New Strahler numbers for rooted plane trees

D. Auber,M. Delest,J.P. Domenger,P. Duchon and J.M. Fédou

Abstract. In this paper, we present an extension of Strahler numbers to rooted plane trees. Several asymptotic properties are proved others are conjectured. We also describe several applications of this extension.

1. Introduction

The Strahler number of binary trees was introduced by the hydrogeologist Horton [12], then refined by Strahler [19] and rediscovered in computer science by Ershov [7]. Numerous results have been obtained from which one can point out explicit expressions for their distribution [9], [13] and links with biology [24], [6]. In enumerative combinatorics, the research mostly involves bijections preserving Strahler numbers between types of objects in order to understand their generating series : between binary trees and Dyck paths by Francon [11], between binary trees and forests of plane trees by Zeilberger [27] and more recently between forests of plane trees and Dyck paths by Viennot [23]. The common construction among these results is to redefine the Strahler number as a *pruning number* on rooted plane trees.

A natural question arises, is there a *natural* definition of Strahler numbers for rooted plane trees that extends Ershov's interpretation [7]? This paper proposes such an extension and describes some related properties. This extension is different from the one described and studied on general tries by Nerbel [17] or Bourdon and *al.* [5].

Apart from biology and hydrogeology fields, Strahler numbers have been used in computer graphics to give synthetic images of trees and landscapes (see Viennot, Eyrolles, Janey and Arques [22]), in Physics by Vannimenus and Viennot [20]. A survey paper can be found in [21]. Thus, Strahler numbers have been proved efficient in many fields. It is natural to apply them to information visualisation systems that deal with trees and graphs. Here, we present two applications of our definition : navigation [2] and research of "quasi-similar" subtrees between two trees [4].

After some definitions, we prove that the computation of these Strahler numbers for all vertices of a tree of size n can be performed in $\Theta(n)$ in time and space. Then we proves that the branching ratio of trees for any family of finite degree (in the sense of meir and Moon [15, 14] is 4 like it is for binary trees [18, 16]. We end with a brief description of the applications to information visualization systems.

2. Definitions and notations

A tree is a connected acyclic graph. A rooted tree is defined from a tree by choosing a vertex called the root, its edges are oriented such that there exists a path from the root to each vertex of the tree. A vertex with no outgoing edge is called a leaf, other vertices are called internal vertices. A successor of a vertex is called a child of this vertex. The degree of a vertex is the number of its children. In this paper, we will consider rooted plane trees i.e. rooted trees where a total order is imposed on the children of each vertex. A binary tree is a rooted tree such that each vertex has zero or two successors. A d-ary tree is such that the degree of its vertices is 0 or (d). Thus a binary tree is a 2-ary one. Let D a finite set of non negative integers containing 0. We consider the set of simply generated trees, $\mathcal{T}_D$, in the sens of Meir and Moor [15, 14] : each vertex of a tree has its arity in D.

The Strahler numbers σ_b were first introduced on binary trees in some work about the morphological structure of river [12, 19]. The recursive definition associates an integer value to each vertex of a binary tree. These values give quantitative information about the complexity of each sub-tree of the original tree. Let s be a vertex of a tree :

- If s is a leaf then $\sigma_b(s) = 1$,
- Else s has two children s_1 and s_2 and

$$\sigma_b(s) = \begin{cases} \max(\sigma_b(s_1), \sigma_b(s_2)) \text{ if } \sigma_b(s_1) \neq \sigma_b(s_2) \\ \sigma_b(s_1) + 1 \text{ otherwise} \end{cases}.$$

The Strahler number of a tree is the integer value associated to its root. If T is a tree, $\sigma_b(T)$ will denote the Strahler number of T.

Definition 2.1. Let T be a tree whose vertices are valued by Strahler numbers. A *branch of Strahler order* k is a maximal path $(s_0, s_1, \ldots, s_p)$ in T, such that for each $i \in [0..n]\,, \sigma_b(s_i) = k$.

We denote by $\beta_k(T)$, the total number of branches in T of Strahler order k.

Theorem 2.2. [18, 16] *Let $B_{k,n}$ be the total number of branches of Strahler order k in the set of binary trees with size $2n+1$, then*

$$\lim_{n\to+\infty} \left(\frac{B_{k,n}}{B_{k+1},n}\right) = 4.$$

This ratio is called the branching ratio of binary trees. Horton has shown [12] that for real rivers this ratio is between 3 and 5. A generalization of Strahler numbers to rooted trees is suggested by the nice interpretation by Ershov [7] who proved that the Strahler number of the root of the binary tree is exactly the minimal number of registers needed to compute an arithmetic expression given by the tree (output by the syntactical analysis). Thus, one can define the Strahler numbers σ on general trees by:

- If s is a leaf then $\sigma(s) = 1$,

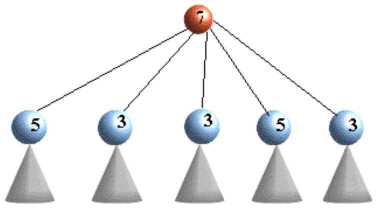


FIGURE 1. Strahler number on one vertex.

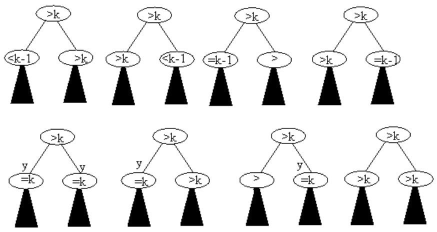


FIGURE 2. Construction of trees of $G_{\{2\}}$.

- Else s has $k+1$ children s_i such that $\sigma(s_i) \leq \sigma(s_j)$, if $i \leq j$ (i.e. children are reordered for the Strahler computation) and
$$\sigma(s) = \underset{0 \leqslant i \leqslant k}{Max}(\sigma(s_i) + i)\,.$$

See an example in Figure 1. As for binary trees, the Strahler number of a tree is the Strahler number of its root. Note that on the set of binary trees, $\mathcal{T}_{\{0,2\}}$, σ and σ_b coincide. The definition of the branching ratio according to the generalization stays unchanged. We conjecture that the theorem 2.2 holds for the class of plane rooted trees. Note that if an internal vertex has a Strahler value s then it has at most s children.

3. Complexity of Strahler number computation

Strahler numbers are useful in Information visualization as we will show in Paragraph 6. In this field, the effective size of the trees is large : 500000 vertices is average. Thus, the algorithm that computes Strahler numbers must be efficient. In this paragraph we prove that our algorithm is linear. In fact this theorem is a consequence of the following lemma

Lemma 3.1. *Let T be a tree and $E(T)$ denote the set of strahler values of vertices in T. Then*

$$\sum_{k \in E(T)} k \leqslant |T| \tag{3.1}$$

Proof. We prove the lemma by induction on the depth of trees. If T has depth 1, then its root has some number k of children, and Strahler number k. Thus equality occurs in (3.1), unless $k = 1$, in which case the inequality is strict.

Now assume (3.1) holds for all trees with depth at least h, and consider a tree T with depth $h+1$. Assume the root of T has p children, which are the roots of subtrees $T_1, \ldots, T_p$ (with decreasing Strahler values). If $\sigma(T) = \sigma(T_1)$, then there is nothing to prove since $E(T) = \bigcup_{i=1..p} E(T_i)$. Thus, we assume $\sigma(T) = k > \sigma(T_1)$. By the recursive definition of the Strahler numbers, this ensures there is some integer i, $2 \leq i \leq p$, such that $\sigma(T_1) \geq k - i + 1$. Thus, we have

$k - 1 \geq \sigma(T_1) \geq \ldots \geq \sigma(T_i) \geq k - i + 1$, so that at least two of the subtrees T_j, $j \leq i$, have the same Strahler value $k' \geq k - p + 1$. Also, the value 1 must certainly appear in each of the sets $E(T_i)$, since all leaves have Strahler value 1.

- if $k' > 1$, then we have $\sum_{1 \leqslant i \leqslant p} \sum_{l \in E(T_i)} l \geqslant k' + p - 1 + \sum_{l \in \bigcup_{1 \leqslant i \leqslant p} E(T_i)} l$ which, by induction, proves that inequality in (3.1) is strict,
- if $k' = 1$, then $k = p$, and we have $\sum_{1 \leqslant i \leqslant p} \sum_{l \in E(T_i)} l \geqslant 1 + \sum_{l \in \bigcup_{1 \leqslant i \leqslant p} E(T_i)} l$ which also proves(3.1) by induction; equality occurs if and only if it occurs in each subtree T_i.

□

From this lemma and the fact that any set of k non negative integers has a sum at least $\binom{k}{2}$, we deduce :

Corollary 3.2. *Let T be a tree then $|E(T)| \leq \sqrt{2|T|}$.*

A bijective proof of this corollary can be found in [1]. The number of different values of Strahler numbers in a given tree T has a sub-linear upper bound and moreover these values are in $[0..N]$. Thus the complexity of the sorting the children of all vertices becomes linear and we have

Theorem 3.3. *Let T be a tree of size n. The valuation σ on all vertices of T can be computed in linear time and space.*

4. Branching ratio for binary trees

In this paragraph, we will prove the Theorem 2.2 [18, 16] in order to present the method that we use. Because D is fixed, we will omit it in all the formulas. Let us define the following sets :

$$\mathcal{S}_k = \{t \in \mathcal{T}_D : \sigma(t) = k\}, \mathcal{L}_k = \{t \in \mathcal{T}_D : \sigma(t) < k\}, \mathcal{G}_k = \{t \in \mathcal{T}_D : \sigma(t) > k\}.$$

Note that the following relations hold

$$\mathcal{T}_D = \mathcal{S}_k \cup \mathcal{L}_k \cup \mathcal{G}_k, \mathcal{S}_k \cup \mathcal{G}_k = \mathcal{G}_{k-1}, \mathcal{L}_k \cup \mathcal{G}_{k-1} = \mathcal{T}_D.. \tag{4.1}$$

Moreover, there is no branch of order k in L_k and only one reduced to one vertex in S_k. Then, consider the series

$$T_D(x) = \sum_{t \in \mathcal{T}_D} x^{|t|}, S_k(y,x) = y \sum_{t \in \mathcal{S}_k} x^{|t|}, L_k(y,x) = \sum_{t \in \mathcal{L}_k} x^{|t|}, G_k(y,x) = \sum_{t \in \mathcal{G}_k} y^{\beta_k(t)} x^{|t|}.$$

The method classically consists in the following steps.

- Write out the equation for $G_k(y, x)$.
- Consider the derivative of $G_k(y, x)$ according to y, we have

$$\frac{\partial G_k(y,x)}{\partial y}|_{y=1} = \sum_{t \in \mathcal{G}_k} \beta_k(t) x^{|t|},$$

Thus for the class of the family $\mathcal{T}_D$,

$$B_{k,n} = \left(\frac{\partial G_k(y,x)}{\partial y} |_{y=1} \ , x^n \right).$$

- Find an asymptotic expression for

$$\frac{\partial G_k(y,x)}{\partial y} |_{y=1} \ .$$

- Compute the limit of the branching ratio.

In what follow, we will omit the x and y variables in the series as soon as there is no ambiguity. Let us now see what happens on the set of binary trees. In this case, $D = \{0, 2\}$. In the first step, write out the equation.

$$G_k = 2L_{k-1}G_k + 2S_{k-1}G_k + y^2 S_k^2 + 2yS_kG_k + G_k^2.$$

This equation means that one can construct trees with Strahler numbers higher than k in several manners which corresponds to the monomials (see figure 2). We add an $''y''$ in the monomial each time that a branch of order k ends. Thus, at the end of the process, there is as many $''y''$ in a monomial coding a given tree than the number of branch of Strahler order k. Thus, in a second step, we derivate and get

$$\frac{\partial G_k(y,x)}{\partial y} |_{y=1} \ = \frac{G_{k-1}(1,x)\,(G_{k-1}(1,x) - G_k(1,x))}{1 - 2T_D(x)}. \tag{4.2}$$

Note that the main singularity is $x = x_0$ such that $T_D(x) = 1/2$. Thus if we want to compute an asymptotic value for $b_{k,n}$, we need to compute the value of $G_k(1, x_0)$, this is the step 3. Here, we have

$$G_k(1,x) = \frac{G_{k-1}(1,x)^2}{1 - 2T_D(x) + 2G_{k-1}(1,x)}.$$

Using relations 4.1, the step 4 is straightforward and thus the branching ratio is of binary trees is 4.

5. Branching ratio for trees of $\mathcal{T}_D$ with finite D

In this section, we assume that D is any finite set of integers that contains 0 and at least one integer larger than 1. The family of trees we are interested in is the set $\mathcal{T}_D$ of plane rooted trees where the arity of each node lies in D. The generating function for all trees in $\mathcal{T}_D$ thus satisfies the polynomial equation

$$F(x) = x \sum_{d \in D} F(x)^d = x\Phi_D(F(x)). \tag{5.1}$$

It is a classical result [8, 10] that this series converges as an analytic function inside the complex disc $|x| \leq \rho_D$, where $\rho_D = \tau/\Phi_D(\tau)$ and τ is the unique positive real solution to the equation $\Phi_D(\tau) = x\Phi'_D(\tau)$. Furthermore, $F(x)$ has a single[1] dominant singularity of the square root type at ρ_D, and this translates into an asymptotic expression $a_n \sim C.\rho^{-n} n^{-3/2}$ for the number of D-trees of size n.

For each integer $k > 1$, $\mathcal{T}_D$ can be partitioned into 3 sets $\mathcal{S}_k, \mathcal{G}_k, \mathcal{L}_k$ with respective generating functions $S_k(x)$, $G_k(x)$ and $L_k(x)$. We will prove that S_k and L_k are both *rational* power series. Since rational functions only have poles as their singuldegrees, and F remains finite at its dominant singuldegrees, this proves that all poles of S_k and L_k must have moduli strictly larger than ρ_D, so that *each* G_k *has a square root singularity at* ρ_D, with the same amplitude as F. This can be interpreted as meaning that, for finite k, *almost all large trees have Strahler number more than* k.

5.1. Generating functions for trees of high Strahler value

In this paragraph, we write explicit equations for $G_k(x)$. If a vertex in a tree has d children, then, if we only want to discriminate its Strahler number so as to be able to decide whether it is strictly larger than some value k, we only need to discriminate whether the Strahler value of each child is larger than k, or one of the values between $k-d+2$ and k, or smaller than $k-d+2$ – all values lower than $k-d+2$ are equivalent in this regard, because even d children each with Strahler value $k-d+1$ will only result in a Strahler value k for the root node. This means that the generating function for all trees where the root has exactly d children and Strahler value larger than k, can be expressed as x times a *polynomial* $\mathcal{Q}_d$ in the variables $G_k, S_k, \ldots, S_{k-d+2}, L_{k-d+2}$. This polynomial $\mathcal{Q}_d$ can be obtained by expanding (formally) the expression $(G_k + S_k + \cdots + S_{k-d+2} + L_{k-d+2})^d$ into a sum of monomials, and, in the resulting sum, selecting only those monomials that lead to a Strahler number strictly over k. Thus , we would obtain a polynomial expression for S_k in terms of itself and a finite set of S_i and L_i; since this polynomial is linear in S_k, this would solve to a rational expression in terms of those lower order series, which proves that each $S_k(x)$ is a rational series.

By performing the global change of variables $S_i = G_{i-1} - G_i$, $L_i = F - G_{i-1}$, we can also express $\mathcal{Q}_d$ as a polynomial $\mathcal{P}_d$ in the variables $G_k, \ldots, G_{k-d+1}, F$; this will later yield somewhat simpler equations. Note that, contrary to $\mathcal{Q}_d$, $\mathcal{P}_d$ does not have nonnegative coefficients. Once this is done, summing over all possible degrees for the root of a D-tree yields the following equation for the series $G_k(x)$:

$$G_k(x) = x \sum_{d \in D} \mathcal{P}_d(F(x), G_k(x), \ldots, G_{k-d+1}(x)) \tag{5.2}$$

(with the provision that $G_i(x) = F(x)$ whenever $i \leq 0$)

We will now prove a recurrence relation between the polynomials $\mathcal{P}_d$ themselves, which will enable us to prove asymptotic results on the branching ratios

---

[1]In fact, there is a singularity at each complex value of the form $\rho_D.e^{2ik\pi/d'}$, where d' is the greatest common divisor of elements of D; when $d' > 1$, the series $F(x)/x$ is invariant upon the change of variable $x \mapsto xe^{2i\pi/d'}$, and D-trees always have a size of the form $d'k+1$ for some k. The proofs in this paper assume that $d' = 1$, so as to make notations easier; they can be extended to the general case easily.

of any simply generated family of trees with a *finite* set of degrees allowed. We conjecture that similar results should hold for infinite sets of degrees (or at least a wide class of them), but were not able to prove them.

To avoid confusion between the generating functions $F(x)$, $G_k(x)$, $S_k(x)$ and $L_k(x)$ and the variables of polynomials $\mathcal{Q}_d$ and $\mathcal{P}_d$, we will use lowercase letters for the latter, and write

$$\mathcal{Q}_d = \mathcal{Q}_d(g, s_0, \dots, s_{d-2}, \ell), \mathcal{P}_d = \mathcal{P}_d(f, g_0, \dots, g_{d-1})$$

We have :

$$\mathcal{Q}_1(g, \ell) = g, \mathcal{Q}_2(g, s_0, \ell) = (g + s_0)^2 + 2g.\ell,$$

or, equivalently,

$$\mathcal{P}_1(f, g_0) = g_0, \mathcal{P}_2(f, g_0, g_1) = g_1^2 - 2g_0g_1 + 2g_0.f,$$

For a given value of k, we obtained $\mathcal{P}_k$ through a computer algebra system. These suggest a recurrence for polynomials :

Lemma 5.1. *For any $d \geq 2$,*

$$\mathcal{P}_d(f, g_0, \dots, g_{d-1}) = g_{d-1}^d + d \int_{g_{d-1}}^{f} \mathcal{P}_{d-1}(t, g_0, \dots, g_{d-2}) dt \tag{5.3}$$

and, equivalently,

(5.4)

$$\mathcal{Q}_d(g, s_0, \dots, s_{d-2}, \ell) = (g + s_0 + \dots + s_{d-2})^d + d \int_{s_{d-2}}^{s_{d-2}+\ell} \mathcal{Q}_{d-1}(g, s_0, \dots, s_{d-3}, t) dt$$

Proof. We prove (5.4); (5.3) is equivalent under the previously mentioned change of variables. Recall that $\mathcal{Q}_d$ is the enumeration polynomial for the ways a tree with Strahler value higher than k can be constructed with a root and d subtrees whose Strahler values are higher than k (counted by the variable g), or any value between k (counted by s_0) and $k-d+2$ (counted by s_{d-2}), or lower than $k-d+2$ (counted by ℓ).

Now consider the set of words $\mathcal{L}_d$ of length d over the alphabet $\{g, k, k-1, \dots, k-d+2, \ell\}$ (where g stands for "higher than k", and ℓ stands for "lower than $k-d+2$"), that, if they are interpreted as the sequence of Strahler values for the d children of a node, result in this node having Strahler value higher than k. $\mathcal{Q}_d$ is none other than the commutative image of $\mathcal{L}_d$ when each letter $k-i$ is mapped to the variable s_i. What we need to do is provide a description of $\mathcal{L}_d$ in terms of $\mathcal{L}_{d-1}$ that, under commutation, can be interpreted as (5.4).

First, note that all words of length d where ℓ does not appear, belong to $\mathcal{L}_d$ (having d children, each with Strahler value at least $k-d+2$, is enough for a vertex to have Strahler value at least $k+1$). This justifies the term $(g + s_0 + \dots + s_{d-2})^d$ in (5.4).

Now consider a word $w \in \mathcal{L}_d$, with $j+1$ total occurrences of ℓ or $k-d+2$ ($j \geq 0$), and the (multi-)set of $j+1$ words of length $d-1$ obtained by first replacing each $k-d+2$ with an ℓ, then removing one of the ℓ. It is easy to see that each of these words belongs to $\mathcal{L}_{d-1}$. Inversely, for any word $w' \in \mathcal{L}_{d-1}$ with j occurrences of ℓ, the (multi-)set of $d(2^j-1)$ obtained by first inserting an additional ℓ in any of d positions in w', then replacing each ℓ with either itself or $k-d+2$ – but leaving at least one occurrence of ℓ, since words in $\mathcal{L}_d$ without ℓ are already accounted for - will produce only words of $\mathcal{L}_d$, each word with $j+1$ total occurrences of $k-d+2$ or ℓ being obtained $j+1$ times. Letting all letters commute, and summing over all words of $\mathcal{L}_{d-1}$, we see that each monomial $M\ell^j$ in $\mathcal{Q}_{d-1}$ becomes $\frac{d}{k+1}M((\ell+s_{d-2})^{j+1}-s_{d-2}^{j+1})$ in $\mathcal{Q}_d$, which is exactly symbolic integration with respect to ℓ on the interval $[s_{d-2}, s_{d-2}+\ell]$. Summing all contributions, we get (5.4). □

Corollary 5.2. *For any $d \geq 1$,*

$$\frac{\partial \mathcal{Q}_d}{\partial g}(g, s_0, \ldots, s_{d-2}, \ell) = d.\,(g+s_0+\cdots+s_{d-2}+\ell)^{d-1} \tag{5.5}$$

We now turn to the asymptotics of the series $G_k(x)$, for an arbitrary (finite) set of degrees D. Recall that the generating function $G_k(x)$, counting all D-trees with Strahler value larger than k according to their sizes, satisfies the general equation 5.2 and the initial condition $\mathcal{P}_0 = 0$.

It is easy to check that, for any $d > 0$, $\mathcal{P}_d$ is homogenous in its $d+1$ variables, with total degree d, degree $d-1$ in the variable f, and degree 1 in the variable g_0. This is important, because solving for $G_k(x)$ will thus result in a *rational* expression in the variables x, $F(x)$, and $G_{k'}(x)$ with $k-\max(D) < k' < k$; more precisely, if we write

$$\mathcal{P}_d = g_0\mathcal{D}_d + \mathcal{N}_d$$

where $\mathcal{D}_d$ and $\mathcal{N}_d$ are both polynomials not involving the variable g_0, then

$$G_k(x) = \frac{x\sum_{d\in D}\mathcal{N}_d(F(x), G_{k-1}(x), \ldots, G_{k-d+1}(x))}{1-x\sum_{d\in D}\mathcal{D}_d(F(x), G_{k-1}(x), \ldots, G_{k-d+1}(x))} \tag{5.6}$$

Now, as k goes to infinity, $G_k(x)$ goes to 0, uniformly in the variable x for $|x| \leq \rho_D$ (since, for a tree to have Strahler value larger than k, it is necessary for it to have at least $k+1$ vertices; the uniformity in x comes from the fact that these are power series in x with nonnegative coefficients, that converge for $x = \rho_D$). The only generating function that does not vanish in (5.6) when k goes to infinity is $F(x)$, so, to obtain asymptotic information on $G_k(x)$, we need to extract the dominant coefficients (with respect to the variable f) in $\mathcal{N}_d$ and $\mathcal{D}_d$. For reasons that will become clear later, we actually want the two dominant terms in $\mathcal{D}_d$, while the single dominant term in $\mathcal{N}_d$ will be sufficient.

From the recurrence (5.3) and the initial conditions $\mathcal{N}_2 = g_1^2$, $\mathcal{D}_2 = 2f - 2g_1$, it is easy to check the following lemma:

Lemma 5.3. *For any $d \geq 2$,*

$$\mathcal{N}_d(f, g_1, \ldots, g_{d-1}) = g_{d-1}^d + d\int_{g_{d-1}}^{f} \mathcal{N}_{d-1}(t, g_1, \ldots, g_{d-2})dt \tag{5.7}$$

$$\mathcal{D}_d(f, g_1, \ldots, g_{d-1}) = d\int_{g_{d-1}}^{f} \mathcal{D}_{d-1}(t, g_1, \ldots, g_{d-2})dt \tag{5.8}$$

$$\mathcal{N}_d = \binom{d}{2} g_1^2 f^{d-2} + O\left(f^{d-3}\right) \tag{5.9}$$

$$\mathcal{D}_d = df^{d-1} - d(d-1)g_1 f^{d-2} + O\left(f^{d-3}\right) \tag{5.10}$$

When plugging these expressions into (5.6), the numerator will always vanish as k goes to infinity (because the numerator always has positive total degree in the g_i variables), whereas the denominator has a term of null degree in the variables g_i, namely, $1 - x\sum_{d\in D} df^{d-1}$. Now notice that this constant term, where f is replaced by $F(x)$, will vanish exactly when $x = \rho_D$ (this is the equation defining the singular behaviour of the generating function $F(x)$). Thus, the "leading" terms (in powers of f) will cancel each other, and we turn to the second leading terms. But, for *each* value of d, the leading term in $\mathcal{N}_d$ and the second leading term in $-\mathcal{D}_d$ are in a ratio that is *exactly* $g_1/2$.

This suggests that $G_k(\rho_D)/G_{k-1}(\rho_D)$ will tend to $1/2$ as k goes to infinity. In fact, all that remains to do in order to prove this, is to prove that the remaining terms in both the numerator and denominator of the right-hand side of (5.6) (all of which have higher total degree in the variables g_i than the dominant terms we have extracted) will never become significant compared to those of higher degree in f. This is the point of the next lemma, for which we unfortunately were unable to prove a replacement when D is an infinite set of degrees.

Lemma 5.4. *Let D be a finite set of allowed degrees, $0 \in D$, $D \not\subset \{0, 1\}$, and let $d' = \max D$. Then, for any $k \geq 1$,*

$$G_k(\rho_D) \geq \frac{1}{d'} G_{k-1}(\rho_D).$$

Corollary 5.5. *For any finite D, let γ_k (respectively, σ_k) denote the singular value of G_k: $\gamma_k = G_k(\rho_D)$ (respectively, $\sigma_k = S_k(\rho_D)$). Then,*

$$\lim_{k\to\infty} \frac{\gamma_k}{\gamma_{k-1}} = \frac{1}{2}, \lim_{k\to\infty} \frac{\sigma_k}{\gamma_k} = 1, \lim_{k\to\infty} \frac{\sigma_k}{\sigma_{k-1}} = \frac{1}{2}$$

5.2. The branching ratio

We now turn to the task of evaluating the branching ratio for the Strahler numbers; that is, we must compare the total numbers, in D-trees of size n, of branches of Strahler order k and $k-1$. We will prove that, for any integer $k \geq 2$, this ratio tends to 4 as n tends to infinity. To do this, we consider the bivariate generating function

$$F_k(x, y) = \sum_{T\in\mathcal{T}_D} x^{|T|} y^{\beta_k(T)}.$$

Clearly, $F_k(x,1) = F(x)$. Using standard generating function manipulations and notations, the total number of branches of order k in D-trees of size n is exactly

$$[x^n]\frac{\partial F_k}{\partial y}(x,1).$$

We have $F_k(x,y) = L_k(x) + yS_k(x) + G_k(x,y)$. Thus

$$G_k(x,y) = \sum_{d\in D} \mathcal{Q}_d(G_k(x,y), yS_k(x), S_{k-1}(x), \ldots, S_{k-d+2}(x), L_{k-d+2}(x)). \tag{5.11}$$

Note that (5.11) proves that $G_k(x,y)$, and by extension $F_k(x,y)$, are *algebraic* series in their two variables, since the various S_i and L_i series are all rational. If we differentiate (5.11) with respect to y and set $y = 1$, we get an equation which we can immediately solve for $\Gamma_k(x) = \partial F_k/\partial y(x,1)$:

$$\Gamma_k(x) = xS_k(x)\frac{\sum_{d\in D}\frac{\partial \mathcal{Q}_d}{\partial s_0}(G_k(x,1), S_k(x), \ldots, S_{k-d+2}(x), L_{k-d+2}(x))}{1 - x\sum_{d\in D}\frac{\partial \mathcal{Q}_d}{\partial g}(G_k(x,1), S_k(x), \ldots, S_{k-d+2}(x), L_{k-d+2}(x))} \tag{5.12}$$

The denominator in (5.12) is exactly $1-x\sum_{d\in D} dF(x)^{d-1} = F(x)/(xF'(x))$. Since this vanishes at each dominant singularity of $F(x)$, and the numerator in (5.12) has a finite, nonzero limit, using a *transfer lemma* [8, 10], we get the following estimates :

- $B_{k,n} \sim b_k \rho_D^{-n} n^{-1/2}$ for some positive constant b_k;
- the branching ratio b_k/b_{k+1} is equal to C_k/C_{k+1}, which is the limit, when x goes to ρ_D, of the ratio $\Gamma_k(x)/\Gamma_{k+1}(x)$.

As a by-product, we also learn that the expected number of branches of Strahler order k, in a random D-tree of size n, is asymptotically $\alpha_k.n$; the branching ratio is just the ratio of α_k/α_{k+1}.

Thus, our task is reduced to that of estimating the limit, when k goes to infinity, of $\lim_{x\to\rho_D^-}\frac{\Gamma_k(x)}{\Gamma_{k+1}(x)}$ which has two factors. Each one has 2 as its limit then this concludes the proof the following theorem

Theorem 5.6. *The branching ratio of trees for any family of finite degree D is 4.*

6. Applications to information visualization

Originally, Strahler numbers were designed to highlight river shapes. The extension that we have given is in the same spirit : highlightening information from tree structures. We show here two results. One concerns navigation tools [2] and, the other file system retrieval [4]. When one visualize large trees on a screen, displaying the elements (vertices and edges) can take more than fifty milliseconds (δ_0). In order to have a high performance visualization system, one needs to accept events coming from the user (mouse movements, keyboard actions) within (δ_0). In order to increase performance, the method proposed by Wills [25] consists in predicting the number of elements that can be displayed during δ_0. Thus the system incrementally displays the tree (by part associated to δ_0) and ensures that the time reaction of the system is less than δ_0 without any request to the clock.

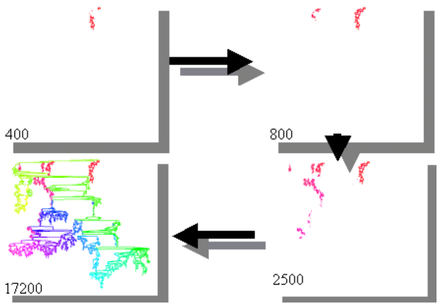


FIGURE 3. Display without Strahler.

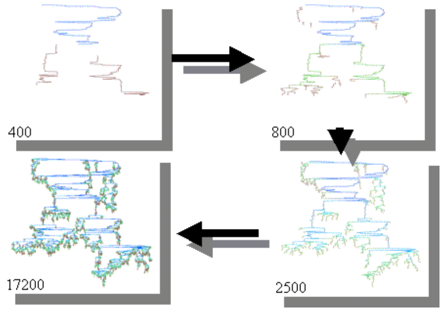


FIGURE 4. Display using Strahler.

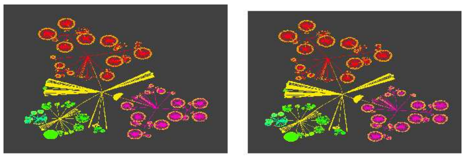
The problem now is to select the *best* elements to display so that the user should not be lost in his data. In our method, we propose to display the elements in decreasing Strahler order of the vertices in decreasing order. The figures show the result for 520000 vertices : the figure 3 shows what happens when *smart* choice is done, and figure 4 shows the result when applying our method.

An other field of interest is to find *similar* structures in an information system. *Similar* means here not the same but quite the same : the user may chooce a "degree of similarity". One real question assigned at theInfovis'03 contest [4] was on file system (that are trees) : given two views S_1 and S_2 of a same file system taken at two different time, show to the user what parts have changed. Work has already been done based on arity by Zemlyachenko [28] and more recently by Dinitz and *alt.* [26]. Their algorithm gives a partition of subtrees into isomorphism equivalence classes. It is proved to be linear. However, it only detect isomorphism and does not provide a measure of similarity for subtrees. In a recent work [evat], we used a combination of three parameters (number of vertices, arity, Strahler number) in order to detect *quasi-isomorphic* subtrees. Figure 5 to Figure 7 show a file system and part of the file system displayed on a screen with colors that suggest to the user the part that are the same in S_1 and S_2 : these systems have about 80000 vertices. In the figure 5, the two systems look similar. Thus roughly, with the "isomorphism level" choosen by the user, he can consider that no big changes have occurred. Applying our algorithm recursively with the same level, one can see in the figure 6 some smooth change, and finally see in Figure 7, just looking at the colors, what has changed. All this methods have been implemented in Tulip [3] and tested on real examples.

References

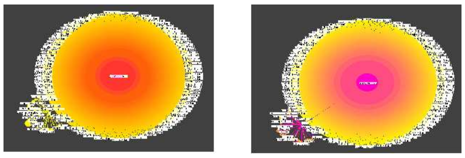
[1] D. Auber. Outils de visualisation de larges structures de données. Dissertation phd, Université Bordeaux 1, 2002.

[2] D. Auber. Using strahler numbers for real time visual exploration of huge graphs. In *International Conference on Computer Vision and Graphics*, volume 1-3 of *Journal of WSCG*, pages 56–69, 2002.



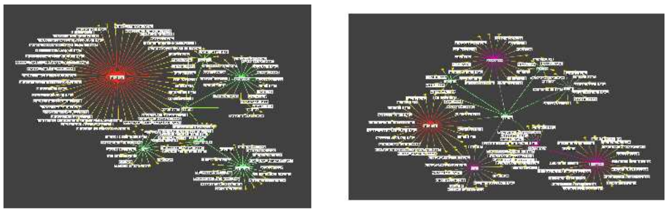
File system S1 File system S2

FIGURE 5. Two system files : overview.



File system S1 File system S2

FIGURE 6. Two system files : first zoom.



File system S1 File system S2

FIGURE 7. Two system files : changes.

[3] D. Auber. Tulip : A huge graph visualisation framework. In P. Mutzel and M. Jünger, editors, *Graph Drawing Softwares*, Mathematics and Visualization, pages 105–126. Springer-Verlag, 2003.

[4] D. Auber, M. Delest, J.P. Domenger, P. Ferraro, and R. Strandh. EVAT : Environment for visualization and analysis of trees. In *IEEE Symposition on Information Visualisation Contest*, volume www.cs.umd.edu/hcil/iv03contest/, pages 124–126, 2003.

[5] J. Bourdon, M. Nebel, and B. Vallée. On the stack-size of general tries. *Theoretical Informatics and Applications*, 35(4):163–185, 2000.

[6] M. Vauchaussade de Chaumont and X.G. Viennot. Enumeration of RNA's secondary structures by complexity. In V. Capasso, E. Grosso, and S.L. Paven-Fontana, editors, *Mathematics in medecine and Biology*, volume 57 of *Lecture Notes in Biomaths*, pages 360–365. Springer-Verlag, 1985.

[7] A.P. Ershov. On programming of arithmetic operations. *Communication of the A.C.M*, 1(8):3–6, 1958.

[8] P. Flajolet and A. Odlyzko. Singularity analysis of generating functions. *SIAM J. Discrete Math.*, 3:216–240, 1990.

[9] P. Flajolet, J.C. Raoult, and J. Vuillemin. The number of registers required for evaluating arithmetic expressions. *Theoretical Computer Science*, 9:99–125, 1979.

[10] P. Flajolet and R. Sedgewick. The average case analysis of algorithms: Complex asymptotics and generating functions. Rapport de recherche 2026, INRIA, 1993.

[11] J. Francon. Sur le nombre de registres nécessaires à l'évaluation d'une expression arithmétique. *RAIRO Informatique théorique*, 18:355–364, 1984.

[12] R.E. Horton. Erosioned development of systems and their drainage basins, hydrophysical approach to quantitative morphomology. *Bulletin Geological Society of America*, 56:275–370, 1945.

[13] R. Kemp. The average number of registers needed to evaluate a binary tree optimally. *Acta Informatica*, 11:363–372, 1979.

[14] A. Meir and J. W. Moon. Erratum: "On an asymptotic method in enumeration". *J. Combin. Theory Ser. A*, 52(1):163, 1989.

[15] A. Meir and J. W. Moon. On an asymptotic method in enumeration. *J. Combin. Theory Ser. A*, 51(1):77–89, 1989.

[16] W.J. Moon. An extension of Horton's law of stream numbers. Math. Colloq. Univ. cape Town, 1980.

[17] Markus E. Nebel. A unified approach to the analysis of Horton-Strahler parameters of binary tree structures. *Random Struct. Algorithms*, 21(3-4):252–277, 2002.

[18] R. Shreve. Statistical law of stream numbers. *J. Geol.*, 74:178–186, 1966.

[19] A.N. Strahler. Hypsomic analysis of erosional topography. *Bulletin Geological Society of America*, 63:1117–1142, 1952.

[20] J. Vannimenus and X.G. Viennot. Combinatorial analysis of ramified patterns. *J. Stat. Phys.*, 54:1529–1538, 1989.

[21] G. Viennot. Trees everywhere. In A. Arnold, editor, *Colloquium on Trees in Algebra and Programming*, Lecture Notes in Computer Science 431, pages 18–41. Springer-Verlag, 1990.

[22] G. Viennot, G. Eyrolles, N. Janey, and D.Arques. Combinatorial analysis of ramified patterns and computer imagery of trees. In *SIGGRAPH Conference*, volume 23 of *Computer Graphics*, pages 31–40, 1989.

[23] X.G. Viennot. A Strahler bijection between Dyck paths and planar trees. In R. Cori and O. Serra, editors, *11th Formal Power Series and Algebraic Combinatorics*, pages 573–584, 1999.

[24] M.S. Waterman. Secondary structure of singled stranded nucleic acids. *Adv. in Maths Suppl. Studies*, I:167–212, 1978.

[25] G.J. Wills. NicheWorks : Interactive visualization of very large graphs. In Giuseppe Di Battista, editor, *5th Symp. Graph Drawing*, volume 1353 of *Lecture Notes in Computer Science*, pages 403–414. Springer-Verlag, 1997.

[26] M. Rodeh Y. Dinitz, A. Itai. On an algorithm of Zemlyachenko for subtree isomorphism. *Information Processing Letters*, 703:141–146, 1999.

[27] D. Zeilberger. A bijection from ordered trees to binary trees that sends the pruning order to the Strahler number. *Discrete Math.*, 82:89–92, 1990.

[28] V.N. Zemlyachenko. Determining tree isomorphism. *Seminar on Combinatorial Mathematics*, pages 54–60, 1971.

LaBRI-Université Bordeaux 1, 351 Cours de la Libération, F33405 Talence
E-mail address: {auber,maylis,domenger,duchon}@labri.fr

I3S UMR 6070 2000, route des Lucioles, BP 121, F06903 Sophia-Antipolis Cedex
E-mail address: fedou@unice.fr