The size of the biggest Caterpillar subtree in binary rooted planar trees

Filippo Disanto*

April 24, 2012

Abstract

We study particular patterns in planar rooted binary trees. In particular we will consider those subtrees having the *caterpillar* property. The size of the biggest caterpillar subtree becomes then a new parameter with respect to which we find several enumerations.

1 Introduction

In this work we want to study particular patterns in planar rooted binary trees. More precisely we will consider what seems to be a new statistic on this well known class of 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 population genetic framework, when trees are used to represent ancestry relations among individuals, the presence of a caterpillar subtree often correspond to interesting phenomena such as *natural selection*.

The problem of considering subtrees structures is not new, see for example [1] and [5]. Up to our knowledge Caterpillars have not yet been considered in the context of planar rooted binary trees. Their study here can be also considered as an introductory step to further works concerning the realization of caterpillars in *non-planar* rooted binary trees. Indeed we believe possible to extend the main approach of this paper to the more difficult non-planar case.

After giving some basic definitions, we will provide the enumeration for the number of planar rooted binary trees of a given size having the biggest caterpillar subtree of size less than (resp. greater than, equal

*Institut für Genetik, Universität zu Köln

to) a fixed integer k. Furthermore we will provide the expected value of the size of the biggest caterpillar subtree when trees of size n are uniformely distributed.

Finally, in Section 5 we will see how caterpillars subtrees correspond to patterns extracted from 132-avoiding permutations. The resulting characterization seems quite interesting and should deserve further studies.

2 Definitions

Planar rooted binary trees are enumerated with respect to the size, i.e. number of leaves, by the well known sequence of *Catalan* numbers corresponding to entry $A000108$ in [6]. The respective generating function $C(x)$ is the following

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

The class of planar rooted binary trees will be denoted by $\mathcal{T}$ while $\mathcal{T}_n$ will represent the subset of $\mathcal{T}$ made of those elements having size n. In what follows we will use the term tree referring to planar binary rooted trees.

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. See for example Fig. 1 (a) (b).

Caterpillars can be also 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 $t_l(i)$ (resp. $t_r(i)$) be the left (resp. 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*'s index (see [3]) is defined as

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

The Colless's index is considered as a measure of tree imbalance (see [3]). Its value ranges between 0 and 1, where 0 corresponds to a completely balanced tree while 1 to an unbalanced one.

From the previous definitions it turns out that a tree of size $n > 2$ is a caterpillar if and only if its Colless's index is 1.

If $t \in \mathcal{T}_n$ we define $\gamma(t)$ as the size of the biggest caterpillar which can be seen as a subtree of t. We observe that, if $n > 1$, then $\gamma(t)$ is at least equal to two. In Fig. 2 we have depicted a tree having $\gamma = 5$.

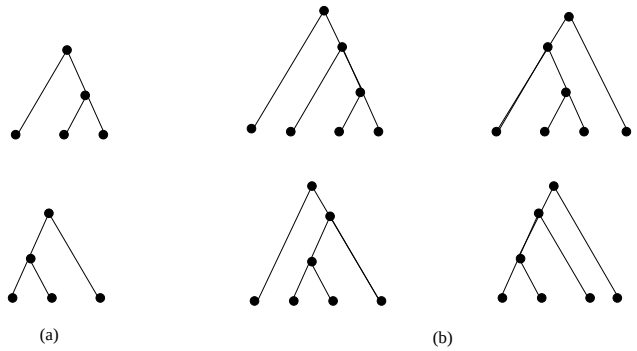

Figure 1: (a) caterpillars of size 3; (b) caterpillars of size 4.

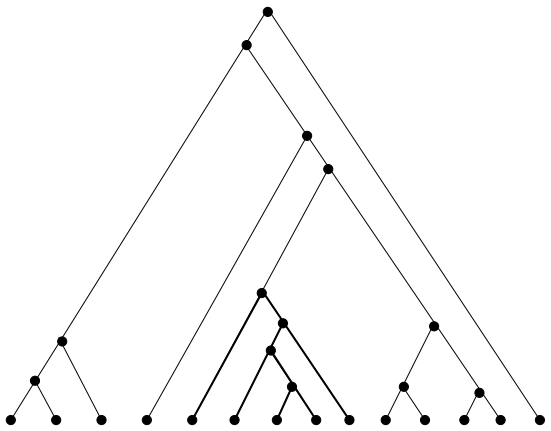

Figure 2: A tree having γ parameter equal to 5. The biggest caterpillar is highlighted.

3 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 the γ parameter *at most* equal to $k \geq 2$.

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

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

Indeed a tree t with $\gamma(t) \leq k$ has either size one or it is made of two trees t_1 and t_2 attached to the root such that $\gamma(t_1) \leq k$ and $\gamma(t_2) \leq k$. We must exclude the case in which one between 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 (1) we obtain

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

Then considering $F_k^+ = C(x) - F_{k-1}^-(x)$ one has the number of trees having $\gamma \geq k$ while taking $F_k = F_k^-(x) - F_{k-1}^-(x)$ one can compute the number of trees of a given size having $\gamma = k$. 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, there is no tree of size $k+1$ having the γ parameter equal to k.

We conclude this section observing that none of the sequences corresponding to F_k^-, F_k^+ and F_k seems to be present in [6].

3.1 Asymptotic growth of trees with no pitchforks

The function $F_k(x)$ is analytic except when x is a solution 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

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

Indeed this can be shown by considering the polynomial

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

which satisfies $y(1/4) > 0$ and $y(2/5) < 0$. Furthermore y is also decreasing between 0 and $1/4$ as it can be seen by solving the equation $y'(x) = 0$ which gives $x = \sqrt[k]{4/(2^{k+1}(k+1))} > 1/4$. 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 (2) yields next

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

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, considering that $(4/5)^{k+1} \sim 0$, gives

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

Thus

$$\rho_k - \frac{1}{4} - \frac{1}{2^{k+3}} < \frac{1}{4}(k+1)\left(\frac{2}{5}\right)^{k+1} \sim \frac{1}{10}k\left(\frac{2}{5}\right)^{k},$$

which means

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

In the following table we show the first approximated values of 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

For a given constant a we can always write

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

then, substituting the solution ρ_k to a we have

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

Defining

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

and by standard asymptotic calculations (see [2]) we have

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

where $n \to \infty$.

We can apply the result in (3) 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 it satisfies asymptotically the following relation:*

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

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

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

4 The average size of the biggest caterpillar subtree

In this section we want to determine $E_n(\gamma)$ which denotes the average value of the parameter $\gamma(t)$ when $t \in \mathcal{T}_n$.

As showed in Section 3, when $k > 0$, $F_k^-(x)$ gives the number of trees having γ at most k. Indeed, also in the case $k = 1$, we have $F_1^- = (1 - \sqrt{1 - 4x + 4x^2})/2 = x$. Where x represents the unique caterpillar of size 1.

Furthermore consider $f_k^{(n)} = [x^n]F_k^-(x)$ and analogously we denote by $C^{(n)} = [x^n]C(x)$ the n-th catalan number. Then we can express the desired average value as follows:

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

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

It is sufficient now to find the n-th term of the function

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

In what follows we want to find a function $\tilde{U}$ which estimates U near the dominant singularity $1/4$. According to [2], the n-th term of the Taylor expansion of $\tilde{U}$ will provide an approximation of $[x^n]U(x)$.

Let us fix x near $1/4$ and let us consider the threshold function

$$k_0 = \log_2 \frac{1}{|1-4x|}.$$

Then, supposing $k \geq k_0$, we have that

$$\sqrt{1 + \frac{2^{k+1}x^{k+1}}{1-4x}} \sim \sqrt{1 + \frac{1}{2^{k+1}(1-4x)}} \sim 1 + \frac{1}{2^{k+2}(1-4x)},$$

while if we suppose $k < k_0$ we will use the approximation

$$\sqrt{1 + \frac{2^{k+1}x^{k+1}}{1-4x}} \sim \sqrt{1 + \frac{1}{2^{k+1}(1-4x)}} \sim \sqrt{\frac{1}{2^{k+1}(1-4x)}}.$$

For the fixed x near $1/4$ we estimate $U(x)$ as follows:

$$\begin{aligned}
U(x) &\sim \frac{\sqrt{1-4x}}{2\sqrt{1-4x}}\sum_{k\geq 1}^{k_0-1}\sqrt{\frac{1}{2^{k+1}}}-\frac{\sqrt{1-4x}}{2}\sum_{k\geq 1}^{k_0-1}1 \\
&\quad +\frac{\sqrt{1-4x}}{2(1-4x)}\sum_{k\geq k_0}\frac{1}{2^{k+2}} \\
&= \frac{1}{2\sqrt{2}}\sum_{k\geq 1}^{k_0-1}\sqrt{\frac{1}{2^k}}-\frac{\sqrt{1-4x}}{2}(k_0-1)+\frac{1}{8\sqrt{1-4x}}\sum_{k\geq k_0}\frac{1}{2^k} \\
&= \frac{1}{2\sqrt{2}}\frac{-\sqrt{2}+2^{1-\frac{k_0}{2}}}{-2+\sqrt{2}} \\
&\quad -\frac{\sqrt{1-4x}}{2}\left(\log_2\left(\frac{1}{|1-4x|}\right)-1\right)+\frac{2^{1-k_0}}{8\sqrt{1-4x}} \\
&= \frac{1}{2}+\frac{1}{2\sqrt{2}}+\sqrt{1-4x}\left(-\frac{1}{\sqrt{2}}+\log_2(\sqrt{1-4x})+\frac{1}{4}\right).
\end{aligned}$$

Using the previous calculation we have the following result.

Proposition 2 *Let us denote*

$$\tilde{U}(x)=\frac{1}{2}+\frac{1}{2\sqrt{2}}+\sqrt{1-4x}\left(-\frac{1}{\sqrt{2}}+\log_2(\sqrt{1-4x})+\frac{1}{4}\right),$$

then

$$E_n(\gamma)\sim\frac{[x^n]\tilde{U}(x)}{C^{(n)}}.$$

As a test one can consider the following table where, for several values of n, we compare the true $E_n(\gamma)$ with the approximation given by Proposition 2.

n	10	20	50	100	200	500	1000
$E_n(\gamma)$	4.535	5.120	6.202	7.107	8.052	9.334	10.318
$\frac{[x^n]\tilde{U}(x)}{C^{(n)}}$	4.032	5.109	6.47	7.490	8.498	9.824	10.825

We can go a step further in our approximation considering the following statement.

Corollary 1 *When $n\to\infty$ we have*

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

Proof. We use the result of Proposition 2 and the well known asymptotic behaviour of Catalan numbers:

$$C^{(n)} \sim \frac{4^{n-1}}{\sqrt{\pi n^3}}.$$

Furthermore, by standard technique (see again [2]), we also calculate the behaviour of

$$\sqrt{1-4x}\log_2(\sqrt{1-4x}) \sim \frac{4^{n-1}\log_2(n)}{\sqrt{\pi n^3}}$$

and

$$-\sqrt{1-4x} \sim \frac{4^n}{2\sqrt{\pi n^3}}.$$

Finally we have

$$\begin{aligned} E_n(\gamma) &\sim \left(\frac{1}{\sqrt{2}} \times \frac{4^n}{2\sqrt{\pi n^3}} - \frac{1}{4} \times \frac{4^n}{2\sqrt{\pi n^3}} + \frac{4^{n-1}\log_2(n)}{\sqrt{\pi n^3}}\right) \times \frac{\sqrt{\pi n^3}}{4^{n-1}} \\ &\sim \log_2(n). \end{aligned}$$

□

For $n = 1000$ in the previous table we have $E_n(\gamma) = 10.318$ while $\log_2(n) = 9.96578$ which is quite close to the true value.

5 Caterpillars in permutations $Av(132)$

In Section 2 we have introduced caterpillars as objects related to planar rooted binary 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 classicaly denotes the class of permutations of size n which are avoiding 132. In particular, in what follows, we will use a bijection $\phi : \mathcal{T}_{n+1} \to Av_n(132)$ which works as described below.

Take $t \in \mathcal{T}_{n+1}$ and visit it according to the pre-order traversal labelling each node of outdegree two in decreasing order starting with the label n for the root. After this first step one has a tree labelled with integers at its nodes of outdegree two. Each leaf now collapses to its direct ancestor which takes a new label receiving on the left (resp. right) the label of its left (resp. right) child. We go on collapsing leaves until we achieve a tree made of one node which is labelled with a permutation of size n. See Fig. 3 for an instance of this mapping.

Through ϕ we can see how caterpillars can be interpreted inside permutations without the pattern 132. In order to do this we need the following definition.

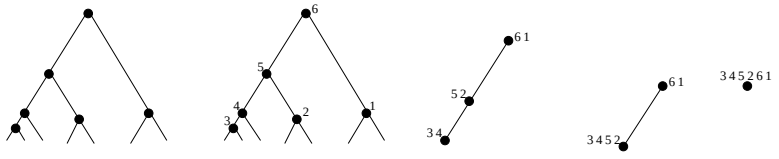

Figure 3: The mapping ϕ.

Let $\pi = \pi_1\pi_2\dots\pi_n$ 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 the entries in π which are between π_k and π_i are less than or equal to π_i.

If $\pi = \pi_1\pi_2\dots\pi_n$ is a permutation, then we define $\tilde{r}_\pi(\pi_i)$ as the permutation one obtains extracting from π the elements belonging to $r_\pi(\pi_i)$ respecting the order. The set of permutations $(\tilde{r}_\pi(\pi_i))_{i=1\dots n}$ will then be denoted by $\tilde{r}_\pi$.

As an example one can consider the permutation π which is depicted in Fig. 4. In this case $\tilde{r}_\pi$ is made of

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

Next proposition describes how caterpillars are realized inside permutations avoiding the pattern 132. It is interesting to see that the presence of such particular subtrees is connected to the property of avoiding the pattern 231.

Proposition 3 *If $t \in \mathcal{T}_{n+1}$ and $\phi(t) = \pi = \pi_1\pi_2\dots\pi_n$, then the following hold:*

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

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

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

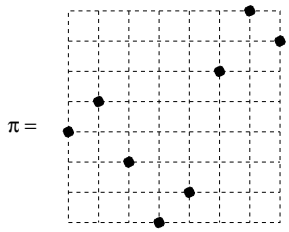

Figure 4: The permutation $\pi = (45312687)$.

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 $r_\pi(m)$. So 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 3, from the previous sections we can derive some properties of the permutations in $\tilde{r}_\pi$ when π avoids the pattern 132. These are stated in the next two corollaries.

Corollary 2 *The number of permutations $\pi \in Av(132)$ such that all elements in $\tilde{r}_\pi$ whose size is 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 $\pi = \pi_1 ... \pi_2$ we say that π_i is a *valley* when π_{i-1} and π_{i-1} (if they exist) are greater than π_i. Analogously π_i is said to be a *peak* if both π_{i-1} and π_{i+1} exist and $\pi_{i-1} < \pi > \pi_{i+1}$. In this sense, the permutations π considered in Corollary 2 can be characterized, among those in $Av(132)$, by the fact that each entry π_i is either a valley or it is such that $\tilde{r}_\pi(\pi_i)$ contains at least one peak. We also observe that sequence A025266 of [6] provides the same list of numbers of the previous corollary as those integers enumerating Motzkin paths with some constraints.

Finally we state the following result which can be deduced from Corollary 1.

Corollary 3 *If $\pi \in Av(132)$ has size n, the expected size of the biggest permutation in $Av(231) \cap \tilde{r}_\pi$ is asymptotic to* $\log_2(n)$.

6 Further works

In the present paper we have focused our attention on the presence of caterpillars subtrees in planar rooted binary trees. As a second step we would like to investigate the case of non-planar rooted binary trees. We think that the approach we have used here could be refined in order to solve the non-planar case enumeration.

Furthermore, we think that the realization of $\tilde{r}_\pi$ for a given permutation π corresponds to an interesting combinatorial object which should deserve further studies.

References

[1] M. Dairyko, L. Pudwell, S. Tyner, C. Wynn, *Non-Contiguous Pattern avoidance in Binary Trees*, available at: `http://faculty.valpo.edu/lpudwell/papers.html`.

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

[3] M. Kirkpatrick, M. Slatkin, *Searching for evolutionary patterns in the shape of a phylogenetic tree*, Evolution, 47(4), 1993, pp.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*, J. Combin. Theory, ser. A 117 (2010) 741-758.

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