研究内节点带标号的 Schröder tree

主要内容：symbolic method 导出函数方程，参数的渐进近似，极限分布情况，双射，随机生成算法

感觉是比较标准的 AC 的整个套路

Ranked Schröder Trees[*]

Olivier Bodini[†] Antoine Genitrini[‡] Mehdi Naima[†]

Abstract

进化树

In biology, a phylogenetic tree is a tool to represent the evolutionary relationship between species. Unfortunately, the classical Schröder tree model is not adapted to take into account the chronology between the branching nodes. In particular, it does not answer the question: how many different phylogenetic stories lead to the creation of n species and what is the average time to get there? In this paper, we enrich this model in two distinct ways in order to obtain two ranked tree models for phylogenetics, i.e. models coding chronology.

年表

For that purpose, we first develop a model of (strongly) increasing Schröder trees, symbolically described in the classical context of increasing labeling trees. Then we introduce a generalization for the labeling with some unusual order constraint in Analytic Combinatorics (namely the weakly increasing trees).

Although these models are direct extensions of the Schröder tree model, it appears that they are also in one-to-one correspondence with several classical combinatorial objects. Through the paper, we present these links, exhibit some parameters in typical large trees and conclude the studies with efficient uniform samplers.

Keywords: Phylogenetic tree; Ranked tree; Analytic Combinatorics; Permutations; Ordered Bell numbers; Uniform sampling.

1 Introduction

In biology a *phylogenetic tree* is a classical tool to represent the evolutionary relationship among species. At each bifurcation, or multifurcation, of the tree, the descendant species from distinct branches have distinguished themselves in some manner.

One of the first illustrations of an evolutionary tree was made by Darwin in his book *On The Origin of Species* [6]. His idea was to represent the divergence of characters and species. Multifurcations represent a well-marked variety of a certain kind and this process then continues on the new varieties and so on. Interest grew in tree evolutions as these models give insight on how species evolved. Different tree models were proposed with the idea of finding trees that fits best nowadays observations and data sets. These models of graphs include rooted, unrooted, labeled, unlabeled, bifurcating or multifurcating trees or networks. By defining some metrics between these models, people develop algorithms focusing on state space exploration or on tree inference. For details on tree models in phylogenetics and inference algorithms see the book of Felsenstein [10] and the one of Steel [18] for a more recent survey with combinatorial aspects also. Thanks to the development of bioinformatics many tools have thus emerged, in order to build automatically such tree diagrams. Some examples of programs are *PHYLIP*, a tool for inferring phylogenetic trees [9] or *PAML* that is phylogenetic analyser based on the maximum likelihood [21]. In order to develop these new tools several structural studies have been realized to model correctly the fundamental parameters defined by biologists.

In 1870 Schröder presented an original model published into the paper *Vier combinatorische Probleme* [17]. The fourth problem presents a phylogenetic 系统发生 tree model enumerating trees by their number of leaves. See for example [8] for the phylogenetic interpretation.

While it has been highlighted that this first model is not adapted to take into account the chronology between branching nodes belonging to two distinct fringe subtrees, other approaches have been developed to consider such a history of the evolution process. In particular in the context of binary trees, we can mention the stochastic model of Yule [22] and its generalization by Aldous [1]. Such tree models, including history evolution, are usually called *ranked tree models* in phylogenetics. But these new models are not based on the original Schröder tree model. To the best of our knowledge, there seems to have been no attempt to enrich Schröder's original model so as to encode the chronology of evolution.

So, the main goal of this paper consists in designing ranked tree models based on the classical Schröder structure. In Figure 1 we have represented the same phylogenetic tree on the left handside as a classical Schröder tree, and on the right handside as a *strongly increasing Schröder tree*, the first model we develop in this paper.

[*] This research is partially supported by the ANR project *MetACOnc*, ANR-15-CE40-0014.

[†] Laboratoire d'Informatique de Paris-Nord, CNRS UMR 7030 - Institut Galilée - Université Paris-Nord, 99, avenue Jean-Baptiste Clément, 93430 Villetaneuse, France. Email: {Olivier.Bodini,Mehdi.Naima}@lipn.univ-paris13.fr.

[‡] Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6 - LIP6 - UMR 7606, F-75005 Paris, France. Email: Antoine.Genitrini@lip6.fr.

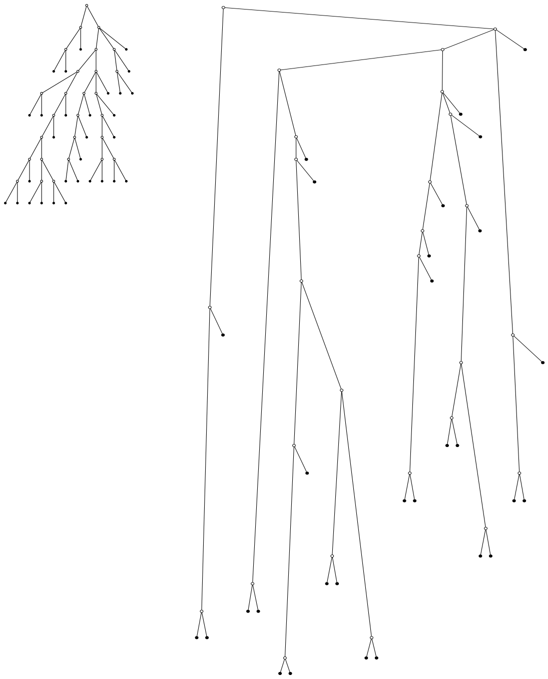


Figure 1: A Schröder tree: without chronological evolution (on the left handside), and with chronological evolution (on the right handside)

A first natural idea in this direction consists in considering the model of a *recursive tree*. Such a structure is a rooted labeled tree, whose root is labeled by 1 and the successors of a given node, with label ν, have a label greater than ν. Each integer between 1 and the total number of nodes is present once in the tree. Many variations of this model have been presented in the literature: see [7] and the references therein. In this context, we are able to define a simple *evolution process* that allows to build very efficiently large trees with simple iterative rules. Furthermore, usually the history of construction is naturally kept in the final large tree through the increasing labeling. It is also important to note that apparently minor changes on the growth rules induce drastic differences in the typical properties of the considered models. See for example the book of Drmota [7] that presents many extensions of the classical model (e.g. plane oriented recursive trees, fixed arity – or out-degree – recursive trees) and details several quantitative studies for different fundamental parameters like the profile of such tree models.

Let us recall the sample of a recursive tree (uniformly for all trees of the same size, i.e. the same number of nodes): *start with the single size-*1 *tree*, reduced to a root, and iterate: *at step* $n \in \{2, 3, \dots\}$ *choose uniformly a node in the tree under construction (labeled with an integer between* 1 *and* $n-1$*) and attach to it a new node labeled by* n.

递归树的随机生成算法
感觉核心在于如何证明分布是均匀的

While many variations on these models have been studied, it is very interesting to note that the increasing version of Schröder trees seems not to have been analyzed. Our model is also very natural due to its similarities to the probabilistic model of Yule trees (cf. e.g. [19]) that take into account the chronological mutations of species.

We develop in this paper two distinct models for phylogenetic trees satisfying in priority two new constraints: (1) to take into account the chronological evolution and (2) to be efficient to simulate. Both models are based on some *increasingly labeling* of Schröder tree structures.

In this paper, we are focusing on the distinct histories possible for a fixed number of final species. From a graph model point of view, it consists in the quantitative study of the number of structures of a given size. Furthermore, beyond some characteristics shared by our model and recursive trees, or increasing fixed arity trees, we will point out several relations to other classical combinatorial objects, in particular permutations, Stirling numbers. Due to the many links to combinatorial objects, increasing Schröder trees are thus interesting in themselves as combinatorial structures.

The paper is organized as follows. In Section 2 we introduce formally our first ranked phylogenetic tree model and introduce a non classical point of view for the tree specification. We present the enumeration of the trees and relate them to permutations. Then we compute important parameters of the model. We conclude this section with the presentation of a linear algorithm for the uniform sampling of trees. Section 3 is devoted to our second model for ranked phylogenetic trees. It is based on a non-classical way of increasingly labeling a tree structure. The section is composed like the first one: after the enumeration of the trees, we relate them to classical combinatorial objects, derive some tree parameters and we finally conclude the section with an efficient unranking algorithm for the uniform sampling of our trees.

枚举计数

参数

均匀分布生成算法

Some technical proofs are detailed in the appendix due to obtain a clear paper structure.

2 Strongly Increasing Schröder trees

The first model we develop is based on a almost classical notion of increasing labeling in Analytic Combinatorics.

2.1 The model and its context

The tree structure associated to strongly increasing Schröder tree corresponds to *Schröder trees*, i.e. the combinatorial class of *rooted plane*[1] *trees whose internal nodes have arity at least* 2. The reader can refer to [11, p. 69] for some details. The size of a Schröder tree is the number of leaves in the tree. Note that in the tree structure neither the internal nodes, nor the leaves are labeled. The combinatorial class $\mathcal{S}$ of Schröder trees is specified as $\mathcal{S} = \mathcal{Z} \cup \textsc{Seq}_{\geq 2}\,\mathcal{S}$, that translates, via the classical *symbolic method* presented by Flajolet and Sedgewick [11], into the following equation, $S(z) = z + \frac{S(z)^2}{1-S(z)}$, satisfied by its ordinary generating function $S(z) = \sum_{n\geq 1} s_n z^n$ where s_n is the number of structures of size n (i.e. with n leaves).

In this section, we are interested in an increasingly labeled variation of Schröder trees.

Definition 2.1. *A* strongly increasing Schröder tree *has a tree structure that is a Schröder tree and moreover its internal nodes are labeled with the integers between* 1 *and* ℓ *(where* ℓ *is the number of internal nodes), in such a way that all labels are distinct and the sequence of labels in each path from the root to a leaf is increasing.*

Note, in the Analytic Combinatorics context, such a labeling of trees is called *increasing labeling* (without the term *strongly*). In order to distinguish clearly this first model from the second one presented in Section 3 we have added this term. But from here, inside this section we will use the classical denotation increasing tree. Trees that are increasingly labeled can be in a certain extent specified with the Greene's operator $\Box_\star$ (cf. for example [11, p. 139]). Then the specification is translated into an equation satisfied by the exponential generating function. But in our context, the size of a tree is the number of leaves (which corresponds to the final number of species), and the increasingly labeling constraint is related to the internal nodes. We specify this class by using a second variable u to mark the internal nodes.

$$S^*(z,u) = \sum_{n,\ell} s_{n,\ell} z^n \frac{u^\ell}{\ell!} = z + \int_{v=0}^{u} \frac{S^*(z,v)^2}{1-S^*(z,v)} \mathrm{d}v.$$

While the integral equation could be analyzed further, we prefer, in the following, to introduce an alternative way to define our objects. This new approach is easier to handle and it also naturally extends to define our second model of trees developed in Section 3, namely the weakly increasing Schröder trees.

In Figure 2 we have represented an increasing Schröder tree of size 30 with 27 internal nodes. This increasing tree is the same tree as the one represented in Figure 1 with the chronological evolution, where the internal node labeled by ℓ is laid on level $\ell - 1$, for all $\ell \in \{1, \ldots, 27\}$.

AC 里确实有一些介绍

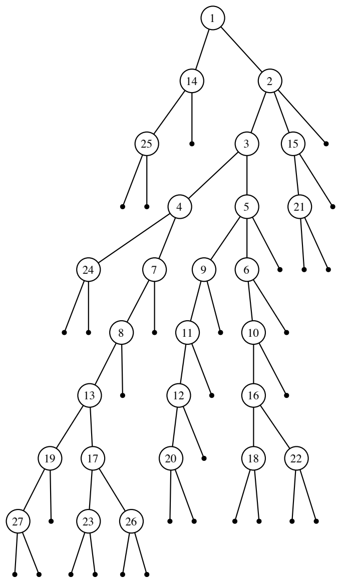


Figure 2: A strongly increasing Schröder tree

In order to describe the building of increasing Schröder trees, we introduce an *evolution process*. It consists in an iterative way that substitutes a leaf by an internal node attached to several leaves. More formally:

- *Start with a single (unlabeled) leaf;*
- Iterate the following process: *at step* ℓ *(for* $\ell \geq 1$*), select a leaf and replace it by an internal node with label* ℓ *attached to a sequence of at least two leaves.*

这个迭代还是比较容易理解的

Remark that the increasing labeling corresponds to the chronology of the tree building.

2.2 Exact enumeration and relationship with permutations

Let us denote by $\mathcal{T}$ the class of increasing Schröder trees. By using the evolution process we exhibit a specification for $\mathcal{T}$ as follows:

$$\mathcal{T} = \mathcal{Z} \cup \left(\Theta\mathcal{T} \times \textsc{Seq}_{\geq 1}(\mathcal{Z})\right). \tag{2.1}$$

In this specification, $\mathcal{Z}$ stands for the leaves, and the operator Θ is the classical pointing operator (cf. in [11, p. 86] for details). The specification is a direct rewriting of the evolution process. A tree is either reduced to a leaf or at each step an atom (i.e. a leaf) is pointed in the tree under construction and is replaced by an internal node (whose labeling is deterministic: it corresponds to

[1] A plane tree is such that the children of a node are ordered.

the step number) attached to a sequence of at least two leaves (the one that has been pointed is reused as the leftmost child, it is the reason why the operator $\text{SEQ}_{\geq 1}$ does not contain the empty sequence and starts with sequences containing one element).

The symbolic method translates this specification into a functional equation satisfied by the generating series associated to the combinatorial class. Note that the functional equation is satisfied by the *ordinary generating series* associated to $\mathcal{T}$: $T(z) = \sum_{n\geq 1} t_n\, z^n$. The increasing labeling is here transparent and thus the objects seems not labeled (in fact, the leaves, marked by $\mathcal{Z}$ are really unlabeled):

一阶线性微分方程，但得到通解后无法待定常系数可能是因为在0处不解析

$$T(z) = z + \frac{z^2}{1-z}\, T'(z). \tag{2.2}$$

By extracting the coefficients of the series, we derive the two following recurrences.

$$\left\{\begin{array}{l} t_1 = 1, \quad t_2 = 1, \\ \text{and for } n > 2, \\ t_n = n \cdot t_{n-1}. \end{array}\right. \qquad \left\{\begin{array}{l} t_1 = 1, \\ \text{and for } n > 1, \\ t_n = \sum_{k=1}^{n-1} k \cdot t_k. \end{array}\right. \tag{2.3}$$

Both recurrences are computed thanks to equation (2.2). The direct extraction $[z^n]\ T(z)$ exhibits the rightmost recurrence. This recurrence exhibits that the calculation of the n-th term is of quadratic complexity (in the number of arithmetic operations). The leftmost recurrence is obtained by extracting $[z^n]\ (1-z) \cdot T(z)$ and then by simplifying the resulted equation. Here the calculation of the n-th term is of linear complexity.

Thus we directly prove $t_n = n!/2$ for all $n \geq 2$. The sequence $(t_n)_n$ appears under the reference `OEIS A001710`[2]. Observing the growth rate of $(t_n)_n$ proves that the ordinary generating series $T(z)$ is formal: its radius of convergence is 0.

收敛半径为0

2.3 Analysis of typical parameters Here we are interested in the quantitative study of four distinct parameters of increasing Schröder trees. The first one corresponds to the number of internal (labeled) nodes of a size-n tree. This fundamental parameter corresponds to the number of steps in the evolution process that are necessary to build the given tree. Recall the arity of internal nodes is at least two, thus this parameter is not deterministic. The second and the third parameters are related to the root node. We study its arity and the number of leaves attached to it in a typical tree of size n. But in a tree of size n (tending to infinity) all internal nodes whose labels are independent from n have the same characteristics than the root: thus these two parameters are also important for the global quantitative aspects of a large tree. Finally, the fourth parameter corresponds to the typical number of binary nodes in a large tree. This study becomes natural once we have seen the typical value of the number of internal nodes of a large tree.

Quantitative analysis of the number of iteration steps A fundamental parameter characterizing the increasing Schröder trees is their number of internal nodes. This parameter is interesting in itself, but furthermore it corresponds to the maximal label value in the tree, and thus it is also the number of steps of the building process.

To study both the number of internal nodes and the number of leaves, we enrich the specification (2.1) with an additional parameter $\mathcal{U}$ marking the internal nodes.

$$\mathcal{T} = \mathcal{Z} \cup \big(\mathcal{U} \times \Theta_{\mathcal{Z}}\mathcal{T} \times \text{SEQ}_{\geq 1}(\mathcal{Z})\big)\,;$$
$$T(z,u) = z + \frac{u\, z^2}{1-z}\, \partial_z T(z,u). \tag{2.4}$$

The operator $\Theta_{\mathcal{Z}}$ consists in pointing an element marked by $\mathcal{Z}$. The partial differentiation according to z is written as $\partial_z \cdot$. With the notation $T(z,u) = \sum_{n\geq 1} t_n(u) z^n$, the equation (2.4) gives two recurrences satisfied either by $(t_n(u))$, or by $(t_{n,k})$, where $t_{n,k}$ is the number of trees with n leaves and k internal nodes (that are increasingly labeled):

$$\left\{\begin{array}{l} t_1(u) = 1, \qquad t_2(u) = u, \\ \text{and if } n > 2, \\ t_n(u) = (1 + (n-1)u) t_{n-1}(u); \end{array}\right.$$
$$\left\{\begin{array}{lll} t_{n,k} = t_{n-1,k} + (n-1)\, t_{n-1,k-1} & & \text{if } 0 < k < n \\ t_{1,0} = 1, & t_{n,1} = 1 & \text{if } n > 1 \text{ and} \\ t_{i,j} = 0 & & \text{otherwise.} \end{array}\right. \tag{2.5}$$

Remark that the extremal conditions are trivially obtained through our construction in particular the sequence $(t_{n,n-1})_n$ is enumerating increasing binary trees. Once again, these efficient recurrences are obtained thanks to the extraction of $[z^n](1-z) \cdot T(z,u)$. In Figure 3, for the tree size-n from 1 to 6, we present the distribution

1 ,					
0 ,	1,				
0 ,	1,	2,			
0 ,	1,	5,	6,		
0 ,	1,	9,	26,	24,	
0 ,	1,	14,	71,	154,	120,

Figure 3: Distribution of $t_{n,k}$ for size-n trees, $n \in \{1, 2, \ldots, 6\}$, of the number of internal nodes $k \in \{0, 1, \ldots, n-1\}$

[2] Throughout this paper, a reference `OEIS A···` points to Sloane's Online Encyclopedia of Integer Sequences `www.oeis.org`.

of the number of trees according to their number k of internal nodes.

The *Borel transform*, denoted as $\mathcal{B}\cdot$, translates an ordinary generating series into its analog exponential generating series. For example, we obtain $\mathcal{B}T(z) = \sum_{n\geq 1} t_n \ \frac{z^n}{n!}$. In particular, due to the growth of the coefficients $(t_n)_n$ we directly observe that $\mathcal{B}T(z)$ is analytic around 0 (with radius of convergence 1).

Proposition 2.1. *The Borel transform on $T(z,u)$ relatively to the variable z gives*

$$\mathcal{B}T(z,u) = \sum_{n\geq 1}\sum_{k=0}^{n-1} t_{n,k}\ u^k \ \frac{z^n}{n!} = \frac{u(1-zu)^{-\frac{1}{u}} - u + z}{1+u}.$$

Here we just present the key-ideas of the proof, but details are given in Appendix A.

Proof. [Key-ideas] Applying the Borel transform on equation (2.4) and then classical properties of the Borel transform for the function $z \cdot f(z)$ and for the derivative $f'(z)$ yields the result.

Let us come back to the polynomial $t_n(u) = \sum_{k=0}^{n-1} t_{n,k}\ u^k$. It corresponds almost to the sequence `OEIS A145324` related to Stirling numbers.

Corollary 2.1. *Let $n \geq 2$. The distribution of the number of internal nodes in increasing Schröder trees of size n is*

$$t_n(u) = \sum_{k=0}^{n-1} t_{n,k}\ u^k = u\prod_{\ell=2}^{n-1}(1+\ell u).$$

The proof relies on a direct rewriting of the first recurrence in equation (2.5). The generating function corresponds to the n-th row in the triangle presented in Figure 3. Although the sequence $(t_n(u))$ is stored in `OEIS` we exhibit here another link with a very classical triangle. By reading each row of the triangle from right to left, we obtain a shifted version of the triangles `OEIS A136124,A143491`. It corresponds almost to the generating function of Stirling Cycle numbers [11, p. 735]: $SC_n(u) = \prod_{i=1}^{n-1}(u+i)$. The associated sequence enumerates size-n permutations that decompose into k cycles, defined as Stirling numbers of the first kind. More formally we prove:

Proposition 2.2. *Defining $\hat{t}_n(u) = \sum_{k=1}^{n} t_{n,k}\ u^{n-k}$, we obtain $\hat{t}_n(u) = \dfrac{u}{1+u} SC_n(u)$.*

Let $\mathcal{X}_n$ be the random variable that maps increasing Schröder trees of size n to their numbers of internal nodes. We want to establish a limit law for the distribution $(\mathbb{P}_{\mathcal{T}_n}(\mathcal{X}_n = k))_k$. But let us first compute its mean and its standard deviation so that we will then study the convergence of the *normalized* random variable $\mathcal{X}_n^\star = \frac{\mathcal{X}_n - \mathbb{E}(\mathcal{X}_n)}{\sqrt{\mathbb{V}(\mathcal{X}_n)}}$. We follow here the classical approach presented, for example, in [11, p. 157]. Since we consider the uniform distribution among trees of a given size n, we obviously get $\mathbb{P}_{\mathcal{T}_n}(\mathcal{X}_n = k) = \frac{t_{n,k}}{t_n}$.

Proposition 2.3. *Let $n \geq 2$, the mean value of $\mathcal{X}_n$ is equal to*

$$\mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n) = n - H_n + \frac{1}{2} = n - \ln n - \gamma + \frac{1}{2} + O\left(\frac{1}{n}\right),$$

with H_n the n-th harmonic number and γ the Euler constant ($\gamma \approx 0.57721\ldots$). Furthermore,

$$\mathbb{V}_{\mathcal{T}_n}[\mathcal{X}_n] = \ln n + \gamma - \frac{\pi^2}{6} - \frac{5}{4} + O\left(\frac{\log n}{n}\right).$$

Recall that the ordinary generating function for the Harmonic numbers sequence is $H(z) = \frac{1}{1-z}\ \ln\frac{1}{1-z}$ (see e.g. [11, p. 388]), then the result is proved by a direct computation. The proof is presented in Appendix A.

This proposition allows us to exhibit the limit law of the distribution $(X_n^\star)$ and proves then that the sequence $(\mathcal{X}_n)$ converges in distribution to a Gaussian law.

Theorem 2.1. *Let $\mathcal{X}_n$ be the random variable describing the distribution of the number of internal nodes in increasing Schröder trees of size n, or equivalently the number of building steps to get a size-n tree, we have* $\dfrac{\mathcal{X}_n - \mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n)}{\sqrt{\mathbb{V}_{\mathcal{T}_n}(\mathcal{X}_n)}} \xrightarrow{d} \mathcal{N}(0,1).$

The proof is obtained via an adaptation of Flajolet and Sedgewick's approach for the limit Gaussian law of Stirling Cycle numbers [11, p. 644]: see Appendix A. Observing the mean value $\mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n)$ we remark that only the second order in the asymptotic behavior permits to conclude that some internal nodes are not binary.

Quantitative characteristics of the root node

The next parameters we are interested in are related to the root of the increasing Schröder trees. Concerning this particular node, we want to understand

1	0					
0	0	1				
0	0	2	1			
0	0	8	3	1		
0	0	40	15	4	1	
0	0	240	90	24	5	1

Figure 4: Distribution of $t_{n,k}$ for size-n trees, $n \in \{1,2,\ldots,6\}$, of the root arity $k \in \{0,1,\ldots,n\}$

first its typical arity, and then the number of leaves attached to it in a large tree.

To avoid the description of several notations in the paper, we have chosen to reuse the previous notations for this new sequence. Thus here the variable $\mathcal{U}$ marks the arity of the root. The specification is direct: either the root-leaf is modified in the evolution process, or it is not the root that is substituted.

$$\begin{aligned}\mathcal{T} =&\mathcal{Z} \cup \big(\mathcal{U} \times \Theta_{\mathcal{Z}}(\mathcal{Z}) \times \textsc{Seq}_{\geq 1}(\mathcal{U} \times \mathcal{Z})\big)\\ &\cup \big(\Theta_{\mathcal{Z}}(\mathcal{T} \setminus \mathcal{Z}) \times \textsc{Seq}_{\geq 1}(\mathcal{Z})\big)\,.\end{aligned}$$

We directly obtain the translation

$$T(z,u) = z + \frac{u^2\ z^2}{1-uz} + \frac{z^2}{1-z}\ \partial_z\left(T(z,u) - z\right).$$

In the same way as before we prove

$$\begin{cases} t_1(u) = 1, \qquad t_2(u) = u^2,\\ \text{and if } n > 2,\\ t_n(u) = u^{n-1}(u-1) + n\ t_{n-1}(u); \end{cases}$$

$$\begin{cases} t_{n,k} = n\ t_{n-1,k} & \text{if } 1 < k < n-1\\ t_{1,0} = 1, \quad t_{2,2} = 1 & \\ t_{n,n-1} = n-1 \quad t_{n,n} = 1 & \text{if } n > 2 \text{ and}\\ t_{i,j} = 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

These sequences are related to `OEIS A094112,A092582`, that define properties on permutations (either some avoiding pattern, or with some fixed size initial run).

Corollary 2.2. *For $n \geq 2$ and $2 \leq k \leq n-1$, we get* $t_{n,k} = n!\dfrac{k}{(k+1)!}$.

A proof by induction is direct.

Theorem 2.2. *Let $\mathcal{X}_n$ be the random variable describing the distribution of the number of children of the root in increasing Schröder trees of size n, we have, for $n \geq 2$ and $2 \leq k \leq n-1$,*

$$\mathbb{P}_{\mathcal{T}_n}(\mathcal{X}_n = k) = \frac{2k}{(k+1)!}, \quad \textit{and} \quad \mathbb{P}_{\mathcal{T}_n}(\mathcal{X}_n = n) = \frac{2}{n!}.$$

The second characteristics for the root node is the number of leaves that are attached to it. Here the specification and thus the ordinary differential equation are more involved. In particular, the operators needed for the specification are not so classical so we prefer to explain directly the differential equation. Once again, let $T(z,u) = \sum_{n,k} t_{n,k}\ u^k z^n$ be the bivariate generating with $t_{n,k}$ the number of size-n increasing Schröder trees with k leaves as children of the root. Then,

$$T\,(z,u) = z + \frac{u^2\ z^2}{1-uz} + \frac{z^2}{1-z}\ \partial_z T\,(z,u) + \frac{z\,(1-u)}{1-z}\ \partial_u T\,(z,u)\,.$$

Let us give the details to understand the construction. A tree is either reduced to a leaf or a single internal node with some leaves: $z + \frac{u^2\ z^2}{1-uz}$. Or in the iterative process a leaf attached to the root is selected, then replaced by an internal node with at least two leaves (that are not anymore attached to the root of the whole tree): $\frac{z}{1-z}\ \partial_u T\,(z,u)$. Or, during the iterative process, a leaf that is not attached to the root is selected and replaced by an internal node attached to at least two leaves: $\frac{z^2}{1-z}\ \partial_z T\,(z,u) - \frac{u\ z}{1-z}\ \partial_u T\,(z,u)$. The second term removes the trees built in the first one where we have selected a leaf attached to the root (and also marked by z).
Again by denoting $T(z,u) = \sum_n t_n(u) z^n$, we can extract the following recurrence, for all $n \geq 4$,

$$\begin{aligned}t_n(u) =&(n+u)\ t_{n-1}(u) + u(1-n)\ t_{n-2}(u)\\ &+ (1-u)\ t'_{n-1}(u) + (u^2-u)\ t'_{n-2}(u),\end{aligned}$$

with $t_1(u) = 1$, $t_2(u) = u^2$ and $t_3(u) = u^3 + 2u$.

Theorem 2.3. *Asymptotically, the mean and the variance of the number $\mathcal{X}_n$ of leaves attached to the root are $\mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n) = \frac{2e}{n} + O\left(\frac{1}{n!}\right)$ and $\mathbb{V}_{\mathcal{T}_n}(\mathcal{X}_n) = \frac{2e}{n} + O\left(\frac{1}{n^2}\right)$.*

Let us remark the second term in the expansion of $\mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n)$ is extremely small in front of the main term.

Quantitative analysis of the number of binary nodes Here the specification is easier to exhibit, and its translation via the symbolic method is direct ($\mathcal{U}$ is marking the binary nodes):

$$\begin{aligned}\mathcal{T} &= \mathcal{Z} \cup \big(\Theta_{\mathcal{Z}}\mathcal{T} \times \big(\mathcal{U} \times \mathcal{Z} \cup \textsc{Seq}_{\geq 2}(\mathcal{Z})\big)\big)\,;\\ T(z,u) &= z + \left(u\ z^2 + \frac{z^3}{1-z}\right)\partial_z T(z,u).\end{aligned}$$

Let us again extract the recurrence $t_n(u)$, for all $n \geq 4$:

$$t_n(u) = (1 + u(n-1))t_{n-1}(u) + (1-u)(n-2)t_{n-2},$$

with $t_1(u) = 1, t_2(u) = u$ and $t_3(u) = 1 + 2u^2$. Note that due to this recurrence, the probability distribution $p_n(u) = \frac{2}{n!}t_n(u)$ also exhibits a simple recurrence (cf. [11, p. 157]). Thus we easily compute the mean and the second factorial moment of the number of binary nodes in size-($n \geq 3$) trees:

$$\mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n) = \frac{7}{3} + n - 2 \cdot \sum_{k=1}^{n} \frac{1}{k} - \frac{1}{n}, \quad \text{and}$$

$$\begin{aligned}\mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n(\mathcal{X}_n - 1)) = \sum_{k=3}^{n} &\frac{2k-2}{k}\mathbb{E}_{\mathcal{T}_{k-1}}(\mathcal{X}_{k-1})\\ &- \frac{2k-4}{k(k-1)}\mathbb{E}_{\mathcal{T}_{k-2}}(\mathcal{X}_{k-2}).\end{aligned}$$

THEOREM 2.4. *Asymptotically, the mean and the variance of the number $\mathcal{X}_n$ of binary internal nodes are* $\mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n) = n - 2\ln(n) + \frac{7}{3} - 2\gamma + O\left(\frac{1}{n}\right)$ *and* $\mathbb{V}_{\mathcal{T}_n}(\mathcal{X}_n) = 4\ln(n) + O(1)$. *Furthermore there is a limiting distribution satisfying* $\frac{\mathcal{X}_n - \mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n)}{\sqrt{\mathbb{V}_{\mathcal{T}_n}(\mathcal{X}_n)}} \xrightarrow{d} \mathcal{N}(0,1)$.

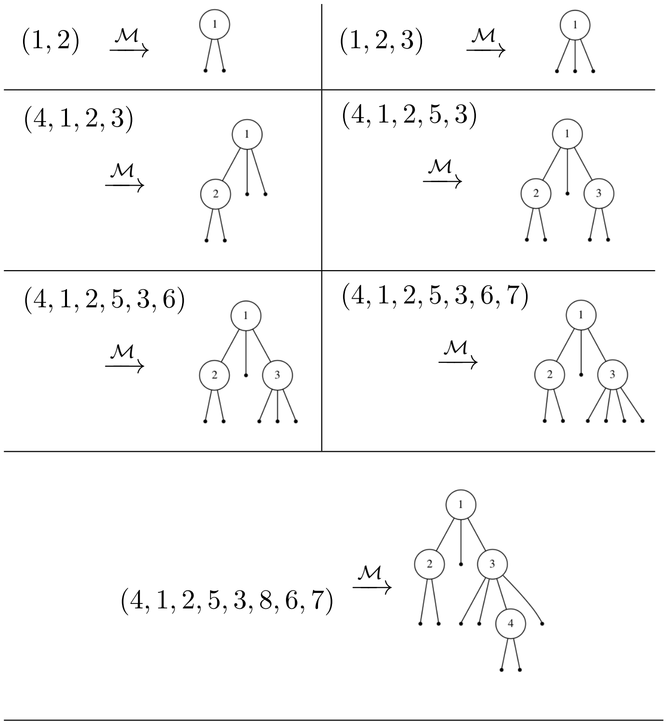
Proof. [Key-ideas] The recurrences give the closed form formulas for $\mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n)$ and $\mathbb{V}_{\mathcal{T}_n}(\mathcal{X}_n)$. Then it is important to notice that $t_n(u)$ can be approximated by $\tilde{t}_n(u)$ verifying $\tilde{t}_n(u) = (1 + u(n-1))\tilde{t}_{n-1}(u)$. The latter recurrence is the same recurrence as the one exhibited for the number of internal nodes (equation (2.5)). Thus, by the same arguments, the sequence of distributions $(t_n(u))$ converges in distribution to a Gaussian law.

构造排列与树之间的双射

2.4 Bijection with permutations Observing the exact value $t_n = n!/2$ enhances the chances of finding some relation between our model of increasing trees and a subclass of permutations. Let us start with this goal. First, for a size-n permutation σ denoted by $(\sigma_1, \ldots, \sigma_n)$, we define σ_i to be its i-th element (the image of i), and $\sigma^{-1}(k)$ to be the preimage of k (the position of k in the permutation). We are now ready to define the recursive map $\mathcal{M}$ between $\mathcal{HP}$, the class of permutations such that 1 appears before 2 and the class $\mathcal{T}$ of increasing Schröder trees. The base case is the permutation $(1,2)$ which corresponds to the root labeled by 1 attached to two unlabeled leaves. Let σ be a size-n permutation in $\mathcal{HP}$, with $n \geq 3$. We observe its the greatest element: if $\sigma_n = n$ then we add a new rightmost leaf to the last added internal node (the one with the largest label); otherwise let $k = \sigma^{-1}(n)$, we create a new binary node ν labeled with a new integer (the smallest as possible) and attached to two new leaves, then we replace the k-th leaf by this new tree rooted at ν, in the tree under construction based on $\sigma_{\setminus n}$, i.e. that is σ without the greatest element n. Remark that during the tree construction we must traverse the leaves, we can take an arbitrary traversal.

THEOREM 2.5. *The map $\mathcal{M}$ is a one-to-one correspondence between $\mathcal{HP}$ and $\mathcal{T}$.*

Proof. The mapping is size preserving: at each iteration we remove exactly one element from the permutation and add exactly one leaf to the tree by either adding a leaf to the last exiting node or by killing one leaf and adding to new ones. The mapping is injective since by induction at each iteration we remove the greatest element of the permutation and its following its index the actions are performed on the resulting tree in a non-ambiguous manner. Finally the mapping is based on two classes with the same number of elements (of each size).



$(1,2)$
$\xrightarrow{\mathcal{M}}$
1
$(1,2,3)$
$\xrightarrow{\mathcal{M}}$
1
$(4,1,2,3)$
$\xrightarrow{\mathcal{M}}$
1
2
$(4,1,2,5,3)$
$\xrightarrow{\mathcal{M}}$
1
2
3
$(4,1,2,5,3,6)$
$\xrightarrow{\mathcal{M}}$
1
2
3
$(4,1,2,5,3,6,7)$
$\xrightarrow{\mathcal{M}}$
1
2
3
$(4,1,2,5,3,8,6,7)$
$\xrightarrow{\mathcal{M}}$
1
2
3
4

Figure 5: A size-8 example of the mapping $\mathcal{M}$

In the Figure 5 we present the mapping on an example. Remark that we have ordered the steps reversely to understand the process in a easiest way.

2.5 Uniform random sampling Obviously, through the latter bijection we are able to obtain an uniform random sampler. It suffices to uniformly sample permutations and to use the bijection to build the associated increasing Schröder tree. While there exists fast algorithms to sample permutations, see for example [2], using the bijection efficiently is not obvious.

However, through the bijection a direct *probabilistic construction* of increasing Schröder trees can be obtained. Such a probabilistic construction presents two main advantages. Firstly, it simplifies the implementation of a sampler and, secondly, and more importantly, it gives a purely probabilistic approach to the original combinatorial class, which we are interesting in. This probabilistic approach can be used to compare to other probabilistic tree models or to exhibit important characteristics of trees on average using probabilities rather than combinatorics.

We introduce in this section an algorithm to uniformly sample increasing Schröder trees of a given size n. A first remark is that the uniform sampling of structures with increasing labeling constraints is not so classical in the context of Analytic Combinatorics. There are some studies by Martínez and Molinero [13, 14] in the context of the recursive method and some other about Boltzmann sampling either directly for the method [5] or focusing on a specific application [3].

For the uniform sampling of our evolution process, we are focusing on two goals. Our fundamental goal consists in controlling the probability distribution used for the sampling. In fact, we may extract some statistical information based on the samplings, thus the probability distribution is central. We choose to sample uniformly trees of the same size, because then we can bias our generator (and tune the bias) to construct other probability distributions. Secondly our algorithmic framework must be very efficient to sample large trees (with several thousands of leaves). Thus a detailed complexity analysis is necessary to be sure that the algorithm cannot be easily improved.

Our approach is based on the combinatorics underlying the very efficient recurrence $t_n = n \cdot t_{n-1}$: a tree of size n can be built from a tree of size $n-1$ in n different ways. We exhibit a construction based on this recurrence. This leads to the following iterative algorithm of random sampling.

这个算法就是根据上一节的双射来构造的

Algorithm 1 Increasing Schröder Tree Builder

```
function TreeBuilder($n$)
    if $n = 1$ then
        return the single leaf
    $T$ := the root labeled by 1 and attached to two leaves
    $\ell := 2$
    for $i$ from 3 to $n$ do
        $k := rand\_int(1, i)$
        if $k = i$ then
            Add a new leaf to the last added internal node in $T$
        else
            Create a new binary node at position $k$ in $T$
                    with label $\ell$ and attached to two leaves
            $\ell := \ell + 1$
    return T
```

The function $rand\_int(a, b)$ returns uniformly at random an integer in $\{a, a+1, \ldots, b\}$.

THEOREM 2.6. *The function* TreeBuilder(n) *in Algorithm 1 is a uniform sampling algorithm for size-n trees. Asymptotically, it operates in $O(n)$ operations on trees and necessitates $O(n \ln n)$ random bits.*

The correctness of the algorithm is a direct consequence of the mapping $\mathcal{M}$. it gives the probabilistic construction of trees of $\mathcal{T}$. Using the adequate data structures, as for example by keeping an array of pointers to all leaves and another one to the last inserted internal node, each insertion in the tree under construction is done in constant time.

3 Weakly Increasing Schröder trees

In this section we aim at developing another model for ranked trees based on Schröder structures. In fact we relax somehow the labeling constraint.

3.1 The model and its context Weakly increasing Schröder trees are a generalization of strongly increasing Schröder trees. The tree structure is still an unlabeled Schröder tree. But the labeling is different. Internal nodes are labeled between 1 to ℓ in such a way that the sequence of labels in each path from the root to a leaf is also increasing. The difference here is that different nodes can have the same label. This model is also built iteratively.

- *Start with a single (unlabeled) leaf;*
- Iterate the following process: *at step ℓ (for $\ell \geq 1$), select a subset of leaves and replace each of them by an internal node with label ℓ attached to a sequence of at least two leaves.*

节点的标号可以有重复

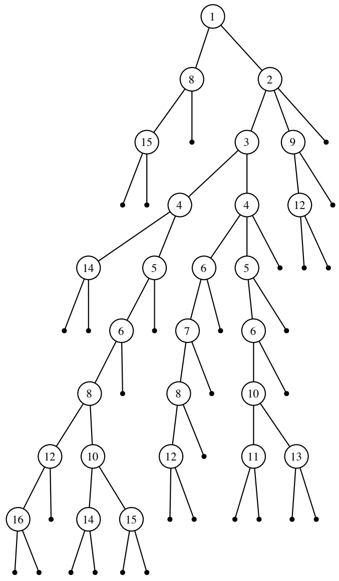


Figure 6: A weakly increasing Schröder tree

In Figure 6 we present a weakly increasing tree of size 30 with 16 distinct labels..

3.2 Exact enumeration and relationship with ordered Bell numbers We can specify the process through the symbolic method. But once again the labeling is transparent and does not appear in the specification.

$$G(z) = z + G\left(\frac{z^2}{1-z} + z\right) - G(z). \tag{3.7}$$

At each iteration and for each leaf we can either leave it as it is or expand it into a new internal node with at least 2 leaves. The configuration where no leaf is expanded is forbidden, thus we remove $G(z)$ in equation (3.7). From this equation we extract the recurrence

$$g_n = \begin{cases} 1 & \text{if } n = 1 \\ \sum\limits_{k=1}^{n-1} \binom{n-1}{k-1} g_k. & \text{otherwise.} \end{cases} \tag{3.8}$$

The first coefficients correspond to a shift of the sequence of Ordered Bell numbers (also called Fubini numbers) referenced as `OEIS A000670`.

$$g_n = 0, 1, 1, 3, 13, 75, 541, 4683, 47293, 545835, 7087261, \ldots$$

By following the approach developed by Pippenger in [16] for the derivation of the exponential generating function for ordered Bell numbers we obtain, by starting from our equation (3.7), $(\mathcal{B}G)'(z) = 1/(2 - e^z)$. Thus, after integration $\mathcal{B}G(z) = \frac{1}{2}\left(z - \ln\left(2 - e^z\right)\right)$. Usually ordered Bell numbers are specified by $B = \text{SEQ}(\text{SET}_{\geq 1}(z))$. Obviously this gives the exponential generating function $B(z) = 1/(2 - e^z)$. Thus, we have proved that our sequence is a shift of the one of ordered Bell numbers. As a by-product, we have exhibited a new way for specifying ordered Bell numbers.

Recall the n-th ordered Bell number, denoted by B_n, counts the total number of partitions of a set of size n where additionally we consider an order over the subsets of the partition.

$$B_n = \sum_{k=0}^{n} k! \left\{ n \atop k \right\} \sim \frac{n!}{2\left(\ln 2\right)^{n+1}}, \tag{3.9}$$

where $\left\{ n \atop k \right\}$ stands for the Stirling partition numbers (also called Stirling numbers of the second kind). The number B_n corresponds to the number g_{n+1} of weakly increasing Schröder trees of size $n + 1$.

3.3 Bijection between ordered Bell numbers and weakly increasing Schröder trees In ordered partitions, the subsets are ordered but the elements inside a subset are not. In the following let us denote by $p = [p_1, p_2, \ldots, p_\ell]$ an ordered partition such that p_i is the subset of the partition at position i. For example if $p = [\{3, 4\}, \{1, 5, 7\}, \{2, 6\}]$, then $p_1 = \{3, 4\}$. We denote by $|p_i|$ the size of the i-th subset: $|p_1| = 2$. The total size (i.e. number of elements) of the partition is denoted by $|p|$ Thus the elements of an ordered partition range from 1 to $|p|$.

For the exhibition of the correspondence we will use a canonical order inside the subsets, consisting in enumerating the elements increasingly.

Let $p = [p_1, p_2, \ldots, p_\ell]$ be an ordered partition, and $p_i = \{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ (with $r \geq 1$), such that $\alpha_1 < \alpha_2 < \cdots < \alpha_r$. We define a *run* in p_i to be a maximal sequence $\alpha_i, \alpha_{i+1}, \ldots, \alpha_j$ equal to $\alpha_i, \alpha_i{+}1, \ldots, \alpha_i{+}j{-}i$. It is maximal in the sense that $\alpha_{i-1} < \alpha_i - 1$ and $\alpha_{j+1} > \alpha_j + 1$. We define the map *runs* that lists all the runs of a subset.

For instance, in our example p, in p_1 there is a single run: $3, 4$ and in p_2, there are 3 runs.

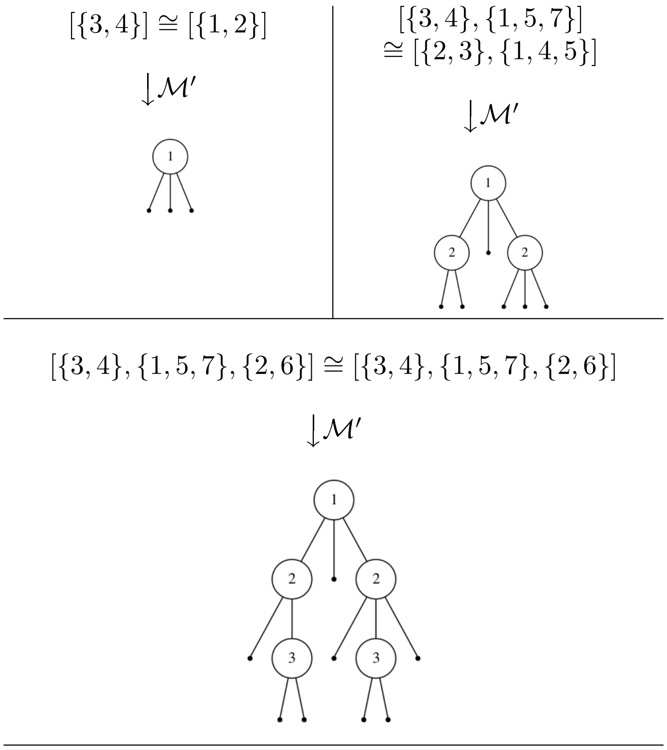
The mapping deals with incomplete ordered partitions (in the sense that some integers are not present in the partition). We define a normalization of a partition, denoted by *norm*, that maps an incomplete ordered partition of size k into the corresponding ordered partition of size k whose elements are $\{1, \ldots, k\}$ and that keeps the relative order between the elements. For example by taking the first two subsets from p as $p' = [p_1, p_2]$, then p' is an incomplete ordered partition of size 5 and we get $\mathit{norm}(p') = [\{2, 3\}, \{1, 4, 5\}]$.

From the ordered partition p, the mapping $\mathcal{M}'$ builds the corresponding tree by processing the subsets of the ordered partition successively. We start by creating a new ordered partition that contains only p_1, $p' = \mathit{norm}([p_1])$. The size of p_1 determines the arity of the root: it equals $|p_1| + 1$. The root label is 1. Then at each step i with $i \in \{2, \ldots, \ell\}$, we process the subset p_i as follows. Normalize the incomplete ordered partition $[p_1, p_2, \ldots, p_i]$. In the normalized ordered partition the corresponding subset of p_i is denoted by p'_i. The number of new internal nodes is $|\mathit{runs}(p'_i)|$, all labeled by i. Suppose $\mathit{runs}(p'_i) = [r_1, r_2, \ldots, r_j]$ (with each r_ℓ a set of successive integers and possibly a single one). Take an order for the leaves in the tree under construction (the postorder one for example) and iterate the process: For ℓ from 1 to j, take the leaf whose index is the first element of r_ℓ and replace it with an internal node with label i of arity $|r_\ell| + 1$.

In Figure 7 the mapping $\mathcal{M}'$ is applied on our example $p = [\{3, 4\}, \{1, 5, 7\}, \{2, 6\}]$. The resulting weakly increasing tree is of size 9.

3.4 Analysis of typical parameters

Quantitative analysis of the number of iteration steps In our classical iterative equation, we add a new



$[\{3,4\}] \cong [\{1,2\}]$
$\downarrow \mathcal{M}'$
1
$[\{3,4\},\{1,5,7\}]$
$\cong [\{2,3\},\{1,4,5\}]$
$\downarrow \mathcal{M}'$
1
2
2
$[\{3,4\},\{1,5,7\},\{2,6\}] \cong [\{3,4\},\{1,5,7\},\{2,6\}]$
$\downarrow \mathcal{M}'$
1
2
2
3
3

Figure 7: Weakly increasing tree of size 8

variable u to mark each iteration.

$$G(z,u) = z + u\, G\left(\frac{z^2}{1-z} + z, u\right) - u\, G(z,u). \tag{3.10}$$

Which leads to the following recurrence,

$$g_{n,k} = \begin{cases} 1 & \text{if } n = 1, k = 0, \\ \sum\limits_{j=1}^{n-1} \binom{n-1}{j-1} g_{j,k-1} & \text{otherwise }. \end{cases} \tag{3.11}$$

This recurrence is analogous to the one relating ordered Bell numbers and Stirling partition numbers.

0 ,						
1 ,						
0 ,	1,	2 ,				
0 ,	1,	6,	6,			
0 ,	1,	14,	36,	24,		
0 ,	1,	30,	150,	240,	120,	
0 ,	1,	62,	540,	1560,	1800,	720

Figure 8: Distribution of $g_{n,k}$ for $n \in \{0,1,2,\ldots,6\}$

THEOREM 3.1. *The distribution of the the number of building steps in weakly increasing Schröder trees of size n satisfies*

$$g_{n,k} = k! \left\{ {n+1 \atop k} \right\}.$$

Let $\mathcal{X}_n$ be the random variable describing this distribution, we have

$$\frac{\mathcal{X}_n - \mathbb{E}_{\mathcal{G}_n}(\mathcal{X}_n)}{\sqrt{\mathbb{V}_{\mathcal{G}_n}(\mathcal{X}_n)}} \xrightarrow{d} \mathcal{N}(0,1),$$

with $\mathbb{E}_{\mathcal{G}_n}(\mathcal{X}_n) \sim \frac{1}{2\ln 2}\, n$ and $\mathbb{V}_{\mathcal{G}_n}(\mathcal{X}_n) \sim \frac{1-\ln 2}{(2\ln 2)^2}\, n$.

Proof. The one-to-one correspondence between weakly increasing Schröder trees and ordered Bell numbers gives the combinatorial proof of the distribution for $(g_{n,k})$.

The analysis of the limiting distribution is classical in the quasi-powers framework. See for example [11, p. 653].

Quantitative analysis of the number of internal nodes In this model the number of iteration steps does not correspond to the number of the internal nodes as at each iteration any subset of leaves can be expanded into internal nodes with new leaves. The specification marking both internal nodes and leaves is

$$G(z,u) = z + G\left(\frac{uz^2}{1-z} + z, u\right) - G(z,u). \tag{3.12}$$

We recall that the substitution $G(\frac{uz^2}{1-z} + z)$ means that for each iteration each leaf can be left as it is z or expanded into an internal node of unbounded arity with new leaves $\frac{z^2}{1-z}$. It is in the second part that an internal node will be created and thus we mark it with u.

THEOREM 3.2. *The average number of internal nodes $\mathbb{E}_{\mathcal{G}_n}(\mathcal{X}_n)$ in size n weakly increasing trees verifies*

$$\mathbb{E}_{\mathcal{G}_n}(\mathcal{X}_n) = n - \ln n + o(1).$$

The main ideas of the proof are in Appendix B.

3.5 Uniform random sampling We introduce in this section an algorithm to uniformly sample weakly increasing Schröder trees of a given size n directly, without an intermediate step of generating uniformly an ordered partition.

The global approach for our algorithmic framework deals with the *recursive generation method* adapted to the Analytic Combinatorics point of view in [12]. But in our context, we note that we can obtain for free (from a complexity view) an *unranking algorithm.* This kind of algorithm has been developed in the 70's by Nijenhuis and Wilf [15] and then has been introduced to the context of Analytic Combinatorics by Martínez and Molinero [13]. Here the idea is not to draw uniformly an object, but first to define a total order over the objects under consideration (here weakly increasing Schröder trees) and then an integer (named the rank) is sampled to build deterministically the associated object. Such an approach gives also a way to do exhaustive generation (refer to the paper [4] for an example of both methods: recursive generation and unranking).

For both types of algorithms (recursive generators and unranking ones), there is a first step of pre-computations (done only once before the sampling of many objects). We must compute (and store) the numbers of trees of sizes from 1 to n. Here this phase can be done with a quadratic complexity (in the number of arithmetic operations) because of the recursive formula for g_n (cf. equation (3.11)).

The second (and last) step for the sampling consists in the recursive construction of the tree of rank r that corresponds to an uniformly sampled integer in $\{0, 1, \dots, g_n - 1\}$. For this purpose, we come back to the original recursive equation (3.8), and in particular, we look at the sum over decreasing k:

$$g_n = \binom{n-1}{n-2} g_{n-1} + \binom{n-1}{n-3} g_{n-2} + \cdots + \binom{n-1}{0} g_1.$$

The latter recurrence is combinatorially easy to understand. Through the evolution process, to build a size n tree, we take a size $k \in \{1, \dots, n-1\}$ tree constructed with exactly one less iteration. The binomial coefficient $\binom{n-1}{k-1}$ corresponds to the number of composition of n in k parts. Then we traverse the tree, and each time we see a leaf, we do the following rule: if the next part is of value 1, we leave the leaf unchanged otherwise for a value $\ell > 1$, we replace the leaf by an internal node (well labeled with the single new value valid for this step) and attached to it ℓ leaves. We then take the next part of the composition into consideration and continue the tree traversal.

In the latter sum, the first term is much bigger than the second one, that is must bigger than the third one and so on. This approach, focusing first on the dominant terms corresponds to the *Boustrophedonic order* presented in [12]. It allows to improve essentially the average complexity of the random sampling algorithm. In our case of weakly increasing Schröder trees that do not follow a standard specification (cf. [12]), the complexity gain is even better.

THEOREM 3.3. *The function* UNRANKTREE *is an unranking algorithm and calling it with the parameters n and an uniformly sampled integer in $\{0, \dots, g_n - 1\}$, it is an uniform sampler for size-n weakly increasing Schröder trees.*

Proof. [Key-ideas] The total order for weakly increasing Schröder trees is the following. Let α and β be two trees. If the size of α is strictly smaller than the one of β, we define $\alpha < \beta$. Let us suppose that both sizes are equal to n. In the recursive construction, let $\tilde{\alpha}$ (and $\tilde{\gamma}_1$ be the tree (resp. the composition for the leaf substitution) building the tree α (and respectively $\tilde{\beta}$ and $\tilde{\gamma}_2$ the ones associated to β). If the size of $\tilde{\alpha}$ is strictly greater than the one of $\tilde{\beta}$, we define $\alpha < \beta$. Let us now suppose that both sizes of $\tilde{\alpha}$ and $\tilde{\beta}$ are equal. By using an arbitrary order for the composition unranking, we can order α and β.

This total order over the trees is satisfied by our algorithm: thus this latter is correct.

Algorithm 3 Composition unranking

```
 1: function UnrankTree($n, s$)
 2:     if $n = 1$ then
 3:         return the tree reduced to a single leaf
 4:     $k := n$
 5:     $r := s$
 6:     while $r >= 0$ do
 7:         $k := k - 1$
 8:         $r := r - \binom{n-1}{k-1} \cdot g_k$
 9:     $r := r + \binom{n-1}{k-1} \cdot g_k$
10:     $k := k + 1$
11:     $s' := r \mod g_k$
12:     $T :=$UnrankTree($k, s'$)
13:     $C :=$UnrankComposition($n, k, r // g_k$)
14:     Substitute in $T$ some leaves according to $C$
15:     return the tree $T$
```

The sequence $(g_k)_{k \leq n}$ and $(\ell!)_{\ell \in \{1,\dots,n\}}$ have been precomputed and stored.

Line 13: The operation // is the Euclidean division.

```
 1: function UnrankComposition($n, k, s$)
 2:     if $n = 1$ and $k = 1$ and $s = 0$ then
 3:         return [1]
 4:     $s' := s$
 5:     if $s' < \binom{n-2}{k-1}$ then
 6:         $C :=$UnrankComposition($n - 1, k, s'$)
 7:         $C[len(C)] := C[len(C)] + 1$
 8:         return $C$
 9:     else
10:         $s' := s' - \binom{n-2}{k-1}$
11:         $C :=$UnrankComposition($n - 1, k, s'$) $\cup$[1]
```

THEOREM 3.4. *Once the pre-computations have been done, the function* UNRANKTREE *necessitates $O(n^2)$ arithmetic operations to construct any tree of size n.*

Due to the fact that usually the difference between n and k is very small, a detailed analysis of the average case, or an more adapted composition unranking should give a better complexity analysis. In fact as we have seen before, in a large typical tree, there are in average $n - \ln n$ internal nodes and thus most of them must be of arity 2 and are given by the first term in the latter sum defining g_n.

Proof. [Proof-ideas] The main idea is the following: during a call to UNRANKTREE, there are exactly the same number of new leaves in the tree under construction to the number of loops in the `while` instruction on Line 6. Outside this `while` block, the number of arithmetic operations is essentially due to the unranking algorithm for compositions. The actual version of this algorithm induces a quadratic complexity in the number of arithmetic operations. The unranking algorithm for the composition is based on the classical result about the composition of $n > 0$ in $k \in \{1, \dots, n\}$ parts:

$$\begin{aligned} C_{n,k} &= \binom{n-1}{k-1} \\ &= C_{n-1,k} + C_{n-1,k-1}. \end{aligned}$$

References

[1] D. Aldous. Probability distributions on cladograms. In D. Aldous and R. Pemantle, editors, *Random Discrete Structures*, pages 1–18. Springer New York, 1996.

[2] A. Bacher, O. Bodini, H.-K. Hwang, and T.-H. Tsai. Generating random permutations by coin tossing: Classical algorithms, new analysis, and modern implementation. *ACM Trans. Algorithms*, 13(2):24:1–24:43, 2017.

[3] O. Bodini, M. Dien, X. Fontaine, A. Genitrini, and H.-K. Hwang. Increasing Diamonds. In *LATIN 2016: Theoretical Informatics - 12th Latin American Symposium*, pages 207–219, 2016.

[4] O. Bodini, M. Dien, A. Genitrini, and A. Viola. Beyond series-parallel concurrent systems: the case of arch processes. In *29th International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*, page to appear, 2018.

[5] O. Bodini, O. Roussel, and M. Soria. Boltzmann samplers for first-order differential specifications. *Discrete Applied Mathematics*, 160(18):2563–2572, 2012.

[6] C. Darwin. *On the Origin of Species by Means of Natural Selection.* Murray, London, 1859.

[7] M. Drmota. *Random trees.* Springer, Vienna-New York, 2009.

[8] J. Felsenstein. The number of evolutionary trees. *Systematic Zoology*, 27(1):27–33, 1978.

[9] J. Felsenstein. Phylip (phylogeny inference package), version 3.5 c, 1993.

[10] J. Felsenstein. *Inferring phylogenies.* Sinauer Associates, 2003.

[11] P. Flajolet and R. Sedgewick. *Analytic combinatorics.* Cambridge University Press, Cambridge, 2009.

[12] P. Flajolet, P. Zimmermann, and B. Van Cutsem. A calculus for the random generation of labelled combinatorial structures. *Theoretical Computer Science*, 132(1-2):1–35, 1994.

[13] C. Martínez and X. Molinero. Generic algorithms for the generation of combinatorial objects. In *28th International Symposium on Mathematical Foundations of Computer Science (MFCS)*, pages 572–581. Springer Berlin Heidelberg, 2003.

[14] X. Molinero. *Ordered Generation of Classes of Combinatorial Structures.* Phd thesis, Universitat Politècnica de Catalunya, 2005.

[15] A. Nijenhuis and H. S. Wilf. *Combinatorial algorithms.* Computer science and applied mathematics. Academic Press, New York, NY, 1975.

[16] N. Pippenger. The hypercube of resistors, asymptotic expansions, and preferential arrangements. *Mathematics Magazine*, 83(5):331–346, 2010.

[17] E. Schröder. Vier Combinatorische Probleme. *Z. Math. Phys.*, 15:361–376, 1870.

[18] M. Steel. *Phylogeny - discrete and random processes in evolution*, volume 89 of *CBMS-NSF regional conference series in applied mathematics.* SIAM, 2016.

[19] M. Steel and A. McKenzie. Properties of phylogenetic trees generated by yule-type specification models. *Mathematical Biosciences*, 170(1):91–112, 2001.

[20] H. S. Wilf. *Generatingfunctionology.* A. K. Peters, Ltd., Natick, MA, USA, 2006.

[21] Z. Yang. Paml 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

[22] G. U. Yule. A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87, 1925.

A Appendix related strongly increasing Schröder trees: Section 2

The Borel transform consists in the following transformation on ordinary generating series:

$$\mathcal{B}\left(\sum_{n\geq 0} a_n z^n\right) = \sum_{n\geq 0} a_n \frac{z^n}{n!}.$$

Lemma A.1. *Using Borel transform formula on formal series, we easily derive the following identities:*

(i) $\mathcal{B}(zf)(z) = \int_0^z \mathcal{B}f\mathrm{d}t$*;*

(ii) $\mathcal{B}(f')(z) = (\mathcal{B}f)'(z) + z(\mathcal{B}f)''(z)$*.*

We are now ready to prove Proposition 2.1.

Proof. [Proof of Proposition 2.1] Applying Borel on equation (2.4) and using properties (i) and (ii) we obtain

$$T(z,u) = zuT(z,u) + (1-u)\cdot\int_0^z T(z,u)dz - \frac{z^2}{2} + z.$$

Then by differentiating by z

$$\frac{\partial(1-zu)T(z,u)}{\partial z} = \frac{\partial(1-u)\cdot\int_0^z T(z,u)dz - \frac{z^2}{2} + z}{\partial z}.$$

Thus, after simplifications

$$(1-zu)\frac{\partial T(z,u)}{\partial z} = T(z,u) - z + 1 \quad \text{with } T(0,0) = 1.$$

Solving the differential equation gives the stated result.

Let us denote by $\mathcal{X}_n$ the random variable corresponding to the to number of internal nodes in increasing Schröder trees of size n. Proposition 2.3 aims at proving the mean value and the variance of $\mathcal{X}_n$.

Proof. [Proof of Proposition 2.3] Recall that the mean and variance can be computed mechanically from the bivariate generating function

$$\mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n) = \frac{[z^n]\partial_u T(z,u)_{|u=1}}{[z^n]T(z,1)}, \text{ and}$$

$$\mathbb{E}_{\mathcal{T}_n}(\mathcal{X}_n^2) = \frac{[z^n]\partial_u^2 T(z,u)_{|u=1}}{[z^n]T(z,1)} + \frac{[z^n]\partial_u T(z,u)_{|u=1}}{[z^n]T(z,1)},$$

where $\cdot_{|u=1}$ stands for the substitution of u by 1.

$$\begin{aligned}\mathbf{E}_{T_n}[\mathcal{X}] &= \frac{[z^n]\partial_u T(z,u)\mid_{u=1}}{[z^n]T(z,1)} \\ &= [z^n]\left(\frac{z}{(1-z)^2} - \frac{1}{1-z}\ln\left(\frac{1}{1-z}\right)\right) \\ &\quad + \frac{1}{2}\left(\frac{1}{1-z} - z - 1\right) \\ &= n - H_n + \frac{1}{2}\end{aligned}$$

Let $n \geq 2$, the mean value of $\mathcal{X}_n$ is equal to

$$\mathbb{E}_{\mathcal{T}_n}[\mathcal{X}_n^2] = n(n-1) - 2n(H_n - 1) + \sum_{k=1}^{n-1}\frac{1}{n-k}H_k,$$

and thus

$$\begin{aligned}\mathbb{E}_{\mathcal{T}_n}[\mathcal{X}_n^2] =& n(n-1) - 2n\,\ln n - 2(\gamma-1)n + \ln^2 n \\ &+ 2\gamma\,\ln n + \gamma^2 - \frac{\pi^2}{6} + O\left(\frac{\log n}{n}\right).\end{aligned}$$

In the same vein, when n tends to infinity, we get

$$\mathbb{V}_{T_n}[\mathcal{X}] = \ln n + \gamma - \frac{\pi^2}{6} - \frac{5}{4} + O\left(\frac{\log n}{n}\right).$$

We are now ready to prove the limit distribution for $\mathcal{X}_n$.

Proof. [Proof of Theorem 2.1] This proof is an adaptation on Flajolet and Sedgewick's proof on the limit Gaussian law of Stirling Cycle numbers [11, p. 644]

We take the probability generating function of $\hat{T}_n(u)$ it is obvious that if $\frac{t_{n,k}}{t_n}$ is a limit Gaussian law then so is $\frac{\hat{t}_{n,k}}{t_n} = \frac{t_{n,n-k}}{t_n}$. We will just get the mirror of the probability the standard deviation will not change $\hat{\sigma_n} = \sigma_n$ and the mean will be the mirror mean so $\hat{\mu}_n = n - \mu_n$.

$$\hat{p}_n(u) = \frac{2u(u+2)(u+3)\dots(u+n-1)}{n!}.$$

Thus we have

$$p_n(u) = \frac{2\Gamma(u+n)}{(u+1)\Gamma(u)\Gamma(n+1)}.$$

Near $u = 1$ we find an estimate of $p_n(u)$ using Stirling formula for the Gamma function

$$p_n(u) = \frac{n^{u-1}}{\Gamma(u)}\left(1 + O\left(\frac{1}{n}\right)\right) = \frac{(e^{u-1})^{\log n}}{\Gamma(u)}\left(1 + O\left(\frac{1}{n}\right)\right).$$

Now we can study the standardized random variable $\hat{X}_n^\star = \frac{\hat{\mathcal{X}} - \hat{\mu}_n}{\hat{\sigma}_n}$. The standardization of a random variable can be translated directly on the characteristic function.

$$\phi_{X_n^\star}(t) = e^{-it\frac{\mu}{\sigma}}\phi_{X_n}(\frac{t}{\sigma}).$$

$$\begin{aligned}\phi_{\hat{X}_n^\star}(t) =& e^{-it\frac{\log n - \gamma + \frac{1}{2}}{\sqrt{\log n + \gamma - \frac{\pi^2}{6} - \frac{5}{4}}}} \\ &\cdot\frac{\left(exp(\log n(e^{\frac{it}{\sqrt{\log n + \gamma - \frac{\pi^2}{6} - \frac{5}{4}}}} - 1))\right)}{\Gamma(u)}\left(1 + O\left(\frac{1}{n}\right)\right).\end{aligned}$$

For a fixed t and as $n \to \infty$,

$$\log \phi_{\hat{X}_n^\star}(t) = -\frac{t^2}{2} + O(\frac{1}{\log n}).$$

This last result is obtained by limited development of $\log n \cos \frac{t}{\sqrt{\log n + \gamma - \frac{\pi^2}{6} - \frac{5}{4}}}$. Finally we have

$$\phi_{\hat{X}_n^\star}(t) \sim e^{-\frac{t^2}{2}},$$

which is the characteristic function of the Gaussian law.

B Appendix related to weakly increasing Schröder trees: Section 3

Derivation for the exponential generating function for (g_n): We have,

$$g_n = \delta_n + \sum_{k=1}^{n-1} \binom{n-1}{k-1} g_k.$$

Where δ_l is 1 for $l = 1$ and 0 otherwise. Adding g_n to both sides gives

$$2g_n = \delta_n + \sum_{k=1}^{n} \binom{n-1}{k-1} g_k.$$

Finally multiplying both sides by $\frac{z^l}{l!}$ and summing over all $l \geq 0$

$$2G(z) = z + \sum_{l \geq 0} \frac{z^l}{l!} \sum_{k=1}^{n} \binom{n-1}{k-1} g_k$$

Deriving this last equation yields to the equation of Ordered Bell number. which has been studied by different authors. See [16] for a derivation of the exponential generating function,

$$G(z)' = \frac{1}{2 - e^z}.$$

Finally we have,

$$G(z) = \frac{1}{2}\left(z - \ln\left(2 - e^z\right)\right).$$

Proof of the theorem 3.2 We define $f_n = \sum\limits_{k=0}^{n-2} \binom{n-2}{k}(k+1) g_{k+1}$

Lemma B.1. $\frac{f_n}{g_n} \sim C$ *where* $C \approx 1.38$ *is a constant*

Proof. Wilf has given an approximation of the error term of Ordered Bell numbers in [20] which we can use,

$$g_n = \frac{(n-1)!}{2\ln(2)^n} + O(\gamma^{n-1}(n-1)!),$$

with $\gamma = \frac{1}{\sqrt{ln(2)^2 + 4\pi^2}} \approx 0.16\ldots$. Remark that $\frac{1}{ln(2)} \approx 1.44 \cdots > \gamma$. Now,

$$\begin{aligned}
\frac{f_n}{g_n} &= \frac{\sum\limits_{k=0}^{n-2} \binom{n-2}{k}(k+1)\left(\frac{k!}{2\ln(2)^{k+1}} + O(\gamma^k k!)\right)}{\frac{(n-1)!}{2\ln(2)^n} + O(\gamma^{n-1}(n-1)!)} \\
&= \frac{\sum\limits_{k=0}^{n-2} \binom{n-2}{k}(k+1)\left(\frac{k!}{2\ln(2)^{k+1}}\right)}{\frac{(n-1)!}{2\ln(2)^n} + O(\gamma^{n-1}(n-1)!)} + O(\frac{1}{n}) \\
&= \frac{\sum\limits_{k=0}^{n-2} (k+1)\left(\frac{1}{\ln(2)^{k+1}}\right)}{\frac{(n-1)}{\ln(2)^n} + O(\gamma^{n-1}(n-1)!)} + O(\frac{1}{n}) \\
&= \frac{n(\sum\limits_{k=1}^{n-1} \frac{1}{(n-1-k)!\ln 2^k)}) - (\sum\limits_{k=1}^{n-1} \frac{n-k}{(n-1-k)!\ln 2^k})}{\frac{(n-1)}{\ln(2)^n} + O(\gamma^{n-1}(n-1)!)} + O(\frac{1}{n}) \\
&= \frac{n(\sum\limits_{k=1}^{n-1} \frac{1}{(n-1-k)!\ln 2^k)} - (\sum\limits_{k=1}^{n-1} \frac{n-k}{(n-1-k)!\ln 2^k})}{\frac{(n-1)}{\ln(2)^n} + O(\gamma^{n-1}(n-1)!)} + O(\frac{1}{n}) \\
&= c + O(\frac{1}{n})
\end{aligned}$$

Then for large n, we can show the result by induction. Taking $Gu_n = \frac{(n-1)!}{2\ln(2)^n}(n - \ln n) + O(\gamma^{n-1}(n-1)!)$.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{G}_n}(\mathcal{X}_n) = \frac{Gu_n}{g_n} \\
&= c + O(\frac{1}{n}) + \frac{\sum\limits_{k=1}^{n-1} \binom{n-1}{k-1}\left(\frac{(k-1)!}{2\ln(2)^k}(k - \ln k + O(\gamma^{k-1}(k-1)!)\right)}{\frac{(n-1)!}{2\ln(2)^n} + O(\gamma^{n-1}(n-1)!)} \\
&= c + O(\frac{1}{n}) + \frac{\sum\limits_{k=1}^{n-1} \binom{n-1}{k-1}\frac{(k-1)!}{2\ln(2)^k} k - \sum\limits_{k=1}^{n-1} \binom{n-1}{k-1}\frac{(k-1)!}{2\ln(2)^k}\ln k}{\frac{(n-1)!}{2\ln(2)^n} + O(\gamma^{n-1}(n-1)!)} \\
&\quad + \frac{\sum\limits_{k=1}^{n-1} \binom{n-1}{k-1} O(\gamma^{k-1}(k-1)!)}{\frac{(n-1)!}{2\ln(2)^n} + O(\gamma^{n-1}(n-1)!)} \\
&= c + O(\frac{1}{n}) + n + c' + \frac{-\sum\limits_{k=1}^{n-1} \binom{n-1}{k-1}\frac{(k-1)!}{2\ln(2)^k}\ln n}{\frac{(n-1)!}{2\ln(2)^n} + O(\gamma^{k-1}(k-1)!)} \\
&\quad + \frac{\sum\limits_{k=1}^{n-1} \binom{n-1}{k-1} O(\gamma^{k-1}(k-1)!)}{\frac{(n-1)!}{2\ln(2)^n} + O(\gamma^{k-1}(k-1)!)} \\
&= c + O(\frac{1}{n}) + n + c' - \ln n + \frac{\sum\limits_{k=1}^{n-1} \binom{n-1}{k-1} O(\gamma^{k-1}(k-1)!)}{\frac{(n-1)!}{2\ln(2)^n} + O(\gamma^{n-1}(n-1)!)} \\
&= c + n + c' - \ln n + O(\frac{1}{n}) \\
&\sim n - \ln n.
\end{aligned}$$