New Strahler numbers for rooted plane trees

D. Auber, J.P. Domenger, M. Delest, P. Duchon, 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 [11], then refined by Strahler [19] and rediscovered in computer science by Ershov [6] who interpreted it as the minimum number of registers allowing the computation of an arithmetic expression. Numerous results have been obtained from which one can point out explicit expressions for their distribution [8], [12]. In enumerative combinatorics, the research mostly involves bijections mapping Strahler numbers to other known statistics on other types of objects in order to understand their generating series : between binary trees and Dyck paths by Françon [10], between binary trees and forests of plane trees by Zeilberger [25] 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 [6]? 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 Nebel [16] or Bourdon *et 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 [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 $\Theta(n)$ time and space. Then we prove that the branching ratio r_k (to be defined later) of any simply generated family of trees with finite degree (in the sense of Meir and Moon [14, 13], tends to 4; for binary trees, it was known [18, 15] that $r_k = 4$ for all k. We end with a brief description of 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; 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 (or arity) of a vertex is the number of its children. In

this paper, we consider rooted plane trees, i.e. rooted trees where a total order is imposed on the children of each vertex.

Let D be a set of nonnegative integers containing 0. We consider the set of simply generated trees, $\mathcal{T}_D$, in the sense of Meir and Moor [14, 13] : each vertex of a tree has its arity in D. In this setting, d-ary trees correspond to $D = \{0, d\}$; $d = 2$ for binary trees.

The Strahler number σ_b was first introduced on binary trees in some work about the morphological structure of rivers [11, 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, 15] 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*

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

This ratio is called the branching ratio of binary trees. Horton has shown [11] 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 [6] who proved that the Strahler number of the root of a 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 number σ on general trees by:

- if s is a leaf then $\sigma(s) = 1$;
- otherwise 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

$$\sigma(s) = \max_{0 \leqslant i \leqslant k} (\sigma(s_i) + i).$$

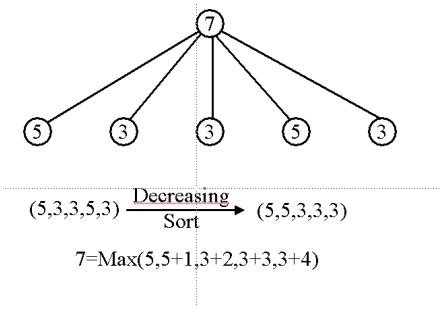

Figure 1: Strahler number on one vertex.

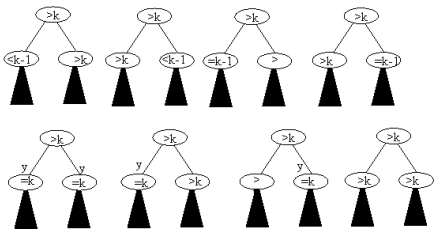

Figure 2: Construction of trees of $G_{\{2\}}$.

Equivalently, without requiring the children to be reordered, we have

$$\sigma(s) = \max\{i + j : \#\{\ell : \sigma(s_\ell) > i\} \geq j\}.$$

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 a weaker (asymptotic) version of Theorem 2.2 holds for any family $\mathcal{T}_D$, and prove it for finite D. 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 Section 6. In this field, the effective size of the trees is large : 500000 vertices is average. Thus, the algorithms that compute Strahler numbers must be efficient. In this section we describe a linear time algorithem, and prove that the nodes of any tree can be sorted according to their Strahler numbers in linear time; this is useful for the kind of applications we describe in Section 6.

3.1 Computing the Strahler valuation in linear time

To compute the Strahler number of a node, one has to sort its children in ascending or descending order of their own Strahler numbers. This sorting requirement, together with the *a priori* possibility that some nodes can have a large number of children, make it non obvious that a linear time algorithm exists to compute the Strahler valuation on a tree.

The key to our algorithm is the following observation : if each node of the tree is informed of the Strahler numbers of its children in increasing order, it only needs two counters to compute its own Strahler number; also, if the tree has N nodes, its Strahler number cannot be larger than N.

Thus, we setup an array S of size N; the k-th cell of the array will contain a list of vertices with Strahler number k, before their Strahler number is transmitted to their father. Initially, the first list contains all the leaves of the tree. Then, until all lists are empty, a vertex is extracted from the first nonempty list, and its

Strahler number is transmitted to its father; if the father now knows the Strahler numbers of all its children, its Strahler number is then computed, and it is inserted in the appropriate list.

Each node in the tree is examined a bounded number of times overall, so that the total complexity of the algorithm is $\Theta(N)$.

3.2 The set of Strahler values of a tree

In some of the applications described in Section 6, one may have to select a subtree of a given tree, for which the Strahler valuation has already been computed, and sort its nodes according to their Strahler numbers. Thus, it is interesting to know that this sorting of N nodes can be performed in $\Theta(N)$ time. This will be a consequence of the following lemma, which, incidentally, is the best counterpart to the property of Strahler numbers on binary trees stating that a (binary) tree with Strahler number k must have at least $2^k - 1$ vertices.

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

$$\sum_{k \in E(T)} k \leqslant |T| \tag{1}$$

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 (1), unless $k = 1$, in which case the inequality is strict.

Now assume (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 $E(T) = \cup E(T_i)$. 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 (1) is strict,
- if $k' = 1$, then $k = p$, and we have $\sum_{1 \leqslant i \leqslant p} \sum_{l \in E(T_i)} l \geqslant 1 + \sum_{l \in \cup_i E(T_i)} l$ which also proves (1) by induction.

□

From this lemma and the fact that any set of k positive integers has a sum at least $\binom{k+1}{2}$, we deduce :

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

A bijective proof of this result 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 these values are in $[1, N]$. Thus the complexity of sorting all nodes becomes linear (say, by using a table similar to the one we described for the Strahler computation, while simultaneously maintaining a balanced tree of the values appearing; each value will have to be inserted only once in this tree, resulting in a total complexity of $O(\sqrt{N}\log N)$ for creating and maintaining it), and we have

Theorem 3.3 *The nodes of any tree of size N, can be sorted according to their Strahler numbers in $\Theta(N)$ time and space.*

4 Branching ratios for binary trees

In this section, we prove a weaker (asymptotic) version of Theorem 2.2 [18, 15] so as to demonstrate the method that we use for more general families. Because D is fixed, we will omit it in all the formulas. We 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 β_k is identically 0 on $\mathcal{L}_k$, and identically 1 on $\mathcal{S}_k$. Next, consider the generating functions

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

Thus, the variable x always counts size, while y corresponds to the number of segments of Strahler value k. Throughout the rest of this paper, whenever we have a bivariate generating function $F(x,y)$, $F(x)$ denotes the corresponding univariate generating function $F(x,1)$.

The method consists in the following steps.

- Write out an equation for $G_k(x,y)$;
- consider the derivative of $G_k(x,y)$ with respect to y, $\frac{\partial G_k}{\partial y}(x,1) = \sum_{t\in\mathcal{G}_k} \beta_k(t)x^{|t|}$;
- find an asymptotic expression for $\frac{\partial G_k}{\partial y}(x,1)$, allowing us to derive asymptotic expressions of the form $B_{k,n} = c_k n^{-1/2}\rho^{-n}(1+o(1))$;
- compute the branching ratio as c_k/c_{k+1}.

In what follows, 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.

$$G_k = x(2L_{k-1}G_k + 2S_{k-1}G_k + S_k^2 + 2S_kG_k + G_k^2) \tag{2}$$

Each monomial in the right-hand part of (2) corresponds to one of the possible configurations of Strahler values for the two children of a node of Strahler value higher than k (see figure 2). In a second step, we replace $S_k(x)$ by $yS_k(x)$ to account for the number of segments of Strahler value k, and differentiate to get

$$\frac{\partial G_k}{\partial y}(x,1) = \frac{2x\left(G_{k-1}(x,1) - G_k(x,1)\right)^2}{1-2F(x)}. \tag{3}$$

The series $F(x)$ is an algebraic series (this is the well-known Catalan series) with square root type dominant singularities, and each $S_k(x)$ and $L_k(x)$, as a rational series, has only poles as singularities. This implies that the radius of convergence of F is strictly less than that of each L_k or L_k, and, in turn, that $G_k(x) = F(x) - L_{k+1}(x)$ has the same singularity structure as $F(x)$. Furthermore, the denominator in (3) vanishes at the singularities, while the numerator takes a finite, nonzero value.

As a result, the singularities of $\partial G_k/\partial y(x,1)$ are of the inverse square root type, so that the coefficients $B_{k,n}$ have an asymptotic of the form

$$B_{k,n} = c_k n^{-1/2} 4^n (1 + o(1)),$$

where c_k is proportional to $(G_{k-1}(1/4,1) - G_k(1/4,1))^2$.

Step 4 consists in proving that $\lim_{k\to\infty} c_{k-1}/c_k = 4$; this is straightforward once one notices that (2) can be rewritten as

$$\frac{G_k(x,1)}{G_{k-1}(x,1)} = 2x \frac{S_k(x,1) - G_k(x,1)}{1 - 2xF(x)},$$

from where elementary singularity analysis entails that the ratio converges to $1/2$ as x goes to $\pm 1/2$.

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.

Due to space constraints, we only sketch the proofs; the method is very similar to what was used in the previous section.

The generating function for all trees in $\mathcal{T}_D$ thus satisfies the polynomial equation

$$F(x) = x \sum_{d\in D} F(x)^d = x\Phi_D(F(x)). \tag{4}$$

It is a classical result [7, 9] that this series converges as an analytic function inside the complex disc $|x| \le \rho_D$, 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 $L_k(x)$. Similarly to the binary situation, S_k and L_k are both *rational* power series, so that their poles must all have moduli strictly larger than ρ_D. Thus, *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.*

[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.

5.1 Generating functions for trees of high Strahler value

Our goal in this paragraph is to write an equation 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.

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 $\mathcal{Q}_d$, $\mathcal{P}_d$ has negative coefficients. Once this is done, summing over all possible degrees for the root of a D-tree yields the following equation for the series $G_k(x)$:

$$G_k(x) = x \sum_{d \in D} \mathcal{P}_d(F(x), G_k(x), \ldots, G_{k-d+1}(x)) \tag{5}$$

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

Our main tool is a recurrence relation between the polynomials $\mathcal{P}_d$ themselves, which enables us to obtain asymptotic results on the branching ratios 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

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

We have :

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

or, equivalently,

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

For the first values of d, we obtained $\mathcal{P}_d$ through a computer algebra system. These suggest a recurrence for polynomials :

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

$$\mathcal{P}_d(f, g_0, \ldots, g_{d-1}) = g_{d-1}^d + d \int_{g_{d-1}}^{f} \mathcal{P}_{d-1}(t, g_0, \ldots, g_{d-2}) dt \tag{6}$$

and, equivalently,

$$\mathcal{Q}_d(g, s_0, \ldots, s_{d-2}, \ell) = (g + s_0 + \cdots + s_{d-2})^d + d \int_{s_{d-2}}^{s_{d-2}+\ell} \mathcal{Q}_{d-1}(g, s_0, \ldots, s_{d-3}, t) dt \tag{7}$$

Proof. We prove (7); (6) 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 $\mathcal{L}_d$ of words of length d over the alphabet $\{g, k, k-1, \ldots, k-d+2, \ell\}$ (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 (7).

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 $(g+s_0+\cdots+s_{d-2})^d$ in (7), and we now turn to words in $\mathcal{L}_d$ where the letter ℓ appears.

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 $d(2^j-1)$ words 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 (7). □

The following corollary is an easy consequence of the previous lemma and of the expressions for the first polynomials $\mathcal{P}_d$ and $\mathcal{Q}_d$:

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

$$\frac{\partial \mathcal{Q}_d}{\partial g}(g, s_0, \ldots, s_{d-2}, \ell) = d.\,(g+s_0+\cdots+s_{d-2}+\ell)^{d-1} \tag{8}$$

Furthermore, $\mathcal{P}_d$ is homogenous of degree d in its variables, and has $d-1$ in f and degree 1 in g_0. Thus, $\mathcal{P}_d = g_0\mathcal{D}_d + \mathcal{N}_d$, where

$$\mathcal{N}_d(f, g_1, \ldots, g_{d-1}) = g_{d-1}^d + d\int_{g_{d-1}}^{f} \mathcal{N}_{d-1}(t, g_1, \ldots, g_{d-2})dt \tag{9}$$

$$\mathcal{D}_d(f, g_1, \ldots, g_{d-1}) = d\int_{g_{d-1}}^{f} \mathcal{D}_{d-1}(t, g_1, \ldots, g_{d-2})dt \tag{10}$$

$$\mathcal{N}_d = \binom{d}{2} g_1^2 f^{d-2} + O\left(f^{d-3}\right) \tag{11}$$

$$\mathcal{D}_d = df^{d-1} - d(d-1)g_1 f^{d-2} + O\left(f^{d-3}\right) \tag{12}$$

In light of this, (5) solves to

$$G_k(x) = \frac{x \sum_{d\in D} \mathcal{N}_d(F(x), G_{k-1}(x), \ldots, G_{k-d+1}(x))}{1 - x \sum_{d\in D} \mathcal{D}_d(F(x), G_{k-1}(x), \ldots, G_{k-d+1}(x))} \tag{13}$$

Notice that the leading terms (in powers of $F(x)$) in the denominator of (13) exactly cancel the term of 1 when $x = \rho_D$, since

$$x \sum_{d\in D} dF(x)^{d-1} = 1$$

is exactly the equation for the singularities of F. Also, note that (from (11-12) the remaining leading terms (in powers of $F(x)$) in the numerator and denominator are in a ratio that is exactly $G_{k-1}(x)$. Since, for any x, all variables of the polynomials $\mathcal{N}_d$ and $\mathcal{D}_d$ tend to 0 as k tends to $+\infty$, this suggests that the ratio $G_k(\rho_D)/G_{k-1}(\rho_D)$ converges to $1/2$ as k goes to $+\infty$. In fact, the only ingredient missing to complete the proof is to justify using the powers of the various G_{k-i} variables to select the dominant terms in (13). This is the reason for the following lemma.

Lemma 5.3 *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 $k \geq 1$,*

$$G_{k-1}(\rho_D) \geq G_k(\rho_D) \geq \frac{1}{d'} G_{k-1}(\rho_D). \tag{14}$$

Proof. *(sketch)* The first part is an obvious consequence of the fact that G_{k-1} sums over more trees than G_k; the second part is proved by defining $\mathcal{E}_d = -\mathcal{D}_d + df^{d-1}$, and using (6) to prove, by induction on d, that

$$\mathcal{N}_d(f, g_1, \ldots, g_{d-1}) \geq \frac{1}{d'} g_{d-1} \mathcal{E}_d(f, g_1, \ldots, g_{d-1}) \geq 0 \tag{15}$$

holds as soon as $0 \leq g_1 \leq \cdots \leq g_{d-1} \leq f$. The lemma is then proved by a convex combination of (15), observing that

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

□

Corollary 5.4 *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 $k-1$. Let r_k denote the limit, as n tends to infinity, of this

ratio $B_{k,n}/B_{k+1,n}$. We will prove that r_k tends to 4 as k tends to infinity – a somewhat weaker result than the previously known fact (for binary trees) that $r_k = 4$. 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 $F_k(x,y) = L_k(x) + yS_k(x) + G_k(x,y)$. Using the notations of the previous paragraph,

$$G_k(x,y) = \sum_{d \in D} \mathcal{Q}_d(G_k(x,y), yS_k(x), S_{k-1}(x), \ldots, S_{k-d+2}(x), L_{k-d+2}(x)). \quad (16)$$

Note that (16) 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 (16) with respect to y and set $y = 1$, we get an equation which we can immediately solve for $\Gamma_k(x) = \partial F_k/\partial y(x,1)$:

$$\Gamma_k(x) = S_k(x) \frac{x \sum_{d \in D} \frac{\partial \mathcal{Q}_d}{\partial s_0}(G_k(x,1), S_k(x), \ldots, S_{k-d+2}(x), L_{k-d+2}(x))}{1 - x \sum_{d \in D} \frac{\partial \mathcal{Q}_d}{\partial g}(G_k(x,1), S_k(x), \ldots, S_{k-d+2}(x), L_{k-d+2}(x))} \quad (17)$$

The denominator in (17) is exactly $1 - x \sum_{d \in D} dF(x)^{d-1} = F(x)/(xF'(x))$. Since this vanishes at the dominant singularity $x = \rho_D$, and the numerator in (17) has a finite, nonzero limit, using a *transfer lemma* [7, 9], 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 for branches of Strahler number k is

$$\frac{b_k}{b_{k+1}} = \frac{\sigma_k}{\sigma_{k+1}} \frac{\sum_{d \in D} \frac{\partial \mathcal{Q}_d}{\partial s_0}(\gamma_k, \sigma_k, \ldots, \sigma_{k-d+2}, L_{k-d+2}(\rho_D))}{\sum_{d \in D} \frac{\partial \mathcal{Q}_d}{\partial s_0}(\gamma_{k+1}, \sigma_{k+1}, \ldots, \sigma_{k-d+3}, L_{k-d+3}(\rho_D))}$$

The first factor tends to 2 by Corollary 5.4. In the second factor, the variables of the polynomial in the numerator and denominator are in an asymtotic factor of 2 (also by Corollary 5.4), and tend to 0 as k tends to infinity, except the last variable $L_j(\rho_D)$ which tends to $F(\rho_D)$. Since all involved polynomials are homogenous with a nonzero term of degree 1 in the variable ℓ, the whole second factor also tends to 2 as k tends to infinity.

Theorem 5.5 *The branching ratio of trees of any family of finite degree D is asymptotically 4.*

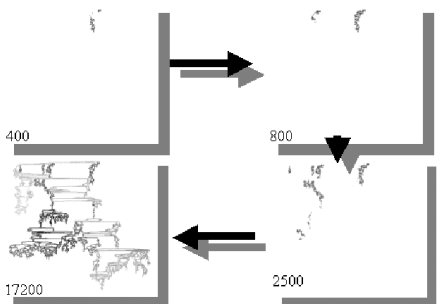

Figure 3: Display without Strahler

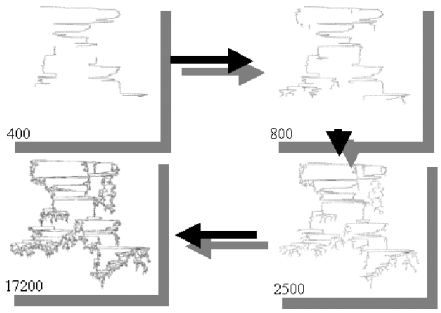

Figure 4: Strahler-guided display

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 visualizes 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 [24] consists in predicting the number of elements that can be displayed during δ_0. Thus the system incrementally displays the tree (by parts associated to δ_0) and ensures that the time reaction of the system is less than δ_0 without any request to the clock.

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. The figures show the result for 520000 vertices : 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* here means not the same but quite the same : the user may choose a "degree of similarity". One real question assigned at the Infovis'03 contest [4] was on file systems (that are trees) : given two views S_1 and S_2 of a same file system taken at two different times, show to the user what parts have changed. Work has already been done based on arity by Zemlyachenko [26] and more recently by Dinitz *et al.* [17]. Their algorithm gives a partition of subtrees into isomorphism equivalence classes. It is proved to be linear. However, it only detects isomorphism and does not provide a measure of similarity for subtrees. In a recent work [4], we used a combination of three parameters (number of vertices, arity, Strahler number) in order to detect *quasi-isomorphic* subtrees. The Figure 5 shows the trees representing the same file system at two different times (S_1 and S_2). Each tree has about 80000 vertices. S_1 and S_2 look similar. Thus roughly, with the "isomorphism level" chosen in this view by the user, he can consider that no big changes have occurred during the chosen time. Applying our algorithm recursively, we detect subtrees that are different. One can see in the figure 6 that some smooth changes have occured on the labels even if the drawing of the tree is the same. This is

Figure 5: Two file systems: overview

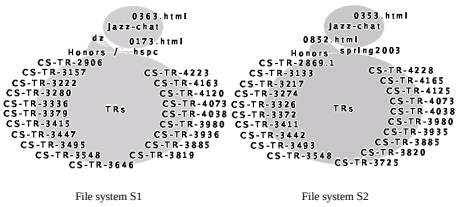

Figure 6: Two file systems: first zoom

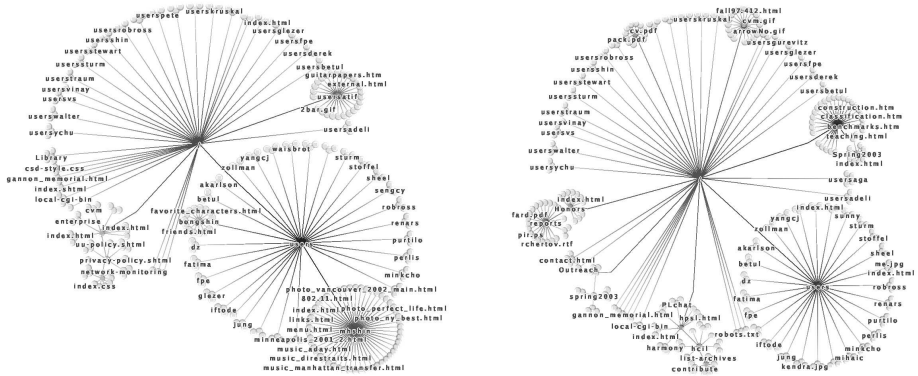

Figure 7: A final zoom on both file systems

due to the navigation tool that we have described above. In the real version[4], we use only colors because same colors suggest same subtrees. The similar big disks in Figure 6 are different and have 951 and 755 nodes, respectively. Figure 7 was obtained by suppressing the two disks after inspection. We have only displayed a similar part of the two subtrees. All these methods have been implemented in Tulip [3] and work on real examples.

References

[1] D. Auber. Outils de visualisation de larges structures de données. PhD dissertation, 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.

[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 In-*

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

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

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

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

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

[11] 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.

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

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

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

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

[16] M.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.

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

[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] 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.

[25] 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.

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

Auber, David
LaBRI, Université Bordeaux 1
auber@labri.fr
Domenger, Jean-Philippe
LaBRI, Université Bordeaux 1
domenger@labri.fr
Delest, Maylis
LaBRI, Université Bordeaux 1
maylis@labri.fr
Duchon, Philippe
LaBRI, Université Bordeaux 1
duchon@labri.fr
Fédou, Jean-Marc
I3S, Université de Nice-Sophia Antipolis
fedou@unice.fr