Some families of trees arising in permutation analysis

Mathilde Bouvel*
Institute of Mathematics
University of Zurich, Switzerland
mathilde.bouvel@math.uzh.ch

Marni Mishna†
Department of Mathematics
Simon Fraser University, Canada
mmishna@sfu.ca

Cyril Nicaud‡
Laboratoire d'Informatique Gaspard Monge (LIGM)
Univ Gustave Eiffel, CNRS, ESIEE Paris
Marne-la-Vallée, France
cyril.nicaud@u-pem.fr

Submitted: Sep 30, 2016; Accepted: Apr 2, 2020; Published: May 15, 2020

Abstract

We extend classical results on simple varieties of trees (asymptotic enumeration, average behavior of tree parameters) to trees counted by their number of leaves. Motivated by genome comparison of related species, we then apply these results to strong interval trees with a restriction on the arity of prime nodes. Doing so, we describe a filtration of the set of permutations based on their strong interval trees. This filtration is also studied from a purely analytical point of view, thus illustrating the convergence of analytic series towards a non-analytic limit at the level of the asymptotic behavior of their coefficients.

Mathematics Subject Classifications: 05A15, 05A16

1 Introduction

Permutations can be realized in many different forms using a variety of structures. The idea of viewing permutations as enriched trees has been around for several decades in different research communities. For example, in the foundational enumerative study [1] of pattern-avoiding permutations, the properties of the corresponding decomposition trees play a crucial role. Very classically in the complexity analysis of sorting algorithms,

*Research supported by ANR project Magnum 2010-BLAN-0204 and SNF Marie Heim-Vögtlin grant PMPDP2_151254
†Research supported by NSERC Discovery grant RGPIN-04157
‡Research supported by ANR project Magnum 2010-BLAN-0204

parameters of an increasing tree structure are linked to the complexity of the sorting algorithm. Additionally, *PQ-trees* [6] appear in the context of graph algorithms and *strong interval trees* arise in comparative genomics [7] (and references therein, for instance).

The focus of this article is the study of strong interval trees, and the average value of certain parameters. The parameters are motivated by algorithm analysis in comparative genomics. The expected shape of these trees under a uniform distribution of permutations is described in [7]. They turn out to be extremely flat and rather degenerate, on average. This fact impacts the companion algorithm analysis. In the particular context of genome rearrangements, the permutations that arise in the biological data (for example in the study of mammalian genomes) do not appear under a uniform distribution, and are rarely flat in the same way. Consequently, Bouvel *et al.* [7] considered a subclass of strong interval trees – selected because they represent what is known as *commuting scenarios* [3] – that correspond to the class of *separable permutations*. This was framed a first step towards a more relevant model of permutations which arise in genome comparison.

By studying asymptotic enumeration and parameter formulas for separable permutations, they proved that the complexity of the algorithm of [3] solving the *perfect sorting by reversals* problem is polynomial-time on separable permutations, whereas this problem is NP-complete in general. Thus, this subclass better mirrored experimental data. Furthermore, they determined some average-case properties of the perfect sorting scenarios for separable permutations. We refer the reader to either of these sources for a more detailed description of the biological problem considered, and the implication of the parameter analysis.

Ideally, a clear understanding of the properties possessed by the strong interval trees that represent the comparison of actual genomes might tell us something about the evolutionary process. Bouvel *et al.* [7] conclude their study on separable permutations with a suggestion for the next class to study: strong interval trees with degree restrictions on certain internal nodes. These trees offer a very controlled way to introduce bias in the distribution of strong interval trees. This is precisely what we do in this work: we study strong interval trees where the so-called *prime nodes* (defined in Section 3) have a bounded number of children. We give a complete analysis of these restricted sets of trees. They can be completely understood combinatorially and analytically, and so we have access to formulas for enumeration and the average values of some tree parameters that are ultimately related to computing perfect sorting scenarios. These parameter values can help tailor the next generation of genome comparison models.

Although our initial motivation comes from genomics, our study has ramifications of independent interest in analytic combinatorics. Indeed, our work reveals a very lush substructure of permutations whose study from an analytical point of view allows us to formulate new questions on the convergence of sequences of combinatorial series.

Specifically, we define a sequence of families of trees (which are almost simple varieties of trees) whose combinatorial limit is the set of all strong interval trees. Simple varieties of trees are amenable to a standard set of tools for enumeration, parameter analysis, and random generation. However, the complete class of strong interval trees is not a simple variety of trees, so these techniques are not applicable to study the full class. Can we

understand, at the analytical level, the convergence of the generating functions of these sub-classes, which are all algebraic, towards the non-algebraic generating function of the full class? As we explain in details in our work, this question is naturally asked for strong interval trees, but it could also be considered for other classes such as k-regular graphs [13] or λ-terms of bounded unary height [5].

The organization of this article is as follows. First, in Section 2 we present some very general theorems for asymptotic enumeration and parameter analysis in families of trees counted by leaves, that are widely applicable, and use standard results on analytic inversion. In Section 3 we describe strong interval trees, a decomposable combinatorial class of trees counted by leaves, in bijection with permutations. We introduce a filtration of strong interval trees in Section 4 – the filtration is obtained by bounding the arity of so-called *prime* nodes. This restriction is motivated by algorithmic problems in genome rearrangements. Theorem 5 and Corollary 6 present asymptotic formulas for these trees. They witness an intriguing analytic phenomenon: the convergence of a sequence of (well-behaved and algebraic) families of trees towards the (transcendental and non-analytic) class of permutations. Section 5 establishes some first results in the exploration of this phenomenon.

A preliminary version of this work appeared in the extended abstract [8].

2 When the size of a tree is the number of leaves

There are many works which study the average case behavior of tree parameters, where the size of a tree is the number of internal nodes or of both internal nodes and leaves. The generating functions of these trees satisfy a functional equation of the form $T(z) = z \cdot \Phi(T(z))$. Such a class of trees is said to be a *simple variety of trees*. When Φ satisfies certain conditions, such as analyticity, then there are formulas for inversion, resulting in explicit enumerative results. The subject is exhaustively treated in Section VII.3 of [12], and the references listed therein.

If, instead, we define the size of a tree as the number of leaves, the generating function satisfies a relation of the form $T(z) = z + \Lambda(T(z))$. The generating function equation is a direct translation of a combinatorial decomposition according to the root node. The same general theorems on inversion still work, and it suffices to apply them and unravel the results. Even though they are less frequent, these have also been studied in the literature, and the applicability of the inversion lemmas is noted in Example VII.13 of [12], and in Theorem 2 of [15]. In this section we do this explicitly, to have the arguments easily at hand. In this work, when referring to simple varieties of trees, we mean a family of trees where the size is defined as the number of leaves, and whose study falls into the scope of the results of the present section.

Table 1 summarizes the results of this section. We determine asymptotic formulas for the number of trees, and several key parameters. The shape of the formulas are, unsurprisingly, not unlike those that arise in the study of trees counted by internal nodes.

Asymptotic number of trees with n leaves	$\sqrt{\frac{\rho}{2\pi\Lambda''(\tau)}} \cdot \frac{\rho^{-n}}{n^{3/2}}$
The average number of nodes of arity κ in trees with n leaves	$\frac{\lambda_\kappa \tau^\kappa}{\rho} \cdot n$
The average number of internal nodes in trees with n leaves	$\frac{\Lambda(\tau)}{\rho} \cdot n = \frac{\tau-\rho}{\rho} \cdot n$
The average subtree size sum in trees with n leaves	$\sqrt{\frac{\pi}{2\rho\Lambda''(\tau)}} \cdot n^{3/2}$

Table 1: A summary of parameters of trees given by $T = z + \Lambda(T)$ for Λ as in Theorem 1. The value τ is the unique solution to $\Lambda'(\tau) = 1$ between 0 and $R_\Lambda < 1$, and $\rho = \tau - \Lambda(\tau)$.

2.1 Asymptotic number of trees

Our entire analysis is roughly a consequence of the Analytic Inversion Lemma and Transfer Theorems. The version to which we appeal is given and proved in [12]. Citations to original sources may be found therein. The following theorem is a slight adaptation of Proposition IV.5 and Theorem VI.6 of [12] to combinatorial equations of the form $\mathcal{T} = \mathcal{Z} + \Lambda(\mathcal{T})$ instead of $\mathcal{T} = \mathcal{Z} \times \Phi(\mathcal{T})$. Recall [12, p.266] that $T(z)$ has *span* d if $T(z) = z^r G(z^d)$ for some $r \geqslant 0$. Its *period* is its largest span. If $T(z)$ has period 1, then it is *aperiodic*.

Theorem 1. *Let Λ be a function analytic at 0, with the following series expansion*

$$\Lambda(z) = \sum_{n \geqslant 2} \lambda_n z^n,$$

where the λ_n's are non-negative real numbers. Let R_Λ be the radius of convergence of this series. Under the condition $\lim_{x \to R_\Lambda^-} \Lambda'(x) > 1$, there exists a unique solution $\tau \in (0, R_\Lambda)$ of the equation $\Lambda'(\tau) = 1$. Then, the formal solution $T(z)$ of the equation

$$T(z) = z + \Lambda(T(z)) \tag{1}$$

is analytic at 0, its unique positive, real valued dominant singularity is at $\rho = \tau - \Lambda(\tau)$. The function $T(z)$ has a singular expansion valid for points near ρ in an appropriate Δ-domain given by

$$T(z) = \tau - \sqrt{\frac{2\rho}{\Lambda''(\tau)}} \left(1 - \frac{z}{\rho}\right)^{1/2} + \mathcal{O}\left(1 - \frac{z}{\rho}\right). \tag{2}$$

Moreover, if $\gcd\{n - 1 \mid \lambda_n \neq 0\} = 1$, then T is aperiodic, ρ is the only dominant singularity and

$$[z^n]T(z) \sim \sqrt{\frac{\rho}{2\pi\Lambda''(\tau)}} \cdot \frac{\rho^{-n}}{n^{3/2}}. \tag{3}$$

If the value of $\gcd\{n - 1 \mid \lambda_n \neq 0\} = d \neq 1$, then we can show that $T(z) = zH(z)$ with H d-periodic. In this case, $T(z)$ will have d singularities on the circle of convergence,

by the Daffodil lemma (see [12, Lemma IV.1] for a nice proof). In such a situation the asymptotic formula is composed of a sum of terms, each resembling the right hand side of Equation (3).

Proof. The conditions on Λ imply that both $\Lambda(x)$ and $\Lambda'(x)$ are increasing continuous functions for x in the real interval $(0, R_\Lambda)$. Since $\Lambda'(0) = 0$ and since $\lim_{x \to R_\Lambda^-} \Lambda'(x) > 1$, there exists $R' \in (0, R_\Lambda)$ such that $\Lambda'(R') > 1$. Hence there exists a unique $\tau \in (0, R')$, and thus on $(0, R_\Lambda)$, such that $\Lambda'(\tau) = 1$.

Now observe that Equation (1) admits a unique formal power series solution $T(z)$, which has non-negative coefficients, by bootstrapping the coefficients. By Analytic Inversion [12, Lemma IV.2], this solution is analytic at $z = 0$ and with $T(0) = 0$: Equation (1) can be rewritten $\Psi(T(z)) = z$, with $\Psi(x) = x - \Lambda(x)$, and $\Psi'(0) \neq 0$.

Let r be the radius of convergence of $T(z)$. As $T(z)$ is analytic, we have that $r > 0$. We can justify that r is finite as follows. Let $T(r) \in (0, +\infty]$ be defined by $T(r) = \lim_{x \to r^-} T(x)$. Following almost exactly the proof of Proposition IV.5 in [12, p. 278], we deduce that $T(r) = \tau$. As τ is finite, this implies in particular that r is finite. By Pringsheim's Theorem r is a dominant singularity of $T(z)$. Moreover, since T and Ψ are inverse functions, we can determine a form for this dominant singularity, which we henceforth refer to as ρ: $\rho = \tau - \Lambda(\tau)$.

The remainder of the proof closely follows the proof Theorem VI.6 in [12, p. 405], using our specific equations to obtain Equation (2). In the aperiodic case, the Daffodil Lemma and the Analytic Inversion Lemma ensure that there is no other singularity than ρ on the circle of radius ρ, and that $T(z)$ can be analytically continued in a Δ-domain at ρ. Applying the Transfer Theorem yields Equation (3), concluding the proof. □

2.2 Parameter Analysis

In the case of trees counted by internal nodes, the study of recursively defined parameters is very straightforward, starting from generating function equations. We can describe analogous versions for trees counted by leaves. In particular, we consider additive parameters, and describe a modified iteration lemma, adapted to our notion of size. We illustrate the lemma on the number of internal nodes, the subtree size sum and the number of nodes of a given arity.

2.2.1 General additive parameters

Our focus is on tree parameters that can be computed additively by parameters of subtrees. More precisely, we consider a parameter $\xi(t)$ for trees $t \in \mathcal{T}$ which satisfies the relation

$$\xi(t) = \eta(t) + \sum_{j=1}^{\deg(t)} \sigma(t_j),$$

where $\deg(t)$ is the arity of the root, t_j are its children sub-trees, η is a simpler tree parameter, and σ is either ξ or a simpler tree parameter. Let $\Xi(z)$, $H(z)$ and $\Sigma(z)$ be the

associated cumulative generating functions of ξ, η and σ. That is,

$$\Xi(z) = \sum_{t\in\mathcal{T}} \xi(t) z^{|t|}, \quad H(z) = \sum_{t\in\mathcal{T}} \eta(t) z^{|t|} \quad \text{and} \quad \Sigma(z) = \sum_{t\in\mathcal{T}} \sigma(t) z^{|t|}.$$

Lemma VII.1 in [12] has an analogue for trees counted by their leaves, and it is proved in a very similar way.

Lemma 2 (Iteration Lemma for trees counted by their leaves)**.** *Let $\Lambda(T) = \sum \lambda_k T^k$ and let $\mathcal{T}$ be a class of trees satisfying $\mathcal{T} = \mathcal{Z} + \Lambda(\mathcal{T})$. The cumulative generating functions are related by the equation*

$$\Xi(z) = H(z) + \Lambda'(T(z))\,\Sigma(z).$$

In particular, if $\sigma \equiv \xi$, one has $\Xi(z) = \frac{H(z)}{1-\Lambda'(T(z))} = H(z)\cdot T'(z)$.

Proof. Unraveling the definition of $\xi(t)$, we have

$$\Xi(z) = H(z) + \widetilde{\Xi}(z) \;\text{ with }\; \widetilde{\Xi}(z) = \sum_{t\in\mathcal{T}} z^{|t|} \sum_{j=1}^{\deg(t)} \sigma(t_j).$$

Splitting the sum defining $\widetilde{\Xi}(z)$ according to the value r of the degree of the root of t, we get:

$$\begin{aligned}
\widetilde{\Xi}(z) &= \sum_{r\geqslant 1} \sum_{t_1,\dots,t_r\in\mathcal{T}} \lambda_r z^{|t_1|+\cdots+|t_r|}(\sigma(t_1)+\cdots+\sigma(t_r)) \\
&= \sum_{r\geqslant 1} \sum_{t_1,\dots,t_r\in\mathcal{T}} \lambda_r \left(\sigma(t_1) z^{|t_1|} z^{|t_2|+\cdots+|t_r|} + \cdots + \sigma(t_r) z^{|t_r|} z^{|t_1|+\cdots+|t_{r-1}|}\right) \\
&= \sum_{r\geqslant 1} \lambda_r \times r \times \Sigma(z) T(z)^{r-1} = \Lambda'(T(z))\,\Sigma(z).
\end{aligned}$$

In the case $\sigma \equiv \xi$, $\Xi(z) = \frac{H(z)}{1-\Lambda'(T(z))}$ is derived immediately. The last equality is a consequence of $T'(z)(1-\Lambda'(T(z))) = 1$, which is obtained by differentiating $T(z) = z + \Lambda(T(z))$ with respect to z. □

Note that if $\sigma \equiv \xi$, the parameter is said to be *recursive*. Most basic parameters are recursive, and in what follows we shall use this case only.

Note also that when analytic treatment applies, $T(z)$ has a square-root singularity (see Theorem 1), so that $T'(z)$ has an inverse[1] square-root singularity (by analytic derivation, as we shall see). Therefore, whenever $H(z)$ tends to a positive real when $z \to \rho$ (under some analytic conditions), then the Transfer Theorem yields an asymptotic equivalent of the mean value of the parameter of the form $c \cdot n$. The scenario is rather classic, and the hypotheses about the domain follow in a straightforward way. This situation is the case for the number of nodes of fixed arity and the number of internal nodes, as shown below.

[1] A multiplicative inverse, not functional inverse.

2.2.2 Three applications

Number of nodes with exactly κ children. We "mark" nodes of arity κ by setting

$$\eta(t) = \begin{cases} 1 & \text{if the root of } t \text{ is of arity } \kappa, \\ 0 & \text{otherwise.} \end{cases}$$

Hence if $\kappa \geqslant 2$, $H(z) = \sum_{t\in\mathcal{T}} \eta(t) z^{|t|} = \sum_{t_1,\dots,t_\kappa\in\mathcal{T}} \lambda_\kappa z^{|t_1|+|t_2|+\cdots+|t_\kappa|}$ so that $H(z) = \lambda_\kappa\, T(z)^\kappa$. There can be no unary nodes in a proper specification, so $\kappa \neq 1$. If $\kappa = 0$, then $H(z) = z$ which is not interesting since it is simply counting the number of leaves *i.e.* the size of the tree.
By Lemma 2, for any $\kappa \geqslant 2$ one has $\Xi(z) = \lambda_\kappa T(z)^\kappa \cdot T'(z)$. Since the singular expansion of $T(z)$ near ρ is

$$T(z) = \tau - \gamma\sqrt{1 - z/\rho} + o\left(\sqrt{1 - z/\rho}\right), \text{with } \gamma = \sqrt{\frac{2\rho}{\Lambda''(\tau)}} \tag{4}$$

then near ρ, one has $T(z)^\kappa = \tau^\kappa + \mathcal{O}\left(\sqrt{1 - z/\rho}\right)$. Using the Singular Differentiation Theorem [12, Theorem VI.8, p. 419] we have

$$T'(z) = \frac{\gamma}{2\rho\sqrt{1 - z/\rho}} + o\left(\frac{1}{\sqrt{1 - z/\rho}}\right), \text{ so that } \Xi(z) = \frac{\lambda_\kappa\gamma\tau^\kappa}{2\rho\sqrt{1 - z/\rho}} + o\left(\frac{1}{\sqrt{1 - z/\rho}}\right),$$

from which we get the asymptotics of the coefficients of the cumulative generating function

$$[z^n]\Xi(z) \sim \frac{\lambda_\kappa\gamma\tau^\kappa\rho^{-n-1}}{2\sqrt{\pi n}}.$$

The asymptotics of the average value across all trees of size n is then (by the Transfer theorem)

$$\frac{[z^n]\Xi(z)}{[z^n]T(z)} \sim \frac{\lambda_\kappa\gamma\tau^\kappa\rho^{-n-1}}{2\sqrt{\pi n}} \cdot \sqrt{\frac{2\pi\Lambda''(\tau)}{\rho}} \frac{n^{3/2}}{\rho^{-n}} \sim \frac{\lambda_\kappa\tau^\kappa}{\rho} \cdot n,$$

as reported in Table 1.

Number of internal nodes. For this parameter, just take the following definition for η:

$$\eta(t) = \begin{cases} 0 & \text{if } t \text{ is just one leaf,} \\ 1 & \text{otherwise.} \end{cases}$$

One has $H(z) = \sum_{t\in\mathcal{T}} \eta(t) z^{|t|} = T(z) - z$, and therefore (with the γ of Equation (4))

$$\Xi(z) = (T(z) - z)\; T'(z) = \frac{\gamma(\tau - \rho)}{2\rho\sqrt{1 - z/\rho}} + o\left(\frac{1}{\sqrt{1 - z/\rho}}\right).$$

It follows that

$$[z^n]\Xi(z) \sim \frac{\gamma(\tau - \rho)\rho^{-n-1}}{2\sqrt{\pi n}} \quad \text{and} \quad \frac{[z^n]\Xi(z)}{[z^n]T(z)} \sim \frac{\tau - \rho}{\rho} \cdot n.$$

Subtree size sum. We are interested in the subtree size sum parameter, defined by $\eta(t) = |t|$. This implies that $H(z) = zT'(z)$, so that

$$\Xi(z) = zT'(z)^2 = \frac{\gamma^2}{4\rho(1-z/\rho)} + o\left(\frac{1}{1-z/\rho}\right) \quad \text{and} \quad [z^n]\Xi(z) \sim \frac{\gamma^2}{4\rho} \cdot \rho^{-n}.$$

Unlike the two previous examples, this is not an inverse of square-root singularity. In this case, for the average value of the subtree size sum, we find

$$\frac{[z^n]\Xi(z)}{[z^n]T(z)} \sim \frac{\gamma^2}{4\rho}\rho^{-n} \cdot \sqrt{\frac{2\pi\Lambda''(\tau)}{\rho}}\frac{n^{3/2}}{\rho^{-n}} \sim \sqrt{\frac{\pi}{2\rho\Lambda''(\tau)}} \cdot n^{3/2},$$

that is, an asymptotic equivalent in $n^{3/2}$. This behavior is typical for such path length related parameters.

There are many other tree parameters that we could consider in a similar fashion, particularly additive parameters, as the generating function manipulations are similar. Pitman and Rizzolo find the distribution of the height of a random leaf [15]. This is a parameter that we might consider. Although we are illustrating only average case computations here, we could also obtain information about higher moments upon consideration of higher derivatives.

3 Strong Interval Trees

Our main application is the study of parameters of *strong interval trees* which encode permutations. They have been introduced in the early 2000's in a bioinformatics context [14, 4], as they are a very effective data structure for algorithms in reconstruction of genome evolution scenarios, as we briefly mentioned in the introduction. Under a different name, and roughly at the same time, these objects also made their appearance in combinatorics, in the study of permutation patterns: strong interval trees (rather called (substitution) decomposition trees) are a tree representation of the block decomposition of permutations described by Albert and Atkinson [1]. Although the proper definition of strong interval trees is relatively recent, it can be traced to older notions of decomposition (of graphs in particular): it is a close relative of the modular decomposition trees of permutation graphs [4] and even has origins in the PQ-trees of Booth and Lueker [6].

In this section, we review the definition of strong interval trees and the bijection with permutations. Then, we turn to a presentation of these objects as a constructible combinatorial class, in the flavor of what is done in Section 2.

3.1 Definition and bijection with permutations

Strong interval trees are most often defined via the bijection that relates them to permutations. Different presentations of this bijection can be found for instance in [1, 4, 7]. For the reader who is not familiar with these objects, we review the definition of strong interval trees, and the correspondence with permutations below.

Here, we consider the one-line notation of permutations, and hence we view a permutation of size n as a word containing exactly once each symbol in $\{1, 2, \ldots, n\}$.

An *interval* of a permutation σ is a factor of σ, such that the underlying set of symbols is an interval of integers. For instance, 7 9 10 11 13 8 12 and 3 1 5 4 2 are intervals of the permutation

$$6\ 7\ 9\ 10\ 11\ 13\ 8\ 12\ 3\ 1\ 5\ 4\ 2,$$

but 10 11 13 is not (12 is missing). For every permutation σ of size n, the singletons i (for $1 \leqslant i \leqslant n$) and σ itself are intervals of σ. They are called *trivial* intervals of σ.

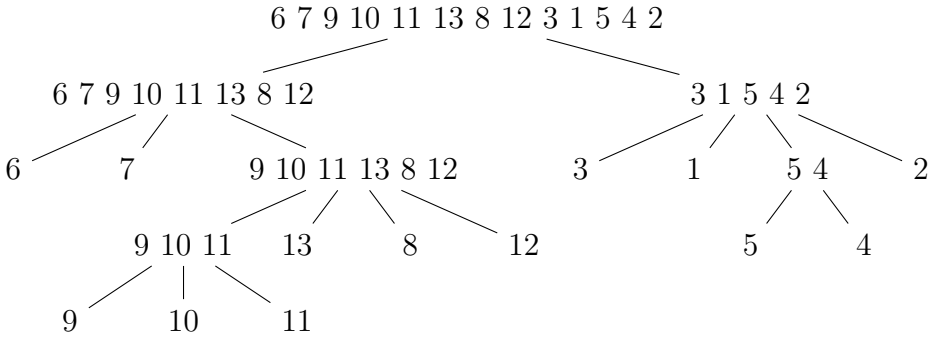
A permutation is said to be *simple* when its only intervals are the trivial ones. Note that our convention will be that 1, 1 2 and 2 1 are not simple permutations, although they satisfy the above definition. It is immediate to check that there is no simple permutation of size 3 (each permutation of size 3 contains an interval of size 2), and that there are 2 simple permutations of size 4, namely 2 4 1 3 and 3 1 4 2. A larger simple permutation is for instance 3 5 7 1 4 2 6. We will go back to the enumeration of simple permutations in the next subsection.

Two intervals of σ *overlap* when their intersection is neither empty nor equal to one of them. Returning to our example of 6 7 9 10 11 13 8 12 3 1 5 4 2, the intervals 6 7 and 7 9 10 11 13 8 12 overlap (their intersection is 7), but 10 11 and 5 4 do not. A *strong* interval of σ is an interval that does not overlap any other interval of σ. The trivial intervals are obviously strong. On our running example, the non-trivial strong intervals are

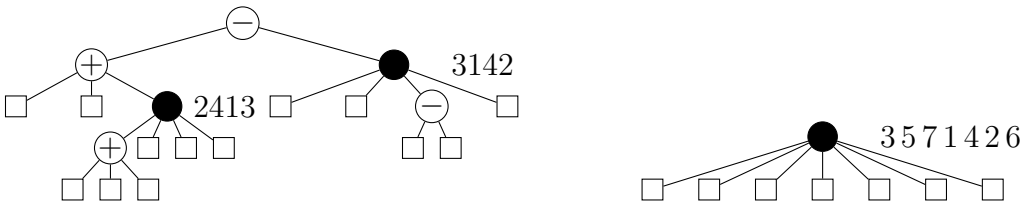
$$5\ 4\ ; \quad 3\ 1\ 5\ 4\ 2\ ; \quad 9\ 10\ 11\ ; \quad 9\ 10\ 11\ 13\ 8\ 12 \quad \text{and} \quad 6\ 7\ 9\ 10\ 11\ 13\ 8\ 12.$$

From their definition, it follows immediately that the inclusion order on the set of strong intervals of a permutation σ induces a tree structure, where the leaves are the singletons, and the root is the σ itself. This is the *strong interval tree* of σ.

From there, and depending on the context, the definition of the strong interval tree may vary. For us, these trees are *embedded in the plane*, by imposing the order of the leaves. Namely, from left to right, the leaves (corresponding to singletons of σ) are required to appear in the same order as in σ. The strong interval tree of our running example would then be:



From this tree, there is a last step that we perform before obtaining what we refer to as the strong interval tree in our work. It relies on an important remark, proved for



(a) $\sigma_1 = 6\ 7\ 9\ 10\ 11\ 13\ 8\ 12\ 3\ 1\ 5\ 4\ 2$ (b) $\sigma_2 = 3\ 5\ 7\ 1\ 4\ 2\ 6$

Figure 1: Two permutations and their associated strong interval trees

instance in [9]. To state it, we first need to describe how to associate a permutation of size k to each node of the strong interval tree with k children.

Note first that given several disjoint strong intervals, the natural order on integers induces an order among them: the smaller the elements contained in the interval, the smaller the interval itself. On our running example, we have for instance that 5 4 is smaller than 8 which is itself smaller than 9 10 11. Consider now a non-singleton strong interval, corresponding to an internal node of the strong interval tree. Its children (assume there are k of them) are also strong intervals, and they are disjoint. So they can be ordered as described above. To this node of the tree, we associate a permutation τ of size k built as follows: $\tau_i = j$ if the i-th child from the left is the j-th smallest one. For instance, the permutation τ associated with the node labeled 9 10 11 13 8 12 in our running example is 2 4 1 3 since 8 is smaller than 9 10 11, itself smaller than 12 and 13.

The permutations labeling the node enjoy a remarkable property (see [9], or in a somewhat different presentation [1]): they are either increasing $(1\ 2 \dots k)$ or decreasing $(k \dots 2\ 1)$ or simple. In the remainder of this article, when speaking about strong interval trees, we mean the plane tree whose structure has been described above, but whose internal nodes are only labeled by $\oplus$, $\ominus$ (corresponding to increasing or decreasing permutations respectively), or by a simple permutation. In particular, the leaves carry no label. Nodes labeled by $\oplus$ or $\ominus$ are called *linear*, whereas those labeled by simple permutations are called *prime*. Figure 1-(a) shows the strong interval tree of our running example, and Figure 1-(b) represents the strong interval tree of a simple permutation. What can be observed on this example is true in general: the trees corresponding to simple permutations consist of a single prime node labeled by the permutation itself, with pending leaves.

In a strong interval tree, it is impossible for a node labeled by $\oplus$ (resp. $\ominus$) to have a child carrying the same label. This property appears (although in disguise) in [1], but is simply proved by contradiction: assuming that a parent and a child both carry the label $\oplus$ (resp. $\ominus$) contradicts that the child is a strong interval (indeed, it overlaps an interval resulting from the union of one of its own children with one of its siblings).

With this in mind, strong interval trees are now just plane trees, where internal nodes are of arity at least 2 and carry labels $\oplus$, $\ominus$ or α for any simple permutation α, with the additional conditions that a node labeled by $\oplus$ (resp. $\ominus$) does not have a child carrying the same label, and that the number of children of a node labeled by a simple permutation α is exactly the size of α.

It turns out (and the proof follows immediately from [1]) that any such tree is the strong interval tree of a permutation. Moreover, the above construction provides a bijection between permutations and strong interval trees. Note that the bijection is completely constructive, and that it can be computed in linear time, although this is quite difficult to achieve, see [4].

3.2 Strong interval trees as a constructible class

From now on, we denote by $\mathcal{P}$ the class of strong interval trees. As we have seen above, this class is a set of trees where some internal nodes are signed and others are enriched with a simple permutation. More precisely, the characterization of strong interval trees given above can be rephrased as shown in the following theorem.

Theorem 3 (Reformulated from [1])**.** *The class of permutations is in a size-preserving bijection with the combinatorial class $\mathcal{P}$ of strong interval trees. These are enriched trees defined by the following combinatorial system, where size is given by the number of leaves:*

$$\begin{aligned}\mathcal{P} &= \mathcal{Z}_{\square} + \mathcal{N}_{\oplus} \cdot \mathrm{Seq}_{\geqslant 2}\mathcal{U}_{\oplus} + \mathcal{N}_{\ominus} \cdot \mathrm{Seq}_{\geqslant 2}\mathcal{U}_{\ominus} + \mathcal{N}_{\bullet} \cdot\ S(\mathcal{P}),\\ \mathcal{U}_{\oplus} &= \mathcal{Z}_{\square} + \mathcal{N}_{\ominus} \cdot \mathrm{Seq}_{\geqslant 2}\mathcal{U}_{\ominus} + \mathcal{N}_{\bullet} \cdot\ S(\mathcal{P}),\\ \mathcal{U}_{\ominus} &= \mathcal{Z}_{\square} + \mathcal{N}_{\oplus} \cdot \mathrm{Seq}_{\geqslant 2}\mathcal{U}_{\oplus} + \mathcal{N}_{\bullet} \cdot\ S(\mathcal{P}).\end{aligned} \tag{5}$$

Above, the class $\mathcal{Z}$ is an atomic class with a single element of size 1, the $\mathcal{N}$ classes are all epsilon classes containing a single element of size 0, marking internal nodes, and the function $S(z) = \sum_{j\geqslant 4} s_j z^j$ is the generating function for simple permutations.

Notice that $\mathcal{U}_{\oplus}$ and $\mathcal{U}_{\ominus}$ define combinatorial classes which are in obvious size-preserving bijection. In the following, in order to deal with one class instead of two, we replace them by the equivalent class $\mathcal{U} = \mathcal{Z}_{\square} + \mathcal{N}_{\circ} \cdot \mathrm{Seq}_{\geqslant 2}\mathcal{U} + \mathcal{N}_{\bullet} \cdot\ S(\mathcal{P})$. Doing so, we change the labels of the linear nodes having a linear parent (replacing them by $\circ$). This does not affect the enumeration of the class. Indeed, these labels are determined since a linear node and its linear parent have different labels.

It is not hard to view $\mathcal{U}$ as a family of trees not unlike those studied in Section 2:

Corollary 4. *The following combinatorial equivalences are true:*

$$\mathcal{P} \equiv \mathrm{Seq}_{\geqslant 1}\mathcal{U} \qquad \text{and} \qquad \mathcal{U} \equiv \mathcal{Z} + \mathrm{Seq}_{\geqslant 2}\mathcal{U} + S(\mathrm{Seq}_{\geqslant 1}\mathcal{U}).$$

Consequently, $\mathcal{U}$ is in bijection with a class of Λ-trees, or in other words its generating function $U(z)$ satisfies $U(z) = z + \Lambda(U(z))$, for $\Lambda(x) = \frac{x^2}{1-x} + \sum_{j\geqslant 4} s_j \left(\frac{x}{1-x}\right)^j$, where s_j is the number of simple permutations of size j.

Proof. This equivalence is derived from Equation (5), the fact that $\mathcal{U} \equiv \mathcal{U}_{\oplus} \equiv \mathcal{U}_{\ominus}$, and the intermediary equivalence $\mathcal{P} \equiv \mathcal{U} + \mathrm{Seq}_{\geqslant 2}\mathcal{U}$. □

There is however an important difference between $\mathcal{U}$ and the set of classes that are covered by Theorem 1: the function Λ defined by $\Lambda(x) = \frac{x^2}{1-x} + \sum_{j\geqslant 4} s_j \left(\frac{x}{1-x}\right)^j$ has zero

radius of convergence. The divergence comes from the coefficients s_j. This is immediate from the asymptotic formulas for these values, given by Albert *et al.* [2]. We make heavy use of their estimates, and recall them now.

The sequence enumerating simple permutations, $(s_n)_n$, has label A111111 in the On-Line Encyclopedia of Integer Sequences [16]. This sequence is not P-recursive, but it does satisfy a simple functional inversion formula (see [2]), and we have calculated exact values of s_n for $n < 800$. Albert *et al.* [2] determined the following bounds:

$$\frac{n!}{e^2}\left(1-\frac{4}{n}\right) \leqslant s_n \leqslant \frac{n!}{e^2}\left(1-\frac{4}{n}+\frac{2}{n(n-1)}\right). \tag{6}$$

Here are the first few terms in the generating function for simple permutations:

$$S(z) = 2z^4 + 6z^5 + 46z^6 + 338z^7 + 2926z^8 + 28146z^9 + 298526z^{10} + 3454434z^{11} + \dots$$

Because $S(z)$, and hence $\Lambda(x)$, are not analytic at the origin, neither $\mathcal{P}$ nor $\mathcal{U}$ are simple varieties of trees whose analysis is covered by Section 2. However, we can express $\mathcal{U}$ as the combinatorial limit of a sequence of simple varieties of trees, and properties of $\mathcal{U}$ can next easily be translated into properties of $\mathcal{P}$ via the relation $\mathcal{P} \equiv \mathrm{Seq}_{\geqslant 1}\mathcal{U}$. Now, $\mathcal{P}$ is in bijection with permutations, so we have a strong understanding of the enumeration by several other methods. However, this does give us the template for a general strategy of studying limit families, and the solution has some insightful subtleties which we explore next.

4 Prime-Degree Restricted Strong Interval Trees

We filter the class of trees $\mathcal{P}$ by a parameter that bounds the maximal arity of prime nodes. The class of trees wherein this parameter is bounded is a simple variety of trees, and the results of Section 2 are applicable. Our motivation for studying this restriction of strong interval trees is twofold.

First, as indicated above, the full class of trees is in bijection with permutations, and hence the ordinary generating function is neither analytic at the origin, nor algebraic. We give a natural way to express the generating function as the limit of a sequence of analytic, algebraic generating functions. We believe this example is instructive and opens the way to adapting this strategy to study other combinatorial classes which are similarly complex, as we discuss in Section 5.

Our second motivation comes from the study of genome rearrangements, specifically in the model of perfect sorting by reversals. Indeed, as shown in [3, 7], the algorithmic complexity of finding an evolutionary scenario in this model depends heavily on the maximal arity of prime nodes in the strong interval trees of the permutations that encodes the genomes (recording the order of the genes): the smaller this maximal arity, the more efficient the algorithm. Based on biological data for mammalian genomes [11], it appears that prime nodes occur relatively rarely, and are of small arity. In [7], the combinatorics of strong interval trees without any prime nodes was investigated, resulting in a better

understanding of the so-called *commuting scenarios*. Now allowing some prime nodes to occur, but with a bounded arity, we are going a step further in this analysis, while focusing on subclasses of strong interval trees that seem to represent the biological data better than uniform random permutations. We propose the permutations that arise in these classes as a better testing ground for algorithms destined for biological purposes.

4.1 The filtration for permutations

We define the class $\mathcal{P}_k$ as follows, where $S_k(z) = \sum_{j=4}^{k} s_j z^j$:

$$\mathcal{P}_k = \mathcal{Z} + 2\,\mathrm{Seq}_{\geqslant 2}\mathcal{U}_k + S_k(\mathcal{P}_k) \qquad \text{and} \qquad \mathcal{U}_k = \mathcal{Z} + \mathrm{Seq}_{\geqslant 2}\mathcal{U}_k + S_k(\mathcal{P}_k).$$

That is, only prime nodes of arity at most k are allowed. We refer to the classes denoted by $\mathcal{P}_k$, as classes of *prime-degree restricted* strong interval trees.

The containment $\mathcal{P}_k \subset \mathcal{P}_{k+1}$ is obvious. When $k \geqslant n$, the restriction on the prime node degree has no impact, hence the set of trees of size n in $\mathcal{P}_k$ is precisely the set of trees of size n in $\mathcal{P}$ when $k \geqslant n$. From this property we then deduce the limit of combinatorial classes $\lim_{k\to\infty} \mathcal{P}_k = \mathcal{P}$.

Furthermore, by the same manipulations as for the full class, we derive:

$$\mathcal{P}_k \equiv \mathrm{Seq}_{\geqslant 1}\mathcal{U}_k \qquad \text{and} \qquad \mathcal{U}_k \equiv \mathcal{Z} + \mathrm{Seq}_{\geqslant 2}\mathcal{U}_k + S_k(\mathrm{Seq}_{\geqslant 1}\mathcal{U}_k). \tag{7}$$

As in the case of the full class, $\mathcal{U}_k$ is isomorphic to a Λ_k-tree with $\Lambda_k(x) = \frac{x^2}{1-x} + \sum_{j=4}^{k} s_j \left(\frac{x}{1-x}\right)^j$. This class is certainly algebraic and is a simple variety of trees. The enumerative analysis of Section 2 applies directly to these families of trees $\mathcal{U}_k$, then giving access to enumeration and parameter average behavior for $\mathcal{P}_k$ also, even if $\mathcal{P}_k$ is not itself a simple variety of trees. This is done in the remaining part of this section, focusing on applications to the study of genome rearrangements. Also, keeping in mind our next goal of letting k go to infinity to recover the class $\mathcal{P}$, we would like to preserve k as much as possible in the formulas.

4.2 Asymptotic enumeration

The combinatorial specification in Equation (7) allows us to directly apply Theorem 1 to determine asymptotic formulas for the coefficients of the generating functions $P_k(z)$ of the classes $\mathcal{P}_k$.

To state the theorem relatively compactly, we first introduce some notation. Since the finite sum $\sum_{j=4}^{k} s_j(\frac{x}{1-x})^j$ is a polynomial in $\frac{x}{1-x}$, the function $\Lambda_k(x)$ is certainly analytic at 0. The radius of convergence of Λ_k is easily seen to be 1, and $\lim_{x\to 1^-} \Lambda_k'(x) = +\infty$. Hence, Theorem 1 gives the following expansion, valid in an appropriate Δ-domain:

$$U_k(z) = \tau_k - \beta_k \left(1 - \frac{z}{\rho_k}\right)^{1/2} + \mathcal{O}\left(1 - \frac{z}{\rho_k}\right),$$

with τ_k, ρ_k and β_k defined in Theorem 5 below.

Theorem 5. *For any fixed k, the number of prime-degree restricted strong interval trees of size n, denoted $p_k(n)$, grows asymptotically like*

$$p_k(n) \sim \frac{\gamma_k}{(1-\tau_k)^2}\,\rho_k^{-n}\,n^{-3/2}, \qquad \text{as } n \to \infty, \tag{8}$$

where $\Lambda_k(x) = \dfrac{x^2}{1-x} + \displaystyle\sum_{j=4}^{k} s_j \left(\frac{x}{1-x}\right)^j$, *$\tau_k$ satisfies* $1 - \Lambda_k'(\tau_k) = 0$, $\rho_k = \tau_k - \Lambda_k(\tau_k)$, $\beta_k = \sqrt{\dfrac{2\rho_k}{\Lambda_k''(\tau_k)}}$ *and* $\gamma_k = \dfrac{\beta_k}{2\sqrt{\pi}}$.

Proof. By the first relation in Equation (7), $P_k(z) = \frac{U_k(z)}{1-U_k(z)}$. By Theorem 1, the value of $U_k(z)$ at its dominant singularity ρ_k is τ_k. Moreover, τ_k is less than the radius of convergence of Λ_k, *i.e.*, $\tau_k < 1$. So the composition $P_k(z) = \frac{U_k(z)}{1-U_k(z)}$ is subcritical (see [12, paragraph VI.9]): this implies that the dominant singularity of $P_k(z)$ is also ρ_k. We unravel the relation to get a singular expansion of $P_k(z)$ at ρ_k in terms of the components of the expression for $U_k(z)$, valid in a similar region:

$$P_k(z) = \frac{U_k(z)}{1-U_k(z)} = \frac{\tau_k}{1-\tau_k} - \frac{\beta_k}{(1-\tau_k)^2}\left(1-\frac{z}{\rho_k}\right)^{1/2} + O\left(1-\frac{z}{\rho_k}\right).$$

Again, the result follows from the classic Transfer theorem. □

For the parameter analysis it is useful to now compare the first term of the expansions for $u_k(n) = [z^n]U_k(z)$ and $p_k(n)$:

$$u_k(n) \sim \gamma_k\,\rho_k^{-n}\,n^{-3/2} \qquad p_k(n) \sim \frac{\gamma_k}{(1-\tau_k)^2}\rho_k^{-n}\,n^{-3/2}.$$

They differ only by a multiplicative factor of $(1-\tau_k)^{-2}$. Note that along the proof of Theorem 5, we have seen that $\tau_k < 1$, an inequality that will be useful in Section 5 to bound the asymptotic estimate of Equation (8).

Corollary 6. *For any fixed k, as n tends to infinity, $p_k(n)$ behaves like $\frac{\gamma_k}{(1-\tau_k)^2}\rho_k^{-n}\,n^{-3/2}$. When k goes to infinity, this estimate is no larger than*

$$\frac{1}{(1-\frac{e}{k})^2}\sqrt{\frac{e}{4k\pi}}\left(\frac{k}{e}\right)^n\left(1+\frac{5}{2}\frac{\log k}{k}+\mathcal{O}\left(\frac{1}{k}\right)\right)^n n^{-3/2}. \tag{9}$$

Proof. The estimate for $p_k(n)$ has been proved in Theorem 5 above. To obtain the announced upper bound on this estimate, we rely on strong bounds on τ_k and ρ_k proved in Proposition 8 and Theorem 11 (see Section 4.7). Assuming that k is large enough, these two technical results yield:

- $\tau_k \leqslant \frac{\tau_k}{1-\tau_k} \leqslant \frac{e}{k}$, hence $\frac{1}{(1-\tau_k)^2} \leqslant \frac{1}{(1-\frac{e}{k})^2}$;

k	τ_k	ρ_k	k	τ_k	ρ_k
4	0.2258458016	0.1454726242	9	0.1463252500	0.1102193554
5	0.2043553556	0.1364583031	10	0.1375961304	0.1057725121
6	0.1841224072	0.1277948168	11	0.1300393555	0.1017629085
7	0.1689470150	0.1210046262	12	0.1234001218	0.09810173382
8	0.1565912704	0.1152312243	13	0.1174959122	0.09472586497

Table 2: Computed approximate values for ρ_k and τ_k for small values of k, using Maple code available at https://github.com/marnijulie/strong-interval-trees-maple.

- $\rho_k \leqslant \frac{e}{k}$ and $\Lambda_k''(\tau_k) \geqslant 2$, hence $\gamma_k \leqslant \sqrt{\frac{e}{4k\pi}}$;
- and $\rho_k^{-n} = \left(\frac{e}{k}\right)^{-n}\left(1 - \frac{5}{2}\frac{\log k}{k} + \mathcal{O}\left(\frac{1}{k}\right)\right)^{-n} = \left(\frac{k}{e}\right)^{n}\left(1 + \frac{5}{2}\frac{\log k}{k} + \mathcal{O}\left(\frac{1}{k}\right)\right)^{n}$. □

Table 2 contains numeric approximations for τ_k and ρ_k in the range $k = 4 \ldots 13$. Using these estimates gives good asymptotic approximations and the enumerative formulas given in Equation (8) converge quickly for fixed k. For example, when $k = 8$, our asymptotic formula is within 2% of the correct value at $n = 10$.

4.3 When k is a function of n

As pointed out by a referee, our estimate of $p_k(n)$ and its upper bound (Theorem 5 and Corollary 6) should extend to the case where k is allowed to depend on n, in some controlled way to be determined. Indeed, more terms in the Taylor expansion of $\Lambda_k(u)$ near τ_k (see Equation (2)) are easily obtained using Newton iteration, which result in more terms in the approximation of $U_k(z)$ and $P_k(z)$ near ρ_k in the powers of $\sqrt{1 - z/\rho_k}$. The formulas so obtained should generalize our result.

However, to complete the estimate requires some precise computation of the error terms to establish that this method is valid, and it appears to be necessary to go back to the proofs of the analytic tools we use. For instance, the Taylor development of $\Lambda_k(u)$ near τ_k is

$$\Lambda_k(u) = \Lambda_k(\tau_k) + \sum_{i \geqslant 1} \frac{\Lambda_k^{(i)}(\tau_k)}{i!}(u - \tau_k)^i, \tag{10}$$

where $\Lambda_k^{(i)}$ denotes the i-th derivative of Λ_k. From the bounds on $\tilde{\tau}_k$ provided in Proposition 8, we can show that $\Lambda_k^{(i)}(\tau_k) = \Theta(k^{2i-2})$ for all $i \geqslant 2$. In a formula with an error term, such as $\Lambda_k(u) = \Lambda_k(\tau_k) + u - \tau_k + \mathcal{O}((u - \tau_k)^2)$, the constant within the $\mathcal{O}$ notation is super-polynomial in n for $k = n^\alpha$ with $\alpha > \frac{1}{2}$.

Secondly, there is a subtle issue in the usage of the Transfer Theorem, as ρ_k now tends to 0 when n tends to infinity: all constants in the proof of the theorem now depend on n. Every step of the proof should therefore be revisited carefully.

Though it can certainly be done to establish a range for $k := k(n)$ where our statement can be extended, it is beyond the scope of this article to do such an involved analysis, and we leave this question open.

4.4 Towards Stirling's approximation

We can get a sense of the impact of such a dependency by considering the case $k = n$. In this case our convergence formulas do not hold because the error terms are no longer negligible. That said, it is close in this case: purely formal manipulations do reconcile the limit of the tree asymptotics and Stirling's formula, in some sense, up to a constant factor.

Indeed, our analysis of $\mathcal{P}$ has brought together two classic asymptotic facts. The asymptotic growth of each $\mathcal{P}_k$ is of the form $p_k(n) \sim \gamma\rho^{-n}n^{-3/2}$ for some real valued ρ and γ. (Note that although $\mathcal{P}_k$ is not a simple variety of trees, the asymptotic behavior of $p_k(n)$ is of the same form as for such families.) But for the full class $\mathcal{P}$, the classical Stirling's approximation of $n!$ gives $p(n) \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$. Subtle analysis is required to reconcile these two estimates, and our upper bound on the asymptotic estimate of $p_k(n)$ allows us to take a first step in this direction.

For any n, the strong interval tree of a permutation of size n contains no prime node of arity larger than n. Thus, if $k \geqslant n$, all of the trees corresponding to permutations of size n are contained in $\mathcal{P}_k$, and hence $p_k(n) = n!$ for $k \geqslant n$. Now, forget for a moment that the estimates for $p_k(n)$ as $n \to \infty$ is valid only for fixed k, and consider the expression in (9) with $k = n$. It simplifies as follows:

$$\frac{1}{(1-\frac{e}{n})^2}\sqrt{\frac{e}{4n\pi}}\left(\frac{n}{e}\right)^n\left(1+\frac{5}{2}\frac{\log n}{n}+\mathcal{O}\left(\frac{1}{n}\right)\right)^n n^{-3/2} = \sqrt{\frac{e}{4\pi}}\left(\frac{n}{e}\right)^n\sqrt{n}\cdot(1+o(1)).$$

This is a constant times Stirling's formula (the constant being $\sqrt{\frac{e}{8\pi^2}}$).

The formula falls apart disastrously for $k > n$, since $p_k(n) = n!$ in this case, but this is not accounted for in the formula, and the quantity in (9) gains an unwanted factor of 2^n. This does not contradict the correctness of our asymptotic form for appropriate k.

4.5 Asymptotic parameter analysis

The average shape of general strong interval trees was described in [7]. This study is essentially based on Equation (6), which shows that simple permutations make up about 1/9 of all permutations. As a consequence, general strong interval trees have a very flat shape (i.e. height 1 or 2) with probability tending to 1, and this shape governs the average case behavior of any tree parameter. However, the prime-degree restricted trees are much richer in this regards, and the asymptotic formulas for parameters are given in Table 1.

We focus here on some parameters which are related to the average case complexity analysis of perfect sorting scenarios for σ, that is, to parsimonious evolutionary scenarios in the model of perfect sorting by reversals (see [7] for a detailed explanation of this connection). We will be specifically interested in the number of internal nodes (which is

Asymptotic number of trees with n leaves	$\sqrt{\frac{\rho_k}{2\pi\Lambda_k''(\tau_k)}}(1-\tau_k)^{-2}\cdot\rho_k^{-n}\,n^{-3/2}$
The average number of internal nodes	$(\tau_k-\rho_k)\rho_k^{-1}\,n$
The average number of prime nodes	$S_k(\frac{\tau_k}{1-\tau_k})\rho_k^{-1}\,n$
The average subtree size sum	$\sqrt{\frac{\pi}{2\Lambda_k''(\tau_k)}}\,n^{3/2}$

Table 3: A summary of asymptotic behavior for trees in $\mathcal{P}_k$.

related to the number of reversals in a scenario), the number of prime nodes (since the complexity of computing a parsimonious scenario depends on it) and the average subtree sum size (which has a tight connection to the average reversal size). These parameters give important insight into the average case analysis of perfect sorting by reversals.

We have seen above that the generating function of $\mathcal{U}_k$ satisfies $U_k(z) = z + \Lambda_k(U_k(z))$ with $\Lambda_k(x) = \frac{x^2}{1-x} + \sum_{j=4}^{k} s_j \left(\frac{x}{1-x}\right)^j$. Consequently, $\mathcal{U}_k$ is a simple variety of trees, and this allows to apply directly the results of Section 2 for the average number of internal nodes or the average subtree size sum in $\mathcal{U}_k$ trees. The average number of prime nodes in $\mathcal{U}_k$ trees can also be derived using the general framework developed in Section 2. Then, the behavior of these parameters in $\mathcal{P}_k$ trees is deduced from the already observed identity

$$\mathcal{P}_k = \mathcal{U}_k + \circ \times \mathrm{Seq}_{\geqslant 2}\mathcal{U}_k. \tag{11}$$

Here $\circ$ is an element of size 0. We track the number of $\circ$s in a tree using an auxiliary variable and a bivariate generating function. Note that even though $\mathcal{P}_k$ is not a simple variety of trees, the behavior of the studied parameters are of the same order as in such families of trees.

The results proved in this section are summarized in Table 3.

4.5.1 Number of internal nodes

Let $U_k^{(k)}(z,y)$ (resp. $P_k(z,y)$) be the bivariate generating function of $\mathcal{U}_k$ trees (resp. $\mathcal{P}_k$ trees), where z counts the size (*i.e.*, the number of leaves) and y counts the number of internal nodes ($\circ$). It follows from Equation (11) that

$$P_k(z,y) = U_k(z,y) + y\cdot\frac{U_k(z,y)^2}{1-U_k(z,y)}.$$

Consequently, we have

$$\begin{aligned}\frac{\partial}{\partial y}P_k(z,y)\Big|_{y=1} = \quad & \frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1} + \frac{2U_k(z,1)}{1-U_k(z,1)}\frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1} \\ & + \frac{U_k(z,1)^2}{1-U_k(z,1)} + \frac{U_k(z,1)^2}{(1-U_k(z,1))^2}\frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1}.\end{aligned} \tag{12}$$

Near ρ_k, we know that $U_k(z,1) = U_k(z) = \tau_k - \beta_k\sqrt{1-z/\rho_k} + \mathcal{O}(1-z/\rho_k)$. The weaker estimate $U_k(z) = \tau_k + \mathcal{O}(\sqrt{1-z/\rho_k})$ gives $\frac{1}{1-U_k(z)} = \frac{1}{1-\tau_k} + \mathcal{O}(\sqrt{1-z/\rho_k})$, and these are enough to estimate all rational fractions in $U_k(z)$ that appear in Equation (12). Moreover, the generating function $\frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1}$ counts $\mathcal{U}_k$ trees weighted by their number of internal nodes. As seen in Section 2.2.2, it then follows from Lemma 2 that, near ρ_k,

$$\frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1} = \frac{\beta_k(\tau_k-\rho_k)}{2\rho_k\sqrt{1-z/\rho_k}} + o\left(\frac{1}{\sqrt{1-z/\rho_k}}\right).$$

Combining these asymptotic estimates gives, near ρ_k,

$$\frac{\partial}{\partial y}P_k(z,y)\Big|_{y=1} = \frac{\beta_k(\tau_k-\rho_k)}{2\rho_k(1-\tau_k)^2}\frac{1}{\sqrt{1-z/\rho_k}} + o\left(\frac{1}{\sqrt{1-z/\rho_k}}\right).$$

Recalling the identity $\gamma_k = \frac{\beta_k}{2\sqrt{\pi}}$ and the asymptotic behavior of $[z^n]P_k(z,1) = p_k(n)$ given in Theorem 5, we deduce that the average number of internal nodes in $\mathcal{P}_k$ trees is

$$\frac{[z^n]\frac{\partial}{\partial y}P_k(z,y)\Big|_{y=1}}{[z^n]P_k(z,1)} \sim_{n\to\infty} \frac{\beta_k(\tau_k-\rho_k)}{2\rho_k(1-\tau_k)^2\sqrt{\pi n}}\rho_k^{-n}\cdot\frac{(1-\tau_k)^2}{\gamma_k\rho_k^{-n}}n^{3/2} = \frac{(\tau_k-\rho_k)}{\rho_k}\cdot n.$$

4.5.2 Number of prime nodes

Like before, let us denote by $U_k(z,y)$ (resp. $P_k(z,y)$) the bivariate generating function of $\mathcal{U}_k$ trees (resp. $\mathcal{P}_k$ trees), counted by size (for z) and number of prime nodes (for y). We know an asymptotic estimate of $U_k(z,1) = U_k(z)$ near ρ_k, and we now apply the method of Section 2 to compute one for $\frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1}$.

For any $\mathcal{U}_k$ tree t, let $\sigma(t) = \xi(t)$ denote the number of prime nodes in t and let $\eta(t)$ be 1 if the root of t is a prime node, 0 otherwise. With the notation of Lemma 2, we have

$$H(z) = \sum_{t\in\mathcal{U}_k}\eta(t)z^{|t|} = S_k\left(\frac{U_k(z)}{1-U_k(z)}\right)$$

$$\text{and}\quad \frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1} = \Xi(z) = H(z)\cdot\frac{\partial}{\partial z}U_k(z) = S_k\left(\frac{U_k(z)}{1-U_k(z)}\right)\cdot\frac{\partial}{\partial z}U_k(z),$$

where $S_k(u) = \sum_{j=4}^k s_j u^j$ as before.

The asymptotic estimate of $U_k(z)$ near ρ_k is $U_k(z) = \tau_k - \beta_k\sqrt{1-z/\rho_k} + \mathcal{O}(1-z/\rho_k)$, from which we deduce that $U_k(z)^j = \tau_k^j + o(1)$. Moreover, singular differentiation gives, near ρ_k,

$$\frac{\partial}{\partial z}U_k(z) = \frac{\beta_k}{2\rho_k\sqrt{1-z/\rho_k}} + o\left(\frac{1}{\sqrt{1-z/\rho_k}}\right).$$

Consequently, we obtain that near ρ_k,

$$\frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1} = S_k\left(\frac{\tau_k}{1-\tau_k}\right)\cdot\frac{\beta_k}{2\rho_k}\cdot\frac{1}{\sqrt{1-z/\rho_k}} + o\left(\frac{1}{\sqrt{1-z/\rho_k}}\right).$$

Now turning to $\mathcal{P}_k$ trees, Equation (11) implies that

$$P_k(z,y) = U_k(z,y) + \frac{U_k(z,y)^2}{1-U_k(z,y)} = \frac{U_k(z,y)}{1-U_k(z,y)}.$$

Differentiation gives

$$\frac{\partial}{\partial y}P_k(z,y)\Big|_{y=1} = \left(\frac{1}{(1-U_k(z,1))^2}\right)\frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1}.$$

The asymptotic estimates obtained above give that, near ρ_k,

$$\frac{\partial}{\partial y}P_k(z,y)\Big|_{y=1} = S_k\left(\frac{\tau_k}{1-\tau_k}\right)\cdot\frac{\beta_k}{2\rho_k(1-\tau_k)^2}\cdot\frac{1}{\sqrt{1-z/\rho_k}} + o\left(\frac{1}{\sqrt{1-z/\rho_k}}\right).$$

Finally, we deduce that the average number of prime nodes in $\mathcal{P}_k$ trees is

$$\frac{[z^n]\frac{\partial}{\partial y}P_k(z,y)\Big|_{y=1}}{[z^n]P_k(z,1)} \sim_{n\to\infty} \frac{S_k\left(\frac{\tau_k}{1-\tau_k}\right)\cdot\beta_k}{2\rho_k(1-\tau_k)^2\sqrt{\pi n}}\rho_k^{-n}\cdot\frac{(1-\tau_k)^2}{\gamma_k\rho_k^{-n}}n^{3/2} = \frac{S_k\left(\frac{\tau_k}{1-\tau_k}\right)}{\rho_k}\cdot n.$$

4.5.3 Subtree size sum

Again, we denote by $U_k(z,y)$ (resp. $P_k(z,y)$) the bivariate generating function of $\mathcal{U}_k$ trees (resp. $\mathcal{P}_k$ trees), counted by size (for z) and subtree size sum (for y). In this case, Equation (11) gives

$$P_k(z,y) = U_k(z,y) + \frac{U_k(zy,y)^2}{1-U_k(zy,y)}.$$

As before, we have $U_k(z,1) = U_k(z)$. Note also that $\frac{\partial}{\partial z}U_k(z,y)\Big|_{y=1} = \frac{\partial}{\partial z}U_k(z)$. It follows that

$$\begin{aligned}\frac{\partial}{\partial y}P_k(z,y)\Big|_{y=1} =& \frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1}\\ &+\left(\frac{2U_k(z,1)}{1-U_k(z,1)} + \frac{U_k(z,1)^2}{(1-U_k(z,1))^2}\right)\left(z\frac{\partial}{\partial z}U_k(z) + \frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1}\right)\end{aligned}$$

and we proceed like in the previous cases. Near ρ_k, the asymptotic estimate of $\frac{\partial}{\partial z}U_k(z)$ is

$$\frac{\partial}{\partial z}U_k(z) = \frac{\beta_k}{2\rho_k\sqrt{1-z/\rho_k}} + o\left(\frac{1}{\sqrt{1-z/\rho_k}}\right),$$

and we have seen in Section 2 that

$$\frac{\partial}{\partial y}U_k(z,y)\Big|_{y=1} = \frac{\beta_k^2}{4\rho_k(1-z/\rho_k)} + o\left(\frac{1}{1-z/\rho_k}\right),$$

since this function counts $\mathcal{U}_k$ trees weighted by their subtree size sum. Consequently, the asymptotic estimate of $\frac{\partial}{\partial y}P_k(z,y)\Big|_{y=1}$ near ρ_k is

$$\frac{\partial}{\partial y}P_k(z,y)\Big|_{y=1} = \frac{\beta_k^2}{4\rho_k(1-\tau_k)^2(1-z/\rho_k)} + o\left(\frac{1}{1-z/\rho_k}\right).$$

We conclude that the average value of the subtree size sum in $\mathcal{P}_k$ trees is

$$\frac{[z^n]\frac{\partial}{\partial y}P_k(z,y)\Big|_{y=1}}{[z^n]P_k(z,1)} \sim_{n\to\infty} \frac{\beta_k^2}{4\rho_k(1-\tau_k)^2}\rho_k^{-n}\cdot\frac{(1-\tau_k)^2}{\gamma_k\rho_k^{-n}}n^{3/2} = \frac{\beta_k^2}{4\rho_k\gamma_k}\cdot n^{3/2}.$$

4.6 Random generation

Equation (7) gives immediate access to random sampling of trees in $\mathcal{P}_k$. Thinking of the classes $\mathcal{P}_k$ as possible models for the biological data collected in [11], it is interesting to generate random trees in $\mathcal{P}_k$, to compare them with the trees obtained from the data. In this context, our interest is the global shape of the trees, and not the particulars of the internal nodes. It is straightforward to produce a random generator which generates trees in $\mathcal{P}_k$ of size approximately 10000 for $k \leqslant 800$ up to generating the simple permutation labels (prime and linear nodes are however distinguished).

Figure 2 illustrates a tree of size 300 drawn uniformly at random from $\mathcal{P}_6$ using Maple's combstruct package. The white nodes are prime, the red are linear with $\oplus$ sign, and the blue linear with $\ominus$ sign. The average size of a subtree is 14.2, which is close to the expected average of approximately 13.99. It has 34 prime nodes, which is close to the expected number of approximately 34.7.

One of our long term goals on the biological side is to identify the very specific traits which arise in permutations which encode mammalian genome comparisons, and to provide more adequate models. Chauve, McCloskey and Mishna [11] have taken some preliminary steps in this direction, and a reasonable model should use $\mathcal{P}_k$ trees as subtrees.

4.7 Estimates and bounds on ρ_k and τ_k

In this section we prove the technical bounds that allow us to prove the upper bound for the asymptotic estimate of $p_k(n)$ given in Corollary 6. In particular, this section gives estimates on τ_k, ρ_k and $\Lambda_k(\tau_k)$ as functions of k. The first ingredient is a more explicit bound for s_n, the number of simple permutations of size n.

Lemma 7 (Bound for s_n)**.** *For every $n \geqslant 4$, $s_n \leqslant \sqrt{2\pi}\, n^{n+1/2}\, e^{-n-2}$.*

Figure 2: A tree from $\mathcal{P}_6$ generated uniformly at random

Proof. This inequality follows from Equation (6), stating that $s_n \leqslant \frac{n!}{e^2}\left(1-\frac{4}{n}+\frac{2}{n(n-1)}\right)$, and the following upper bound on $n!$: $n! \leqslant \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}\,e^{\frac{1}{12n}}$. Combining these two, our claim will follow if we prove that $(1-\frac{4}{n}+\frac{2}{n(n-1)})e^{\frac{1}{12n}} \leqslant 1$ for $n \geqslant 4$. This is equivalent to $(1-\frac{4}{n}+\frac{2}{n(n-1)}) \leqslant e^{-\frac{1}{12n}}$. And since $1-x \leqslant e^{-x}$, it is sufficient to prove that $1-\frac{4}{n}+\frac{2}{n(n-1)} \leqslant 1-\frac{1}{12n}$, *i.e.*, that $4-\frac{2}{n-1} \geqslant \frac{1}{12}$. This obviously holds for $n \geqslant 4$, concluding the proof. □

From this estimate, the derivations of the bounds on τ_k and ρ_k are relatively straightforward, but technical. Working with the value $\tilde{\tau}_k = \frac{\tau_k}{1-\tau_k}$ simplifies the expressions. To derive those bounds, it is essential to keep in mind this sequence of inequalities, which follow from $\tau_k < 1$ and $\rho_k = \tau_k - \Lambda_k(\tau_k)$:

$$0 < \rho_k < \tau_k < \tilde{\tau}_k < 1.$$

Proposition 8 (Bounds for $\tilde{\boldsymbol{\tau}}_{\boldsymbol{k}}$)**.** *For any $\alpha < \frac{e-2}{e-1}$, there exists $k(\alpha)$ such that for $k > k(\alpha)$*

$$\left(\frac{\alpha}{ks_k}\right)^{\frac{1}{k-1}} < \tilde{\tau}_k < \left(\frac{1}{ks_k}\right)^{\frac{1}{k-1}}.$$

Consequently,

$$\frac{e}{k}\left(\frac{\alpha e^3}{\sqrt{2\pi}\,k^{5/2}}\right)^{\frac{1}{k-1}} < \tilde{\tau}_k < \frac{e}{k}\left(\frac{e^3}{\sqrt{2\pi}\,k^{3/2}(k-4)}\right)^{\frac{1}{k-1}} < \frac{e}{k}.$$

Computational evidence suggests that $k(\alpha) = 4$, for all α near $\frac{e-2}{e-1}$.

Proof. The starting point is the equation $\Lambda_k'(x) = 1$, satisfied by τ_k. Because $\Lambda_k(x) = \frac{x^2}{1-x} + \sum_{j=4}^k s_j (\frac{x}{1-x})^j$, it is convenient to consider this equation under the change of variables $y = \frac{x}{1-x}$, *i.e.*, $x = \frac{y}{1+y}$. Notice that it implies $\frac{1}{(1-x)^2} = (1+y)^2$.

Derivation gives $\Lambda_k'(x) = \frac{1}{(1-x)^2} - 1 + \frac{1}{(1-x)^2} \sum_{j=4}^k j s_j (\frac{x}{1-x})^{j-1}$, so that the equation $\Lambda_k'(x) = 1$ can be rewritten as

$$(1+y)^2 - 1 + (1+y)^2 \sum_{j=4}^k j s_j y^{j-1} = 1 \quad \text{which implies} \quad \frac{2-(1+y)^2}{(1+y)^2} = \sum_{j=4}^k j s_j y^{j-1}. \tag{13}$$

The next step towards proving the stated inequalities is the fact that for $0 < y < 1$, $1 - 5y < \frac{2-(1+y)^2}{(1+y)^2} < 1$, which is immediately proved by simple manipulations of inequalities. Indeed, we now observe that (by definition of $\tilde{\tau}_k$), Equation (13) is satisfied at $y = \tilde{\tau}_k$. Consequently, these inequalities yield an upper and a lower bound for $\sum_{j=4}^k j s_j \tilde{\tau}_k^{j-1}$:

$$1 - 5\tilde{\tau}_k < \sum_{j=4}^k j s_j \tilde{\tau}_k^{j-1} < 1. \tag{14}$$

From Equation (14), we get $k s_k \tilde{\tau}_k^{k-1} \leqslant \sum_{j=4}^k j s_j \tilde{\tau}_k^{j-1} < 1$, from which the upper bound $\tilde{\tau}_k < \left(\frac{1}{k s_k}\right)^{\frac{1}{k-1}}$ follows. From there, deriving $\tilde{\tau}_k < \frac{e}{k} \left(\frac{e^3}{\sqrt{2\pi}\, k^{3/2}(k-4)}\right)^{\frac{1}{k-1}}$ is then a routine exercise using $\frac{k!}{e^2}\left(1 - \frac{4}{k}\right) \leqslant s_k$ (see Equation (6)) and Stirling's inequality $\left(\frac{k}{e}\right)^k \sqrt{2\pi k} \leqslant k!$. This quantity is no larger than $\frac{e}{k}$ as soon as $k \geqslant 5$. This concludes the part of the proof about upper bounds.

For the lower bounds, we start again from Equation (14) above. We use the inequality $1 - 5\tilde{\tau}_k - \sum_{j=4}^{k-1} j s_j \tilde{\tau}_k^{j-1} < k s_k \tilde{\tau}_k^{k-1}$, and combine it with the bound $0 < \tilde{\tau}_k \leqslant e/k = o(1)$, and an upper bound on $\sum_{j=4}^{k-1} j s_j \tilde{\tau}_k^{j-1}$ obtained below. We split this sum as

$$\sum_{j=4}^{k-1} j\, s_j\, \tilde{\tau}_k^{j-1} = \underbrace{\sum_{j=4}^{k-\iota_k-1} j\, s_j\, \tilde{\tau}_k^{j-1}}_{A(k)} + \underbrace{\sum_{j=k-\iota_k}^{k-1} j\, s_j\, \tilde{\tau}_k^{j-1}}_{B(k)}.$$

where $\iota_k = \lfloor k^{\frac{1}{3}} \rfloor$. Note that ι_k a non-decreasing integer function of k that tends to infinity and such that $\iota_k = o(\sqrt{k})$. Lemmas 9 and 10 below prove that

$$A(k) = \sum_{j=4}^{k-\iota_k-1} j s_j \tilde{\tau}_k^{j-1} = \mathcal{O}\left(\frac{1}{k^3}\right) \qquad \text{and that} \qquad B(k) = \sum_{k-\iota_k}^{k-1} j s_j \tilde{\tau}_k^{j-1} = \frac{1}{e-1} + o(1).$$

It follows that

$$k s_k \tilde{\tau}_k^{k-1} > 1 - 5\tilde{\tau}_k - \sum_{j=4}^{k-1} j s_j \tilde{\tau}_k^{j-1} = 1 - \frac{1}{e-1} + o(1) = \frac{e-2}{e-1} + o(1).$$

Hence for any $\alpha < \frac{e-2}{e-1}$, there exists $k(\alpha)$ such that for any $k \geqslant k(\alpha)$, we have $k\, s_k\, \tilde{\tau}_k^{k-1} > \alpha$, and therefore $\tilde{\tau}_k > \left(\frac{\alpha}{k\, s_k}\right)^{\frac{1}{k-1}}$. To conclude the proof, we plug in the upper bound on s_k from Lemma 7. □

Lemma 9. *The quantity* $A(k) = \sum_{j=4}^{k-\iota_k-1} j s_j \tilde{\tau}_k^{j-1}$ *defined in the proof of Proposition 8 satisfies* $A(k) = \mathcal{O}\left(\frac{1}{k^3}\right)$.

Proof. It is convenient to define $b_j = j\, s_j$. For any $k \geqslant 5$, since $\tilde{\tau}_k < e/k$, $a_j = b_j\, e^{j-1}\, k^{1-j}$ is an upper bound on $j\, s_j\, \tilde{\tau}_k^{j-1}$, so that $A(k) \leqslant \sum_{j=4}^{k-\iota_k-1} a_j$. In what follows, we prove that $\sum_{j=4}^{k-\iota_k-1} a_j = \mathcal{O}\left(\frac{1}{k^3}\right)$ (which is enough to conclude, since $A(k) > 0$).

We claim that for some integer j_0, the sequence $(b_j)_{j \geqslant j_0}$ is log-convex. Indeed, for any $j \geqslant 6$, $\frac{b_j^2}{b_{j-1} b_{j+1}} = \frac{j^2}{(j-1)(j+1)} \frac{s_j^2}{s_{j+1} s_{j-1}}$, and Equation (6) then gives

$$\frac{b_j^2}{b_{j-1} b_{j+1}} \leqslant \frac{j^2}{(j-1)(j+1)} \frac{j!^2 \left(1 - \frac{4}{j} + \frac{2}{j(j-1)}\right)^2}{(j-1)!(j+1)! \left(1 - \frac{4}{j+1}\right)\left(1 - \frac{4}{j-1}\right)} = 1 - \frac{1}{j} + \mathcal{O}\left(\frac{1}{j^2}\right).$$

In particular, there exists an integer $j_0 \geqslant 6$ such that for any $j \geqslant j_0$, $\frac{b_j^2}{b_{j-1} b_{j+1}} < 1$ and therefore the sequence $(b_j)_{j \geqslant j_0}$ is log-convex.

Now, note that for any k, $(a_j)_{j \geqslant j_0}$ is also log-convex, since $\frac{a_j^2}{a_{j-1} a_{j+1}} = \frac{b_j^2}{b_{j-1} b_{j+1}}$ for all j. The reason for considering the sequence (b_j) instead of (a_j) in the first place is to ensure that j_0 does not depend on k, although the definition of $a_j = j\, s_j\, e^{j-1}\, k^{1-j}$ depends on k.

Log-convex sequences are decreasing down to a given minimum then increasing, and therefore are bounded from above by the values reached at the extremities. Thus for all $j \in \{j_0, \dots, k - \iota_k - 1\}, a_j \leqslant \max\{a_{j_0}, a_{k-\iota_k-1}\} \leqslant a_{j_0} + a_{k-\iota_k-1}$. Consequently,

$$\sum_{j=4}^{k-\iota_k-1} a_j = \sum_{j=4}^{j_0-1} a_j + \sum_{j=j_0}^{k-\iota_k-1} a_j \leqslant \sum_{j=4}^{j_0-1} a_j + k\, a_{j_0} + k\, a_{k-\iota_k-1},$$

and the result will follow if we find adequate upper bounds on each on these three terms, which we now do.

For any $j \in \{4, \dots j_0 - 1\}$, we have $a_j = j\, s_j\, e^{j-1}\, k^{1-j} \leqslant j j!\, e^{j-1}\, k^{1-j} \leqslant j_0 j_0!\, e^{j_0-1}\, k^{-3}$ so that $\sum_{j=4}^{j_0-1} a_j \leqslant j_0^2 j_0!\, e^{j_0-1}\, k^{-3} = \mathcal{O}(k^{-3})$.

For the term $k\, a_{j_0}$, we have $k\, a_{j_0} = j_0\, s_{j_0}\, e^{j_0-1}\, k^{2-j_0} = \mathcal{O}(k^{-3})$ since $j_0 \geqslant 6$. Using Lemma 7 and the fact that for all $x \in (0,1)$, $\log(1-x) < -x$, we obtain the bound for the last term. More precisely, we have:

$$\begin{aligned} k\, a_{k-\iota_k-1} &\leqslant k^2 \cdot s_{k-\iota_k-1} \cdot e^{k-\iota_k-2} \cdot k^{2-k+\iota_k} \\ &\leqslant k^2 \cdot \sqrt{2\pi} \cdot (k - \iota_k - 1)^{k-\iota_k-1/2} \cdot e^{-k+\iota_k-1} \cdot e^{k-\iota_k-2} \cdot k^{2-k+\iota_k} \\ &\leqslant \frac{\sqrt{2\pi}}{e^3} k^{7/2} \cdot \left(1 - \frac{\iota_k + 1}{k}\right)^{k-\iota_k-1/2} \end{aligned}$$

$$\leqslant \frac{\sqrt{2\pi}}{e^3} k^{7/2} \cdot \exp\left((k - \iota_k - 1/2) \log\left(1 - \frac{\iota_k + 1}{k}\right)\right)$$
$$\leqslant \frac{\sqrt{2\pi}}{e^3} k^{7/2} \cdot \exp\left(-\frac{(k - \iota_k - 1/2)(\iota_k + 1)}{k}\right).$$

The quantity in the exponential is asymptotically equivalent to $-\iota_k = -\lfloor k^{\frac{1}{3}} \rfloor$. Hence $k\, a_{k-\iota_k-1}$ decreases super-polynomially fast toward 0, and is therefore a $\mathcal{O}(k^{-3})$ too. □

Lemma 10. *The quantity $B(k) = \sum_{k-\iota_k}^{k-1} j s_j \tilde{\tau}_k^{j-1}$ defined in the proof of Proposition 8 satisfies $B(k) = \frac{1}{e-1} + o(1)$.*

Proof. With the change of variable $i = k - j$, we can write $B(k) = \sum_{i=1}^{\iota_k} (k-i)\, s_{k-i}\, \tilde{\tau}_k^{k-i-1}$. By Lemma 7 and the upper bound on $\tilde{\tau}_k$ proved in Proposition 8, we have

$$(k-i)\, s_{k-i} \leqslant \frac{\sqrt{2\pi}(k-i)^{k-i+3/2}}{e^{k-i+2}},$$
$$\tilde{\tau}_k^{k-i-1} \leqslant \frac{e^{k-i-1}}{k^{k-i-1}} \left(\frac{e^3}{\sqrt{2\pi}\, k^{3/2}\, (k-4)}\right) \cdot \left(\frac{e^3}{\sqrt{2\pi}\, k^{3/2}\, (k-4)}\right)^{\frac{-i}{k-1}}.$$

Therefore $(k-i)\, s_{k-i}\, \tilde{\tau}_k^{k-i-1} \leqslant \left(1 - \frac{i}{k}\right)^{k-i+3/2} \left(e^{-3}\sqrt{2\pi}\, k^{3/2}\, (k-4)\right)^{\frac{i}{k-1}} \cdot \frac{1}{1-\frac{4}{k}}$. Since $i \leqslant \iota_k$ and $e^{-3}\sqrt{2\pi}\, k^{3/2}\, (k-4) \geqslant 1$ as soon as $k \geqslant 5$, we obtain that for $k \geqslant 5$,

$$(k-i)\, s_{k-i}\, \tilde{\tau}_k^{k-i-1} \leqslant \left(1 - \frac{i}{k}\right)^{k-i+3/2} \left(e^{-3}\sqrt{2\pi}\, k^{3/2}\, (k-4)\right)^{\frac{\iota_k}{k-1}} \cdot \underbrace{\frac{1}{1-\frac{4}{k}}}_{1+\mathcal{O}(\frac{1}{k})}.$$

Using again that $\log(1-x) < -x$ for $x \in (0,1)$, we have

$$(1 - i/k)^{k-i+3/2} = e^{(k-i+3/2)\log\left(1-\frac{i}{k}\right)} \leqslant e^{-\frac{(k-i+3/2)i}{k}} = e^{-i+\frac{i^2}{k}-\frac{3i}{2k}}.$$

Recalling that $i \leqslant \iota_k = \lfloor k^{\frac{1}{3}} \rfloor$, this gives $(1 - i/k)^{k-i+3/2} \leqslant e^{-i} \exp(k^{-1/3}) = e^{-i}\,(1 + o(1))$. Proceeding similarly, the middle term satisfies

$$\left(e^{-3}\sqrt{2\pi}\, k^{3/2}\, (k-4)\right)^{\frac{\iota_k}{k-1}} = 1 + o(1).$$

Therefore, we obtain $(k-i)\, s_{k-i}\, \tilde{\tau}_k^{k-i-1} \leqslant e^{-i}\,(1 + o(1))$, where the function hidden in the $o(1)$ notation depends on k but not on i. Consequently, summing over i, we obtain

$$B(k) \leqslant \left(\sum_{i=1}^{\iota_k} e^{-i}\right)(1 + o(1)) \leqslant \left(\sum_{i=1}^{\infty} e^{-i}\right)(1 + o(1)) = \frac{1 + o(1)}{e-1},$$

as claimed. □

Theorem 11 (Bounds for $\boldsymbol{\rho_k}$)**.** *There exists a constant β such that for any $\alpha < \frac{e-2}{e-1}$, there exist $k(\alpha,\beta)$ such that for any $k \geqslant k(\alpha,\beta)$,*

$$\frac{e}{k}\left(\frac{\alpha e^3}{\sqrt{2\pi}\,k^{5/2}}\right)^{\frac{1}{k-1}}\left(1-\frac{\beta}{k}\right) < \rho_k < \frac{e}{k}\left(\frac{e^3}{\sqrt{2\pi}\,k^{3/2}(k-4)}\right)^{\frac{1}{k-1}}.$$

Consequently, $\rho_k = \frac{e}{k}\left(1-\frac{5}{2}\frac{\log k}{k}+\mathcal{O}(\frac{1}{k})\right)$.

Proof. The upper bound is immediate from the bound $\rho_k < \tilde{\tau}_k$ and Proposition 8.
For the lower bound, we start from $\rho_k = \tau_k - \Lambda_k(\tau_k)$. The definitions of $\tilde{\tau}_k$ and Λ_k give $\rho_k = \tilde{\tau}_k\left(1-\frac{2\tilde{\tau}_k}{1+\tilde{\tau}_k}-\sum_{j=4}^{k} s_j\tilde{\tau}_k^{j-1}\right)$. Our main step is to deduce from this equality that $\rho_k \geqslant \tilde{\tau}_k(1-\beta/k)$ for some constant β. The lower bound will then follow from Proposition 8.

As in the proof of Proposition 8, we leverage upper bounds on $\tilde{\tau}_k$ to build a lower bound on $1-\frac{2\tilde{\tau}_k}{1+\tilde{\tau}_k}-\sum_{j=4}^{k} s_j\tilde{\tau}_k^{j-1}$. In this case, we use $\frac{2\tilde{\tau}_k}{1+\tilde{\tau}_k} \leqslant 2\tilde{\tau}_k \leqslant 2\frac{e}{k}$, and we will bound the summation by splitting the sum at the same place:

$$\sum_{j=4}^{k} s_j\,\tilde{\tau}_k^{j-1} = \sum_{j=4}^{k-\iota_k-1} s_j\,\tilde{\tau}_k^{j-1} + \sum_{j=k-\iota_k}^{k-1} s_j\,\tilde{\tau}_k^{j-1} + s_k\,\tilde{\tau}_k^{k-1}.$$

Even though it is not the same summation, we can re-use the bounds from Lemmas 9 and 10. Indeed,

$$\sum_{j=4}^{k-\iota_k-1} s_j\,\tilde{\tau}_k^{j-1} \leqslant \sum_{j=4}^{k-\iota_k-1} j\,s_j\,\tilde{\tau}_k^{j-1} = A(k) = \mathcal{O}\left(\frac{1}{k^3}\right)$$

$$\text{and}\quad \sum_{j=k-\iota_k}^{k-1} s_j\,\tilde{\tau}_k^{j-1} \leqslant \sum_{j=k-\iota_k}^{k-1} \frac{j}{k-\iota_k} s_j\,\tilde{\tau}_k^{j-1} = \frac{B(k)}{k-\iota_k} = \mathcal{O}\left(\frac{1}{k}\right).$$

Finally, Proposition 8 ensures that $k\,s_k\,\tilde{\tau}_k^{k-1} \leqslant 1$, and we obtain $\frac{2\tilde{\tau}_k}{1+\tilde{\tau}_k}+\sum_{j=4}^{k} s_j\tilde{\tau}_k^{j-1} = \mathcal{O}\left(\frac{1}{k}\right)$. It follows that for some β, there exists $k(\beta)$ such that when $k \geqslant k(\beta)$ we have:

$$\frac{2\tilde{\tau}_k}{1+\tilde{\tau}_k}+\sum_{j=4}^{k} s_j\tilde{\tau}_k^{j-1} \leqslant \frac{\beta}{k} \quad \text{and hence} \quad \rho_k \geqslant \tilde{\tau}_k\left(1-\frac{\beta}{k}\right),$$

which together with Proposition 8 proves the lower bound.

To obtain the claimed asymptotic estimate of ρ_k, it is enough to observe that both the upper and the lower bound behave like $\frac{e}{k}\left(1-\frac{5}{2}\frac{\log k}{k}+\mathcal{O}(\frac{1}{k})\right)$. More precisely,

$$\begin{aligned}\left(\frac{\alpha e^3}{\sqrt{2\pi}\,k^{5/2}}\right)^{\frac{1}{k-1}} &= \exp\left(\frac{\log(k^{-5/2})+\mathcal{O}(1)}{k-1}\right) = \exp\left(-\frac{5}{2}\frac{\log k}{k-1}+\frac{\mathcal{O}(1)}{k-1}\right)\\ &= 1-\frac{5}{2}\frac{\log k}{k}+\mathcal{O}\left(\frac{1}{k}\right),\end{aligned}$$

So that

$$\left(\frac{\alpha e^3}{\sqrt{2\pi}\, k^{5/2}}\right)^{\frac{1}{k-1}} \left(1-\frac{\beta}{k}\right) = 1 - \frac{5}{2}\frac{\log k}{k} + \mathcal{O}\left(\frac{1}{k}\right)$$

and

$$\begin{aligned}\left(\frac{e^3}{\sqrt{2\pi}\, k^{3/2}(k-4)}\right)^{\frac{1}{k-1}} &= \exp\left(\frac{\log(k^{-5/2}) + \log(\frac{1}{1-4/k}) + \mathcal{O}(1)}{k-1}\right)\\ &= 1 - \frac{5}{2}\frac{\log k}{k} + \mathcal{O}\left(\frac{1}{k}\right).\end{aligned}$$

This concludes the proof. □

It was known in [10] that $\rho_k = \frac{e}{k}(1+o(1))$, but we are able to produce a more precise estimate. We require this precision when we consider the limit as $k \to \infty$.

Looking at the asymptotic estimate of $p_k(n)$ provided by Theorem 5, and aiming at obtaining an upper bound on this estimates, the only missing piece is a lower bound on $\Lambda_k''(\tau_k)$. The definition of Λ_k (see Theorem 5) gives

$$\begin{aligned}\Lambda_k''(x) &= \frac{2}{(1-x)^3}\left(1 + \sum_{j=4}^{k} j s_j \left(\frac{x}{1-x}\right)^{j-1} + \frac{1}{2(1-x)}\sum_{j=4}^{k} j(j-1)s_j\left(\frac{x}{1-x}\right)^{j-2}\right)\\ &\geqslant \frac{2}{(1-x)^3} \text{ for all } x \in (0,1),\end{aligned}$$

and therefore the series expansion of $(1-x)^{-3}$ ensures that $\Lambda_k''(\tau_k) \geqslant 2 + 6\tilde{\tau}_k$. We could expand this expression further, and use lower bounds on $\tilde{\tau}_k$, but it turns out that for our purposes, the bound $\Lambda_k''(\tau_k) \geqslant 2$ is sufficient.

This completes the required set of elements for the bounds of Corollary 6.

5 Studying a combinatorial class *via* its filtration

The study thus far illustrates a strategy to enumerate classes $\mathcal{C}$ of trees whose generating functions satisfy $C(z) = z + \Lambda(C(z))$, in particular in the case where Λ is not analytic.

Specifically, we have considered a sequence of analytic Λ_k such that as formal power series, $\lim_{k\to\infty} \Lambda_k = \Lambda$, and studied first the sets $\mathcal{C}_k$ of Λ_k-trees. One such example is truncations at order k.

The main goal is to obtain results about $\mathcal{C}$ from what is known about the classes $\mathcal{C}_k$. More specific goals are: to determine conditions so that the limit of the asymptotics of the subclasses tends to the asymptotics of the whole class; and to determine which parameter formulas are valid under the limit. A first step towards this is to consider the case when Λ is analytic. We show that in this case, we obtain the correct asymptotic formula when taking the limit as k tends to infinity, *i.e.* that limits in n and k commute.

Consider a series $\Lambda(x) = \sum_{i\geqslant 2} \lambda_i x^i$ with non-negative coefficients, analytic at 0. And for all $k \geqslant 2$, define $\Lambda_k(x) = \sum_{i=2}^{k} \lambda_i x^i$. We denote respectively by $\mathcal{C}$ and $\mathcal{C}_k$ the classes of trees whose generating functions satisfy

$$C(z) = z + \Lambda(C(z)) \qquad \text{and} \qquad C_k(z) = z + \Lambda_k(C_k(z)).$$

Suppose that Λ has radius of convergence R. There is a unique solution $\tau \in (0, R)$ to the equation $\Lambda'(x) = 1$. It follows from Theorem 1 that $C(z)$ is analytic at 0 and has a unique dominant singularity $\rho = \tau - \Lambda(\tau)$. Under the further assumption that $C(z)$ is aperiodic, we conclude that the coefficients of this series behave asymptotically like $[z^n]C(z) \sim \sqrt{\frac{\rho}{2\pi\Lambda''(\tau)}} \cdot \frac{\rho^{-n}}{n^{3/2}}$.

Lemma 12. *For all $k \geqslant 2$, there exists a unique $\tau_k \in (0, +\infty)$ such that $\Lambda_k'(\tau_k) = 1$. Moreover, the sequence $(\tau_k)_{k\geqslant 2}$ is decreasing and converges to τ as k goes to infinity.*

Proof. Fix some $k \geqslant 2$. From the definition of $\Lambda_k(x) = \sum_{i=2}^{k} \lambda_i x^i$, it follows that $\Lambda_k'(x)$ is a polynomial with non-negative coefficients, increasing from 0 to $+\infty$ when x varies from 0 to $+\infty$. Moreover, its derivative Λ_k'' being nowhere zero on $(0, +\infty)$, Λ_k' is strictly increasing. Therefore, there is a unique positive solution to $\Lambda_k'(x) = 1$, that we denote τ_k.

The fact that the sequence $(\tau_k)_{k\geqslant 2}$ is decreasing is immediate from

$$1 = \Lambda_k'(\tau_k) = \sum_{i=2}^{k} i\lambda_i \tau_k^{i-1} \leqslant \sum_{i=2}^{k+1} i\lambda_i \tau_k^{i-1} = \Lambda_{k+1}'(\tau_k)$$

and the fact that Λ_{k+1}' is increasing.

The sequence $(\tau_k)_{k\geqslant 2}$ being decreasing and non-negative, it admits a limit, that we denote ℓ. We want to prove that $\ell = \tau$, *i.e.*, that $\Lambda'(\ell) = 1$. First, for all k, $\ell \leqslant \tau_k$, so that $\Lambda_k'(\ell) \leqslant 1$. Moreover, the sequence $(\Lambda_k'(\ell))_k$ is increasing (we keep adding non-negative terms), and thus converges towards a limit that is no larger than 1. This limit being $\Lambda'(\ell)$, we obtain that $\Lambda'(\ell) \leqslant 1$. Now as Λ' is increasing, and $\Lambda'(\tau) = 1$ it follows that $\ell \leqslant \tau$. As $\tau < R$ and $(\tau_k)_k$ is decreasing, this is sufficient to get that Λ is defined for τ_k, for sufficiently large k. For any such large k, we have

$$1 = \Lambda_k'(\tau_k) = \sum_{i=2}^{k} i\lambda_i \tau_k^{i-1} \leqslant \sum_{i\geqslant 2} i\lambda_i \tau_k^{i-1} = \Lambda'(\tau_k),$$

and taking the limit in k gives $\Lambda'(\ell) \geqslant 1$, by continuity of Λ'. □

We have a similar result for the sequence of radii of convergence.

Lemma 13. *For all $k \geqslant 2$, define $\rho_k = \tau_k - \Lambda_k(\tau_k)$. The sequence $(\rho_k)_{k\geqslant 2}$ converges to ρ as k goes to infinity.*

Proof. It is enough to prove that $(\Lambda_k(\tau_k))_k$ converges to $\Lambda(\tau)$. Like in the previous proof, since $(\tau_k)_{k\geqslant 2}$ is decreasing towards $\tau < R$, we get that for k large enough, Λ is defined in τ_k. For such large k, we have

$$\sum_{j=2}^{k} \lambda_k \tau^k \leqslant \sum_{j=2}^{k} \lambda_k \tau_k^k \leqslant \sum_{j\geqslant 2} \lambda_k \tau_k^k \quad \text{that is to say } \Lambda_k(\tau) \leqslant \Lambda_k(\tau_k) \leqslant \Lambda(\tau_k).$$

Because $(\Lambda_k(\tau))_k$ and $(\Lambda(\tau_k))_k$ share the same limit $\Lambda(\tau)$, by the squeeze theorem, as k goes to infinity, we obtain that $\lim_{k\to+\infty} \Lambda_k(\tau_k) = \Lambda(\tau)$. □

From Theorem 1, we obtain

$$[z^n]C_k(z) \sim \sqrt{\frac{\rho_k}{2\pi\Lambda_k''(\tau_k)}} \cdot \frac{\rho_k^{-n}}{n^{3/2}},$$

and the two lemmas above ensure that taking the limit in k in this estimates give $\sqrt{\frac{\rho}{2\pi\Lambda''(\tau)}} \cdot \frac{\rho^{-n}}{n^{3/2}}$. In addition, $\lim_{k\to+\infty} C_k(z) = C(z)$ from which we get $[z^n] \lim_{k\to+\infty} C_k(z) \sim \sqrt{\frac{\rho}{2\pi\Lambda''(\tau)}} \cdot \frac{\rho^{-n}}{n^{3/2}}$. In other words, taking the limit in k in the estimate of the number of trees of size n in $\mathcal{C}_k$ gives the estimates of the number of trees of size n in $\mathcal{C}$.

Acknowledgements

We are indebted to Cedric Chauve and Carine Pivoteau for guidance and ideas. We thank Rosemary McCloskey wrote the code for the Boltzmann generator, amongst other extremely useful things. MM is particularly grateful to both LIGM and LaBRI for hosting her during the course of this work. We are also thankful to the anonymous referees for their observations and careful verifications.

References

[1] M.H. Albert and M.D. Atkinson. Simple permutations and pattern restricted permutations. *Discrete Math.*, 300:1–15, 2005.

[2] M.H. Albert, M.D. Atkinson, and M. Klazar. The enumeration of simple permutations. *J. Integer Seq.*, 6:03.4.4, 2003.

[3] S. Bérard, A. Bergeron, C. Chauve, and C. Paul. Perfect sorting by reversals is not always difficult. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 4:4–16, 2007.

[4] A. Bergeron, C. Chauve, F. de Montgolfier, and M. Raffinot. Computing common intervals of K permutations, with applications to modular decomposition of graphs. In *Proc. 13th Annual European Symposium on Algorithms*, volume 3669 of *Lecture Notes in Comput. Sci.*, pages 779–790. Springer, 2005.

[5] O. Bodini, D. Gardy, and B. Gittenberger. Lambda-terms of bounded unary height. In *Proceedings of the Eighth Workshop on Analytic Algorithmics and Combinatorics*, pages 23–32. SIAM, 2011.

[6] K. S. Booth and G. S. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *J. Comput. System Sci.*, 13(3):335–379, 1976.

[7] M. Bouvel, C. Chauve, M. Mishna, and D. Rossin. Average-case analysis of perfect sorting by reversals. *Discrete Math. Algorithms Appl.*, 3(3):369–392, 2011.

[8] M. Bouvel, M. Mishna, and C. Nicaud. Some simple varieties of trees arising in permutation analysis. In *Proc. FPSAC 2013*, pages 825–836. DMTCS proceedings, 2013.

[9] B. M. Bui Xuan, M. Habib, and C. Paul. Revisiting Uno and Yagiura's Algorithm. In *Proc. ISAAC 2005* volume 3827 of *Lecture Notes in Comput. Sci.*, pages 146–155. Springer, 2005.

[10] G. Chapuy, A. Pierrot, and D. Rossin. On growth rate of wreath-closed permutation classes. Talk at the conference *Permutation Patterns*, 2011.

[11] C. Chauve, R. McCloskey, and M. Mishna. Personal communication, 2011.

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

[13] I. Gessel. Symmetric functions and P-recursiveness. *J. Combin. Theory Ser. A*, 53(2):257–285, 1990.

[14] S. Heber and J. Stoye. Finding All Common Intervals of k Permutations. In *Proc. CPM 2001* volume 2089 of *Lecture Notes in Comput. Sci.*, pages 207–218. Springer, 2001.

[15] J. Pitman and D. Rizzolo. Schröder's problems and scaling limits of random trees. *Trans. Amer. Math. Soc* 367(1): 6943–6969, 2015.

[16] The On-Line Encyclopedia of Integer Sequences. Published electronically at http://oeis.org.