Séminaire Lotharingien de Combinatoire **68** (2013), Article B68b

UNBALANCED SUBTREES IN BINARY ROOTED ORDERED AND UN-ORDERED TREES

FILIPPO DISANTO*

Abstract. Binary rooted trees, both in the ordered and in the un-ordered case, are well studied structures in the field of combinatorics. The aim of this work is to study particular patterns in these classes of trees. We consider completely unbalanced subtrees, where unbalancing is measured according to the so-called Colless index. The size of the biggest unbalanced subtree becomes then a new parameter with respect to which we find several enumeration formulas.

1. Introduction

The aim of this work is to study particular unbalanced patterns in rooted binary trees, both in the ordered and un-ordered case. More precisely, we consider a new statistic on trees. We are interested in the size of the biggest subtree having the *caterpillar* property.

Caterpillars have already been considered in the case of *coalescent* trees, see for example the interesting work of Rosenberg [4]. In particular, in a genetic population framework, when trees are used to represent ancestry relations among individuals, the presence of a caterpillar subtree often indicates phenomena such as *natural selection*.

Subtree structures have already been considered in [1] and [5]. However, to our knowledge, enumerative properties of caterpillar subtrees have never been investigated. This is what we do here.

In Section 2 we start by giving some basic definitions. We then enumerate ordered rooted binary trees of a given size having the biggest caterpillar subtree of size less than, greater than, or equal to a fixed integer k. Furthermore, we provide the expected value of the size of the biggest caterpillar subtree when ordered trees of size n are uniformly distributed and n is large.

In Section 3 we see how caterpillar subtrees correspond to patterns extracted from 132-avoiding permutations. The resulting characterization is interesting and will represent a starting point for further studies on sub-structures of permutations.

Finally, in Section 4 we study caterpillars realized in un-ordered binary rooted trees. The resulting approach is similar to the one used in the ordered case, and it provides asymptotic formulas for the probability of a tree of a given size with "small" caterpillar subtrees.

2. Caterpillars in ordered rooted binary trees

2.1. **Definitions.** Ordered rooted binary trees are enumerated with respect to the size, i.e., the number of leaves, by the well known sequence of *Catalan numbers*, corresponding to entry $A000108$ in [6]. The generating function for Catalan numbers is denoted

*Email: fdisanto@uni-koeln.de.

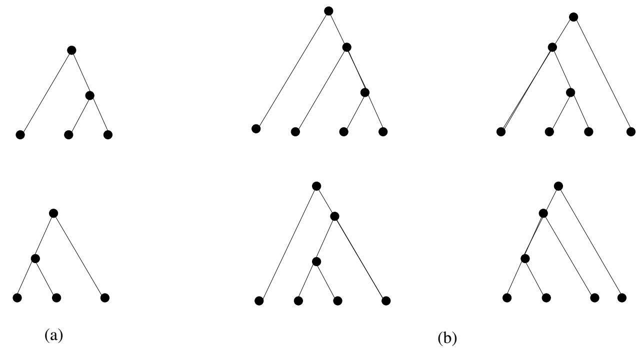

FIGURE 1. (a) caterpillars of size 3; (b) caterpillars of size 4.

by $C(x)$, and it is given explicitly by

$$C(x) = \frac{1-\sqrt{1-4x}}{2}. \tag{1}$$

We denote the class of ordered rooted binary trees by $\mathcal{T}$, while $\mathcal{T}_n$ denotes the subset of $\mathcal{T}$ consisting of those elements which have size n. In all of Section 2, the term "tree" will refer to "ordered binary rooted tree".

We define a tree in $\mathcal{T}_n$ to be a *caterpillar* of size n if each node is a leaf or it has at least one leaf as a direct descendant. For examples see Figure 1(a), (b).

In addition, caterpillars can be characterized by the fact that they are the "most unbalanced" trees. As a measure of tree imbalance, we take the following index. Given a tree t and a node i, let $t_l(i)$ (respectively $t_r(i)$) be the left (respectively right) subtree of t determined by i. We define

$$\Delta_t(i) = |\text{size}(t_l(i)) - \text{size}(t_r(i))|.$$

If $t \in \mathcal{T}_n$, its *Colless* index (see [3]) is defined by

$$\frac{2}{(n-2)(n-1)} \cdot \sum_{i \text{ node of } t} \Delta_t(i).$$

The Colless index is considered as a measure of tree imbalance. Its value ranges between 0 and 1, where 0 corresponds to a completely balanced tree while 1 to an unbalanced one.

Based on the previous definitions, a tree of size $n > 2$ is a caterpillar if and only if its Colless index is 1.

If $t \in \mathcal{T}_n$, we define $\gamma(t)$ as the size of the biggest caterpillar which occurs as a subtree of t. We observe that, if $n > 1$, then $\gamma(t) \geq 2$. In Figure 2 we show a tree with $\gamma = 5$.

2.2. **A recursive construction for the size of the biggest caterpillar subtree.** Let $F_k^-(x)$ be the ordinary generating function which gives the number of trees having γ parameter *at most* equal to $k \geq 2$.

It is easy to see that F_k^- satisfies the equation

$$F_k^- = x + (F_k^-)^2 - 2^{k-1}x^{k+1}. \tag{2}$$

Indeed, a tree t having $\gamma(t) \leq k$ has either size one or it is built by attaching two trees t_1 and t_2, with $\gamma(t_1) \leq k$ and $\gamma(t_2) \leq k$, to the root of t. We must exclude the case in

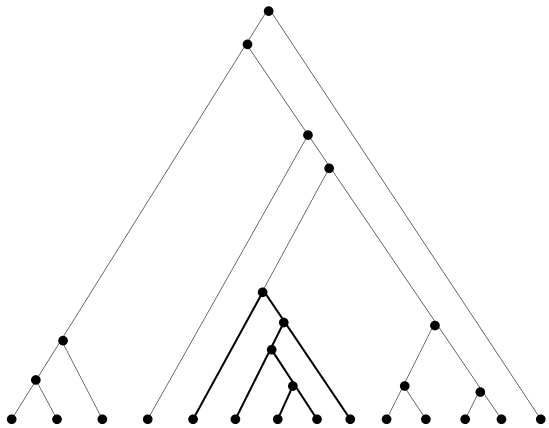

FIGURE 2. A tree with γ parameter equal to 5. The biggest caterpillar is highlighted.

which one among t_1 and t_2 has size 1 and the other one is a caterpillar of size k. Since there are exactly 2^{k-2} caterpillars of size k, the previous formula follows.

From (2) we obtain

$$F_k^-(x) = \frac{1 - \sqrt{1 - 4x + 2^{k+1}x^{k+1}}}{2}. \tag{3}$$

Then, considering $F_k^+ = C(x) - F_{k-1}^-(x)$, one obtains the generating function for the number of trees with $\gamma \geq k$, while, taking $F_k = F_k^-(x) - F_{k-1}^-(x)$, one can compute the generating function for the number of trees with $\gamma = k$. The following table shows the first coefficients of the Taylor expansion of F_k^-, F_k^+ and F_k when $k = 5$.

$k = 5$	1	2	3	4	5	6	7	8	9	10
F_k^-	1	1	2	5	14	26	100	333	1110	3742
F_k^+	0	0	0	0	8	16	48	160	560	1952
F_k	0	0	0	0	8	0	16	64	240	832

Note that the sixth coefficient of F_k is 0. Indeed, as the reader can easily check, for $k = 5$ there is no tree of size $k + 1$ with γ parameter equal to k.

We conclude this section observing that none of the sequences corresponding to F_k^-, F_k^+ and F_k is at the moment present in [6] with other combinatorial interpretations.

2.2.1. *Asymptotic growth.* The function $F_k(x)$ attains its singularities at the solutions of the equation $1 - 4x + 2^{k+1}x^{k+1} = 0$. By *Pringsheim's theorem* (see [2]), we can assume, for our purposes, that the dominant singularity of $F_k(x)$ corresponds to the positive real solution of $1 - 4x + 2^{k+1}x^{k+1} = 0$ which is closer to the origin. Let ρ_k be this solution. We observe that, when k increases, ρ_k approaches $1/4$. In order to prove this claim, we remark that, for $k \geq 2$, we have

$$\frac{1}{4} < \rho_k < \frac{2}{5}. \tag{4}$$

Indeed, this can be shown by considering the polynomial

$$y = 1 - 4x + 2^{k+1}x^{k+1},$$

which satisfies $y(1/4) > 0$, $y(2/5) < 0$, and is such that $y \geq 1-4x > 0$ for $0 \leq x < 1/4$. We now proceed by *bootstrapping* (see [2]). Writing the defining equation for ρ_k as

$$x = \frac{1}{4}(1 + 2^{k+1}x^{k+1}),$$

and making use of (4), we arrive at

$$\frac{1}{4}\left(1 + \frac{1}{2^{k+1}}\right) < \rho_k < \frac{1}{4}\left(1 + \left(\frac{4}{5}\right)^{k+1}\right),$$

which is sufficient to prove that $\rho_k \to 1/4$.

A further iteration of the previous inequality shows that

$$\rho_k < \frac{1}{4}\left(1 + 2^{k+1}\left(\frac{1}{4}\left(1 + \left(\frac{4}{5}\right)^{k+1}\right)\right)^{k+1}\right),$$

which, in view of

$$\left(1 + \left(\frac{4}{5}\right)^{k+1}\right)^{k+1} = 1 + (k+1)\left(\frac{4}{5}\right)^{k+1} + \mathcal{O}\left(k^2\left(\frac{4}{5}\right)^{2k}\right),$$

yields

$$\rho_k < \frac{1}{4}\left(1 + \frac{1}{2^{k+1}} + (k+1)\left(\frac{2}{5}\right)^{k+1} + \mathcal{O}\left(k^2\left(\frac{8}{25}\right)^{k}\right)\right).$$

Thus

$$\rho_k - \frac{1}{4} - \frac{1}{2^{k+3}} < \frac{1}{4}(k+1)\left(\frac{2}{5}\right)^{k+1} + \mathcal{O}\left(k^2\left(\frac{8}{25}\right)^{k}\right),$$

which means that

$$\rho_k = \frac{1}{4} + \frac{1}{2^{k+3}} + \mathcal{O}\left(k\left(\frac{2}{5}\right)^{k}\right).$$

In the following table we display the first values for ρ_k.

ρ_2	0.3090169...
ρ_3	0.2718445...
ρ_4	0.2593950...
ρ_5	0.2543301...
ρ_6	0.2520691...
ρ_7	0.2510085...

Now observe that, for a given constant a, we can always write

$$1 - 4x + 2^{k+1}x^{k+1} = (a - x)(4 - 2^{k+1}\sum_{i=0}^{k} a^i x^{k-i}) + 1 - 4a + 2^{k+1}a^{k+1}.$$

Replacing a by ρ_k, we get

$$1 - 4x + 2^{k+1}x^{k+1} = (\rho_k - x)(4 - 2^{k+1}\sum_{i=0}^{k} \rho_k^i x^{k-i}).$$

If we now set

$$B(x) = 4 - 2^{k+1} \sum_{i=0}^{k} \rho_k^i x^{k-i},$$

then, by standard asymptotic calculations (see [2]), we obtain that, for large n,

$$\begin{aligned} [x^n]F_k^- &\sim \frac{1}{4}\sqrt{\frac{B(\rho_k)\rho_k}{\pi n^3}} \left(\frac{1}{\rho_k}\right)^n \\ &= \frac{1}{4}\sqrt{\frac{4\rho_k - (k+1)2^{k+1}\rho_k^{k+1}}{\pi n^3}} \left(\frac{1}{\rho_k}\right)^n. \end{aligned} \tag{5}$$

We can apply formula (5) to provide the asymptotic behaviour of trees with no caterpillar of size 3. Caterpillars with three leaves are also called *pitchforks* in [4].

Proposition 1. *The number of pitchfork-free trees of size n is given by $[x^n]F_2^-$, and, as $n \to \infty$, it satisfies the asymptotic approximation*

$$[x^n]F_2^-(x) \sim \frac{1}{4}\sqrt{\frac{4R - 24R^3}{\pi n^3}} \left(\frac{1}{R}\right)^n,$$

where $R = \frac{1}{4}(\sqrt{5} - 1) = 0.3090169\ldots$.

When $n = 100$ the ratio between $[x^{100}]F_2^-$ and its approximation is $0.9933\ldots$.

2.3. **The average size of the biggest caterpillar subtree.** In this section we determine the value $E_n(\gamma)$, which denotes the average of $\gamma(t)$ when $t \in \mathcal{T}_n$.

As shown in Section 2.2, when $k > 0$, $F_k^-(x)$ gives the number of trees with γ at most k. Indeed, also in the case $k = 1$, we have $F_1^- = (1 - \sqrt{1 - 4x + 4x^2})/2 = x$ which represents the unique caterpillar of size 1.

Furthermore consider $f_k^{(n)} = [x^n]F_k^-(x)$, and analogously let us denote by $C^{(n)} = [x^n]C(x)$ the n-th Catalan number. We can write the desired average value as

$$\begin{aligned} E_n(\gamma) &= \frac{1 f_1^{(n)} + \sum_{k\geq 1}(k+1)(f_{k+1}^{(n)} - f_k^{(n)})}{C^{(n)}} \\ &= \frac{-f_1^{(n)} - \cdots - f_{n-1}^{(n)} + n f_n^{(n)} + \sum_{k\geq n}(k+1)(f_{k+1}^{(n)} - f_k^{(n)})}{C^{(n)}} \\ &= \frac{-f_1^{(n)} - \cdots - f_{n-1}^{(n)} + nC^{(n)} + \sum_{k\geq n}(C^{(n)} - f_k^{(n)})}{C^{(n)}} \\ &= \frac{\sum_{k=1}^{n-1}(C^{(n)} - f_k^{(n)}) + C^{(n)} + \sum_{k\geq n}(C^{(n)} - f_k^{(n)})}{C^{(n)}} \\ &= \frac{C^{(n)} + \sum_{k\geq 1}(C^{(n)} - f_k^{(n)})}{C^{(n)}} \\ &= 1 + \frac{\sum_{k\geq 1}(C^{(n)} - f_k^{(n)})}{C^{(n)}}. \end{aligned}$$

In the previous calculation we rely on the fact that $f_k^{(n)} = C^{(n)}$ for $k \geq n$.

We now focus our attention on the generating function $U(x)$, which is defined as

$$U(x) = \sum_{k\geq 1}(C(x) - F_k^-(x)) = \frac{1}{2}\sum_{k\geq 1}\left(\sqrt{1-4x+2^{k+1}x^{k+1}} - \sqrt{1-4x}\right).$$

Near the dominant singularity $x = 1/4$, we may replace the term $\sqrt{1-4x+2^{k+1}x^{k+1}}$ by the corresponding root $\sqrt{1-4x+\frac{1}{2^{k+1}}}$. The effect of the substitution in the sum can be measured, expanding up to first order, as

$$\left|\sum_{k\geq 1}\sqrt{1-4x+2^{k+1}x^{k+1}} - \sum_{k\geq 1}\sqrt{1-4x+\frac{1}{2^{k+1}}}\right| \sim \left|x-\frac{1}{4}\right|\sum_{k\geq 1}(k+1)\left(\frac{1}{2^{k-1}}\right)^{1/2},$$

where $\sum_{k\geq 1}(k+1)\left(\frac{1}{2^{k-1}}\right)^{1/2} = 8+5\sqrt{2} \simeq 15.071\ldots$. By the mentioned substitution, we obtain the generating function $V(x)$, explicitly given by

$$V(x) = \frac{1}{2}\sum_{k\geq 1}\left(\sqrt{1-4x+\frac{1}{2^{k+1}}} - \sqrt{1-4x}\right),$$

whose coefficients grow asymptotically like those of $U(x)$. Considering the n-th coefficient of $V(x)$, we define the sequence $(g_n)_{n\geq 1}$ by

$$g_n = \frac{[x^n]V(x)}{C^{(n)}} = \sum_{k\geq 1}\left(1-\left(1+\frac{1}{2^{k+1}}\right)^{-n+1/2}\right),$$

which gives an asymptotic approximation of $E_n(\gamma) - 1$.

By a *Poisson/Mellin-transform* approach (see [7] for details), we can now further investigate the growth of the coefficients g_n when n is large. By a Poisson-transform we reduce the problem to the asymptotic analysis of a *harmonic sum*, which is then studied using Mellin transforms.

Setting

$$C = \sum_{k\geq 2}\left(1-\left(1+\frac{1}{2^k}\right)^{1/2}\right) \simeq -0.24056\ldots,$$

we write

$$g_n = C + \sum_{k\geq 2}\left(1+\frac{1}{2^k}\right)^{1/2}\left(1-\left(1+\frac{1}{2^k}\right)^{-n}\right) = C + h_n.$$

If $H(x)$ is the exponential generating function of the sequence $(h_n)_{n\geq 1}$, we compute the associated Poisson-transform $\tilde{H}(x)$ and obtain

$$\tilde{H}(x) = \sum_{n\geq 0} h_n \frac{x^n}{n!}\exp(-x) = \sum_{k\geq 2}(1+2^{-k})^{1/2}\left(1-\exp\left(\frac{-x}{2^k+1}\right)\right).$$

We are now interested in the behaviour of $\tilde{H}(x)$ when $x \to \infty$. Indeed, for n large, h_n is approximated by $\tilde{H}(n)$. Observe that $\tilde{H}(x)$ is a harmonic sum, i.e., it is of the form

$$\tilde{H}(x) = \sum_k \lambda_k \tilde{h}(\mu_k \cdot x).$$

In the open strip of complex numbers $s = \alpha + i\beta$ such that $-1 < \alpha < 0$, the associated Mellin-transform is

$$\begin{aligned}\mathcal{M}(\tilde{H}; s) =& \sum_{k\geq 2} \frac{\lambda_k}{\mu_k^s} \cdot \int_0^\infty \tilde{h}(x) x^{s-1} dx = -\Gamma(s) \sum_{k\geq 2} (1+2^{-k})^{1/2}(1+2^k)^s \\ =& -\Gamma(s) \sum_{k\geq 2} 2^{ks}[(1+2^{-k})^{s+1/2} - 1] - \Gamma(s) \sum_{k\geq 2} (2^s)^k \\ =& -\Gamma(s) \sum_{k\geq 2} 2^{ks}[(1+2^{-k})^{s+1/2} - 1] + \Gamma(s) \frac{4^s}{2^s - 1}. \end{aligned} \tag{6}$$

The behaviour of $\tilde{H}(x)$ for x large can be obtained by the analysis of singular expansions of (6) to the right of the strip $-1 < \Re(s) < 0$. The transform can be analytically continued in $0 < \Re(s) \leq M$ for any $M > 0$, then the poles of interest are just those at $s = 0$ (double pole) and at $s = \chi_k = \frac{2k\pi i}{\log 2}$ for $k \in \mathbb{Z} \setminus \{0\}$.

The singular expansion of $\Gamma(s)$ at $s = 0$ begins $\Gamma(s) \sim \frac{1}{s} - \eta$ (with $\eta \simeq 0.57721\ldots$ being Euler's constant), and similarly one has $\frac{1}{2^s-1} \sim \frac{1}{s\log 2} - \frac{1}{2}$ for s close to 0. Furthermore, we need to consider the expansion $4^s \sim 1 + s\log 4$. Putting all together, we obtain the expansion of (6) near the double pole $s = 0$:

$$\frac{1}{s^2 \log 2} + \frac{1}{s}\left(-\frac{\eta}{\log 2} + \frac{3}{2} + C\right) + \mathcal{O}(1).$$

On the other hand, the expansion of (6) near $s = \chi_k$ reads

$$\frac{\Gamma(\chi_k)}{\log 2 \cdot (s - \chi_k)} + \mathcal{O}(1),$$

given that $4^{\chi_k} = 1$.

Each pole s_0 contributes to the asymptotic of $\tilde{H}(x)$ with a term determined by the rule (see again [7])

$$\frac{d}{(s-s_0)^{k+1}} \to -\frac{(-1)^k d}{k!} \cdot x^{-s_0} (\log x)^k.$$

In this manner, when $x \to \infty$, we find for any $M > 0$

$$\tilde{H}(x) = \log_2(x) + \frac{\eta}{\log 2} - \frac{3}{2} - C - \frac{P(\log_2 x)}{\log 2} + \mathcal{O}(x^{-M}),$$

where

$$P(x) = \sum_{k\neq 0} \Gamma(\chi_k) \exp(-2k\pi i \cdot x) \tag{7}$$

is a function of period 1 with mean zero and minute fluctuations bounded by $\max(|P(x)|) \simeq 10^{-6}$, see Figure 3.

The behaviour of the coefficients h_n can be obtained by Depoissonization, which leads to

$$h_n = \log_2(n) + \frac{\eta}{\log 2} - \frac{3}{2} - C - \frac{P(\log_2 n)}{\log 2} + \mathcal{O}\left(\frac{\log^* n}{n}\right),$$

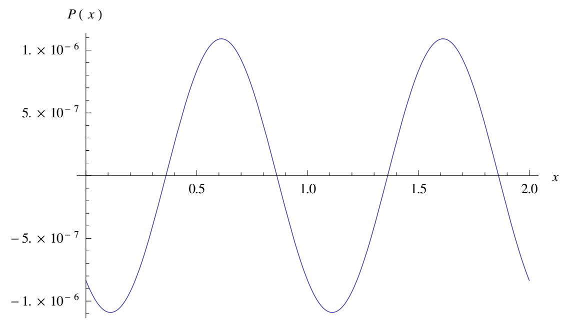

FIGURE 3. Fluctuations determined by $P(x)$.

from which we infer

$$E_n(\gamma) = \log_2(n) + \frac{\eta}{\log 2} - \frac{1}{2} - \frac{P(\log_2 n)}{\log 2} + \mathcal{O}\left(\frac{\log^* n}{n}\right).$$

We summarize the previous calculations in the following proposition.

Proposition 2. *When n is large, the expected size of the biggest caterpillar sub-tree in a tree of size n is given by*

$$E_n(\gamma) = \log_2(n) + \frac{\eta}{\log 2} - \frac{1}{2} - \frac{P(\log_2 n)}{\log 2} + \mathcal{O}\left(\frac{\log^* n}{n}\right), \tag{8}$$

where $\eta \simeq 0.57721\ldots$ is Euler's constant and $P(x)$ is a small periodic fluctuation of mean zero defined by (7)*.*

Evaluating the non-fluctuating term, we obtain the second row of the following table, while in the first row we find, for several values of n, the true $E_n(\gamma)$ computed by generating functions.

n	50	100	200	500	1000
$E_n(\gamma)$	6.202	7.107	8.052	9.334	10.318
(8)	5.976	6.976	7.976	9.298	10.298

As a corollary to Proposition 2, we obtain the following result.

Corollary 3. *For $n \to \infty$, we have*

$$\frac{\log_2(n)}{E_n(\gamma)} \sim 1.$$

3. Caterpillars in permutations in the set $Av(132)$

In Section 2.1 we have introduced caterpillars as objects related to trees. We know that also the class of permutations avoiding the pattern 132 is enumerated by Catalan numbers. Indeed, one can bijectively map the set $\mathcal{T}_{n+1}$ onto the set $Av_n(132)$, where the last symbol denotes the class of permutations of size n avoiding the pattern 132. In particular, we will use a bijection $\phi : \mathcal{T}_{n+1} \to Av_n(132)$ which works as described below.

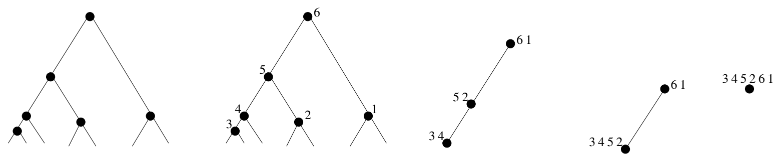

FIGURE 4. The mapping ϕ.

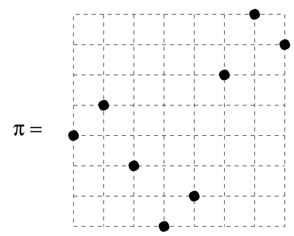

FIGURE 5. The permutation $\pi = (45312687)$.

Bijection ϕ. Take $t \in \mathcal{T}_{n+1}$ and visit t according to pre-order traversal. At the same time, starting with the label n for the root, label each node of outdegree two in decreasing order. After this first step one obtains a tree with integers associated with the nodes of outdegree two. Each leaf now collapses to its direct ancestor which takes a new label receiving on the left and right the label of its left and right child respectively. We continue collapsing leaves until we obtain a tree made of one node which is labelled with a permutation of size n. See Figure 4 for an instance of this mapping.

Using ϕ, one can see how caterpillars are realized in permutations with no 132 pattern. We need the following definition. Let $\pi = \pi_1\pi_2 \ldots \pi_n$ be a permutation. For a given entry π_i, we define $r_\pi(\pi_i)$ as the set made of those entries π_k such that:

1) $\pi_k \leq \pi_i$;
2) all entries of π which are placed between π_k and π_i are less than or equal to π_i.

Given $\pi = \pi_1\pi_2 \ldots \pi_n$, let $\tilde{r}_\pi(\pi_i)$ be the permutation one obtains by extracting the elements of $r_\pi(\pi_i)$ from π, respecting the order. The set of permutations $\{\tilde{r}_\pi(\pi_i)\}_{i=1,\ldots,n}$ is denoted by $\tilde{r}_\pi$. As an example, consider the permutation π which is shown in Figure 5. In this case, $\tilde{r}_\pi$ is made of

$$
\begin{aligned}
\tilde{r}_\pi(4) &= (1),\\
\tilde{r}_\pi(5) &= (45312),\\
\tilde{r}_\pi(3) &= (312),\\
\tilde{r}_\pi(1) &= (1),\\
\tilde{r}_\pi(2) &= (12),\\
\tilde{r}_\pi(6) &= (453126),\\
\tilde{r}_\pi(8) &= (45312687),\\
\tilde{r}_\pi(7) &= (1).
\end{aligned}
$$

The next proposition describes how caterpillars look like in permutations avoiding the pattern 132. It is interesting to see that the presence of such particular subtrees is linked to the property of avoidance of the pattern 231.

Proposition 4. *If $t \in \mathcal{T}_{n+1}$ and $\phi(t) = \pi = \pi_1\pi_2 \dots \pi_n$, then the following holds:*

i) *caterpillar subtrees of t correspond through ϕ to those permutations in $\tilde{r}_\pi$ which avoid the pattern 231;*

ii) *$\gamma(t) - 1$ corresponds to the size of the biggest permutation in*

$$Av(231) \cap \tilde{r}_\pi.$$

Proof. Label t according to the procedure ϕ. If a node is labelled with m, consider the subtree t_m whose root is m. The nodes belonging to t_m form the subsequence of π made of the elements of $r_\pi(m)$. Therefore, we find the pattern 231 in $\tilde{r}_\pi(m)$ if and only if we can find a node in t_m having two descendants which are not leaves of t. It is now sufficient to observe that t_m is a caterpillar if and only if it does not contain such a node. Summarizing, for every node m of t, t_m is a caterpillar subtree of size $k+1$ if and only if $\tilde{r}_\pi(m) \in Av_k(231)$. □

Using the results of Proposition 4 as well as those contained in previous sections, we can describe some properties of the permutations in $\tilde{r}_\pi$ when π avoids the pattern 132.

Corollary 5. *The generating function of the number of permutations $\pi \in Av(132)$ such that all elements in $\tilde{r}_\pi$ of size greater than one contain the pattern 231 is given by*

$$\frac{F_2^-(x)}{x} - 1 = \frac{1 - 2x - \sqrt{1 - 4x + 8x^3}}{2x}.$$

The first terms of the sequence are:

$$1, 0, 1, 2, 6, 16, 45, 126, 358, 1024, 2954, 8580, 25084, 73760, 218045.$$

Remark. Given $\pi = \pi_1\pi_2 \dots \pi_n$, we say that π_i is a *valley* when π_{i-1} and π_{i+1} (if they exist) are greater than π_i; while π_i is said to be a *peak* if both π_{i-1} and π_{i+1} exist and $\pi_{i-1} < \pi_i > \pi_{i+1}$. In this sense, the permutations π considered in Corollary 5 can be characterized, among those in $Av(132)$, by the fact that each entry π_i either is a valley or for which $\tilde{r}_\pi(\pi_i)$ contains at least one peak. We also observe that sequence A025266 of [6] provides the same list of numbers given in the previous corollary. The mentioned sequence also enumerates Motzkin paths with additional constraints.

Finally we state the following result which can be deduced from Corollary 3.

Corollary 6. *If $\pi \in Av(132)$ has size n, the expected size of the biggest permutation in $Av(231) \cap \tilde{r}_\pi$ is asymptotically $\log_2(n)$ as $n \to \infty$.*

4. Caterpillars in un-ordered rooted binary trees

In the previous sections we have focused our attention on the presence of caterpillar subtrees in ordered rooted binary trees. As a second step we would like to investigate the un-ordered case. Let us start by recalling some basic enumerative properties.

Un-ordered rooted binary trees are enumerated with respect to the size, i.e., number of leaves, by the sequence $w_1, w_2, \dots, w_n, \dots$ of the so-called *Wedderburn–Etherington* numbers. This sequence corresponds to entry $A001190$ of [6]. The corresponding generating function $W(x)$ is defined implicitly by the equation

$$W(x) = x + \frac{1}{2}W(x)^2 + \frac{1}{2}W(x^2).$$

The asymptotic behaviour of $(w_n)_{n>0}$ is given by

$$w_n \sim \frac{\lambda}{2\sqrt{\pi}} n^{-3/2} \rho^{-n}, \tag{9}$$

where

$$\frac{\lambda}{2\sqrt{\pi}} = 0.3187766259\ldots \text{ and } \rho = 0.40269750367\ldots.$$

See for example [2].

The class of un-ordered rooted binary trees is denoted by $\mathcal{W}$, while $\mathcal{W}_n$ represents the subset of $\mathcal{W}$ whose elements have size n. In the present section we use the word "tree" to refer to "un-ordered binary rooted tree".

The definition of a *caterpillar tree* in $\mathcal{W}$ is the same as in the ordered case. We have to pay attention to the fact that now, due ton the un-ordering constraint, the number of different caterpillar trees of fixed size is one (see again Figure 1(a), (b)). Analogously to Section 2, we define the parameter $\gamma(t)$ as the size of the biggest caterpillar subtree of the tree t.

Let $W_k(x)$ be the ordinary generating function for the number of trees with γ *at most* equal to $k > 0$.

One can see that, similarly to the functions F_k^- of Section 2, W_k satisfies the equation

$$W_k(x) = x + \frac{1}{2}W_k(x)^2 + \frac{1}{2}W_k(x^2) - x^{k+1}. \tag{10}$$

The generating function $W_k(x)$ has radius of convergence ρ_k which is at least $\rho = 0.402\ldots$ and (for $k \geq 2$) at most $1/2$. The latter bound can be observed from a comparison with the solution to the functional equation $g = x + x^2/2 + g^2/2 - x^3$. Indeed the solution g counts those trees with no caterpillar of size 3 where each tree of size greater than two such that its left root sub-tree is isomorphic to the right root sub-tree is counted by $1/2$. From (10), we obtain

$$W_k(x) = 1 - \sqrt{1 - 2\left(x + \frac{W_k(x^2)}{2} - x^{k+1}\right)} = 1 - \sqrt{1 - 2\varphi_k(x)}, \tag{11}$$

and we see that ρ_k corresponds to the smallest positive solution of $\varphi_k(x) = 1/2$.

The function $\varphi_k(x)$ is analytic in the disc $|x| < \rho_k^{1/2}$, which then contains the one determined by ρ_k. Expanding $\varphi_k(x)$ near $x = \rho_k$, we obtain

$$\varphi_k(x) = \varphi_k(\rho_k) + \left(1 - \rho_k^k - k\rho_k^k + \rho_k W_k'(\rho_k^2)\right)(x - \rho_k) + \mathcal{O}((x - \rho_k)^2),$$

and, substituting this into (11), we obtain the singular expansion of $W_k(x)$ at $x = \rho_k$:

$$W_k(x) = 1 - \sqrt{2\rho_k - 2\rho_k^{k+1} - 2k\rho_k^{k+1} + 2\rho_k^2 W_k'(\rho_k^2)} \cdot \sqrt{1 - \frac{x}{\rho_k}} + \mathcal{O}((x-\rho_k)^{3/2})$$
$$= 1 - \lambda_k \cdot \sqrt{1 - \frac{x}{\rho_k}} + \mathcal{O}((x-\rho_k)^{3/2}). \tag{12}$$

Starting from (12) and performing a standard singularity analysis, we obtain the asymptotics for the number of trees of size n with gamma parameter at most k. Let us denote the latter number by $w_{n,k}$. Then, as $n \to \infty$, we have

$$w_{n,k} = \frac{\lambda_k}{2\sqrt{\pi}} n^{-3/2} \rho_k^{-n} + \mathcal{O}(n^{-5/2}\rho_k^{-n}). \tag{13}$$

We now proceed as described in [2], where one finds a procedure which can be used to numerically approximate the constants ρ_k and λ_k which are involved in (13). The accuracy of the approximations depends on a parameter m, which is here taken as $m = 10$.

Once we have fixed k, we can compute the numbers $w_{1,k}, w_{2,k}, \ldots, w_{m,k}$ by recursively applying (10). The values for $k = 1, \ldots, 5$ are listed in the table below.

k	$w_{1,k}$	$w_{2,k}$	$w_{3,k}$	$w_{4,k}$	$w_{5,k}$	$w_{6,k}$	$w_{7,k}$	$w_{8,k}$	$w_{9,k}$	$w_{10,k}$
1	1	0	0	0	0	0	0	0	0	0
2	1	1	0	1	1	2	3	6	10	19
3	1	1	1	1	2	4	7	14	27	55
4	1	1	1	2	2	5	9	19	37	78
5	1	1	1	2	3	5	10	21	42	89

Using these entries, we define

$$\tilde{\varphi}_k(x) = x - x^{k+1} + \frac{1}{2}\sum_{i=1}^{m} w_{i,k} x^{2i},$$

and we consider ρ_k approximated by the smallest positive solution $\tilde{\rho}_k$ of $\tilde{\varphi}_k(x) = 1/2$. We can also estimate $W_k'(\rho_k^2)$ and find $\sum_{i=1}^m i \cdot w_{i,k} \cdot \tilde{\rho}_k^{\,2i-2}$, thereby obtaining an approximation $\tilde{\lambda}_k$ for λ_k. We observe that increasing the precision $m > 10$ — as we will see in the next paragraph — does not change the first five (respectively four) digits of $\tilde{\rho}_k$ (respectively $\tilde{\lambda}_k$). It is then reasonable to assume that, up to five (respectively four) digits, $\tilde{\rho}_k = \rho_k$ (respectively $\tilde{\lambda}_k = \lambda_k$).

The results for $k = 2, \ldots, 5$ are listed in the following table. In the last column we find the ratio between the true value of $w_{50,k}$ and the one given by (13) calculated using $\tilde{\rho}_k$ and $\tilde{\lambda}_k$.

k	$\tilde{\rho}_k$	$\tilde{\lambda}_k/(2\sqrt{\pi})$	$w_{50,k}/(13)$
2	0.46745	0.2789	1.008
3	0.42291	0.2991	1.009
4	0.41001	0.3089	1.010
5	0.40550	0.3139	1.011

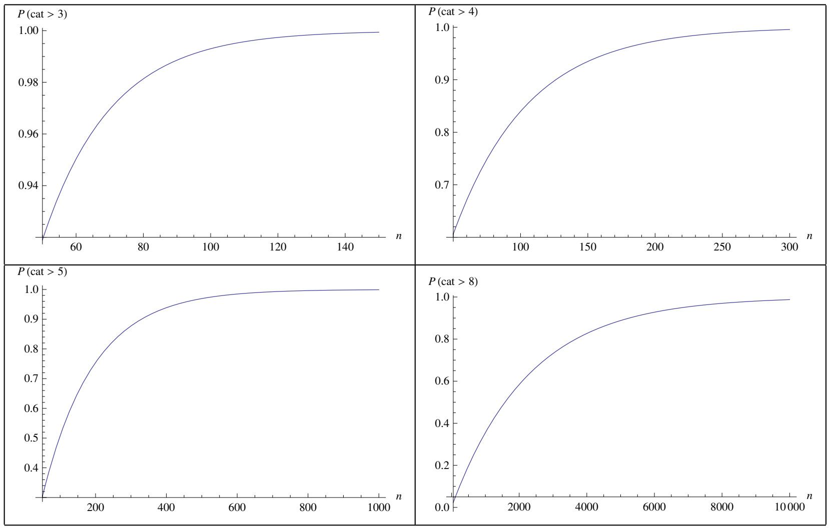

FIGURE 6. The probability of a tree of size n with at least one caterpillar of size $k = 3, 4, 5, 8$.

Formula (13), together with (9), gives the probability of a tree of size $n \gg 1$ having no caterpillar of size greater than k. As an example, take $n = 100$ and $k = 5$. Then we have

$$\frac{w_{100,5}}{w_{100}} \sim \frac{0.3139}{0.3187} \times \left(\frac{0.40550}{0.40269}\right)^{-100} = 0.984 \times (1.006)^{-100} \sim 0.5.$$

Roughly speaking 50% of trees of size 100 have no caterpillar of size greater than 5.

Small caterpillars. In order to use the asymptotic result (13), we list, in this final section, the values of ρ_k and $\lambda_k/(2\sqrt{\pi})$ for $2 \leq k \leq 10$. Furthermore, we do this with more accuracy than before. Indeed we choose $m = 30$ for our approximations, which gives, at least, 10 digits of accuracy with respect to the exact values. In this way, we will be able to find the probabilities $w_{n,k}/w_n$ when n is large and k is small (i.e., less than or equal to 10). The table below shows the values we are interested in.

k	ρ_k	$\lambda_k/(2\sqrt{\pi})$
2	0.4674554078	0.2789408958
3	0.4229139375	0.2991123692
4	0.4100112389	0.3089581337
5	0.4055024052	0.3139472095
6	0.4038017227	0.3164492710
7	0.4031375239	0.3176775180
8	0.4028738458	0.3182668950
9	0.4027683607	0.3185438777
10	0.4027260095	0.3186717321

Using these values, we plot in Figure 6 the probability of a tree with at least one caterpillar of size greater than k, i.e., $1 - (w_{n,k}/w_n)$, for large n and $k = 3, 4, 5, 8$.

Acknowledgements

We would like to thank an anonymous referee for several suggestions that contributed to improve this manuscript. This work was supported by the German Research Foundation (DFG-SPP1590).

References

[1] M. Dairyko, L. Pudwell, S. Tyner, C. Wynn, *Non-Contiguous Pattern avoidance in Binary Trees*, Electronic Journal of Combinatorics **19**(3) (2012), Article P22, 21 pp.

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

[3] M. Kirkpatrick, M. Slatkin, *Searching for evolutionary patterns in the shape of a phylogenetic tree*, Evolution **47** (1993), 1171–1181.

[4] N. A. Rosenberg, *The mean and the variance of the numbers of r-pronged nodes and r-caterpillars in Yule generated genealogical trees*, Annals of Combinatorics **10** (2006), 129–146.

[5] E. S. Rowland, *Pattern avoidance in binary trees*, Journal of Combinatorial Theory Ser. A **117** (2010), 741–758.

[6] N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, available at: `http://oeis.org/`.

[7] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley & Sons, New York, 2001.