# On the shape of binary trees

Mireille Bousquet-Mélou, CNRS, LaBRI, Bordeaux





ArXiv math.CO/0501266 ArXiv math.PR/0500322 (with Svante Janson)

http://www.labri.fr/~bousquet

## A complete binary tree



n internal vertices, called nodes (size n)

n + 1 external vertices, called leaves

# An (incomplete) binary tree



n nodes (size n)

## A plane tree



Each node has a (possibly empty) ordered sequence of children

```
n \text{ edges}, n+1 \text{ nodes } (\text{size } n)
```

What is the shape of a (large) random binary tree?

# The (horizontal) profile of a binary tree



A binary tree of height 4 (or 5...)

## The horizontal profile of a plane tree



A plane tree of height 3.

#### NEW! The vertical profile of a binary tree



A binary tree of right width 3, and vertical profile [2,2;4,2,1,1].

Why study the vertical profile of trees?

## Why study the vertical profile of trees?

Wait and see!

Limit results on the shape of trees: general approach

- Enumerative combinatorics (decompositions of trees, recurrence relations, generating functions...)
- **Singularity analysis** [Flajolet-Odlyzko 90]: extract from the generating functions the asymptotic behaviour of their coefficients

#### **Decomposition and enumeration of binary trees**

Let  $a_n$  be the number of binary trees with n nodes (internal vertices):

$$\begin{cases} a_0 = 1\\ a_n = \sum_{m=0}^{n-1} a_m a_{n-m-1} \end{cases}$$

Let  $A(t) := \sum_{n \ge 0} a_n t^n$  be the associated generating function:

$$A(t) = 1 + tA(t)^2.$$



#### **Decomposition and enumeration of plane trees**

Let  $a_n^*$  be the number of plane trees with *n* edges (n + 1 vertices):

$$\begin{cases} a_0^* = 1 \\ a_n^* = \sum_{m=0}^{n-1} a_m^* a_{n-m-1}^* \end{cases}$$

Let  $A^*(t) := \sum_{n \ge 0} a_n^* t^n$  be the associated generating function:

$$A^*(t) = 1 + tA^*(t)^2.$$



### The Catalan numbers

There are as many binary trees with n nodes as plane trees with n edges:

$$a_n = a_n^*$$

The associated generating function is

$$A(t) = A^*(t) = \sum_n a_n t^n$$

It satisfies

$$A(t) = 1 + tA(t)^{2} \Rightarrow A(t) = \frac{1 - \sqrt{1 - 4t}}{2t} = \sum_{n \ge 0} \frac{1}{n + 1} \binom{2n}{n} t^{n}.$$

Hence

$$a_n = a_n^* = \frac{1}{n+1} \binom{2n}{n}$$

is the *n*th Catalan number.

# Part 1. Height and width of binary trees



#### Counting trees of bounded height

Let  $H_{\leq j}(t) \equiv H_{\leq j}$  be the generating function of binary trees of height at most j:

$$H_{\leqslant j} = \sum_{n \ge 0} h_{n,\leqslant j} t^n$$



Let  $H^*_{\leq j}(t) \equiv H^*_{\leq j}$  be the generating function of plane trees of height at most j, where t counts edges.

$$\begin{aligned} H^*_{\leq 0} &= 1\\ H^*_{\leq j} &= 1 + t H^*_{\leq j-1} H^*_{\leq j} \end{aligned}$$



## Counting trees of bounded (right) width

Let  $W_{\leq j}(t) \equiv W_{\leq j}$  be the generating function of binary trees of right width at most j.



## Trees of bounded height or bounded width: functional equations



## Generating functions for trees of bounded height/width

**Proposition**: The generating function of plane trees of height  $\leq j$  is:

$$H^*_{\leq j} = A \frac{1 - Q^{j+1}}{1 - Q^{j+2}}$$

where A counts plane trees and Q = A - 1 [de Bruijn, Knuth, Rice 72].

#### Generating functions for trees of bounded height/width

**Proposition**: The generating function of plane trees of height  $\leq j$  is:

$$H^*_{\leq j} = A \frac{1 - Q^{j+1}}{1 - Q^{j+2}}$$

where A counts plane trees and Q = A - 1 [de Bruijn, Knuth, Rice 72].

 $\triangleleft \ \triangleleft \ \diamond \ \vartriangleright \ \triangleright$ 

**Proposition** [mbm 06]: The generating function of binary trees of right width  $\leq j$  is:

$$W_{\leq j} = A \frac{(1 - Z^{j+2})(1 - Z^{j+7})}{(1 - Z^{j+4})(1 - Z^{j+5})},$$

where A counts binary trees and  $Z \equiv Z(t)$  is the unique series in t such that

$$Z = t \frac{(1+Z^2)^2}{1-Z+Z^2}$$
 and  $Z(0) = 0.$ 

But. . . why?

## Limit results on the shape of trees: general approach

• Enumerative combinatorics: Using a recursive decomposition of trees, write equations for the relevant generating functions and solve them...

- Use the results (or the technique) of **singularity analysis** [Flajolet-Odlyzko 90] to extract from these series the asymptotic behaviour of their coefficients
  - Results: "If  $B(t) = \sum_{n} b_{n}t^{n}$  behaves like this in the neighborhood of its dominant singularity  $t_{s}$ , then its coefficients  $b_{n}$  behave asymptotically like that."
  - Technique: Cauchy's formula

$$b_n = \frac{1}{2i\pi} \int_{\mathcal{C}_n} B(t) \frac{dt}{t^{n+1}}$$

for a carefully chosen contour  $C_n$ .

#### The average height of binary trees: experimental approach

Consider the average height of plane trees of size n:



#### Convergence in law of the height: experimental approach

The number of trees of size n and height at most j is  $h^*_{n,\leqslant j}.$  For  $x \geqslant 0$  , consider

$$F_n(x) := \mathbb{P}\left(\frac{\mathsf{H}_n^*}{\sqrt{n}} \leqslant x\right) = \frac{1}{a_n} h_{n,\leqslant x\sqrt{n}}^*,$$

the proportion of trees of size n having height at most  $x\sqrt{n}$ . It is the probability that a random tree of size n has height at most  $x\sqrt{n}$ .

Graph of  $F_n(x)$  :



#### Limit results for the height of trees

**Convergence in law**: For all  $x \ge 0$ ,

$$\mathbb{P}\left(\frac{\mathsf{H}_n^*}{\sqrt{n}} \leqslant x\right) \to F(x),$$

where



**Convergence of the moments**: Let  $H_n^*$  be the (random) height of a plane tree with *n* edges. As  $n \to \infty$ ,

$$\frac{\mathbb{E}(\mathsf{H}_n^*)}{\sqrt{n}} \to \sqrt{\pi}, \qquad \mathbb{E}\left(\left(\frac{\mathsf{H}_n^*}{\sqrt{n}}\right)^k\right) \to k(k-1)\Gamma(k/2)\zeta(k) \qquad \text{for } k \ge 2.$$

[Flajolet-Odlyzko 82], [Brown-Schubert 84] + similar results for binary trees

#### Limit results for the right width of binary trees [mbm 06]

**Convergence of the moments**: Let  $W_n$  be the (random) right width of a binary tree with n nodes. Then, as  $n \to \infty$ ,

$$\frac{\mathbb{E}(\mathsf{W}_n)}{n^{1/4}} \to \frac{3\sqrt{\pi}}{\sqrt{2}\Gamma(3/4)}, \qquad \mathbb{E}\left(\left(\frac{\mathsf{W}_n}{n^{1/4}}\right)^2\right) \to 6\sqrt{\pi},$$
$$\mathbb{E}\left(\left(\frac{\mathsf{W}_n}{n^{1/4}}\right)^k\right) = \frac{24\sqrt{\pi}k!\zeta(k-1)}{\sqrt{2^k}\Gamma((k-2)/4)} \quad \text{for } k \ge 3.$$

**Convergence in law:** 

$$\mathbb{P}\left(\frac{\mathsf{W}_n}{n^{1/4}} \ge x\right) \to G(x)$$



# Height and width of binary trees



Part 2. What about the profiles?



## Gallery of (horizontal and vertical) profiles



Horizontal profiles of random binary trees with 1000 nodes.



Vertical profiles of random binary trees with 1000 nodes.

### The average (horizontal and vertical) profiles

Average horizontal profile of plane trees of size 10, 20, 30, 40, 50:



Average vertical profile of binary trees of size 5, 10, 20:



#### One new ingredient: bivariate generating functions

Before: Given j, how many plane trees of size n have height at most j?

$$\Rightarrow$$
 series  $H^*_{\leqslant j}(t) = \sum_n h_{n,\leqslant j} t^n$ 

Now: Given j, how many plane trees of size n have exactly k nodes at height j?

$$\Rightarrow$$
 series  $P_j^*(t,u) = \sum_{n,k} p_{n,k,j} t^n u^k$ 

#### The number of nodes at height j (plane trees)

Let  $P_j^* \equiv P_j^*(t, u)$  be the generating function of plane trees, counted by edges (variable t) and by the number of nodes at height j (variable u):

$$P_0^* = uA(t)$$
  
 $P_j^* = 1 + tP_{j-1}^*P_j^*$ 



## **Proposition** [???]:

$$P_j^* = A \frac{1 - MQ^j}{1 - MQ^{j+1}}$$

where A counts plane trees, Q = A - 1 and

$$M = \frac{A - u - tuA^2}{u + A(1 - u) + tuA^2(1 - A)}.$$

#### The number of nodes at abscissa j (binary trees)

Let  $V_j \equiv V_j(t, u)$  be the generating function of binary trees, counted by nodes (t) and by the number of nodes at abscissa j (variable u).

 $V_0 = 1 + utV_1^2$ For  $j \ge 1$ ,  $V_j = 1 + tV_{j+1}V_{j-1}$ 



**Proposition** [mbm 06]: the series  $V_j$  are algebraic. Moreover,

$$V_j = A \frac{(1 + MZ^j)(1 + MZ^{j+5})}{(1 + MZ^{j+2})(1 + MZ^{j+3})}$$

where A counts binary trees,

$$Z = t \frac{\left(1 + Z^2\right)^2}{1 - Z + Z^2},$$

and  $M \equiv M(t, u)$  is the unique power series in t such that

$$M = (u-1)\frac{Z(1+MZ)^2(1+MZ^2)(1+MZ^6)}{(1+Z)^2(1+Z+Z^2)(1-Z)^3(1-M^2Z^5)}.$$

## The horizontal profile

A tree of size n has, on average, height  $O(\sqrt{n})$ . How many nodes are there at a given (horizontal) level?

???

## The horizontal profile

A tree of size n has, on average, height  $O(\sqrt{n})$ . How many nodes are there at a given (horizontal) level?

About  $\sqrt{n}$ .

## The horizontal profile

A tree of size n has, on average, height  $O(\sqrt{n})$ . How many nodes are there at a given (horizontal) level?

# About $\sqrt{n}$ .

Let  $X_n^*(j)$  be the number of nodes located at height j in a random plane tree with n edges.

We study the quantity

$$\frac{\mathsf{X}_n^*(\lfloor\lambda\sqrt{n}\rfloor)}{\sqrt{n}}.$$

## The vertical profile

A binary tree of size n has, on average, width  $O(n^{1/4})$ . How many nodes are there on a given (vertical) layer?

# About $n^{3/4}$ .

Let  $Y_n(j)$  be the number of nodes located at abscissa j in a random binary tree with n nodes.

We study the quantity

$$\frac{\mathsf{Y}_n(\lfloor \lambda n^{1/4} \rfloor)}{n^{3/4}}.$$

## The horizontal profile of plane trees

• The sequence  $\frac{{\rm X}_n^*(\lfloor\lambda\sqrt{n}\rfloor)}{\sqrt{n}}$  converges in law for each  $\lambda$  (and as a process [Drmota-Gittenberger 97] )

• The first moment:

$$\mathbb{E}\left(\frac{\mathsf{X}_n^*(\lfloor\lambda\sqrt{n}\rfloor)}{\sqrt{n}}\right) \to 2\lambda e^{-\lambda^2}.$$

On average, there are about  $2\lambda e^{-\lambda^2}\sqrt{n}$  nodes at height  $\lfloor \lambda \sqrt{n} \rfloor$  in a plane tree having *n* edges.



In other words, the height of a random node in a random tree, once normalized by  $\sqrt{n}$ , follows a law of density  $2\lambda e^{-\lambda^2}$ .

#### The vertical profile of binary trees

Let  $Y_n(j)$  be the (random) number of nodes at abscissa j in a binary tree having n nodes.

• The sequence  $\frac{Y_n(\lfloor \lambda n^{1/4} \rfloor)}{n^{3/4}}$  converges in law to a random variable  $Y(\lambda)$  described explicitly by its Laplace transform for every  $\lambda$  (and as a process [mbm-Janson 06])

• Moreover,

$$\mathbb{E}\left(\frac{\Upsilon_n(\lfloor\lambda n^{1/4}\rfloor)}{n^{3/4}}\right) \to \frac{1}{\sqrt{2\pi}} \sum_{m \ge 0} \frac{(-\sqrt{2}|\lambda|)^m}{m!} \cos\frac{(m+1)\pi}{4} \Gamma\left(\frac{m+3}{4}\right).$$
This gives the average number of nodes at abscissa  $\lfloor\lambda n^{1/4}\rfloor$  in a random binary tree having  $n$  nodes.

# [mbm 06]

#### An example: nodes at abscissa 0

• The random variable

 $\frac{\mathsf{Y}_n(0)}{n^{3/4}},$ 

which gives the (normalized) number of nodes at abscissa 0, converges in law to a variable Y(0) such that

$$E\left(Y(0)^k\right) = \left(\frac{\sqrt{2}}{3}\right)^k \frac{\Gamma(1+3k/4)}{\Gamma(1+k/2)}.$$

Hence (...)  $Y(0) = \frac{\sqrt{2}}{3\sqrt{T_{2/3}}}$ , where  $T_{2/3}$  follows a unilateral stable law of parameter 2/3.

$$E(e^{-aT_{2/3}}) = e^{-a^{2/3}}$$
 for  $a \ge 0$ .

Merci Alain Rouault !

#### An example: nodes at abscissa 0

• The random variable

 $\frac{\mathsf{Y}_n(0)}{n^{3/4}},$ 

which gives the (normalized) number of nodes at abscissa 0, converges in law to a variable Y(0) such that

$$E\left(Y(0)^k\right) = \left(\frac{\sqrt{2}}{3}\right)^k \frac{\Gamma(1+3k/4)}{\Gamma(1+k/2)}.$$

Hence (...)  $Y(0) = \frac{\sqrt{2}}{3\sqrt{T_{2/3}}}$ , where  $T_{2/3}$  follows a unilateral stable law of parameter 2/3.

$$E(e^{-aT_{2/3}}) = e^{-a^{2/3}}$$
 for  $a \ge 0$ .

Merci Alain Rouault !

• The number of nodes lying at a positive abscissa, normalized by n, converges to U(0,1) [Aldous].

## Part 3. Why study the vertical profile?

• Binary trees (drawn in a canonical way) form a family of embedded trees. There are many other families!





Random increments 0,  $\pm 1$  along edges

- The same questions can be asked for any class of embedded trees:
- what is the largest label? (width)
- How many nodes have label j? (profile)

• They have interesting combinatorial properties (mysterious algebraic series)

• They have interesting combinatorial properties (mysterious algebraic series)

The GF of embedded plane trees with increments  $\pm 1$  having largest label at most j is

$$T_j = T \frac{(1 - Z^{j+1})(1 - Z^{j+5})}{(1 - Z^{j+2})(1 - Z^{j+4})},$$

where  $T = 1 + 2tT^2$  and

$$Z = t \frac{(1+Z)^4}{1+Z^2}.$$

• They have interesting combinatorial properties (mysterious algebraic series)

• Embedded trees with nonnegative labels are related bijectively to planar maps [Cori-Vauquelin 81], [Chassaing-Schaeffer 04], [Del Lungo, Del Ristoro, Penaud 00], [Bouttier-Di Francesco-Guitter 03]



• They have interesting combinatorial properties (mysterious algebraic series)

• Embedded trees with nonnegative labels are related bijectively to planar maps [Cori-Vauquelin 81], [Chassaing-Schaeffer 04], [Del Lungo, Del Ristoro, Penaud 00], [Bouttier-Di Francesco-Guitter 03]



• They are related to ISE (the Integrated SuperBrownian Excursion) [Aldous 93], [Borgs et al. 99]

#### The Integrated SuperBrownian Excursion

The ISE is a (random) probability distribution on  $\mathbb{R}^d$  that occurs (almost) everywhere. At least, as soon as a branching structure (tree) is combined with an embedding of the nodes in the space. The ISE describes how the space is occupied by the nodes [Aldous 93], [Marckert-Mokkadem 03].



• Let  $T_n$  denote a random embedded tree with n nodes. Let  $\mu_n$  be the occupation measure of  $T_n$ :

$$\mu_n = \sum_{v \in T_n} \delta(\ell(v)),$$

where  $\ell(v)$  denotes the label (position) of the node v, and  $\delta(x)$  is the Dirac measure at x.



Histogram of  $\mu_n$ 



• Let  $T_n$  denote a random embedded tree with n nodes. Let  $\mu_n$  be the occupation measure of  $T_n$ :

$$\mu_n = \frac{1}{n} \sum_{v \in T_n} \delta(\ell(v)), \quad ),$$

where  $\ell(v)$  denotes the label (position) of the node v, and  $\delta(x)$  is the Dirac measure at x. Then  $\mu_n$  is a probability distribution (total weight 1).



• Let  $T_n$  denote a random embedded tree with n nodes. Let  $\mu_n$  be the occupation measure of  $T_n$ :

$$\mu_n = \frac{1}{n} \sum_{v \in T_n} \delta(\ell(v) n^{-1/4}),$$

where  $\ell(v)$  denotes the label (position) of the node v, and  $\delta(x)$  is the Dirac measure at x. Then  $\mu_n$  is a probability distribution (total weight 1). It is random since it depends on the random tree  $T_n$ .



• Let  $T_n$  denote a random embedded tree with n nodes. Let  $\mu_n$  be the occupation measure of  $T_n$ :

$$\mu_n = \frac{1}{n} \sum_{v \in T_n} \delta(\ell(v) n^{-1/4}),$$

where  $\ell(v)$  denotes the label (position) of the node v, and  $\delta(x)$  is the Dirac measure at x.

• Thm. As n grows,  $\mu_n \rightarrow \mu_{ise}$ , where  $\mu_{ise}$  is the ISE.

[Aldous 93, Borgs et al. 99, Janson-Marckert 04] This holds for "many" families of random embedded trees!



Histogram of  $\mu_n$ 



## Another occurrence of ISE: Properly embedded trees



In dimension  $d \ge 8$ , the occupation measure of properly embedded trees converges also to the *d*-dimensional ISE. [Derbez-Slade 98]

## Main objective: study ISE (in 1D) via embedded trees



- $\bullet$  Largest label  $\rightarrow$  law of the maximum of the support of the ISE
- Number of nodes at abscissa  $j=\lambda n^{1/4}\to$  law of the density of the ISE at a fixed point  $\lambda$

• Number of nodes at abscissa  $\leqslant j=\lambda n^{1/4}\to$  law of the distribution function of the ISE at a fixed point  $\lambda$ 

## A beautiful slide for Wjcch.

• Let  $M_i$  be the *i*th moment of the ISE ( $M_i$  is a random variable!)

$$M_i = \lim_n \frac{1}{n} \sum_{v \in T_n} \left( \ell(v) n^{-1/4} \right)^i$$

#### A beautiful slide for Wjcch.

• Let  $M_i$  be the *i*th moment of the ISE ( $M_i$  is a random variable!)

$$M_i = \lim_n \frac{1}{n} \sum_{v \in T_n} \left( \ell(v) n^{-1/4} \right)^i$$

• Then  $\mathbb{E}(M_1^{2k}) = 0$  and  $\mathbb{E}(M_1^{2k}) = \frac{a_k \Gamma(1/2)}{2^{k/2} \Gamma((5k-1)/2)}$ , where  $a_0 = -2$  and for k > 0,

$$4a_k = \sum_{i=1}^{k-1} \binom{2k}{2i} a_i a_{k-i} + k(2k-1)(5k-4)(5k-6)a_{k-1}$$

#### A beautiful slide for Wjcch.

• Let  $M_i$  be the *i*th moment of the ISE ( $M_i$  is a random variable!)

$$M_i = \lim_{n} \frac{1}{n} \sum_{v \in T_n} \left( \ell(v) n^{-1/4} \right)^i$$

• Then  $\mathbb{E}(M_1^{2k}) = 0$  and  $\mathbb{E}(M_1^{2k}) = \frac{a_k \Gamma(1/2)}{2^{k/2} \Gamma((5k-1)/2)}$ , where  $a_0 = -2$  and for k > 0,

$$4a_k = \sum_{i=1}^{k-1} \binom{2k}{2i} a_i a_{k-i} + k(2k-1)(5k-4)(5k-6)a_{k-1}$$

• Also,

$$\mathbb{E}(M_1^{2k}M_2^{\ell}) = \frac{a_{k,\ell} \,\Gamma(1/2)}{2^{(k+\ell)/2} \,\Gamma((5k+3\ell-1)/2)},$$

where  $a_{0,0} = -2$  and the  $a_{k,\ell}$  are determined by induction on  $k + \ell$ :

$$a_{k,\ell} = \frac{1}{4} \sum_{(0,0)<(i,j)<(k,\ell)} {\binom{2k}{2i} \binom{\ell}{j} a_{i,j} a_{k-i,\ell-j}} + 2\ell(\ell-1)a_{k+1,\ell-2}$$

$$+\frac{1}{4}k(2k-1)(5k+3\ell-4)(5k+3\ell-6)a_{k-1,\ell}+\frac{1}{2}(4k+1)\ell(5k+3\ell-4)a_{k,\ell-1}.$$

# That's it!



## The Laplace transform of $Y(\lambda)$

Let  $\lambda \ge 0$ . The sequence  $Y_n(\lfloor \lambda n^{1/4} \rfloor)$  converges in distribution to a nonnegative random variable  $Y(\lambda)$  whose Laplace transform is given, for  $|a| < 4/\sqrt{3}$ , by

$$\mathbb{E}\left(e^{aY(\lambda)}\right) = L(\lambda, a)$$

where

$$L(\lambda, a) = 1 + \frac{48}{i\sqrt{\pi}} \int_{\Gamma} \frac{A(a/v^3)e^{-2\lambda v}}{(1 + A(a/v^3)e^{-2\lambda v})^2} v^5 e^{v^4} dv,$$

 $A(x) \equiv A$  is the unique solution of

$$A = \frac{x}{24} \frac{(1+A)^3}{1-A}$$

satisfying A(0) = 0, and the integral is taken over

$$\Gamma = \{1 - te^{-i\pi/4}, t \in (\infty, 0]\} \cup \{1 + te^{-i\pi/4}, t \in [0, \infty)\}.$$

## Did you say "ISE"?

