphs

In the last section, we have explained informally how a limit connected component should be built. We now construct it directly. We first describe the deterministic operation we will in the construction. For a given excursion h, let $A_h = \{(x,y) : 0 \le x \le \sigma, 0 \le y \le h(x)\}$ be the set of points under h and above the x-axis. Let

$$\ell(\xi) = \ell((x,y)) = \sup\{x' \le x : y = h(x')\}$$
$$r(\xi) = r((x,y)) = \inf\{x' \ge x : y = h(x')\}$$

be the points of $[0,\sigma]$ nearest to x for which $h(\ell(\xi)) = h(r(\xi)) = y$ (see Figure 3.3). It is now straightforward to describe how the points of a finite pointset $\mathcal{Q} \subset A_h$ can be used to make vertex-identifications:

for $\xi \in \mathcal{Q}$, we simply identify the images of x and $r(x)$ in the canonical projection from $[0, \sigma]$ onto $\mathcal{T}_h$. (Hereafter we abuse notation by referring to points of $\mathcal{T}_h$ using some points of $[0, \sigma]$.) We write $g(h, \mathcal{Q})$ for the resulting "glued" metric space; the tree metric is altered in the obvious way to accommodate the vertex-identifications.

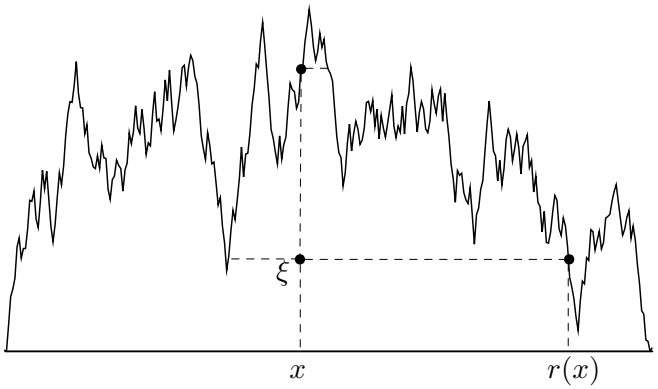

Figure 3.3: *A finite excursion h on $[0, 1]$ coding a compact real tree $\mathcal{T}_h$. Horizontal lines connect points of the excursion which form equivalence classes in the tree. The point $\xi = (x, y)$ yields the identification of the equivalence classes $[x]$ and $[r(x)]$, which are represented by the horizontal dashed lines.*

Given $\tilde{\mathbf{e}}^{(\sigma)}$, write $\mathcal{P}$ for the points of a homogeneous Poisson point process of rate $\frac{1}{2}$ in the plane which fall under the excursion $2\tilde{\mathbf{e}}^{(\sigma)}$. Note that as a consequence of the homogeneity of $\mathcal{P}$, conditional on $\tilde{\mathbf{e}}^{(\sigma)}$, the number of points $|\mathcal{P}|$ has a Poisson distribution with mean $\int_0^\sigma \tilde{\mathbf{e}}^{(\sigma)}(x)dx$.

PROCEDURE 1: VERTEX IDENTIFICATIONS WITHIN A TILTED TREE

1. Sample a tilted excursion $\tilde{\mathbf{e}}^{(\sigma)}$.
2. Sample a set $\mathcal{P}$ containing a $\mathrm{Poisson}\left(\int_0^\sigma \tilde{\mathbf{e}}^{(\sigma)}(x)dx\right)$ number of points uniform in the area under $2\tilde{\mathbf{e}}^{(\sigma)}$.
3. Output $g(2\tilde{\mathbf{e}}^{(\sigma)}, \mathcal{P})$.

The above procedure constructs metric spaces which are distributed as the connected components of critical random graphs; in other words $g(2\tilde{e}^{(\sigma)}, \mathcal{P})$ is the scaling limit of a connected component of $G(n, p)$ conditioned on its size in the following sense:

Theorem 3.4. *Let $\tilde{C}_m^p$ be a connected component of $G(n, p)$ conditioned to have size $m \leq n$, considered as a metric space equipped with the graph distance. Suppose that $mn^{-2/3} \to \sigma \in (0, \infty)$. Then, as $n \to \infty$,*

$$n^{-1/3}\tilde{C}_m^p \to g(2\tilde{\mathbf{e}}^{(\sigma)}, \mathcal{P})$$

in distribution for the Gromov–Hausdorff topology.

3.4 The structural point of view

In this section we adopt a radically different approach closest to the graph theoretic point of view: we see connected graphs as a cycle structure decorated with trees. This permits to exhibit the Brownian continuum random trees hidden in the scaling limit $g(2\tilde{\mathbf{e}}, \mathcal{P})$. This picture also yields some interesting information about the distributions of specific lengths in the connected components.

GRAPHS AND THEIR CYCLE STRUCTURE. The number of surplus edges, or simply *surplus*, of a connected labeled graph $G = (V, E)$ is defined to be $s = s(G) = |E| - |V| + 1$. We say that the connected graph G is *unicylic* if $s = 1$, and *complex* if $s \geq 2$. Define the *core* (sometimes called the *2-core*)

$C = C(G)$ to be the maximum induced subgraph of G which has minimum degree two (so that, in particular, if G is a tree then C is empty). Clearly the graph induced by G on the set of vertices $V \setminus V(C)$ is a forest. So if $u \in V \setminus V(C)$, then there is a unique shortest path in G from u to some $v \in V(C)$, and we denote this v by $c(u)$. We extend the function $c(\,\cdot\,)$ to the rest of V by setting $c(v) = v$ for $v \in V(C)$.

We next define the *kernel* $K = K(G)$ to be the multigraph obtained from $C(G)$ by replacing all paths whose internal vertices all have degree two in C and whose endpoints have degree at least three in C by a single edge [see, e.g., 66]. If the surplus s is at most 1, we agree that the kernel is empty; otherwise the kernel has minimum degree three and precisely $s - 1$ more edges than vertices. It follows that the kernel always has at most $2s$ vertices and at most $3s$ edges. We write $\mathrm{mult}(e)$ for the number of copies of an edge e in K. We now define $\kappa(v)$ to be "the closest bit of K to v", whether that bit happens to be an edge or a vertex. Formally, if $v \in V(K)$ we set $\kappa(v) = v$. If $v \in V(C) \setminus V(K)$ then v lies in a path in G that was contracted to become some copy e_k of an edge e in K; we set $\kappa(v) = e_k$. If $v \in V(G) \setminus V(C)$ then we set $\kappa(v) = \kappa(c(v))$. In this last case, $\kappa(v)$ may be an edge or a vertex, depending on whether or not $c(v)$ is in $V(K)$. The graphs induced by G on the sets $\kappa^{-1}(v)$ or $\kappa^{-1}(e_k)$ for a vertex v or an edge e_k of the kernel K are trees; we call them *vertex trees* and *edge trees*, respectively, and denote them $T(v)$ and $T(e_k)$. In each copy e_k of an edge uv, we distinguish in $T(e_k)$ the vertices that are adjacent to u and v on the unique path from u to v in the core $C(G)$, and thus view $T(e_k)$ as doubly-rooted.

Remark. Before we define the corresponding notions of core and kernel for the limit of a connected graph, it is instructive to discuss the description of a finite connected graph G given in Section 3.3, and to see how the core appears in that picture. Let $G = (V, E)$ be connected and on $V = [m]$ for some $m \geq 1$. Let $T = T[G]$ be the depth-first tree. Let $E^* = E \setminus E(T) \subseteq \mathcal{A}(T)$ be the set of surplus edges which must be added to T in order to obtain G. Let V^* be the set of endpoints of edges in E^*, and let $T_C(G)$ be the union of all shortest paths in $T[G]$ between elements of V^*. Then the core $C(G)$ is precisely $T_C(G)$, together with all edges in E^*, and $T_C(G) = T[C(G)]$.

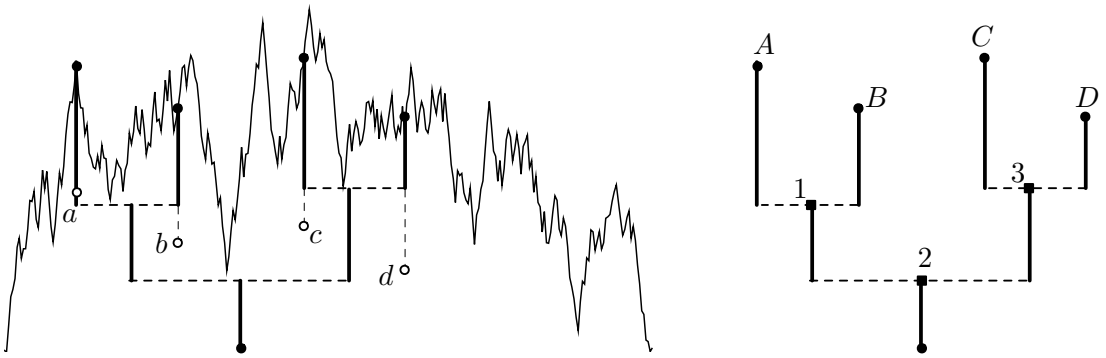

Figure 3.4: *An excursion h and the* reduced tree *which is the subtree $T_R(h, \mathcal{Q})$ of $\mathcal{T}_h$ spanned by the root and the leaves A, B, C, D corresponding to the pointset $\mathcal{Q} = \{a, b, c, d\}$ (which has size $k = 4$). The tree $T_R(h, \mathcal{Q})$ is a combinatorial tree with edge-lengths. It has $2k$ vertices: the root, the leaves and the branch-points $1, 2, 3$. The dashed lines have zero length.*

THE CYCLE STRUCTURE OF SPARSE CONTINUOUS METRIC SPACES. Now consider a real tree $\mathcal{T}_h$ derived from an excursion h, along with a finite pointset $\mathcal{Q} \subset A_h$ which specifies certain vertex-identifications, as described in Section 3.3.2. Let $\mathcal{Q}_x = \{x : \xi = (x, y) \in \mathcal{Q}\}$ and let $\mathcal{Q}_r = \{r(x) : \xi = (x, y) \in \mathcal{Q}\}$, both viewed as sets of points of $\mathcal{T}_h$. We let $T_C(h, \mathcal{Q})$ be the union of all shortest paths in $\mathcal{T}_h$ between vertices in the set $\mathcal{Q}_x \cup \mathcal{Q}_r$. Then $T_C(h, \mathcal{Q})$ is a subtree of $\mathcal{T}_h$, with at most $2|\mathcal{Q}|$ leaves. We define the *core* $C(h, \mathcal{Q})$ of $g(h, \mathcal{Q})$ to be the metric space obtained from $T_C(h, \mathcal{Q})$ by identifying x and $r(x)$ for each $\xi = (x, y) \in \mathcal{Q}$. We obtain the *kernel* $K(h, \mathcal{Q})$ from the core $C(h, \mathcal{Q})$ by replacing each maximal path in $C(h, \mathcal{Q})$ for which all points but the endpoints have degree two by an edge. For an edge uv of $K(h, \mathcal{Q})$, we write $\pi(uv)$ for the path in $C(h, \mathcal{Q})$ corresponding to uv, and $|\pi(uv)|$ for its length.

For each x, let $c(x)$ be the nearest point of $T_C(h, \mathcal{Q})$ to x in $\mathcal{T}_h$. In other words, $c(x)$ is the point of $T_C(h, \mathcal{Q})$ which minimizes $d_h(x, c(x))$. The nearest bit $\kappa(x)$ of $K(h, \mathcal{Q})$ to x is then defined in an analogous way to the definition for finite graphs. For a vertex v of $K(h, \mathcal{Q})$, we define the *vertex tree* $T(v)$ to be the subgraph of $g(h, \mathcal{Q})$ induced by the points in $\kappa^{-1}(v) = \{x : c(x) = v\}$ and the *mass* $\mu(v)$ as the Lebesgue measure of $\kappa^{-1}(v)$. Similarly, for an edge uv of the kernel $K(h, \mathcal{Q})$ we define the *edge tree* $T(uv)$ to be the tree induced by $\kappa^{-1}(uv) = \{x : c(x) \in \pi(uv), c(x) \neq u, c(x) \neq v\} \cup \{u, v\}$ and write $\mu(uv)$ for the Lebesgue measure of $\kappa^{-1}(uv)$. The two points u and v are considered as distinguished in $T(uv)$, and so we again view $T(uv)$ as doubly-rooted. It is easily seen that these sets are countable unions of intervals, so their measures are well-defined. Figures 3.4 and 3.5 illustrate the above definitions.

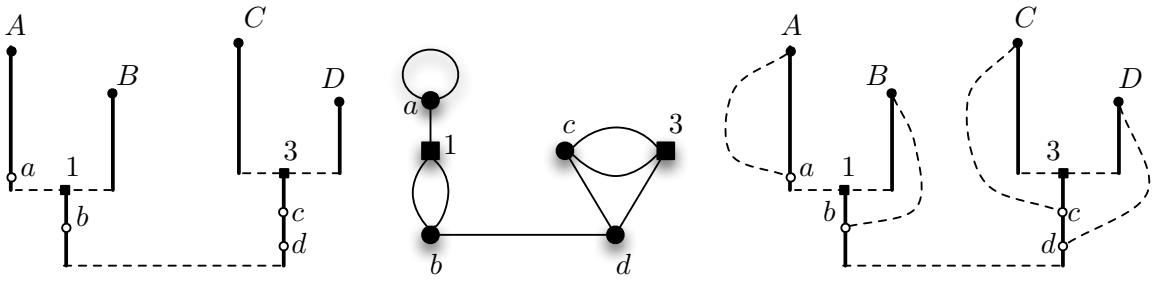

Figure 3.5: *From left to right: the tree $T_C(h, \mathcal{Q})$ from the excursion and pointset of Figure 3.4, the corresponding kernel $K(h, \mathcal{Q})$ and core $C(h, \mathcal{Q})$. The dashed lines indicate vertex identifications.*

SAMPLING A LIMIT CONNECTED COMPONENT. There are two key facts for the first construction procedure. The first is that, for a random metric space $g(2\tilde{\mathbf{e}}, \mathcal{P})$ as above, conditioned on its mass, an edge tree $T(uv)$ is distributed as a Brownian CRT of mass $\mu(uv)$ and the vertex trees are almost surely empty. The second is that the kernel $K(2\tilde{\mathbf{e}}, \mathcal{P})$ is almost surely 3-regular (and so has $2(|\mathcal{P}|-1)$ vertices and $3(|\mathcal{P}|-1)$ edges). Furthermore, for any 3-regular K with t loops,

$$\mathbf{P}\left(K(2\tilde{\mathbf{e}}, \mathcal{P}) = K \mid |\mathcal{P}|\right) \propto \left(2^t \prod_{e \in E(K)} \mathrm{mult}(e)!\right)^{-1}. \tag{3.3}$$

These two facts, together with some additional arguments, justify the validity of the following sampling procedure. Let us condition on $|\mathcal{P}| = k$. As explained before, it then suffices to describe the construction of a component of standard mass $\sigma = 1$.

Theorem 3.5. *The metric space generated by Procedure 2 is distributed as $g(2\tilde{\mathbf{e}}, \mathcal{P})$, conditioned on $|\mathcal{P}| = k$.*

PROCEDURE 2: RANDOMLY RESCALED BROWNIAN CRT'S

- If $k = 0$ then let the component simply be a Brownian CRT of total mass 1.
- If $k = 1$ then let (X_1, X_2) be a Dirichlet$(\frac{1}{2}, \frac{1}{2})$ random vector, let $\mathcal{T}_1, \mathcal{T}_2$ be independent Brownian CRT's of sizes X_1 and X_2, and identify the root of $\mathcal{T}_1$ with a uniform leaf of $\mathcal{T}_1$ and with the root of $\mathcal{T}_2$, to make a "lollipop" shape.
- If $k \geq 2$ then let K be a random 3-regular graph with $2(k-1)$ vertices chosen according to the probability measure in (3.3), above.
  1. Order the edges of K arbitrarily as $e_1, \ldots, e_{3(k-1)}$, with $e_i = u_i v_i$.
  2. Let $(X_1, \ldots, X_{3(k-1)})$ be a Dirichlet$(\frac{1}{2}, \ldots, \frac{1}{2})$ random vector.
  3. Let $\mathcal{T}_1, \ldots, \mathcal{T}_{3(k-1)}$ be independent Brownian CRT's, with tree $\mathcal{T}_i$ having mass X_i, and for each i let r_i and s_i be the root and a uniform leaf of $\mathcal{T}_i$.
  4. Form the component by replacing edge $u_i v_i$ with tree $\mathcal{T}_i$, identifying r_i with u_i and s_i with v_i, for $i = 1, \ldots, 3(k-1)$.

3.5 The stick-breaking construction

In the previous section, we have seen that there are some *Brownian* continuum random trees hidden in the scaling limit $g(2\tilde{\mathbf{e}}, \mathcal{P})$. One of the beguiling features of the Brownian CRT is that it can be constructed in so many different ways, in particular there is a the stick-breaking construction. It is possible to show that $g(2\tilde{\mathbf{e}}, \mathcal{P})$, and in particular all the Brownian continuum random trees it contains may actually *jointly* be constructed using a *single* stick-breaking process. We start by describing shortly the construction for a single Brownian CRT after [4].

STICK-BREAKING CONSTRUCTION OF THE BROWNIAN CRT. Consider an inhomogeneous Poisson process on $[0, \infty)$ with instantaneous rate t at $t > 0$. Let $J_1, J_2, \ldots$ be its inter-jump times, in the order they occur (J_1 being measured from 0). Now construct a sequence of real trees $(\mathcal{A}_n, n \geq 1)$ as follows. First take a (closed) line-segment of length J_1. Then attach another line-segment of length J_2 to a uniform position on the first line-segment. Attach subsequent line-segments at uniform positions on the whole of the structure already created. Finally, take the closure of the object obtained.

Aldous [7] proves that the real tree $\mathcal{A}_n$ is distributed like the subtree of the Brownian CRT spanned by n uniform points and the root. This is the notion of *random finite-dimensional distributions* for continuum random trees [see also 73]. The sequence of these random f.d.d.'s specifies the distribution of the CRT [see 7]. One actually has convergence of $\mathcal{A}_n$ to the Brownian continuum random tree in the strong sense:

Theorem 3.6. *As $n \to \infty$, $\mathcal{A}_n$ converges in distribution to the Brownian CRT in the Gromov–Hausdorff distance* $\mathrm{d_{GH}}$.

This construction for a Brownian CRT may be extended in the following procedure for constructing $g(\tilde{\mathbf{e}}, \mathcal{P})$. In the following, let $\mathrm{U}[0,1]$ denote the uniform distribution on $[0,1]$.

PROCEDURE 3: A STICK-BREAKING CONSTRUCTION

First construct a graph with edge-lengths on which to build the component:

- CASE $k = 0$. Let $\Gamma = 0$ and start the construction from a single point.
- CASE $k = 1$. Sample $\Gamma \sim \text{Gamma}(\frac{3}{2}, \frac{1}{2})$ and $U \sim \text{U}[0,1]$ independently. Take two line-segments of lengths $\sqrt{\Gamma}U$ and $\sqrt{\Gamma}(1-U)$. Identify the two ends of the first line-segment and one end of the second.
- CASE $k \geq 2$. Let $m = 3k-3$ and sample a kernel K according to the distribution (3.3). Sample $\Gamma \sim \text{Gamma}(\frac{m+1}{2}, \frac{1}{2})$ and $(Y_1, Y_2, \ldots, Y_m) \sim \text{Dirichlet}(1, 1, \ldots, 1)$ independently of each other and the kernel. Label the edges of K by $\{1, 2, \ldots, m\}$ arbitrarily and attach a line-segment of length $\sqrt{\Gamma}Y_i$ in the place of edge i, $1 \leq i \leq m$.

Now run an inhomogeneous Poisson process of rate t at time $t > 0$, conditioned to have its first point at $\sqrt{\Gamma}$. For each subsequent inter-jump time J_i, $i \geq 2$, attach a line-segment of length J_i to a uniformly-chosen point on the object constructed so far. Finally, take the closure of the object obtained.

Theorem 3.7. *Procedure 3 generates a component with the same distruction as $g(2\tilde{\mathbf{e}}, \mathcal{P})$ conditioned to have $|\mathcal{P}| = k \geq 1$.*

A few comments are in order to explain Procedure 3. First, this theorem implicitly contains information about the total length of the core of $g(2\tilde{\mathbf{e}}, \mathcal{P})$: remarkably, conditional upon $|\mathcal{P}|$, the total length of the core has precisely the right distribution from which to "start" the inhomogeneous Poisson process. Equivalently, the bias of the entire excursion/tree can be obtained by biasing the lengths to $|\mathcal{P}|$ leaves only.

Also, the joint distribution of the masses of the edge-trees may be explained using this construction. The process may indeed be seen as a continuous urn model, with the m partially-constructed edge trees corresponding to the balls of m different colors in the urn, the probability of adding to a particular edge tree being proportional to the total length of its line segments. Since concentration kicks in as the number of edges of a tree gets large, on may focus of the number of edges of the edge-trees. Let $N_1(n), N_2(n), \ldots, N_m(n)$ be the number of balls at step n of Pólya's urn model started with one ball of each color, and evolving in such a way that every ball picked is returned to the urn along with two extra balls of the same color [see 40]. Then $N_1(0) = N_2(0) = \cdots = N_m(0) = 1$, and the vector

$$\left(\frac{N_1(n)}{m+2n}, \ldots, \frac{N_m(n)}{m+2n} \right)$$

converges almost surely to a limit which has distribution $\text{Dirichlet}(\frac{1}{2}, \ldots, \frac{1}{2})$ [44, Section VII.4], [15, Chapter V, Section 9]. This is also the distribution of the proportions of total mass in each of the edge trees of the component.

Chapter 4

Mean-field minimum spanning trees

In this chapter, we present both the beginning and the end of our story about minimum spanning trees. The beginning consists in pinning down the diameter, and the end – the scaling limit – was only made possible thanks to the work on random graphs presented in Chapter 3. The results on the diameter are based on [P10] written with Louigi Addario-Berry and Bruce Reed, and the scaling limit relies on [P5] which is joint with with Louigi Addario-Berry, Christina Goldschmidt and Grégory Miermont.

4.1 Introduction

Recall the setting described in Section 1.5. We consider a complete graph on n vertices, with edges weighted by independent $[0,1]$-uniform random variables. The uniform distribution is convenient, but not crucial; in particular, any other absolutely continuous distribution would do since as an *unweighted graph*, the minimum spanning tree only depends on the relative ranks of the edges.

In order to estimate distances in the minimum spanning tree M_n, we intend to *track* the relevant information as the tree M_n is constructed by Kruskal's algorithm. Let $F(n,p)$ be the *forest* consisting of the edges of M_n which have weight at most p. The process $(F(n,p), p \in [0,1])$ describes the construction of the minimum spanning tree $M_n = F(n,1)$ as Kruskal's algorithm is performed. Most importantly, $F(n,p)$ is intimately connected to $G(n,p)$ (the version which contains the edges with weight at most p). For instance, the vertex sets of the connected components of $F(n,p)$ and $G(n,p)$ are identical for every p; indeed, Kruskal's algorithm only discard edges that would bind two nodes which are already in the same connected component. This simple observation underlies the entire analysis.

Rather than trying to find a needle in a haystack, we first try to restrict the range of values of p one needs to look at in order to estimate distances.

THE METRIC STRUCTURE OF THE MST AND THE CRITICAL WINDOW. The random graph phase transition immediately suggests that the metric structure of the minimum spanning tree should be built *within or at least close to the critical window.* In the subcritical phase all connected components have size $O(\log n)$, so all distances are at most $O(\log n)$ as well. In the supercritical phase, it suffices to focus on the evolution of the distances in the giant component. Note that every edge added by Kruskal's algorithm is uniformly chosen among all those which do not create a cycle. There, all other connected components have size $O(\log n)$, but the average size is $O(1)$. If all connected components had size one (and in particular were not growing on their own before hooking to the giant component) then every *tree* that is constructed during the supercritical phase would be a *uniform random recursive tree*, a growing tree in which incoming nodes hook up to a uniformly random parent. Such trees of size n have diameter $O(\log n)$, which strongly

suggest that the paths built during the supercritical phase be rather small: the slight variations in the sizes of the small connected components should only modify moderately this picture, and the longest path built in the supercritical phase should have size $O(\log^\beta n)$ for some constant β.

THE DIAMETER OF THE MINIMUM SPANNING TREE. The natural next step consists in addressing the actual order of magnitude of distances in M_n. The above arguments show that one should see most of the construction of the metric structure in $F(n,p)$ as we cross the critical window at $p \sim 1/n$.

For $p = 1/n$ the largest component is a tree of size $\Theta(n^{2/3})$ with probability bounded away from zero [10, 28, 66], and such a tree is uniformly random given its vertex set; this tree *must* appear as a subtree of the minimum spanning tree, so that the diameter of M_n is at least of order $\sqrt{n^{2/3}} = n^{1/3}$ (see also Chapter 3). Most of the work in [P9, P10] consists in establishing the upper bound by showing that the large components hook up nicely among each other, and that no path larger than $O(n^{1/3})$ is ever built. Further arguments about the tail probabilities yield that the diameter of the minimum spanning tree M_n is such that $\mathbf{E}[\mathrm{diam}(M_n)] = \Theta(n^{1/3})$. We explain this in Section 4.2. This provides the scaling factor by which one should divide the distance in order to (hopefully) observe a non-trivial and compact scaling limit for $M_n := F(n,1)$.

THE SCALING LIMIT OF THE MINIMUM SPANNING TREE. Knowing the results about the scaling limit of $G(n,p)$ inside the critical window presented in Chapter 3, one would hope that much more information should be accessible using the connection with Kruskal's algorithm than the mere order of magnitude of the length the longest path. In particular, it is reasonable to expect that for $p = 1/n + \lambda n^{-4/3}$, the scaling limit of $F(n,p)$ should be obtained from that of $G(n,p)$ by removing the cycles that should not be there. In other words, one possible approach to the scaling limit of the minimum spanning tree $M_n = F(n,1)$ would be to

1. choose λ large enough such that $F(n,1)$ and $F(n,p)$ are sufficiently close ($\epsilon n^{1/3}$) in the Gromov–Hausdorff sense (this should be possible since the metric is built close to the critical phase)
2. look at the scaling limit of $G(n,p)$ for this large value of λ, and break down the cycles of the continuum random graph obtained to obtain (at least in distribution) the scaling limit of $F(n,p)$, for this value of λ.

Of course, for any fixed $\lambda \in \mathbb{R}$, $G(n,p)$ and $F(n,p)$ both have multiple large connected components, but there exists a random but finite λ such that the largest one is never again defeated by a smaller challenger for further values of p. This connected component is a good approximation of M_n. Note that again, in the backwards procedure which erases the edges as p decreases, the next edge to be erased is uniformly random: so to obtain (a graph distributed like) $F(n,p)$ from $G(n,p)$ it suffices to remove *uniform random edges* unless doing so would disconnect a connected component. In Section 4.3 we explain how to make this idea formal. These arguments yield the existence of a random compact real tree $\mathcal{M}$ such that, as $n \to \infty$,

$$n^{-1/3} M_n \to \mathscr{M}$$

in distribution in the Gromov–Hausdorff sense, which is the main result of this chapter. The random metric space $\mathscr{M}$ is not the Brownian CRT. In particular, as the scaling $n^{-1/3}$ suggests, its box-counting dimension is three, while that of the Brownian CRT is two.

4.2 Towards a compact object: the diameter

The following theorem answers the question of Frieze and McDiarmid [53, Research Problem 23], proving that the diameter of the minimum spanning tree M_n is actually much smaller than $\sqrt{n}$.

Theorem 4.1. *There exists a constant $C \in (0,\infty)$ such that, for all n large enough,*

$$C^{-1} n^{1/3} \leq \mathbf{E}[\mathrm{diam}(M_n)] \leq C n^{1/3}.$$

As we already mentioned, the rough idea is to control the increase in diameter of $F(n,p)$ as p increases from zero to one. The range of values is more or less easily reduced to a range which is *essentially* the critical window. Two problems arise when trying to make this idea formal:

- we need to cover all bases, and we must control the increase in diameter on a range $(p,1]$, for some point p at which we can bound the diameter, and
- some of the nice properties which allow the necessary control only happen with high probability towards the end of the critical window.

So it seems that the first point requires that p be inside the critical window (to have control at location p) while the second one requires that p be outside (to have control on the increase). To deal with this issue, we start at a random point tailored to provide the best of both worlds. Consider an increasing sequence $1/n < p_0 < p_1 < \ldots < p_t < 1$ of values of p at which we could take a snapshot of the random graph process. Specifically, we fix some large constant f_0, and for $i \geq 1$, we set $f_i = (5/4)^i f_0$, stopping at the first integer t for which $f_t \geq n^{1/3}/\log n$, and choose $p_i = 1/n + f_i/n^{4/3}$. This is similar to Łuczak's method of considering "moments" of the graph process [77].

For each p_i, we consider the largest component $C_1^n(p_i)$ of $G(n,p_i)$. Define D_i to be the diameter of $M_n \cap C_1^n(p_i)$. We intend to control the increase in diameter (more precisely, $D_{i+1} - D_i$) between any two successive p_i and p_{i+1}. Note that D_i might not be the diameter of $F(n,p_i)$, and that $C_1^n(p_i)$ might not be contained in $C_1^n(p_{i+1})$, although we certainly expect that it should happen for i large enough. The increase will be bounded using the following easy lemma. For a graph $G = (V,E)$, and $U \subset V$ we write $G[U]$ to denote the subgraph induced by G on U. We also write $\ell(G)$ for the length of the longest (simple) path in G.

Lemma 4.1. *Let G, G' be graphs such that $G \subset G'$. Let H and H' be connected components of G, G' respectively. Then* $\operatorname{diam}(H') \leq \operatorname{diam}(H) + 2\ell(G'[V - V(H)]) + 2$.

We proceed now in three phases depending on a random index $i^\star \in \{1,2,\ldots,t\}$ which we will soon define: the "early" critical phase $[1/n, p_{i^\star}]$ to decide where to start tracking distances, the late critical phase $[p_{i^\star}, p_t]$, and finally the remainder of the range $[p_t, 1]$.

THE LATE CRITICAL PHASE. For $1 \leq i < t$, we say $G(n,p_i)$ is *well-behaved* if the following events occur:

A_i: $|C_1^n(p_i)| \geq (3/2)n^{2/3}f_i$ and the longest path of $C_1^n(p_i)$ has length at most $f_i^4 n^{1/3}$, and

B_i: the longest path of $G(n,p_{i+1})[V - V(C_1^n(p_i))]$ has length lower than $n^{1/3}/\sqrt{f_i}$.

If $G(n,p_i)$ is well-behaved then by Lemma 4.1, $D_{i+1} - D_i \leq 2n^{1/3}/\sqrt{f_i}$. Let $i^\star$ be the smallest integer for which $G(n,p_j)$ is well-behaved for all $i^\star \leq j < t$ or $i^\star = t$ if $G(n,p_{t-1})$ is not well-behaved. Once the graph is well-behaved, the increase in diameter is easily bounded. By Lemma 4.1, we have deterministically that

$$D_t - D_{i^\star} \;\leq\; 2\sum_{i=i^\star}^{t-1} n^{1/3}/\sqrt{f_i} \;\leq\; 2f_0 n^{1/3}\sum_{i=1}^{t-1}(4/5)^{i/2} \;=\; O(n^{1/3}). \tag{4.1}$$

THE SUPERCRITICAL PHASE. By definition, we have $p_t = 1/n + 1/(n\log n)$, so p_t is not quite supercritical in the sense that $np_t \to 1$, as $n \to \infty$. For such p_t, we cannot prove the poly-logarithmic bound we claimed holds if p_t had been c/n for $c > 1$. However, p_t is far enough from $1/n$ that we are able to prove that $\mathbf{E}[\operatorname{diam}(M_n) - D_t] = O(n^{1/3})$:

Lemma 4.2. *One has* $\mathbf{E}[\operatorname{diam}(M_n)] - \mathbf{E}[D_t] = O(n^{1/6}(\log n)^{7/2})$.

This slight modification in the extents of the phases permits us to even their contributions, and keep p_t within the range $1/n+o(1/n)$, which happens to be crucial to analyze the events A_i and B_i, $i = 1, 2, \dots, t$. It follows that

$$\mathbf{E}[\mathrm{diam}(M_n)] = \mathbf{E}[D_{i^\star}] + O(n^{1/3}). \tag{4.2}$$

SKIPPING THE EARLY CRITICAL PHASE. Finally, it remains to bound $\mathbf{E}[D_{i^\star}]$, which amounts to estimating the distribution of $i^\star$. The key to doing so is to show that for all j between 0 and $t-1$

$$\mathbf{P}(i^\star = j+1) \;\leq\; 6e^{-\sqrt{f_j/8}}. \tag{4.3}$$

Using (4.3) together with (4.1) and the fact that the longest path has length no longer than n yields that

$$\mathbf{E}[D_{i^\star}] \leq f_0^4 n^{1/3} + n\mathbf{P}(i^\star = t) + \sum_{i=1}^{t-1} f_i^4 n^{1/3} \mathbf{P}(i^\star = i),$$

which then yields that $\mathbf{E}[D_{i^\star}] = O(n^{1/3})$. To prove (4.3), we note that if $i^\star = j+1$ and $j > 0$, then one of A_j or B_j must fail. We show that the probability of any of these events happening is small enough:

Lemma 4.3. *The following bounds hold for the events A_j and B_j defined above:*
(a) $\mathbf{P}(A_j \text{ fails}) \leq e^{-\sqrt{f_j}}$
(b) $\mathbf{P}(B_j \text{ fails}) \leq 5e^{-\sqrt{f_j/8}}$.

The proof of Lemma 4.3 is the main technical step, and refines existing estimates which were only asymptotic. The proof goes by analyzing precisely some walks associated to the graph. Even though it is the technical core of the proof, giving even just a decent sketch would require too much space, and dilute the big picture, so we skip it and move on to the description of the scaling limit.

4.3 Rescaling the minimum spanning tree

In this section, we refine the strategy to show that after suitable rescaling of distances and of mass, the minimum tree M_n, viewed as a measured metric space, converges in distribution to a random compact measured metric space $\mathscr{M}$ of total mass measure one, which is a *random real tree*.

The space $\mathscr{M}$ is the scaling limit of the minimum spanning tree on the complete graph. It is binary and its mass measure is concentrated on the leaves. The space $\mathscr{M}$ shares all these features with the Brownian continuum random tree [4, 6, 7, 75]. However, $\mathscr{M}$ is not the Brownian CRT; we rule out this possibility by proving that $\mathscr{M}$ has box-counting dimension three (see Section 4.3.2 for a definition); the CRT has both box-counting dimension two and Hausdorff dimension two.

4.3.1 Description of the results

We consider the minimum spanning tree as an element of $(\mathcal{M}, \mathrm{d_{GHP}})$. In order to do this, we slightly abuse notation and let $n^{-1/3}M_n$ denote the measured metric space obtained from M_n by rescaling distances by $n^{-1/3}$ *and* assigning mass $1/n$ to each vertex; so $n^{-1/3}M_n$ also carries a probability measure, the discrete mass measure. (We will do so in the entire section, but it should always be clear from context in particular the topologies that we are considering measured metric spaces.) The main result of this section is the following theorem.

Theorem 4.2. *There exists a random, compact measured metric space $\mathscr{M}$, such that as $n \to \infty$,*

$$n^{-1/3}M_n \to \mathscr{M}$$

in distribution for the Gromov–Hausdorff–Prokhorov topology. The limit $\mathscr{M}$ is a random real tree. It is almost surely binary, and its mass measure is concentrated on the leaves of $\mathscr{M}$. Furthermore, the laws of $\mathscr{M}$ and of the Brownian CRT are mutually singular.

As mentioned earlier, we approach the study of M_n and of its scaling limit $\mathscr{M}$ via a detailed description of the graph $G(n,p)$ and of the forest $F(n,p)$ for $p = 1/n + \lambda/n^{1/3}$ with $\lambda \in \mathbb{R}$. Write

$$(G_\lambda^{n,i}, i \geq 1)$$

for the components of $G(n, 1/n + \lambda/n^{4/3})$ listed in decreasing order of size. For each $i \geq 1$, we then write $n^{-1/3}G_\lambda^{n,i}$ for the *measured metric space* obtained from $G_\lambda^{n,i}$ by rescaling distances by $n^{-1/3}$ *and* giving each vertex mass $n^{-2/3}$. We likewise define a sequence $(T_\lambda^{n,i}, i \geq 1)$ of graphs, and a sequence $(n^{-1/3}T_\lambda^{n,i}, i \geq 1)$ of measured metric spaces, starting from the forest $F(n, 1/n + \lambda/n^{4/3})$ instead of from the graph $G(n, 1/n + \lambda/n^{4/3})$.

Theorem 3.1 states that for each $\lambda \in \mathbb{R}$, there is a random sequence $(\mathscr{G}_\lambda^i, i \geq 1)$ of compact measured metric spaces, such that

$$(n^{-1/3}G_\lambda^{n,i}, i \geq 1) \xrightarrow{d} (\mathscr{G}_\lambda^i, i \geq 1), \tag{4.4}$$

for the topology of $\mathrm{d}_{\mathrm{GHP}}^4$. (Theorem 3.1 is, in fact, slightly weaker than this because the metric spaces there are considered without their accompanying measures, but it is easily strengthened.) Using the convergence in (4.4) and an analysis of the cycle breaking algorithm (the backward Kruskal procedure), we prove:

Theorem 4.3. *Fix $\lambda \in \mathbb{R}$. Then there exists a random sequence $(\mathscr{T}_\lambda^i, i \geq 1)$ of compact measured metric spaces, in fact compact measured real trees, such that as $n \to \infty$,*

$$(n^{-1/3}T_\lambda^{n,i}, i \geq 1) \xrightarrow{d} (\mathscr{T}_\lambda^i, i \geq 1)$$

for the topology associated to $\mathrm{d}_{\mathrm{GHP}}^4$.

Furthermore, the sequence $(\mathscr{T}_\lambda^i, i \geq 1)$ is constructed from $(\mathscr{G}_\lambda^i, i \geq 1)$ by a continuum analogue of the cycle breaking procedure which samples cut points according to length measure on the core of the real graphs until no more cycle remains. (Recall that the results in Chapter 3 provide precise distributional descriptions of the cores and kernels of the components of $(\mathscr{G}_\lambda^i, i \geq 1)$.) Showing that the continuum analogue of cycle breaking is well-defined and commutes with the appropriate limits is somewhat involved.

Note that, for fixed n, the process tracking the minimum spanning tree of the largest connected component $(n^{-1/3}T_\lambda^{n,1}, \lambda \in \mathbb{R})$ is eventually constant (p actually reaches one, and even passes it), and we write T^n for the space obtained from $\lim_{\lambda\to\infty} T_\lambda^{n,1}$ by giving each vertex mass $1/n$ (this renormalizes the mass, which is $n \times n^{-2/3}$ in $\lim_{\lambda\to\infty} T_\lambda^{n,1}$). Then T^n has the same distribution as $n^{-1/3}M_n$, the measured metric space corresponding to the minimum spanning tree. So proving Theorem 4.2 reduces to proving convergence of T^n.

In order to establish that T^n converges in distribution in the space $(\mathcal{M}, \mathrm{d}_{\mathrm{GHP}})$ as $n \to \infty$, we rely on two ingredients. First, the convergence in Theorem 4.3 implies that the first component $n^{-1/3}T_\lambda^{n,1}$ converges in distribution as $n \to \infty$ to $\mathscr{T}_\lambda^1$ in the space $(\mathcal{M}, \mathrm{d}_{\mathrm{GHP}})$. Second, the results in Section 4.2 entail that

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} \mathbf{P}(\mathrm{d}_{\mathrm{GH}}(n^{-1/3}T_\lambda^{n,1}, T^n) \geq \epsilon) = 0. \tag{4.5}$$

This is enough to prove a version of our main result for the metric spaces without their measures.

The strengthening to the level of measured metric space requires to tweak the measure once again: Let $n^{-1/3}\hat{T}_\lambda^{n,1}$ be the measured metric space obtained from $n^{-1/3}T_\lambda^{n,1}$ by rescaling the measure so that the total mass is one. Then

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} \mathbf{P}(\mathrm{d}_{\mathrm{GHP}}(n^{-1/3}\hat{T}_\lambda^{n,1}, T^n) \geq \epsilon) = 0.$$

Since T^n and $n^{-1/3}M_n$ have the same distribution and $(\mathcal{M}, \mathrm{d}_{\mathrm{GHP}})$ is a complete and separable space, the so-called principle of accompanying laws (Theorem 9.1.13 of [102]) entails that

$$n^{-1/3}M_n \xrightarrow{d} \mathscr{M}$$

in the space $(\mathcal{M}, \mathrm{d_{GHP}})$ for some limiting random measured metric space $\mathscr{M}$ which is thus the scaling limit of the minimum spanning tree on the complete graph. Furthermore, still as a consequence of the principle of accompanying laws, $\mathscr{M}$ is also the limit in distribution of $\mathscr{T}_\lambda^1$ as $\lambda \to \infty$ in the space $(\mathcal{M}, \mathrm{d_{GHP}})$.

4.3.2 Properties of the scaling limit

Finally, we sketch some of the arguments leading to the properties of the scaling limit $\mathscr{M}$, in particular the one which allows us to ensure that $\mathscr{M}$ is not the Brownian CRT.

$\mathscr{M}$ IS BINARY, AND ITS MASS MEASURE IN CONCENTRATED ON THE LEAVES. For fixed $\lambda \in \mathbb{R}$, each component of $\mathscr{T}_\lambda$ is almost surely binary. Since $\mathscr{M}$ is compact and (if the measure is ignored) is an increasing limit of $\mathscr{T}_\lambda^1$ as $\lambda \to \infty$, it will follow that $\mathscr{M}$ is almost surely binary.

To prove that the mass measure is concentrated on the leaves of $\mathscr{M}$, we use a result of Łuczak [77] on the size of the largest component in the barely supercritical regime. This result in particular implies that for all $\epsilon > 0$,

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} \mathbf{P}\left(\left|\frac{|V(T_\lambda^{n,1})|}{2\lambda n^{2/3}} - 1\right| > \epsilon\right) = 0.$$

Since $|V(T^n)|$ has n vertices, it follows that for any $\lambda \in \mathbb{R}$, the proportion of the mass of T^n already present in $T_\lambda^{n,1}$ is asymptotically negligible. But (4.5) tells us that for λ large, with high probability every point of T^n not in $T_\lambda^{n,1}$ has distance $o_{\lambda\to\infty}(1)$ from a point of $T_\lambda^{n,1}$, so has distance $o_{\lambda\to\infty}(1)$ from a leaf of T^n. Passing this argument to the limit, it will follow that $\mathscr{M}$ almost surely places all its mass on its leaves.

THE MEASURED METRIC SPACE $\mathscr{M}$ IS NOT THE CRT. To prove that $\mathscr{M}$ is distinct from the CRT, we look at a simple and natural notion of fractal dimension, the *box-counting* or Minkowski dimension [43]. Given a compact metric space X and $r > 0$, let $N(X, r)$ be the number of balls of radius at most r needed to cover X. We define the lower and upper box-counting dimensions by

$$\underline{\dim}_M(X) = \liminf_{r\downarrow 0} \frac{\log N(X,r)}{\log(1/r)} \qquad \text{and} \qquad \overline{\dim}_M(X) = \limsup_{r\downarrow 0} \frac{\log N(X,r)}{\log(1/r)},$$

respectively. If $\underline{\dim}_M(X) = \overline{\dim}_M(X)$, the common value is called the box-counting dimension and is denoted $\dim_M(X)$.

Proposition 4.1. *As $r \to 0$, we have the following almost sure convergence:*

$$\frac{\log(N(\mathscr{M}, r))}{\log(r^{-1})} \to 3.$$

For the Brownian CRT $\mathcal{T}$, $N(\mathcal{T}, r)$ almost surely has order r^{-2} as $r \to 0$ (see Proposition 5.2 of [38]), so the laws of $\mathscr{M}$ and of the Brownian CRT are mutually singular. To prove Proposition 4.1, we use the asymptotics about the structure of $\mathscr{C}_1^\lambda$, the scaling limit of the largest component of $G(n, 1/n + \lambda n^{-4/3})$, as $\lambda \to \infty$. The intuitive argument goes as follows: the metric space $\mathscr{C}_1^\lambda$ has mass about 2λ and about $\Theta(\lambda^3)$ (the largest excursion of W^λ is essentially the parabola $t\lambda - t^2/2$). Thus, the CRTs which decorate the $\Theta(\lambda^3)$ edges of the kernel of $\mathscr{C}_1^\lambda$ have mass of order λ^{-2}, hence distances of order λ^{-1}. Putting everything together, $N(\mathscr{C}_1^\lambda, \lambda^{-1})$ should be about the number of edges of the kernel, that is $\Theta(\lambda^3)$, hence the box-counting dimension.

Chapter 5

Limit theorems for recursive partitions

In this chapter, we present some results on the asymptotic behaviour of some models of recursive partitions. This has been initially motivated by the estimation of cost of search queries in multidimensional data structures (Section 5.2). The problem of the dual tree of the (self-similar) lamination of the disk (Section 5.3) happened to be amenable to similar ideas. We rely on the documents [P23] which is joint work with Ralph Neininger et Henning Sulzbach and [P25] with Henning Sulzbach.

5.1 Generalities

In this chapter, we present some recent work on the asymptotic behavior of some recursive models which play an important role in computer science via the divide-and-conquer paradigm. The limit processes involved are not completely standard in that they are not Brownian or even Lévy processes. It makes them slightly more complex to capture since the lack of homogeneity or independence between the increments ruins many an approach.

The collection of ideas used in the proofs originate in the now so-called *contraction method* which has been developed mostly in the theoretical computer science literature. The general approach is very natural: one looks for a convergence in distribution for a sequence of random variables; the recursive structure of the divide-and-conquer problems yields equations which bind all the distributions of all these random variables; if the random variables converge in distribution, their law should be a fixed point of a *limit fixed point equation*. Then, the idea is to devise a suitable space of probability measures in which the fixed point equation is a contraction, which ensures by a fixed-point theorem that there is a unique possible limit. The contraction method gives a framework for this approach and general conditions which ensure the convergence to this fixed point.

The collection of results proved using the contraction method has until very recently mostly been focused on scalar or (finite-dimensional) vector-valued random variables. In particular, the method has permitted the analysis of the performance of many algorithms via some parameters such as the path length (sum of the length of paths to the root). Only very recently, Neininger and Sulzbach [90] have developed a general approach for random processes. The results presented in Section 5.2 are a direct application of their theorems. Most of the work consists in proving that the theorem indeed applies which (essentially) requires to construct the limit process, prove that it is continuous, verify that it has the right first moment, and that its supremum has a finite second moment. The results in this section settle open problems which had been left open since the first average-case analysis by Flajolet et al. [51]. Section 5.3 is devoted to a recent results for the related problem of recursive lamination of the disk. We initially intended to take a route which closely follows the one for the quad tree, until we realized that significant short-cuts could

be made; these short-cuts yield both elementary proofs and stronger results, but unfortunately rely on the very specific structure of the recursion at hand.

Before proceeding to the examples, I warn the reader that the presentation here is much more descriptive than in the other chapters, and focuses more on the issues and the objects than on the ideas underlying the proofs.

5.2 Quadtrees and partial match queries

5.2.1 Context and history

Recall the presentation of the model in Section 1.6. To gain a refined understanding of the cost beyond the level of expectations we pursue two directions. First, to quantify the order of typical deviations from the mean we study the order of the variance together with limit distributions. However, deriving higher moments turns out to be subtle. In particular, when the query line is random (like when studying the cost $C_n(\xi)$ at a uniformly random location ξ) although the four subtrees at the root are independent given their sizes, the contributions of the two subtrees that *do hit* the query line are *dependent*. The relative location of the query line inside these two subtrees is again uniform, but unfortunately it is *the same* in both regions. Hence, one cannot easily setup recurrence relations and perform an asymptotic analysis exploiting independence. This issue has not been addressed appropriately in the past, and there is currently no result on the variance or higher moments for $C_n(\xi)$.

The second issue lies in the definition of the cost measure itself: even if the data follow some distribution, should one assume that the query follows the same distribution? In other words, should we focus on $C_n(\xi)$? Maybe not. But then, what distribution should one use for the query line?

One possible approach to overcome both problems is to consider the query line to be fixed and to study $C_n(s)$ for a fixed $s \in [0,1]$. This raises another problem: even if s is fixed at the top level, as the search is performed, the relative location of the queries in the recursive calls varies from one node to another. Thus, in following this approach, one is led to consider the entire stochastic process $(C_n(s))_{s\in[0,1]}$; this is the method we use here.

Write

$$\beta := \frac{\sqrt{17}-3}{2}.$$

Recently Curien and Joseph [33] obtained some results in this direction. They proved that for every fixed $s \in (0,1)$,

$$\mathbf{E}[C_n(s)] \sim K_1(s(1-s))^{\beta/2}n^{\beta}, \qquad \text{and} \qquad K_1 = \frac{\Gamma(2\beta+2)\Gamma(\beta+2)}{2\Gamma(\beta+1)^3\Gamma\left(\beta/2+1\right)^2}. \tag{5.1}$$

On the other hand, Flajolet et al. [51, 52] prove that, along the edge one has $\mathbf{E}[C_n(0)] = \Theta(n^{\sqrt{2}-1})$, so that $\mathbf{E}[C_n(0)] = o(n^{\beta})$ (see also [33]). The behavior about the x-coordinate U of the first data point certainly resembles that along the edge, so that one has $\mathbf{E}[C_n(U)] = o(n^{\beta})$. It suggests that $C_n(s)$ should not be concentrated around its mean, and that $n^{-\beta}C_n(s)$ should converge to a non-trivial random variable as $n \to \infty$.

5.2.2 Main results and implications

We denote by $\mathbb{D}[0,1]$ the space of càdlàg functions on $[0,1]$ and by $\|f\| := \sup_{t\in[0,1]} |f(t)|$ the uniform norm of $f \in \mathbb{D}[0,1]$. Our main contribution is to prove the following convergence result:

Theorem 5.1. *Let $C_n(s)$ be the cost of a partial match query at a fixed line s in a random quadtree. Then, there exists a random continuous function Z such that, as $n \to \infty$,*

$$\left(\frac{C_n(s)}{K_1 n^{\beta}}, s \in [0,1]\right) \xrightarrow{d} (Z(s), s \in [0,1]). \tag{5.2}$$

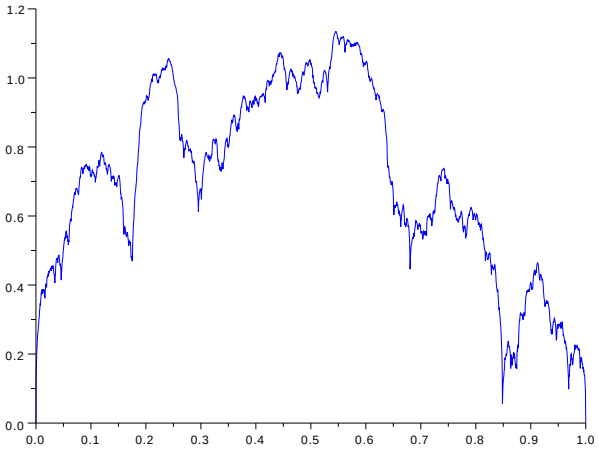

Figure 5.1: *The limit partial match process Z.*

This convergence in distribution holds in $\mathbb{D}[0,1]$ equipped with the Skorokhod topology.

The limit process Z may be characterized as follows (see Figure 5.1 for a simulation).

Proposition 5.1. *The distribution of the random function Z in (5.2) is a fixed point of the following functional recursive distributional equation, as process in $s \in [0,1]$,*

$$Z(s) \stackrel{d}{=} \mathbf{1}_{\{s<U\}}\left[(UV)^\beta Z^{(1)}\left(\frac{s}{U}\right) + (U(1-V))^\beta Z^{(2)}\left(\frac{s}{U}\right)\right] + \mathbf{1}_{\{s\geq U\}}\left[((1-U)V)^\beta Z^{(3)}\left(\frac{s-U}{1-U}\right) + ((1-U)(1-V))^\beta Z^{(4)}\left(\frac{s-U}{1-U}\right)\right], \tag{5.3}$$

where U and V are independent $[0,1]$-uniform random variables and $Z^{(i)}$, $i = 1,\dots,4$ are independent copies of the process Z, which are also independent of U and V. Furthermore, Z in (5.2) is the only continuous solution of (5.3) such that $\mathbf{E}[Z(s)] = (s(1-s))^{\beta/2}$ for all $s \in [0,1]$ and $\mathbf{E}[\|Z\|^2] < \infty$.

The fixed point equation in (5.3), which might at first look awful, is the simple limit which arise from the natural decomposition at the root of the quad tree: either the query falls on the left and there are two contributions, or it falls on the right, and there are also two contributions (see Figure 1.2).

It turns out that the convergence that implies Theorem 5.1 is actually strong enough to guarantee convergence of the variance of the costs of partial match queries. The following theorem for uniform queries ξ is the direct extension of the pioneering work of Flajolet and Puech [49], Flajolet et al. [51] for the expected cost of partial match queries at a uniform line ξ in random two-dimensional trees.

Theorem 5.2. *If ξ is uniformly distributed on $[0,1]$, independent of (C_n) and Z, then*

$$\frac{C_n(\xi)}{K_1 n^\beta} \to Z(\xi),$$

in distribution. Moreover, we have $\mathbf{Var}\left(C_n(\xi)\right) \sim K_4 n^{2\beta}$ where, with K_1 given in (5.1),

$$K_4 := K_1^2 \cdot \mathbf{Var}(Z(\xi)) = K_1^2\left(\frac{2(2\beta+1)}{3(1-\beta)}\mathbf{B}(\beta+1,\beta+1)^2 - \mathbf{B}(\beta/2+1,\beta/2+1)^2\right).$$

Here $\mathrm{B}(a,b) := \int_0^1 t^{a-1}(1-t)^{b-1}dt$ denotes the Eulerian integral for $a, b > -1$. In particular, Theorem 5.2 identifies the asymptotic order of $\mathbf{Var}\,(C_n(\xi))$ which is to be compared with studies that neglected the dependence between the contributions of the subtrees mentioned above [81, 88, 89]. A refined result for the asymptotic order of $\mathbf{Var}\,(C_n(s))$ at a fixed position is

$$\mathbf{Var}\,(C_n(s)) \sim K_1^2 \mathbf{Var}\,(Z(s))\, n^{2\beta},$$

where $s \in (0,1)$.

Another consequence of Theorem 5.1 concerns the order of magnitude of the cost of the worst query $\sup_{s\in[0,1]} C_n(s)$. This guarantees that even the worst query has cost of order n^β.

Theorem 5.3. *Let $S_n = \sup_{s\in[0,1]} C_n(s)$. Then, as $n \to \infty$,*

$$\frac{S_n}{K_1 n^\beta} \to S := \sup_{s\in[0,1]} Z(s)$$

in distribution and with convergence of all moments. In particular,

$$\mathbf{E}[S_n] \sim K_1 n^\beta \mathbf{E}[S], \quad \textit{and} \quad \mathbf{Var}(S_n) \sim K_1^2 n^{2\beta} \mathbf{Var}(S).$$

We also mention the following nice fact: the one-dimensional marginals of the limit process $(Z(s), s \in [0,1])$ are all the same up to a multiplicative constant:

Theorem 5.4. *There is a random variable $Z \geq 0$ such that for all $s \in [0,1]$,*

$$Z(s) \stackrel{d}{=} (s(1-s))^{\beta/2} Z. \tag{5.4}$$

The distribution of Z is characterized by its moments $c_m := \mathbf{E}\,[Z^m]$, $m \in \mathbb{N}$. They are given by $c_1 = 1$ and the recurrence, for $m \geq 2$,

$$c_m = \frac{\beta m + 1}{(m-1)(m+1-3\beta m/2)} \sum_{\ell=1}^{m-1} \binom{m}{\ell} \mathrm{B}(\beta\ell + 1, \beta(m-\ell)+1) c_\ell c_{m-\ell}. \tag{5.5}$$

Convergence of all moments of the supremum $n^{-\beta} S_n$ in Theorem 5.3 implies uniform integrability of any moment of the process $n^{-\beta} C_n$, hence the following result about convergence of all moments.

Theorem 5.5. *For all $s \in [0,1]$, we have*

$$\mathbf{E}\left[\left(\frac{C_n(s)}{K_1 n^\beta}\right)^m\right] \to \mathbf{E}\,[Z(s)^m] = c_m (s(1-s))^{\beta m/2},$$

for all $m \in \mathbb{N}$ as $n \to \infty$ where c_m is given in (5.5). The analogous result holds true for $Z_n(\xi)$ where ξ is uniform on $[0,1]$ and independent of $(Z_n)_{n\geq 0}$ and Z. Moreover, for any natural number $\ell > 0$, positions $0 \leq s_1 < \ldots < s_\ell \leq 1$, and $k_1, \ldots, k_\ell \in \mathbb{N}$ one has

$$\mathbf{E}[C_n^{k_1}(s_1) \cdots C_n^{k_\ell}(s_\ell)] \sim (K_1 n^\beta)^{\sum_{j=1}^\ell k_j} \cdot \mathbf{E}[Z^{k_1}(s_1) \cdots Z^{k_\ell}(s_\ell)].$$

Remark. The results presented here may be extended to the related case of the k-d tree. Details may be found in [P24].

5.3 Recursive laminations of the disk

5.3.1 Context and history

The work of Curien and Le Gall [34] was motivated by the pioneering work of Aldous [8, 9] who studied *uniform random triangulations* of the disk which arise as limiting objects for uniform triangulations of regular n-gons as $n \to \infty$. The recursive nature of the triangulation hides a natural tree, which is dual to the triangulation. Each face of a triangulation is associated with a node and two nodes are connected in the tree if and only if their corresponding faces share an edge in the triangulation. The tree is rooted at a node associated to a fixed edge of the n-gon. In the case of uniform triangulations, the classical bijection between triangulations of the n-gon and rooted binary trees implies that the dual tree is a uniformly random rooted binary tree. Therefore, this tree converges to the Brownian continuum random tree. However, much more is true. By definition, distances in the tree correspond to the number of chords to cross to go from one face to an other. The embedding of the tree inside the n-gon yields an ordering of the nodes which are each associated with one of the n edges of the n-gon. When nodes are listed in this order, the height process of the dual tree converges uniformly to a Brownian excursion after distances have been rescaled by $n^{-1/2}$.

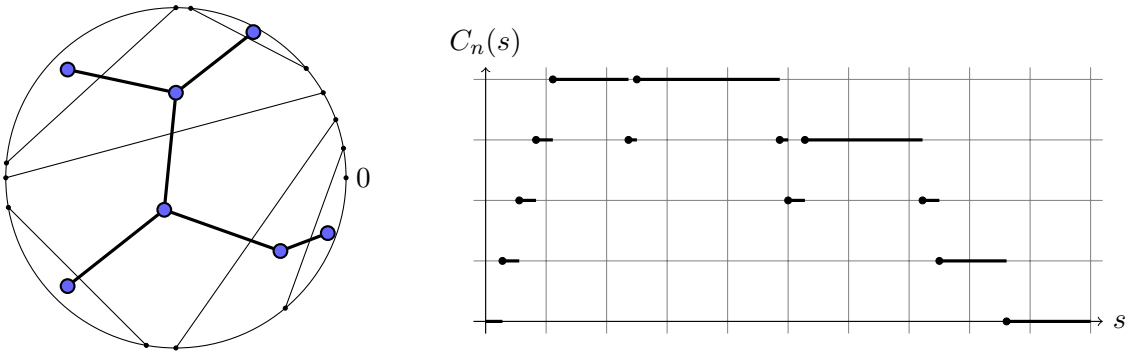

Figure 5.2: *A lamination, its right-continuous height process and the corresponding rooted dual tree. Distances in the tree correspond to the number of chords separating fragments in the laminations.*

5.3.2 The dual tree of the recursive lamination

Before going further, let us introduce some notation. We consider the circle $\mathscr{C} = \{z \in \mathbb{C} : |z| = 1/(2\pi)\}$ as a subset of the complex plane. For convenience, $\mathscr{C}$ is identified with the unit interval where the points 0 and 1 have been glued: we identify $s \in [0,1]$ with the point $\frac{1}{2\pi}\exp(2\pi i s) \in \mathscr{C}$. At some time n, we let $\mathfrak{L}_n$ be the collection of inserted (closed) chords, $C_n(s)$ the number or chords in $\mathfrak{L}_n$ which intersect the straight line going through the points 0 and s of the circle. The value $C_n(s)$ is precisely the height of the node corresponding to the face adjacent to the point s in the dual tree T_n, see Figure 5.2. A priori, for any $n \geq 1$, $C_n(s)$ is not properly defined at endpoints of chords, and we consider a right-continuous version.

Let

$$\beta = \frac{\sqrt{17}-3}{2} = 0.561552\ldots. \tag{5.6}$$

Using the theory of fragmentation processes [19], Curien and Le Gall [34] proved that there exists a random continuous process $\mathscr{M}$ such that for every $s \in [0,1]$ $n^{-\beta/2}C_n(s) \to \mathscr{M}(s)$ in probability, as $n \to \infty$. The process $\mathscr{M}$ encodes the limiting triangulation in the sense of [9] (for a detailed description, see [34, Section 2.3]). Almost surely, for any $\epsilon > 0$, the process $\mathscr{M}$ is $(\beta-\epsilon)$-Hölder continuous, and for any $s \in [0,1]$ we have

$$\mathbf{E}[\mathscr{M}(s)] = \kappa(s(1-s))^{\beta} \tag{5.7}$$

for some constant $\kappa > 0$. Finally, $\mathscr{M}$ inherits the recursive structure of the lamination process and satisfies the following distributional fixed-point equation: let $\mathscr{M}^{(0)}, \mathscr{M}^{(1)}$ denote independent copies of $\mathscr{M}$, let also (U, V) be independent of $(\mathscr{M}^{(0)}, \mathscr{M}^{(1)})$ with density $2\mathbf{1}_{\{0\le u\le v\le 1\}}$ on $[0,1]^2$. Then the process defined by

$$\begin{cases} (1-(V-U))^{\beta}\mathscr{M}^{(0)}\left(\dfrac{s}{1-(V-U)}\right) & \text{if } s < U \\ (1-(V-U))^{\beta}\mathscr{M}^{(0)}\left(\dfrac{U}{1-(V-U)}\right) + (V-U)^{\beta}\mathscr{M}^{(1)}\left(\dfrac{s-U}{V-U}\right) & \text{if } U \le s < V \\ (1-(V-U))^{\beta}\mathscr{M}^{(0)}\left(\dfrac{s-(V-U)}{1-(V-U)}\right) & \text{if } s \ge V, \end{cases} \tag{5.8}$$

is distributed like the initial process $\mathscr{M}$.

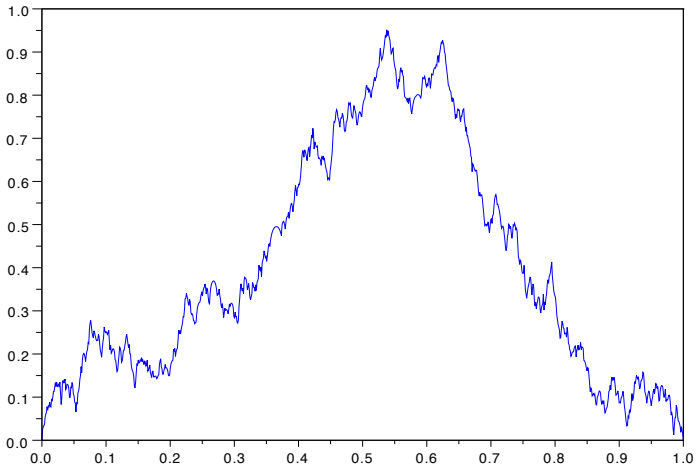

Figure 5.3: *An instance of limit height process $\mathscr{M}$.*

In some sense, the height process or the dual tree is arguably the important object. The proof of convergence of random recursive lamination by Curien and Le Gall [34] relies only on the convergence of the finite dimensional distributions of the height process (no rescaling is needed for the convergence). The main purpose of this section is to show that, after a suitable rescaling of distances, the dual tree T_n of the recursive lamination process converges almost surely to a compact real tree in the Gromov–Hausdorff sense.

Theorem 5.6. *Almost surely as $n \to \infty$,*

$$(T_n, n^{-\beta/2} d_n) \to (\mathcal{T}_{\mathscr{M}}, d_{\mathscr{M}})$$

in the sense of the Gromov–Hausdorff distance between compact metric spaces.

Note that the number of chords N_n inserted by time n is of order $\sqrt{n}$. More precisely, Curien and Le Gall [34] show that $N_n/\sqrt{n} \to \sqrt{\pi}$ almost surely. Thus, in the statement of Theorem 5.6, we may replace $n^{\beta/2}$ by $N_n^{\beta}\pi^{-\beta/2}$ (the scaling of distances in T_n is volume to the β.) The limit metric space $\mathcal{T}_{\mathscr{M}}$ is yet another *natural* random real tree which does not come from a Brownian excursion [4, 6, 7] or an other more general Lévy process [37, 76]. Other examples include the fragmentations trees of Haas and Miermont [57, 58], and the scaling limit of the minimum spanning tree discussed in Chapter 4.

Proposition 5.2. *The real tree $(\mathcal{T}_{\mathscr{M}}, d_{\mathscr{M}})$ is almost surely binary and has its mass concentrated on the leaves. Furthermore, its box-counting dimension equals $1/\beta$ almost surely.*

The proof of Theorem 5.6 relies on Theorem 5.7 below. Here there is some sort of miracle, due to the very specific recursive structure: proving uniform convergence of $n^{-\beta/2}C_n(s)$ for $s \in [0,1]$ simply reduces to proving the convergence at an independent *uniformly random* location. We do not intend to give the formal argument here, but we try to give hint of why this ought to be true:

- there are only two contributing terms (two smaller subregions to examine) when s and 0 are separated by the first chord (case $U \le s < V$ of equation (5.8)), and
- if this happens, then one of these two contributions is completely decoupled: the relative location of the ray in the subregion, $U/(1-V+U)$, is *exactly* uniform and independent of $\{V-U, 1-V+U\}$!

This observation is the key to the entire proof, and explains why we can obtain uniform convergence in all L^p. Indeed, one can treat convergence at a uniformly random location easily (for instance, it is one of the preliminary results in [34]); it then *suggests* that there be actually only one "non-vanishing" contributing subproblem left, regardless of the case, and that there *ought to be* a nice underlying contraction by taking L^p for $p \ge 2$ (even for the uniform norm). Making this idea formal then shows that:

Theorem 5.7. *As $n \to \infty$,*

$$n^{-\beta/2}C_n \to \mathscr{M} \qquad \text{almost surely and in } L^m, \text{ for all } m \in \mathbb{N}. \tag{5.9}$$

Up to a multiplicative constant, the process $\mathscr{M}$ is the unique solution of (5.8) (in distribution) with càdlàg paths continuous at 1 subject to $\int_0^1 \mathbf{E}[\mathscr{M}(s)^2]ds < \infty$.

Note that Theorem 5.7 is actually much stronger than what is needed to prove Gromov–Hausdorff convergence of T_n; indeed, it implies in particular that all the moments of the rescaled height also converge. Note however that, a priori, the process $\mathscr{M}$ is not fully identified because of the free multiplicative constant. (Curien and Le Gall [34] proved that the scaling constant κ in (5.7) exists, but they did need to identify it for the main topic there is the limit lamination which is not affected by this leading constant.) In order to identify $\mathscr{M}$ precisely, we study the asymptotics of $\mathbf{E}[C_n(\xi)]$ for an independent uniform random variable ξ.

Theorem 5.8. *Let $\gamma = \beta/2 + 1$ and $\bar{\gamma} = \frac{-\sqrt{17}+1}{4}$. Then*

$$\mathbf{E}\left[C_n(\xi)\right] = \frac{\sqrt{\pi}}{4}\sum_{k=1}^{n}\binom{n}{k}(-1)^{k+1}\frac{\Gamma(k-\gamma+1)\Gamma(k-\bar{\gamma}+1)}{k!\Gamma(k+3/2)\Gamma(2-\gamma)\Gamma(2-\bar{\gamma})}.$$

Furthermore, as $n \to \infty$,

$$\mathbf{E}[C_n(\xi)] = cn^{\beta/2} + O(1) \qquad \text{with} \qquad c = \frac{\sqrt{\pi}\Gamma(2\gamma-1/2)}{2\Gamma(\gamma)\Gamma^2(\gamma+1/2)} = 1.178226\ldots. \tag{5.10}$$

Theorem 5.8 below is the key to properly identifying the limit dual tree. The first order asymptotics also follows the work of Bertoin and Gnedin [20] on non-conservative fragmentations; the error term may be obtained using their representation with little extra work.

Corollary 5.1. *The process $\mathscr{M}$ in (5.9) is such that*

$$\mathbf{E}[\mathscr{M}(s)] = \kappa(s(1-s))^{\beta}, \qquad \kappa = \frac{c}{\mathrm{B}(\beta+1,\beta+1)} = 3.34443\ldots,$$

where c is given in (5.10), which identifies uniquely the solution of (5.8) among all processes with càdlàg paths continuous at 1 subject to $\int_0^1 \mathbf{E}[\mathscr{M}(s)^2]ds < \infty$.

OTHER SELF-SIMILAR RECURSIVE LAMINATIONS. The lamination process we have introduced is actually an instance of a more general fragmentation process which is also discussed in [34] using a two-stage split procedure: first pick a fragment with probability proportional to its mass to the power $\alpha = 2$ (the Lebesgue measure of the corresponding intersection of the portion of the disk with the circle), then choose the random chord within this fragment by sampling two independent uniform points on the intersection with the circle. In the language of fragmentation theory [19], α is the *index of self-similarity*, and the actual split given the fragment is described by a *dislocation measure*, which is here (essentially) given by the two uniform points conditioned to fall in the same fragment. One may define related fragmentations where the next fragment to split is chosen with probability proportional to its mass to the power $\alpha \in \mathbb{R}$, the cases of interest here are those with $\alpha \geq 0$. When $\alpha \geq 0$, Curien and Le Gall [34] have shown that the limit laminations are all identical (the set of chords are the same); however, and although it encodes the same lamination for every $\alpha \geq 0$, the encoding process (related to the dual trees) depends on whether $\alpha > 0$ or $\alpha = 0$. It is thus a natural question to investigate the dual tree in the case $\alpha = 0$.

5.3.3 The dual tree of the homogeneous lamination

When $\alpha = 0$, the choice of the next fragment is independent of its mass – hence homogeneous – and there is a drastic change in the behaviour of the height process. Note here that every trial yields a new insertion, and the lamination at time n contains n chords. Write $C_n^{(0)}(s)$ for the height in the dual tree of the fragment containing $s \in [0,1]$. Curien and Le Gall [34] prove that for every $s \in (0,1)$ the quantity $n^{-1/3} C_n^{(0)}(s)$ converges in probability as $n \to \infty$, where the point-wise limit $\mathscr{H}(s)$ may be described by a process $\mathscr{H}$ with continuous sample paths which satisfy another, similar but different, fixed-point equation: let $\mathscr{H}^{(0)}$, $\mathscr{H}^{(1)}$ denote independent copies of $\mathscr{H}$ such that $(\mathscr{H}^{(0)}, \mathscr{H}^{(1)})$, (U,V), W are independent, (U,V) has density $2\mathbf{1}_{\{0 \leq u \leq v \leq 1\}}$ and W is uniformly distributed on the unit interval. Then, the process defined by, for every $s \in [0,1]$,

$$\begin{cases} W^{1/3} \mathscr{H}^{(0)}\left(\dfrac{s}{1-(V-U)}\right) & \text{if } s < U \\ W^{1/3} \mathscr{H}^{(0)}\left(\dfrac{U}{1-(V-U)}\right) + (1-W)^{1/3} \mathscr{H}^{(1)}\left(\dfrac{s-U}{V-U}\right) & \text{if } U \leq s < V \\ W^{1/3} \mathscr{H}^{(0)}\left(\dfrac{s-(V-U)}{1-(V-U)}\right) & \text{if } s \geq V. \end{cases} \tag{5.11}$$

is distributed like the original process $\mathscr{H}$. Furthermore, Curien and Le Gall [34] prove that the limit process $\mathscr{H}$, which is distributed like H here, satisfies

$$\mathbf{E}[\mathscr{H}(s)] = \kappa^{(0)} (s(1-s))^{1/2} \tag{5.12}$$

for some unidentified constant $\kappa^{(0)} > 0$.

In this case, the approach used to prove Theorem 5.6 yields the following result: let $T_n^{(0)}$ denote the tree dual to the homogeneous laminations $\mathfrak{L}_n^{(0)}$, and let $d_n^{(0)}$ denote the graph distance in $T_n^{(0)}$.

Theorem 5.9. *Almost surely, as $n \to \infty$,*

$$(T_n^{(0)}, n^{-1/3} d_n^{(0)}) \to (\mathcal{T}_{\mathscr{H}}, d_{\mathscr{H}}),$$

in the Gromov–Hausdorff sense.

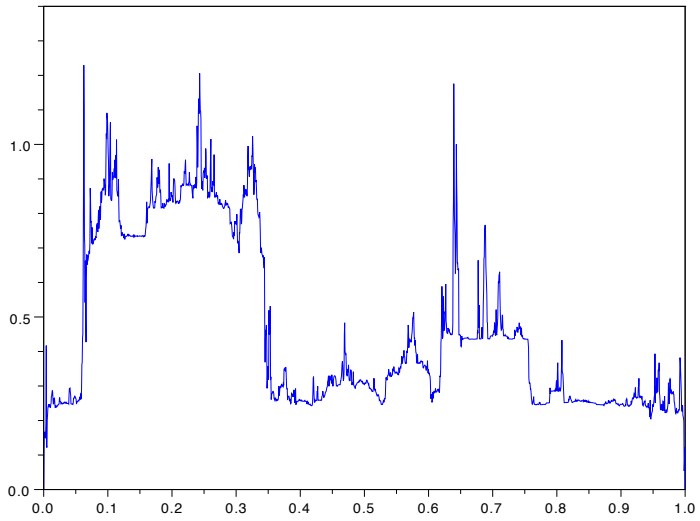

Figure 5.4: *An instance of the limit process $\mathscr{H}$.*

Publications

The documents are available online at http://algo.inria.fr/broutin/publications, via the DOI, or upon request.

[P1] L. Addario-Berry and N. Broutin. Total progeny in killed branching random walk. *Probability Theory and Related Fields*, 151:265–295, 2011. doi:10.1007/s00440-010-0299-2

[P2] L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi. On combinatorial testing problems. *The Annals of Statistics*, 38:3063–3092, 2010. doi:10.1214/10-AOS817

[P3] L. Addario-Berry, N. Broutin, and C. Goldschmidt. Critical random graphs: limiting constructions and distributional properties. *Electronic Journal of Probability*, 15:741–775, 2010.

[P4] L. Addario-Berry, N. Broutin, and C. Goldschmidt. The continuum limit of critical random graphs. *Probability Theory and Related Fields*, 152:367–406, 2012. doi:10.1007/s00440-010-0325-4

[P5] L. Addario-Berry, N. Broutin, C. Goldschmidt, and G. Miermont. The scaling limit of the minimum spanning tree of the complete graph, submitted, 60 p, 2013. arxiv:1301.1664

[P6] L. Addario-Berry, N. Broutin, and C. Holmgren. Cutting down trees with a Markov chainsaw. Submitted, 2011. arxiv:1110.6455

[P7] L. Addario-Berry, N. Broutin, and G. Lugosi. Effective resistance of random trees. *The Annals of Applied Probability*, 19:1092–1107, 2009. doi:10.1214/08-AAP572

[P8] L. Addario-Berry, N. Broutin, and G. Lugosi. The longest minimum weight path in a complete graph. *Combinatorics, Probability and Computing*, 19:1–19, 2010. doi:10.1017/S0963548309990204

[P9] L. Addario-Berry, N. Broutin, and B. Reed. The diameter of the minimum spanning tree of the complete graph. In *Fourth Colloquium on Mathematics and Computer Science Algorithms, Trees Combinatorics and Probability*, volume AG of *DMTCS Proc.*, pages 237–240, 2006.

[P10] L. Addario-Berry, N. Broutin, and B. Reed. Critical random graphs and the structure of a minimum spanning tree. *Random Structures Algorithms*, 35:323–347, 2009. doi:10.1002/rsa.20241

[P11] N. Broutin and L. Devroye. Large deviations for the weighted height of an extended class of trees. *Algorithmica*, 46:271–297, 2006. doi:10.1007/s00453-006-0112-x

[P12] N. Broutin and L. Devroye. The height of list tries and TST. In *International Conference on Analysis of Algorithms*, volume AH of *DMTCS Proceedings*, pages 271–282, 2007.

[P13] N. Broutin and L. Devroye. An analysis of the height of tries with random weights on the edges. *Combinatorics, Probability and Computing*, 17:161–202, 2008. doi:10.1017/S0963548307008796

[P14] N. Broutin, L. Devroye, N. Fraiman, and G. Lugosi. Connectivity threshold for Bluetooth graphs. *Random Structures and Algorithms*, to appear, 2011. doi:10.1002/rsa.20459

[P15] N. Broutin, L. Devroye, and E. McLeish. Weighted height of random trees. *Acta Informatica*, 45: 237–277, 2008. doi:10.1007/s00236-008-0069-0

[P16] N. Broutin, L. Devroye, and E. McLeish. Note on the structure of Kruskal's algorithm. *Algorith-*

mica, 56:141–156, 2010. doi:10.1007/s00453-008-9164-4

[P17] N. Broutin, L. Devroye, E. McLeish, and M. de la Salle. The height of increasing trees. *Random Structures and Algorithms*, 32:494–518, 2008. doi:10.1002/rsa.20202

[P18] N. Broutin and O. Fawzi. Longest distance in random circuits. *Combinatorics, Probability and Computing*, vol. 21, pp. 856–881, 2012. doi:10.1017/S0963548312000260

[P19] N. Broutin and P. Flajolet. The height of random binary unlabelled trees. In *Fifth Colloquium on Mathematics and Computer Science*, volume AI of *DMTCS Proc.*, pages 121–134, 2008.

[P20] N. Broutin and P. Flajolet. The distribution of height and diameter in random non-plane binary trees. *Random Structures and Algorithms*, vol. 41, pp. 215–252, 2012. doi:10.1002/rsa.20393

[P21] N. Broutin and C. Holmgren. The total path length of split trees. *The Annals of Applied Probability*, vol. 22, pp. 1745–1777, 2012. doi:10.1214/11-AAP812

[P22] N. Broutin and J.-F. Marckert. Asymptotics of trees with a prescribed degree sequence and applications. *Random Structures and Algorithms*, to appear, 2012. doi:10.1002/rsa.20463

[P23] N. Broutin, R. Neininger, and H. Sulzbach. Partial match queries in random quadtrees. In Y. Rabani, editor, *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1056–1065, 2012.

[P24] N. Broutin, R. Neininger, and H. Sulzbach. A limit process for partial match queries in random quadtrees. *The Annals of Applied Probability*, to appear, 2012. arxiv:1202.1342

[P25] N. Broutin and H. Sulzbach. The dual tree of a recursive triangulation of the disk, Submitted, 29 pages, 2012. arXiv:1211.1343

References

[1] R. Abraham and J.-F. Delmas. The forest associated with the record process on a Lévy tree. arXiv:1204.2357 [math.PR], 2012.

[2] Louigi Addario-Berry. Tail bounds for the height and width of a random tree with a given degree sequence. *Random Structures & Algorithms*, 41:253–261, 2012.

[3] D. Aldous. The random walk construction of uniform spanning trees and uniform labelled trees. *SIAM Journal on Discrete Mathematics*, 3:450–465, 1990.

[4] D. Aldous. The continuum random tree II: an overview. In M.T. Barlow and N.H. Bingham, editors, *Stochastic Analysis*, pages 23–70. Cambridge University Press, 1991.

[5] D. Aldous. Asymptotic fringe distibutions for general families of random trees. *The Annals of Applied Probability*, 1:228–266, 1991.

[6] D. Aldous. The continuum random tree. I. *The Annals of Probability*, 19:1–28, 1991.

[7] D. Aldous. The continuum random tree III. *The Annals of Probability*, 21:248–289, 1993.

[8] D. Aldous. Recursive self-similarity for random trees, random triangulations and Brownian excursion. *The Annals of Probability*, 22:527–545, 1994.

[9] D. Aldous. Triangulating the circle, at random. *The American Mathematical Monthly*, 101:223–233, 1994.

[10] D. Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *The Annals of Probability*, 25:812–854, 1997.

[11] D. Aldous and J. Pitman. Brownian bridge asymptotics for random mappings. *Random Structures and Algorithms*, 5:487–512, 1994.

[12] D. Aldous and J. Pitman. The standart additive coalescent. *The Annals of Probability*, 26:1703–1726, 1998.

[13] D. Aldous and J. M. Steele. The objective method: probabilistic combinatorial optimization and local weak convergence. In H. Kesten, editor, *Discrete and Combinatorial Probability*, pages 1–72. Springer Verlag, 2003.

[14] D.J. Aldous and B. Pittel. On a random graph with immigrating vertices: Emergence of the giant component. *Random Structures and Algorithms*, 17:79–102, 2000.

[15] K. B. Athreya and P. E. Ney. *Branching Processes*. Springer, Berlin, 1972.

[16] E.A. Bender and E.R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24:296–307, 1978.

[17] N. Berestycki. Recent progress in coalescent theory. *Ensaios Matematicos*, 16:1–193, 2009.

[18] J. Bertoin. A fragmentation process connected to Brownian motion. *Probability Theory and Related Fields*, 117:289–301, 2000.

[19] J. Bertoin. *Random fragmentation and coagulation processes*. Cambridge University Press, Cambridge, 2006.

[20] J. Bertoin and A. Gnedin. Asymptotic laws for nonconservative self-similar fragmentations. *Electronic Journal of Probability*, 9:575–593, 2004.

[21] J. Bertoin and G. Miermont. Asymptotics in Knuth's parking problem for caravans. *Random Structures and Algorithms*, 29:38–55, 2006.

[22] J. Bertoin and J. Pitman. Path transformations connecting Brownian bridge, excursion and meander. *Bull. Sci. Math.*, 118:147–166, 1994.

[23] Jean Bertoin. Fire on trees. arXiv:1011.2308v2 [math.PR], 2011.

[24] Jean Bertoin and Grégory Miermont. The cut-tree of large Galton–Watson trees and the Brownian CRT. 2012.

[25] S. Bhamidi, R. van der Hofstad, and J.S.H. van Leeuwaarden. Novel scaling limits for critical inhomogeneous random graphs. arXiv:0909.1472 [math.PR], 2009.

[26] S. Bhamidi, R. van der Hofstad, and J.S.H. van Leeuwaarden. Scaling limits for critical inhomogeneous random graphs with finite third moments. *Electronic Journal of Probability*, 2012. to appear.

[27] B. Bollobás. The evolution of random graphs. *Transactions of the American Mathematical Society*, 286: 257–274, 1984.

[28] B. Bollobás. *Random Graphs*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, second edition, 2001.

[29] Bela Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1:311–316, 1980.

[30] A. Broder. Generating random spanning trees. In *30th Annual Symposium on Foundations of Computer Science,*, pages 442–447, 1989.

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

[32] P. Chassaing and R. Marchand. Merging costs for the additive Marcus–Lushnikov process, and union-find algorithms. arXiv:math.PR/0406094, 2004.

[33] N. Curien and A. Joseph. Partial match queries in two-dimensional quadtrees: A probabilistic approach. *Advances in Applied Probability*, 43:178–194, 2011.

[34] Nicolas Curien and Jean-François Le Gall. Random recursive triangulation of the disk via fragmentation theory. *The Annals of Probability*, 39:2224–2270, 2011.

[35] N.G. De Bruijn, D. E. Knuth, and S. Rice. The average height of planted plane trees. In R.-C. Read, editor, *Graph Theory and Computing*, pages 15–22, New York, 1972. Academic Press.

[36] M. Drmota and B. Gittenberger. The shape of unlabeled rooted random trees. *European Journal of Combinatorics*, 31:2028–2063, 2010.

[37] T. Duquesne and J.-F. Le Gall. *Random trees, Levy processes and spacial branching processes*, volume 281 of *Asterisque*. 2002.

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

[39] R.T. Durrett and D.L. Iglehart. Functionals of Brownian meander and Brownian excursion. *The Annals of Probability*, 5:130–135, 1977.

[40] F. Eggenberger and G. Pólya. Uber die Statistik verketteter Vorgange. *Zeitschrift fur Angewandte Mathematik und Mechanik*, 3(4):279–289, 1923.

[41] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, 5:17–61, 1960.

[42] S.N. Evans. *Probability and real trees, École d'Été de Probabilités de Saint-Flour XXXV-2005*, volume 1920 of *Lecture Notes in Mathematics*. Springer, 2005.

[43] Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. John Wiley & Sons Ltd., Chichester, 1990.

[44] W. Feller. *An Introduction to Probability Theory and its Applications*, volume II. Wiley, New York, 3rd edition, 1971.

[45] J.A. Fill, N. Kapur, and A. Panholzer. Destruction of very simple trees. *Algorithmica*, 46:345–366, 2006.

[46] R. A. Finkel and J. L. Bentley. Quad trees, a data structure for retrieval on composite keys. *Acta Informatica*, 4:1–19, 1974.

[47] P. Flajolet and A. Odlyzko. The average height of binary trees and other simple trees. *Journal of Computer and System Sciences*, 25(171–213), 1982.

[48] P. Flajolet and A. Odlyzko. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathe-*

matics, 3:216–240, 1990.

[49] P. Flajolet and C. Puech. Partial match retrieval of multidimensional data. *Jounal of the ACM*, 33(2):371–407, 1986.

[50] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, Cambridge, UK, 2009.

[51] P. Flajolet, G. H. Gonnet, C. Puech, and J. M. Robson. Analytic variations on quadtrees. *Algorithmica*, 10: 473–500, 1993.

[52] P. Flajolet, G. Labelle, L. Laforest, and B. Salvy. Hypergeometrics and the cost structure of quadtrees. *Random Structures and Algorithms*, 7:117–144, 1995.

[53] A. Frieze and C. McDiarmid. Algorithmic theory of random graphs. *Random Structures and Algorithms*, 10: 5–42, 1997.

[54] B. Gittenberger and V. Kraus. The degree profile of random Pólya trees. *Journal of Combinatorial Theory, Series A*, 119(7):1528–1557, 2012.

[55] C. Goldschmidt. Critical random hypergraphs: The emergence of a giant set of identifiable vertices. *The Annals of Probability*, 33:1573–1600, 2005.

[56] M. Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*. Birkhauser, 1999.

[57] B. Haas and G. Miermont. The genealogy of self-similar fragmentations with negative index as a continuum random tree. *Electronic Journal of Probability*, 9:57–97, 2004.

[58] B. Haas and G. Miermont. Scaling limits of Markov branching trees with applications to Galton–Watson and random unordered trees. *The Annals of Probability*, 40:2589–2666, 2012.

[59] F. Harary and E.M. Palmer. *Graphical Enumeration*. Academic Press, 1973.

[60] T. E. Harris. *The Theory of Branching Processes*. Springer, Berlin, 1963.

[61] P.C. Hemmer. The random parking problem. *Journal of Statistical Physics*, 57:865–869, 1989.

[62] R. van der Hofstad. Critical behavior in inhomogeneous random graphs. *Random Structures and Algorithms*, 2012. to appear.

[63] S. Janson. Random records and cuttings in complete binary trees. In M. Drmota, P. Flajolet, D. Gardy, and B. Gittenberger, editors, *Mathematics and Computer Science III: Algorithms, Trees, Combinatorics and Probability (Vienna)*, pages 242–253, Basel, 2004. Birkhäuser.

[64] S. Janson. Random cuttings and records in deterministic and random trees. *Random Structures and Algorithms*, 29:139–179, 2006.

[65] S. Janson and J. Spencer. A point process describing the component sizes in the critical window of the random graph evolution. *Combinatorics, Probability and Computing*, 16:631–658, 2007.

[66] S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. Wiley, New York, 2000.

[67] Adrien Joseph. The component sizes of a critical random graph with pre-described degree sequence. ArXiv:1012.2352 [math.PR], 2010.

[68] D.P. Kennedy. The distribution of the maximum Brownian excursion. *Journal of Applied Probability*, 13: 371–376, 1976.

[69] H. Kesten. The incipient infinite cluster in two-dimensional percolation. *Probability theory and related fields*, 73:369–394, 1986.

[70] I. Kortchemski. Invariance principles for Galton–Watson trees conditioned on the number of leaves. *Stochastic Processes and their Applications*, 122:3126–3172, 2012.

[71] M. Kuba and A. Panholzer. Isolating a leaf in rooted trees via random cuttings. *Annals of Combinatorics*, 12: 81–99, 2008.

[72] M. Kuba and A. Panholzer. Isolating nodes in recursive trees. *Aequationes Mathematicae*, 76:258–280, 2008.

[73] J.-F. Le Gall. The uniform random tree in a Brownian excursion. *Probability Theory and Related Fields*, 96: 369–383, 1993.

[74] J.-F. Le Gall. Random trees and applications. *Probability Surveys*, 2:245–311, 2005.

[75] J.-F. Le Gall. Random real trees. *Annales de la faculté des sciences de Toulouse Ser. 6*, 15:35–62, 2006.

[76] J.-F. Le Gall and Y. Le Jan. Branching processes in Levy processes: The exploration process. *The Annals of Probability*, 26:213–252, 1998.

[77] T. Łuczak. Component behavior near the critical point of the random graph process. *Random Structures and Algorithms*, 1(3):287–310, 1990.

[78] T. Łuczak, B. Pittel, and J. C. Wierman. The structure of random graphs at the point of transition. *Transactions of the American Mathematical Society*, 341:721–748, 1994.

[79] J.-F. Marckert and G. Miermont. The CRT is the scaling limit of unordered binary trees. *Random Structures & Algorithms*, 38:467–501, 2011.

[80] J.-F. Marckert and A. Mokkadem. The depth first processes of Galton–watson trees converge to the same Brownian excursion. *The Annals of Probability*, 31:1655–1678, 2003.

[81] C. Martínez, A. Panholzer, and H. Prodinger. Partial match in relaxed multidimensional search trees. *Algorithmica*, 29(1–2):181–204, 2001.

[82] A. Meir and J.W. Moon. Cutting down random trees. *Journal of the Australian Mathematical Society*, 11: 313–324, 1970.

[83] A. Meir and J.W. Moon. Cutting down recursive trees. *Mathematical Biosciences*, 21:173–181, 1974.

[84] A. Meir and J.W. Moon. On the altitude of nodes in random trees. *Canadian Journal of Mathematics*, 30: 997–1015, 1978.

[85] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6:161–179, 1995.

[86] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing*, 7:295–305, 1998.

[87] A. Nachmias and Y. Peres. Critical percolation on random regular graphs. *Random Structures & Algorithms*, 36:111–148, 2010.

[88] R. Neininger. Asymptotic distributions for partial match queries in K-d trees. *Random Structures and Algorithms*, 17:403–427, 2000.

[89] R. Neininger and L. Rüschendorf. Limit laws for partial match queries in quadtrees. *The Annals of Applied Probability*, 11:452–469, 2001.

[90] R. Neininger and H. Sulzbach. On a functional contraction method. Submitted, 2012.

[91] R. Otter. On the number of trees. *Annals of Mathematics*, 49:583–599, 1948.

[92] A. Panholzer. Non-crossing trees revisited: cutting down and spanning subtrees. In *Discrete random walks (Paris 2003)*, volume AC of *DMTCS*, pages 265–276, 2003.

[93] A. Panholzer. Cutting down very simple trees. *Quaestiones Mathematicae*, 29:211–228, 2006.

[94] J. Pitman. Coalescent random forests. *Journal of Combinatorial Theory, Series A*, 85:165–193, 1999.

[95] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer, Berlin, 2006.

[96] G. Pólya. Kombinatorishe Anzahlbestimmungen fúr Gruppen, Graphen und chemishe Verbindungen. *Acta Mathematica*, 68:145–254, 1937.

[97] G. Pólya and R.C. Read. *Combinatorial enumeration of groups, graphs and chemical compounds*. Springer Verlag, New York, 1987.

[98] A. Rényi. On a one-dimensional problem concering random space-filling. *Publ. Math. Inst. Hungar. Acad. Sci.*, 3:109–127, 1958.

[99] A. Rényi and G. Szekeres. On the height of trees. *Journal of the Australian Mathematical Society*, 7:497–507, 1967.

[100] O. Riordan. The phase transition in the configuration model. arXiv:1104.0613 [math.PR], 2011.

[101] D. Rizzolo. Scaling limits of Markov branching trees and Galton-Watson trees conditioned on the number of vertices with out-degree in a given set. arXiv:1105.2528, 2011.

[102] D.W. Stroock. *Probability Theory - An Analytic View*. Cambridge University Press, second edition, 2011.

[103] T.S. Turova. Diffusion approximation for the components in critical inhomogeneous random graphs of rank 1. arXiv:0907.0897, 2009.