MARKUS E. NEBEL

On the Horton-Strahler number for combinatorial tries

Informatique théorique et applications, tome 34, nº 4 (2000), p. 279-296

<http://www.numdam.org/item?id=ITA_2000__34_4_279_0>

Theoretical Informatics and Applications
Theoret. Informatics Appl. **34** (2000) 279–296

ON THE HORTON-STRAHLER NUMBER FOR COMBINATORIAL TRIES

MARKUS E. NEBEL[1]

Abstract. In this paper we investigate the average Horton-Strahler number of all possible tree-structures of binary tries. For that purpose we consider a generalization of extended binary trees where leaves are distinguished in order to represent the location of keys within a corresponding trie. Assuming a uniform distribution for those trees we prove that the expected Horton-Strahler number of a tree with α internal nodes and β leaves that correspond to a key is asymptotically given by

$$\frac{4^{2\beta-\alpha}\log(\alpha)(2\beta-1)(\alpha+1)(\alpha+2)\binom{2\alpha+1}{\alpha-1}}{8\sqrt{\pi}\alpha^{3/2}\log(2)(\beta-1)\beta\binom{2\beta}{\beta}^2}$$

provided that α and β grow in some fixed proportion ρ when $\alpha\to\infty$. A similar result is shown for trees with α internal nodes but with an arbitrary number of keys.

AMS Subject Classification. 05A15, 05C05, 68W40.

1. INTRODUCTION

Let T be a binary tree, *i.e.* a tree where each node has at most two descendants. Then the *Horton-Strahler number* of T denoted $hs(T)$ is recursively defined by

$$hs(T) := \begin{cases} 0 & : \ T \text{ is either a leaf or empty} \\ hs(T.l)+1 & : \ \text{if } hs(T.l)=hs(T.r) \\ \max(hs(T.l),hs(T.r)) & : \ \text{otherwise.} \end{cases}$$

Here, $T.l$ (resp. $T.r$) denotes the left (resp. right) subtree of T. The Horton-Strahler number was originally introduced to classify river systems (see [12] and [22]) but it has also been adopted in computer science, molecular biology, medicine and other disciplines. Ershov [6], for example, has shown that the minimal

[1] Johann Wolfgang Goethe-Universität, Fachbereich Informatik, Senckenbergenlage 31, 60054 Frankfurt am Main, Germany.

number of registers needed to evaluate an arithmetic expression $\mathcal{E}$ with binary operators, which is represented as a binary tree $T(\mathcal{E})$ (the syntax-tree), is given by $1+hs(T(\mathcal{E}))$. If all syntax-trees with n internal nodes (n binary operators) are assumed to be equally likely, then it is known that the average number of registers that are needed to evaluate an expression optimally is given by

$$\frac{1}{2}\log_2(2\pi^2 n) - \frac{\gamma+2}{2\ln(2)} + F(n) + \mathcal{O}(n^{-1/2+\delta}) \tag{1}$$

for all $\delta > 0$, $n \to \infty$, where $\gamma = 0.577215\ldots$ is Euler's constant and F is a periodic, oscillating function of small amplitude (see *e.g.* [7] and [13]). Syntax-trees corresponding to expressions built with unary and binary operators were considered in [9]. Furthermore, the minimum stack-size required for a traversal of a binary tree T is also given by $1 + hs(T)$ (*e.g.* see [7] and [11]). Meir *et al.* [18] investigated the Horton-Strahler number of channel networks with a fixed number of inputs. The combinatorics of the Horton-Strahler analysis has been used in computer graphics for the creation of faithful synthetic images of trees (see [23]). The impact of the Horton-Strahler number on molecular biology comes from theoretical considerations about secondary structures of single-stranded nucleic acids (see [24] and the references given there).

All those applications and studies have in common that they deal with ordinary extended binary trees, *i.e.* trees where each node is either a leaf or has two descendants. All the cited papers which present an average case analysis consider the uniform model, *i.e.* they assume that all trees of a given size are equally likely.

The Horton-Strahler number of tries has been investigated in a recent work by Devroye and Kruszewski [4]. A trie is a binary tree which is used to store the set of keys $K = \{k_1, \ldots, k_n\}$ in the following manner: each key k_i, considered as a string of 0's and 1's due to its binary representation, defines a path in a binary tree T (0 indicates a left turn, 1 a right turn); the trie defined by $k_1, \ldots, k_n$ is the smallest binary tree for which the paths truncated at the leaves of T are all pairwise different. Thus each leaf of T *stores* exactly one of the keys k_i, $1 \le i \le n$. Note that T does not need to be an extended binary tree. T might have internal nodes with only one successor. However, the Horton-Strahler number $hs(T)$ remains unchanged when we turn T into an extended binary tree. Devroye and Kruszewski considered random tries constructed from n i.i.d. sequences of Bernoulli random variables with parameter p, $0 < p < 1$; they have shown that the Horton-Strahler number H_n of those tries fulfils

$$\frac{H_n}{\log n} \to \frac{1}{\log \frac{1}{\min(p,1-p)}}$$

in probability as $n \to \infty$. The presented (Bernoulli-) model of a random trie is very realsitic. For example, if we choose p being $\frac{1}{2}$, this model describes exactly the behavior of tries built from random integer data assuming all integers to be equally likely.

$K = \{k_1, k_2, k_3, k_4\} = \{0001, 0010, 1010, 1101\}$

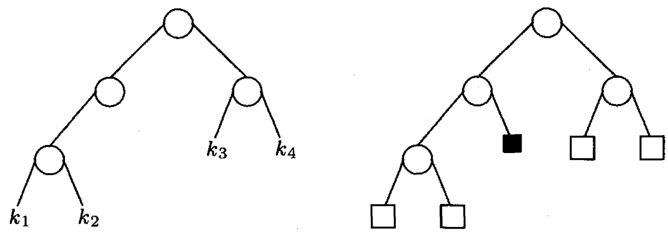

FIGURE 1. An example for a set of keys K, the resulting trie and the corresponding $\mathcal{C}$-trie.

In this paper we will adopt a combinatorial point of view by regarding all tree-structures that might be generated by the trie algorithm. For that purpose we will consider a class of generalized extended binary trees the so-called $\mathcal{C}$-tries (for combinatorial tries) which were introduced in [20]. Let a $\mathcal{C}$*-trie* be an extended binary tree where each leaf is either colored black or white; each black leaf has to be the brother of an internal node. If we now interpret a white leaf as the location of a key and a black leaf as a NIL-pointer the $\mathcal{C}$-tries resemble all the tree-structures which the trie algorithm might generate. An example for that correspondence can be found in Figure 1. It is obvious that all the correspondences given above between the Horton-Strahler number and the parameters such as the number of registers needed for a syntax-tree evaluation or the stack-size for a traversal remain valid even by coloring the leaves. A $\mathcal{C}$-trie with α internal nodes (with α internal nodes and β white leaves) is defined to be of size α (of size (α, β)) and will be called α-trie ((α, β)-trie). Note that $2 \le \beta \le \alpha + 1$ must hold. A $\mathcal{C}$-trie of size $(\alpha, \alpha+1)$ is nothing else but an ordinary extended binary tree. For our investigations we will assume that all $\mathcal{C}$-tries of size α (resp. (α, β)) are equally likely which is a quite different assumption compared to the before mentioned Bernoulli-model where it is more likely that a trie is balanced than being of a linear structure. This is why our investigation is rather of a combinatorial character.

Note that the idea of coloring leaves is not only useful for introducing a combinatorial equivalent for the tree-structures of binary tries. By introducing the number of white leaves as a second parameter we are more flexible to model natural phenomena. By varying the ratio of α and β we can control the average shape of the related trees since a large number of black leaves can only exist if the tree has many linear lists within its inner structure. Thus we hope that the results presented in the rest of this paper will be of interest with respect to geology, molecular biology, synthetic images of trees, channel networks, ... Think for example of a river network modelled by an extended binary tree T and the task to model different lengths of the rivers. By adding a new node with a black leaf as one of its successors in between two arbitrary existing nodes of T (see Fig. 2), we do not change the tree's Horton-Strahler number (as it should be since the

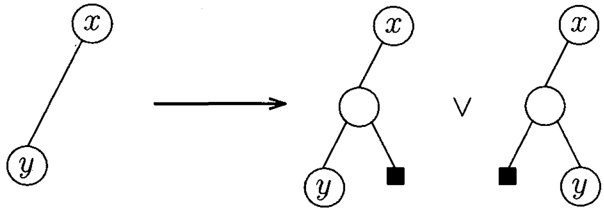

FIGURE 2. Affecting additive parameters by adding a new node in between two existing ones. All trees shown in the figure have the same Horton-Strahler number but different path lengths.

length of a river should not affect the classification of a river network). However, we change length-sensitive parametres like *e.g.* the external path length.

Thus, even if all our results are presented using the term C-trie instead of generalized extended binary tree they should be considered as applicable to many other areas.

2. THE AVERAGE HORTON-STRAHLER NUMBER

The aim of this section is to derive the average Horton-Strahler number for uniform random C-tries as defined in Section 1, *i.e.* the average value of the function hs applied to a set of C-tries of the same size (α, β). We will use generating functions in order to prove our results. The way these generating functions are derived is similar to that in [7], the methodology used to determine asymptotics for the coefficients in question is standard and can be found in [8]. By $[x_1^{n_1} \cdots x_k^{n_k}]f(x_1, \ldots, x_k)$ we denote the coefficient at $x_1^{n_1} \cdots x_k^{n_k}$ in an expansion of $f(x_1, \ldots, x_k)$ at $(x_1, \ldots, x_k) = (0, \ldots, 0)$.

We start our investigations by determining the generating function $H_p(x, y)$ which counts those C-tries that have a Horton-Strahler number of exactly p.

Lemma 1. *Let x mark an internal node and let y mark a white leaf. The generating function $H_p(x, y)$ of C-tries T with $hs(T) = p$ possesses the following closed form representation:*

$$H_p(x, y) = \frac{\sin(\phi)}{\sin(2^{p-1}\phi)} \cdot \frac{xy^2}{1 - 2x - 2xy}, \tag{2}$$

$$\text{where} \qquad \phi = \arccos\left(\frac{1 - 4x(y+1) + 2x^2(2 + y(4+y))}{2x^2y^2}\right). \tag{3}$$

Proof. In order to derive a representation for the generating function in question we have to distinguish the cases shown in Figure 3. For $p \geq 2$ these cases translate

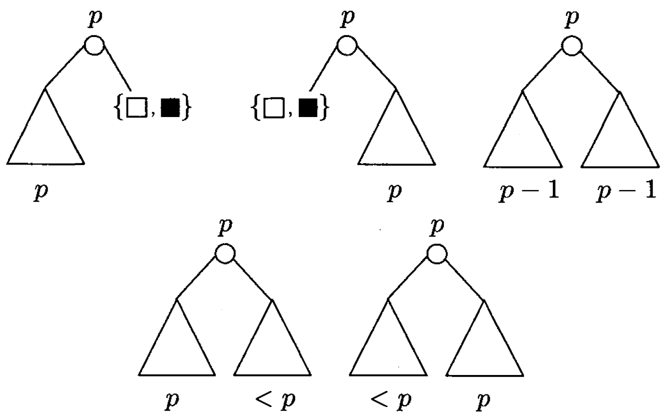

FIGURE 3. All possible decompositions of a $\mathcal{C}$-trie T with $hs(T) = p$. The number below a triangle determines the value of hs of the subtrie represented.

into the following recurrence for $H_p(x,y)$:

$$H_p(x,y) = (2x + 2xy)H_p(x,y) + xH_{p-1}^2(x,y) + 2xH_p(x,y)\sum_{1\le j<p} H_j(x,y). \quad (4)$$

Here, the boundary condition for $p = 1$ is still to be determined yet. It is obvious that a $\mathcal{C}$-trie T with $hs(T) = 1$ has to have a linear structure, *i.e.* either the left or the right subtrie of each internal node has to be a leaf. Thus $H_1(x,y)$ must fulfil $H_1(x,y) = xy^2 + (2x + 2xy)H_1(x,y)$ and therefore

$$H_1(x,y) = \frac{xy^2}{1 - 2x - 2xy}.$$

In order to solve this recurrence we divide both sides of (4) by $xH_p(x,y)$, thus

$$\frac{1}{x} = 2 + 2y + \frac{H_{p-1}^2(x,y)}{H_p(x,y)} + 2\sum_{1\le j<p} H_j(x,y).$$

Subtracting this from the analogous identity obtained for $p+1$ eliminates the summation. We find

$$0 = \frac{H_p^2(x,y)}{H_{p+1}(x,y)} + 2H_p(x,y) - \frac{H_{p-1}^2(x,y)}{H_p(x,y)}. \quad (5)$$

Let $V_p(x,y) := \frac{H_{p-1}(x,y)}{H_p(x,y)}$. Dividing (5) by $H_p(x,y)$ our recurrence can be expressed by means of V_p:

$$V_{p+1}(x,y) = V_p^2(x,y) - 2,\ p \ge 2,$$

with the initial condition $V_2(x,y) = \frac{H_1(x,y)}{H_2(x,y)}$. We can determine $H_2(x,y)$ by using $H_1(x,y)$ and the recurrence (4) which yields

$$V_2(x,y) = \frac{1-4x(y+1)+2x^2(2+y(4+y))}{x^2y^2}.$$

This new recurrence can be solved by a trigonometric change of variables. We set $V_2(x,y) = 2\cos(\phi)$ and generally $V_p(x,y) = 2\cos(\phi_p)$. Since $\cos^2(x) = \frac{1}{2}(1+\cos(2x))$ holds we see that the recurrence translates into

$$2\cos(\phi_{p+1}) = 2\cos(2\phi_p).$$

Therefore, for $p \geq 2$, $\phi_{p+1} = 2\phi_p = 2^{p-1}\phi$ must hold which gives the explicit form

$$\begin{aligned} V_{p+1}(x,y) &= 2\cos(2^{p-1}\phi),\ p \geq 2, \\ V_2(x,y) &= 2\cos(\phi) = \frac{1-4x(y+1)+2x^2(2+y(4+y))}{x^2y^2}. \end{aligned}$$

We can go back to $H_p(x,y)$ by regarding

$$V_p(x,y)V_{p-1}(x,y)\cdots V_2(x) = \frac{H_{p-1}(x,y)}{H_p(x,y)}\frac{H_{p-2}(x,y)}{H_{p-1}(x,y)}\cdots\frac{H_1(x,y)}{H_2(x,y)} = \frac{H_1(x,y)}{H_p(x,y)}.$$

By means of the identity $\sin(2x) = 2\sin(x)\cos(x)$ the product on the left-hand side collapses to $\sin(2^{p-1}\phi)$ when multiplied by $\sin(\phi)$. This completes the proof. □

Next we consider those $\mathcal{C}$-tries that have a Horton-Strahler number of at least p.

Lemma 2. *Let* $S_p(x,y) := \sum_{j\geq p} H_j(x,y)$, $\kappa := \frac{x^2y^2}{(1-2x-2xy)^2}$, $\varepsilon := \sqrt{1-4\kappa}$ *and* $u := \frac{1-\varepsilon}{1+\varepsilon}$. *We have*

$$S_p(x,y) = y\left[\frac{\sqrt{u}(1-u)}{u}\cdot\frac{u^{2^{p-1}}}{1-u^{2^{p-1}}}\right].$$

Proof. In order to prove the lemma we use the identity $\sin(x) = \frac{1}{2i}\left(e^{ix}-e^{-ix}\right)$, $i^2=-1$, which we insert in (2). Together with $t := e^{-i\phi}$ and $r := 2^{p-1}$ we find

$$H_p(x,y) = 2i\sin(\phi)\frac{t^r}{1-t^{2r}}\frac{xy^2}{1-2x-2xy}.$$

Now consider those parts of the representation that depend on p. Summing them up for $j \geq p$ yields

$$\sum_{\substack{j\geq p\\ r=2^{j-1}}}\frac{t^r}{1-t^{2r}} = \sum_{\substack{j\geq p\\ k\geq 0}} t^{2^{j-1}(1+2k)} = \sum_{\substack{m,k\geq 0\\ v=t^{2^{p-1}}}} v^{2^m(2k+1)}.$$

Since the mapping $(m,k) \to 2^m(2k+1)$ is a bijection on $\mathbb{N}^2 \to \mathbb{N}$ the last sum equals $\frac{v}{1-v}$. Therefore $S_p(x,y) = 2i\sin(\phi)\frac{xy^2}{1-2x-2xy}\frac{t^r}{1-t^r}$ holds. Returning to trigonometric functions, *i.e.* setting $t = e^{-i\phi} = \cos(\phi) - i\sin(\phi)$, gives us

$$S_p(x,y) = \left[-i\sin(\phi) + \frac{\sin(\phi)\cos(2^{p-2}\phi)}{\sin(2^{p-2}\phi)}\right]\frac{xy^2}{1-2x-2xy}.$$

For $p > 1$ it is now possible to express the trigonometric functions by means of Chebyshev polynomials. For $p = 1$ we run into trouble since in that case we would refer to the $\frac{1}{2}$-th polynomial which does not exist. Thus, the next step is to express $S_p(x,y)$ by means of $\frac{1}{2}\phi$ instead of ϕ. By applying the identity $2\cos(x)\sin(x) = \sin(2x)$ we find that

$$S_p(x,y) = \frac{-i\sin(\phi)xy^2}{1-2x-2xy} + \frac{\cos(\frac{1}{2}\phi)\sin(\frac{1}{2}\phi)\cos(2^{p-1}\frac{1}{2}\phi)}{\sin(2^{p-1}\frac{1}{2}\phi)}\frac{2xy^2}{1-2x-2xy}$$

holds. From equation (3) we derive closed form expressions for $i\sin(\phi)$ and $\cos(\frac{1}{2}\phi)$ which we insert into the last representation of $S_p(x,y)$. Then, applying the following identities for the Chebyshev polynomial of the first-kind ($T_n(x)$) (see *e.g.* [1] 22.3.15) and the second-kind ($U_n(x)$) (see *e.g.* [1] 22.3.16)

$$T_n(\cos(\phi)) = \cos(n\phi),$$

$$U_n(\cos(\phi)) = \frac{\sin((n+1)\phi)}{\sin(\phi)},$$

yield

$$S_p(x,y) = -\frac{\sqrt{(1-2x)(1-2x-4xy)}}{2x} - \frac{yT_{2^{p-1}}(\hat{\kappa})}{U_{2^{p-1}-1}(\hat{\kappa})}.$$

Here $\hat{\kappa} := \cos(\frac{1}{2}\phi) = -\frac{1-2x-2xy}{2xy}$ holds. Now let $T(x,y)$ be the ordinary generating function of all $\mathcal{C}$-tries. In [21] the following representation can be found:

$$T(x,y) = \frac{1-2x-\sqrt{(1-2x)(1-2x-4xy)}}{2x}.$$

This, together with two further identities for Chebyshev polynomials

$$T_n(x) = U_n(x) - xU_{n-1}(x), \text{ (see } e.g. \text{ [1], 22.5.6)},$$

$$U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x), \text{ (see } e.g. \text{ [15] (B77))},$$

gives us

$$S_p(x,y) = T(x,y) - y + \frac{yU_{2^{p-1}-2}(\hat{\kappa})}{U_{2^{p-1}-1}(\hat{\kappa})},$$

for $U_{-1}(x) := 0$. A closed form representation for $U_n(x)$ is given in [15] (B74). By fundamental algebraic manipulations this representation can be transformed into

$$U_n(x) = -\frac{x^n\left[(1-\sqrt{1-x^{-2}})^{n+1} - (1+\sqrt{1-x^{-2}})^{n+1}\right]}{2\sqrt{1-x^{-2}}}.$$

Now, since $1 - \hat{\kappa}^{-2} = 1 - 4\kappa$ holds, we get

$$S_p(x,y) = T(x,y) - y - 2y\sqrt{\kappa}\frac{(1+\sqrt{1-4\kappa})^{2^{p-1}-1} - (1-\sqrt{1-4\kappa})^{2^{p-1}-1}}{(1+\sqrt{1-4\kappa})^{2^{p-1}} - (1-\sqrt{1-4\kappa})^{2^{p-1}}}.$$

We complete the proof by using the substitutions of the lemma and applying some obvious simplifications. □

Remark. Besides the Horton-Strahler number there is another monotonic marking of binary trees which is related to the evaluation of arithmetic expressions and the traversal. This is the so called *stack-number* of the tree which was investigated in numerous papers (*e.g.* [3, 14, 16, 19, 20] and [21]). It corresponds to the stack-size needed to traverse a tree in preorder using the traditional algorithm (see *e.g.* [17], p. 319ff) and the number of cells on a stack needed to evaluate an arithmetic expression by means of a simple traversal algorithm (see [15] for details). For non-colored extended binary trees we have the following correspondence: the number of trees with α internal nodes, with a stack-number of at most $2^k - 1$, is equal to the number of trees with α internal nodes and a Horton-Strahler number of k (see *e.g.* [15], Th. 5.8). If we inspect the generating functions of the previous lemma and of [20] and [21] we see that such a relation does not exist for C-tries neither of size α nor of size (α, β).

In order to compute the average Horton-Strahler number we need a representation of the generating function $M(x,y) := \sum_{p\geq 1} pH_p(x,y)$. It is not hard to see that $M(x,y) = \sum_{p\geq 1} S_p(x,y)$ holds. Therefore we have

$$\begin{aligned} M(x,y) &= \frac{y\sqrt{u}(1-u)}{u}\sum_{p\geq 1}\frac{u^{2^{p-1}}}{1-u^{2^{p-1}}} \\ &= \frac{y\sqrt{u}(1-u)}{u}\sum_{n\geq 1}(v_2(n)+1)u^n. \end{aligned} \quad (6)$$

Here $v_2(n)$ denotes the *dyadic valuation* of n, *i.e.* the number of positive divisors of n which are a power of two.

Now, everything is prepared to determine an asymptotic equivalent for the average Horton-Strahler number. We use the *Mellin summation method* as described in [10] to evaluate the sum. For that purpose we set $u = \exp(-t)$ and apply the well-known identity

$$\exp(-tj) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \Gamma(s) j^{-s} t^{-s} ds,\ i^2 = -1,$$

for some c in the fundamental strip of the Mellin transform of $\exp(-tj)$ and for $\Gamma(s)$ the complete gamma function. This is how it is possible to express the number-theoretic function $v_2(n)$ by means of the Riemann Zeta function $\zeta(z)$ (see [2] for details). We have (see *e.g.* [15], p. 155)

$$\sum_{n\geq 1} v_2(n) n^{-z} = \sum_{n\geq 1} v_2(2n)(2n)^{-z} = \sum_{\substack{j\geq 1\\ n\geq 1}} (2^j n)^{-z} = \zeta(z)(2^z - 1)^{-1}$$

and therefore with $u = \exp(-t)$

$$\sum_{n\geq 1} (v_2(n)+1) u^n = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \zeta(s) 2^s \Gamma(s) t^{-s} (2^s - 1)^{-1} ds.$$

Now, according to the Mellin summation formula, we have to sum the residues of $\zeta(s)2^s\Gamma(s)t^{-s}(2^s-1)^{-1}$ left to the fundamental strip, *i.e.* the residues with a real part less or equal to one. There are singularities at $s = 1$ and $s = -n$, $n \in \mathbb{N}\cup\{0\}$, but we only have to consider those which are larger than -1 since the others will only imply terms that can be neglected. There are further singularities at $s = \frac{2\pi i k}{\ln(2)} =: \chi_k$, $k \in \mathbb{Z}\backslash\{0\}$, which would imply some oscillation in the lower order terms. As the known methods for multivariate asymptotics only allow to determine the leading term, the singularities χ_k will only be considered later in the univariate case. The sum of the residues for $s = 1$ and $s = 0$ is given by

$$\frac{2}{t} + \frac{2\ln(t) + 2\gamma - 2\ln(\pi) - 3\ln(2)}{4\ln(2)}.$$

Here, $\gamma = 0.5772156649...$ denotes Eulers's constant.

In order to approximate the coefficient of $M(x,y)$ at $x^\alpha y^\beta$ we are interested in an expansion of our generating function at its dominant singularity. For that purpose we assume y being a positive constant not equal to 0 in order to determine the dominant singularity with respect to x, *i.e.* the value of x which has the smallest modulus and which is a singular point of $M(x,y)$. Note that this approach leads to the restriction that our asymptotic will only be valid when α and β grow simultaneously in a fixed proportion. By definition of $M(x,y)$ and properties of the Horton-Strahler number we have the following trivial bounds for $[x^\alpha]M(x,y)$:

$$[x^\alpha]T(x,y) \leq [x^\alpha]M(x,y) \leq \log_2(\alpha+1)[x^\alpha]T(x,y).$$

Under the assumption that y is a constant and since $x = \frac{1}{2+4y}$ is the dominant singularity of $T(x,y)$ the $\mathcal{O}$-transfer method introduced in [8] leads to

$$[x^\alpha]T(x,y) \sim \frac{1+2y}{2} \cdot \frac{(2+4y)^\alpha}{\sqrt{\pi\alpha^3}}.$$

Therefore the Cauchy-Hadamard formula tells us that both the minorant and the majorant of $M(x,y)$ have a radius of convergence of $\frac{1}{2+4y}$ and we infer that $\frac{1}{2+4y}$ is the radius of convergence of $M(x,y)$ itself. Thus, by the theorem of Pringsheim, we can conclude that $x = \frac{1}{2+4y}$ is a dominant singularity of our generating function. It resides to prove that it is the only dominant singularity. For that purpose we consider the representation (6) of $M(x,y)$. Besides the algebraic singularity at $x = \frac{1}{2+4y}$ implied by our substitution, the factor of the sum is only singular for $u = 0$ *i.e.* for $x = 0$. This is why it does not extend the set of dominant singularities. The sum $\sum_{n\geq 1}(v_2(n)+1)u^n$ possesses the minorant $\sum_{n\geq 1} u^n$ and the majorant $\sum_{n\geq 1}(n+1)u^n$ both with a radius of convergence equal to 1. Therefore the set of solutions of $|u| = 1$ might contribute further dominant singularities. However, $|u| = 1$ has only one solution with an appropriate modulus, namely $x = \frac{1}{2+4y}$. Thus we can conclude that there is only one dominant singularity. Since $t = -\log\left(\frac{1-\varepsilon}{1+\varepsilon}\right)$ and ε becomes 0 at our dominant singularity we expand $-\log\left(\frac{1-\varepsilon}{1+\varepsilon}\right)$ about $\varepsilon = 0$ to get $t \sim 2\varepsilon$ and thus $t \sim 2\frac{\sqrt{(1-2x)(1-2x-4xy)}}{1-2x-2xy}$. We conclude that for an expansion at $x = \frac{1}{2+4y}$, t complies with $2\frac{\sqrt{2y(1+2y)}}{y}\sqrt{1-x(2+4y)}$. On the assumption that y is constant and for $u = \exp(-t)$, the factor $\frac{y\sqrt{u}(1-u)}{u}$ possesses the expansion $yt + \mathcal{O}(t^2)$. So, the most significant term of the expansion of $M(x,y)$ around $x = \frac{1}{2+4y}$ is given by

$$\frac{yt\ln(t)}{2\ln(2)} \sim -\frac{\sqrt{2y(1+2y)}}{2\log(2)}\sqrt{1-x(2+4y)}\log((1-x(2+4y))^{-1}). \tag{7}$$

This representation can be used to approximate the coefficients of $M(x,y)$. We find:

Lemma 3. *Let $\rho := \frac{\alpha}{\beta}$ be fixed. The coefficient of $M(x,y)$ at $x^\alpha y^\beta$ is asymptotically given by*

$$\frac{2^{\alpha(1+\frac{1}{\rho})-2}}{\sqrt{\pi\alpha^3}}\binom{\alpha+\frac{1}{2}}{\frac{\alpha}{\rho}-\frac{1}{2}}\log_2(\alpha),$$

$\alpha \to \infty$.

Proof. We use the following well-known expansions

$$\begin{aligned}
\log((1-x(2+4y))^{-1}) &= \sum_{n\geq 1}\frac{x^n}{n}\sum_{k\geq 0}\binom{n}{k}2^{n-k}4^k y^k,\\
\sqrt{1-x(2+4y)} &= \sum_{i\geq 0}\binom{\frac{1}{2}}{i}(-1)^i x^i \sum_{k\geq 0}\binom{i}{k}2^{i-k}4^k y^k,\\
\sqrt{2y+4y^2} &= \sum_{j\geq 0}\binom{\frac{1}{2}}{j}2^{1-j}y^{1-j},
\end{aligned}$$

and extract the coefficient at $x^\alpha y^\beta$ in the resulting expansion of the right-hand side of (7). We find

$$[x^\alpha y^\beta]M(x,y) \sim -\frac{2^{\alpha+\beta-1}}{\log(2)}\binom{\alpha+\frac{1}{2}}{\beta-\frac{1}{2}}\underbrace{\sum_{n\geq 1}\frac{(-1)^{\alpha-n}}{n}\binom{\frac{1}{2}}{\alpha-n}}_{=:\sigma(\alpha)}.$$

By induction on α it is possible to prove the following recursion for $\sigma(\alpha)$:

$$\begin{aligned}
\sigma(0) &= 0,\\
\sigma(\alpha) &= \frac{2\alpha-3}{2\alpha}\sigma(\alpha-1)+\underbrace{\frac{(-1)^{\alpha+1}\sqrt{\pi}}{2\Gamma(\frac{5}{2}-\alpha)\Gamma(\alpha+1)}}_{=:\varsigma(\alpha)}.
\end{aligned}$$

This recursion can be solved by using ordinary generating functions. For $A(z) := \sum_{\alpha\geq 0}\sigma(\alpha)z^\alpha$ we get $A(z) = zA(z) - \frac{3}{2}\int_0^z A(t)dt + \sum_{\alpha\geq 1}\varsigma(\alpha)z^\alpha$. Applying the identity $\sum_{\alpha\geq 1}\varsigma(\alpha)z^\alpha = \frac{2}{3}\sqrt{1-z}(z-1)+\frac{2}{3}$, which for instance Zeilberger's "fast algorithm" (see [25]) finds for you, yields a simple differential equation for $A(z)$ which possesses the solution $A(z) = -\sqrt{1-z}\log(1-z)$ and thus

$$\sigma(\alpha) = [z^\alpha]\sqrt{1-z}\log((1-z)^{-1})$$

holds. By applying the $\mathcal{O}$-transfer method we find the approximation $\sigma(\alpha) \sim -\frac{\log(\alpha)}{2\sqrt{\pi\alpha^3}}$ which proves the lemma. □

In order to get the average value, the coefficient given in Lemma 3 has to be divided by the total number of $\mathcal{C}$-tries of size (α,β). We can proceed in the same way as done in the previous proof in order to approximate the coefficient $[x^\alpha y^\beta]T(x,y)$. Thus we factor $T(x,y)$ into

$$\frac{1-2x}{2x} - \frac{1-2x}{2x}\sqrt{1-4xy(1-2x)^{-1}} \tag{8}$$

which corresponds to an expansion of $T(x,y)$ around the singularity $y = \frac{1-2x}{4x}$ (assuming now that x is constant). Since the leftmost term of (8) only possesses coefficients at y^0 it can be neglected. Furthermore, $\sqrt{1-4xy(1-2x)^{-1}}$ can be expanded using the binomial theorem, which yields

$$\xi(\alpha,\beta) := [x^\alpha y^\beta]\sqrt{1-4xy(1-2x)^{-1}} = \binom{\frac{1}{2}}{\beta}\binom{\alpha-1}{\alpha-\beta}(-1)^\beta 4^\beta 2^{\alpha-\beta}.$$

Now, taking the factor $-\frac{1-2x}{2x}$ into account proves that $[x^\alpha y^\beta]T(x,y) \sim \xi(\alpha,\beta)$ $-\frac{1}{2}\xi(\alpha+1,\beta)$. Thus we conclude that

$$[x^\alpha y^\beta]T(x,y) \sim \frac{2^{\alpha-\beta+1}}{\beta}\binom{2\beta-2}{\beta-1}\binom{\alpha-1}{\alpha-\beta+1}.$$

Dividing the coefficient given in Lemma 3 by the previous quantity provides the following theorem:

Theorem 1. *On the assumption that all (α,β)-tries are equally likely the average Horton-Strahler number of a C-trie of size (α,β) is asymptotically given by*

$$\frac{4^{\alpha(\frac{2}{\rho}-1)}\log(\alpha)(2\frac{\alpha}{\rho}-1)(\alpha+1)(\alpha+2)\binom{2\alpha+1}{\alpha-1}}{8\sqrt{\pi}\alpha^{3/2}\log(2)(\frac{\alpha}{\rho}-1)\frac{\alpha}{\rho}\binom{2\frac{\alpha}{\rho}}{\frac{\alpha}{\rho}}^2},$$

$\rho := \frac{\alpha}{\beta}$ *fixed,* $\alpha \to \infty$.

Remark. The asymptotic given for the number of (α,β)-tries is equal to the exact number of $\mathcal{C}$-tries of this size for $\alpha > 0$. This is due to the fact that we find a factorization of $T(x,y)$ when expanding it around its dominant singularity. Thus, besides some terms at y^0, no terms were neglected when we have developed the leading term and have extracted the coefficients.

Looking at a plot of our average Horton-Strahler number of $\mathcal{C}$-tries (see the last section of this paper) it seems to be hardly dependent on β. This impression is justified when we use Stirling's formula to approximate the binomial coefficient $\binom{2\beta}{\beta}$ within our result. We find that the average Horton-Strahler number of $\mathcal{C}$-tries is asymptotically given by

$$\left(2+\frac{1}{\beta-1}\right)\frac{\log(\alpha)(\alpha+1)(\alpha+2)\sqrt{\pi}}{8\cdot 4^\alpha \alpha^{\frac{3}{2}}\log(2)}\binom{2\alpha+1}{\alpha-1}.$$

Thus, only for very sparse $\mathcal{C}$-tries, *i.e.* $\mathcal{C}$-tries with few white leaves only, we have an influence of β on the average Horton-Strahler number. But for every fixed ρ and $\alpha \to \infty$ also β tends to infinity and thus $\frac{1}{\beta-1}$ becomes zero. Thus it becomes possible to express the average Horton-Strahler number of (α,β)-tries on dependence of α only. We conclude this discussion by noting that

$$\lim_{\alpha\to\infty} \frac{(\alpha+1)(\alpha+2)\sqrt{\pi}}{4^{\alpha+1}\alpha^{\frac{3}{2}}}\binom{2\alpha+1}{\alpha-1} = \frac{1}{2}$$

holds. Thus we have the following corollary:

Corollary 1. *Under the assumption that the number of internal nodes α and the number of white leaves β grow in some fixed proportion the average Horton-Strahler number of (α,β)-tries is asymptotically given by*

$$\frac{\log(\alpha)}{2\log(2)}.$$

Remark. We can also conclude the result of the previous corollary by using the multivariate Darboux-method presented in [5] in order to approximate the coefficient of the leading term in (7) and the number of $\mathcal{C}$-tries of size (α,β). In that case, because of side-conditions given by the method, $\rho = \frac{\alpha}{\beta}$ has to be strictly larger that one. However, it is impossible to derive the more accurate results presented in Lemma 3 and Theorem 1 in that way.

We will now use our generating functions to derive an asymptotic equivalent for the average Horton-Strahler number for $\mathcal{C}$-tries of size α. As methodology is much more developed for univariate generating functions it is possible to derive results of higher precision. Note, that in the uniform model it is not possible to derive a univariate result with respect to the number of white leaves since all the generating functions would count infinitely many $\mathcal{C}$-tries of any given size β. This is due to the fact, that the number of white leaves does not limit the number of internal nodes even if we fix the Horton-Strahler number of the $\mathcal{C}$-tries considered. Thus, we return to (6) and set $y = 1$ in all parts of the generating function. It is obvious that we find the same integral as a representation of $\sum_{n\geq 1}(v_2(n)+1)u^n$ as in the bivariate case since the substitution $u = \exp(-t)$ yields the same result even if we set y to 1. But now it makes sense to consider terms of lower significance also, because the $\mathcal{O}$-transfer method for univariate generating functions makes it possible to translate them into the right contributions for the asymptotics in question. Therefore we sum the residues of the singularities at $s \in \{1, -n, \chi_k\}$, $n \in \mathbb{N}\cup\{0\}$, $k \in \mathbb{Z}\backslash\{0\}$, and multiply them by the expansion of the factor $\frac{y\sqrt{u}(1-u)}{u}|_{y=1}$ which gives us

$$2 + \left(\frac{2\ln(t) + 2\gamma - 2\ln(\pi) - 3\ln(2)}{4\ln(2)}\right)t + \sum_{k\neq 0}\frac{\Gamma(\chi_k)\zeta(\chi_k)}{\ln(2)}t^{1-\chi_k} + \mathcal{O}(t^2). \qquad (9)$$

Again, we are interested in an expansion at the dominant singularity which is $x = \frac{1}{2+4y}|_{y=1} = \frac{1}{6}$. Thus we have to set $t = 2\sqrt{6}\sqrt{1-6x}$ in order to resubstitute

t in (9) which yields the desired expansion:

$$-\frac{\sqrt{6}\sqrt{1-6x}\ln((1-6x)^{-1})}{2\ln(2)}+\frac{\sqrt{6}\sqrt{1-6x}(2\gamma-2\ln(\pi)+\ln(3))}{2\ln(2)}$$
$$+\sum_{k\neq 0}\frac{\Gamma(\chi_k)\zeta(\chi_k)}{\ln(2)}(2\sqrt{6})^{1-\chi_k}(1-6x)^{\frac{1-\chi_k}{2}}+\mathcal{O}(|1-6x|)$$

$$= -\frac{\sqrt{6}\sqrt{1-6x}\ln((1-6x)^{-1})}{2\ln(2)}+\frac{\sqrt{6}\sqrt{1-6x}(2\gamma-2\ln(\pi)+\ln(3))}{2\ln(2)}$$
$$+\frac{2\sqrt{6}}{\ln(2)}\sum_{k\neq 0}\Gamma(\chi_k)\zeta(\chi_k)e^{-\log_2(6)\pi i k}(1-6x)^{\frac{1-\chi_k}{2}}+\mathcal{O}(|1-6x|).$$

Now we can apply the transfer formulae according to [8]. We have

$$[z^n](1-z)^\alpha \sim \frac{n^{-\alpha-1}}{\Gamma(-\alpha)}+\mathcal{O}(n^{-\alpha-2})$$

and

$$[z^n](1-z)^{\frac{1}{2}}\ln((1-z)^{-1}) \sim -\frac{1}{\sqrt{\pi n^3}}\left(\frac{1}{2}\ln(n)+\frac{\gamma+2\log(2)-2}{2}+\mathcal{O}\left(\frac{\ln(n)}{n}\right)\right)$$

which provide the following lemma:

Lemma 4. *The coefficient of $M(x,1)$ at x^α is asymptotically given by*

$$\frac{6^{\frac{1}{2}+\alpha}(\ln(\alpha)-\gamma+2\ln(2)-2+2\ln(\pi)-\ln(3))}{4\cdot\alpha^{\frac{3}{2}}\sqrt{\pi}\ln(2)}$$
$$+\frac{2\sqrt{6}}{\ln(2)}\sum_{k\neq 0}\Gamma(\chi_k)\zeta(\chi_k)e^{-\log_2(6)\pi i k}6^\alpha\alpha^{\frac{\chi_k}{2}-\frac{3}{2}}/\Gamma\left(\frac{\chi_k}{2}-\frac{1}{2}\right)+\mathcal{O}\left(\alpha^{-\frac{5}{2}}\right).$$

Finally, this quantity has to be divided by the asymptotic number of α-tries which is known (see [20]) to be given by $6^{\frac{1}{2}+\alpha}\alpha^{-\frac{3}{2}}/(2\sqrt{\pi})+\mathcal{O}(\alpha^{-\frac{5}{2}})$. After numerous simplifications we find:

Theorem 2. *On the assumption that all C-tries of the same size are equally likely the average Horton-Strahler number of an α-trie is asymptotically given by*

$$\frac{1}{2}\log_2\left(\frac{4}{3}\pi^2\alpha\right)-\frac{\gamma+2}{2\ln(2)}+\Delta\left(\log_2\left(\frac{\alpha}{6}\right)\right)+\mathcal{O}(\alpha^{-1}),$$

$\chi_k = \frac{2\pi ik}{\ln(2)}$, $\alpha \to \infty$. *The function* $\Delta(x)$ *is a periodic function of small modulus (*$|\Delta(x)| < 0.041$*) and possesses the following representation as a Fourier series:*

$$\Delta(x) := \frac{1}{\ln(2)} \sum_{k \neq 0} (\chi_k - 1)\Gamma\left(\frac{\chi_k}{2}\right)\zeta(\chi_k)e^{\pi ikx}.$$

Remark. Note that this is exactly the same result as for non-colored extended binary trees given in (1) when setting α to 6α. The same effect with a different constant can be observed for the average stack-number. In that case we have to set α to $\frac{2}{3}\alpha$ in order to get the same leading term as for non-colored trees. The fact, that there are different constants for different parameters, supports a conjecture stated in [20] which says that it seems to be impossible to conclude the behavior of $\mathcal{C}$-tries with respect to "traversal-parameters" from the well know results for ordinary extended binary trees (*e.g.* by a simple rearrangement together with an appropriate weighting of the trees). The bound $|\Delta(x)| < 0.041$ can be found by means of numerical studies.

3. Visualization and conclusions

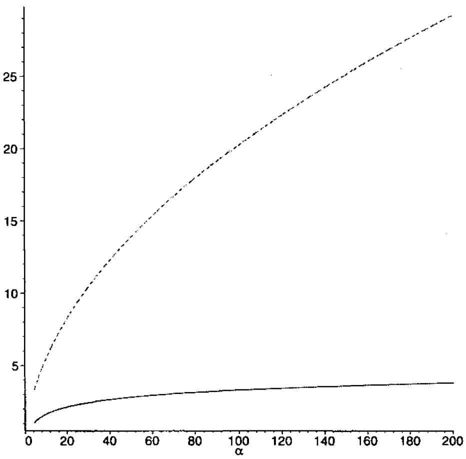

FIGURE 4. The average stack-number (upper graph) and Horton-Strahler number (lower graph) on dependence of the number of internal nodes α.

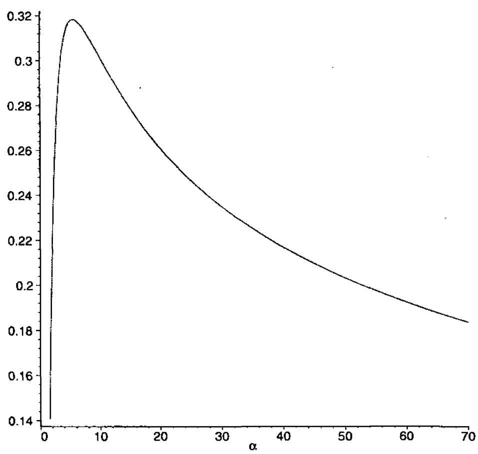

FIGURE 5. The ratio of the average Horton-Strahler number and the average stack-number.

In this section we will provide some plots of the results presented in [20], in [21] and in this paper. This is how we are going to compare the average stack-size with the average Horton-Strahler number which are related in the following way: if we think of applications of the Horton-Strahler number such as a tree traversal or the evaluation of an arithmetic expression, the stack-size of the corresponding tree describes the amount of space needed when we apply a usual preorder traversal or a simple traversal strategy for evaluation. Those methods can be optimized with respect to the amount of space needed by easing the restriction that subtrees must be visited in a fixed order. The space requirement of the resulting strategy is described by the Horton-Strahler number. Therefore, we will speak of an *economy of space* when comparing both parameters.

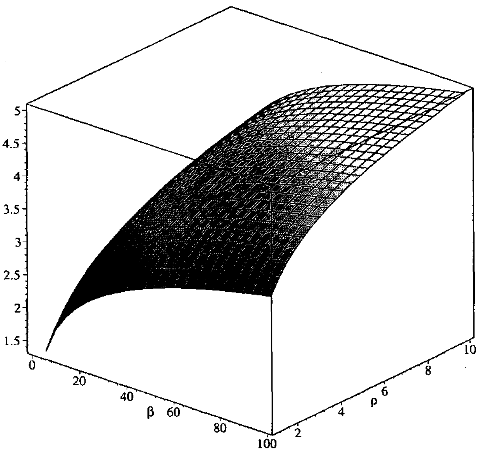

FIGURE 6. A plot of the average Horton-Strahler number on dependence of ρ and β.

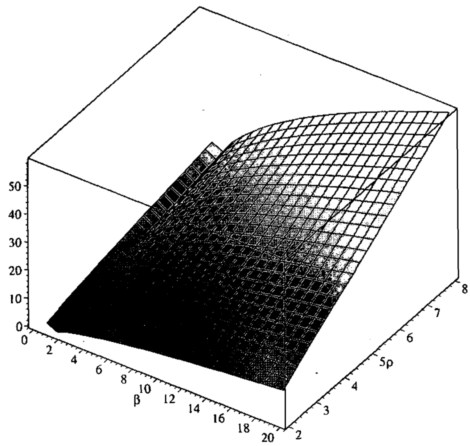

FIGURE 7. The difference of the average stack-number and the average Horton-Strahler number on dependence of ρ and β.

The first plot is presented in Figure 4. It shows the absolute values of the average stack-number and the average Horton-Strahler number. As we can see the order of growth of both graphs is quite different. Even if the total stack-number for $\mathcal{C}$-tries of size α is small, the relative economy of space implied by the application of the optimized algorithms related to the Horton-Strahler number is remarkable (as we can see in Fig. 5). In the bivariate setting a similar behavior can be found. If we take a look at Figure 6 we see that the average Horton-Strahler number for $\mathcal{C}$-tries of size (α, β) grows slowly and is of small value even for *sparse* $\mathcal{C}$-tries, *i.e.* for $\mathcal{C}$-tries with a large internal structure but only a few white leaves. As we would have expected from the univariate case, Figure 7 shows that the economy of space grows the larger the (α, β)-tries become. We can also observe that the advantage of the optimized algorithms gets larger with ρ growing and not only the relative but also the total economy of space get large when the $\mathcal{C}$-tries become sparse. For example on the assumption of $\rho = 8$ the total economy of space is about 60 for (α, β)-tries with only 20 white leaves.

The author would like to thank the anonymous referees for their comments and criticism which helped to improve the quality of the paper. I am also indebted to one of the referees for suggesting the notions α-trie and (α, β)-trie respectively.

References

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions.* Dover (1970).
[2] T.M. Apostol, *Introduction to Analytic Number Theory.* Springer (1976).
[3] N.G. de Bruijn, D.E. Knuth and S.O. Rice, *The Average Height of Planted Plane Trees.* Graph Theory and Computing, edited by R.C. Read. Academic Press (1972).
[4] L. Devroye and P. Kruszewski, On the Horton-Strahler Number for Random Tries. *Theoret. Informatics Appl.* **30** (1996) 443-456.
[5] M. Drmota, Asymptotic Distributions and a Multivariate Darboux Method in Enumeration Problems. *J. Combin. Theory Ser. A* **67** (1994) 169-184.
[6] A.P. Ershov, On Programming of Arithmetic Operations. *Comm. ACM* **1** (1958) 3-6.
[7] P. Flajolet, J.C. Raoult and J. Vuillemin, The Number of Registers required for Evaluating Arithmetic Expressions. *Theoret. Comput. Sci.* **9** (1979) 99-125.
[8] P. Flajolet and A. Odlyzko, Singularity Analysis of Generating Functions. *SIAM J. Discrete Math.* **3** (1990) 216-240.
[9] P. Flajolet and H. Prodinger, Register Allocation for Unary-Binary Trees. *SIAM J. Comput.* **15** (1986) 629-640.
[10] P. Flajolet, X. Gourdon and P. Dumas, Mellin transforms and asymptotics: Harmonic sums. *Theoret. Comput. Sci.* **144** (1995) 3-58.
[11] J. Françon, Sur le nombre de registres nécessaires à l'évaluation d'une expression arithmétique. *Theoret. Informatics Appl.* **18** (1984) 355-364.
[12] R.E. Horton, Erosioned development of systems and their drainage basins, hydrophysical approach to quantitative morphology. *Bull. Geol. Soc. of America* **56** (1945) 275-370.
[13] R. Kemp, The Average Number of Registers Needed to Evaluate a Binary Tree Optimally. *Acta Inform.* **11** (1979) 363-372.
[14] R. Kemp, A Note on the Stack Size of Regularly Distributed Binary Trees. *BIT* **20** (1980) 157-163.
[15] R. Kemp, *Fundamentals of the Average Case Analysis of Particular Algorithms.* Wiley-Teubner Series in Computer Science (1984).
[16] R. Kemp, On the Stack Ramification of Binary Trees. *Random Graphs* **2** (1992) 117-138.
[17] D.E. Knuth, *The Art of Computer Programming*, Volume 1: Fundamental Algorithms, 3rd Ed. Addison-Wesley (1997).
[18] A. Meir, J.W. Moon and J.R. Pounder, On the Order of Random Channel Networks. *SIAM J. Algebraic Discrete Math.* **1** (1980) 25-33.
[19] M.E. Nebel, New Results on the Stack Ramification of Binary Trees. *J. Autom. Lang. Comb.* **2** (1997) 161-175.
[20] M.E. Nebel, The Stack-Size of Tries, A Combinatorial Study. *Theoret. Comput. Sci.* (to appear).
[21] M.E. Nebel, *The Stack-Size of Uniform Random Tries Revisited* (submitted).
[22] A.N. Strahler, Hypsometric (area-altitude) analysis of erosonal topology. *Bull. Geol. Soc. of America* **63** (1952) 1117-1142.

[23] X.G. Viennot, G. Eyrolles, N. Janey and D. Arques, Combinatorial analysis of ramified patterns and computer imagery of trees, in *Proc. SIGGRAPH'89, Computer Graphics* **23** (1989) 31-40.
[24] X.G. Viennot, Trees Everywhere, in *Proc. CAAP'90.* Springer, New York, *Lecture Notes in Comput. Sci.* **431** (1990) 18-41.
[25] D. Zeilberger, The method of creative telescoping, *J. Symbolic Comput.* **11** (1991) 195-204.

Communicated by J. Berstel.
Received February 15, 2000. Accepted October 3, 2000.