Horton-Strahler ordering of random binary trees

1994 J. Phys. A: Math. Gen. 27 285

(http://iopscience.iop.org/0305-4470/27/2/014)

J. Phys. A: Math. Gen. 27 (1994) 285–293. Printed in the UK

Horton–Strahler ordering of random binary trees

Iddo Yekutieli† and Benoit B Mandelbrot‡

† Yale University, PO Box 208283, New Haven, CT 06520-8283, USA
‡ IBM Fellow Emeritus, IBM T J Watson Research Center, Yorktown Heights, NY 10598, USA

Received 13 September 1993

Abstract. We study the Horton–Strahler ordering for random binary trees, which are statistically self-similar branching structures. Extending previously obtained results, we show that near the top of these trees, the expected bifurcation ratios tend strongly to the value 4. But at the root of the tree, the expected bifurcation ratio is less than 4, becoming asymptotically a periodic function of $\log_4 n$.

1. Introduction

The abundance in nature of branching structures has led to extensive studies of their properties. Trees, river networks, and patterns of electric discharge and electrochemical deposition are examples of such structures. Branching processes in chemistry and physics, the hierachical structure of pure states in spin glasses, and the classification of species in biology, can be represented as trees as well. Since the advent of digital computers, binary trees have also been studied in this context. The inherent randomness often apparent in tree-like structures has raised the question of the role it plays in determining their form. This random component may lead to universal features of such branching structures.

One of the simplest models of random branching is the random binary tree. Consider the ensemble of all distinct rooted binary trees of a given magnitude (number of sources, or leaves). Figure 1 shows all five distinct binary trees of magnitude 4. Assigning the same statistical weight to all trees of magnitude n defines the random binary tree model. Many statistical properties of this ensemble can be calculated due to the fact that any binary tree of magnitude n is uniquely decomposed into its two main sub-branches of magnitudes k and $n-k$.

An interesting property of a branching structure has to do with its complexity. In his study of river networks, Horton [1] devised a scheme for indexing the hierarchical structure of the streams. Streams starting from the sources of a river network are assigned the lowest order and, moving downstream, a confluence of streams raises the order of the resulting stream. Strahler [2] slightly modified this ordering scheme, to make it independent of metric or directional properties of the streams. Either scheme applies to all tree-like structures. This work uses the latter of these ordering schemes. In the context of trees, we shall use the terms *leaf, branch* and *root* instead of the corresponding terms for river networks: *source, stream* and *outlet.*

Given a rooted tree structure, Strahler orders the branches recursively:

(i) Each leaf is assigned the order $i = 1$.

0305-4470/94/020285+09$07.50

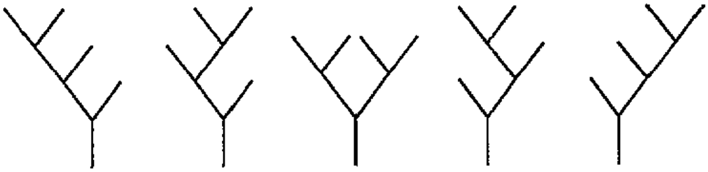

Figure 1. All five distinct binary trees of magnitude 4.

(ii) The order i of a subsequent branch is determined by the orders i_1, i_2 of its two sub-branches: if $i_1 \neq i_2$ then $i = \max\{i_1, i_2\}$; otherwise $i = i_1 + 1$.

When the order of a branch is equal to that of one of its sub-branches, it is considered to be a continuation of this branch, otherwise it is considered a new branch. The order of the whole tree is defined to be the order of the root, its lowest-lying branch; it is a measure of the complexity of the tree. When a branch has more than two sub-branches, only the two of highest orders are considered.

Consider the number of branches N_i of given Strahler order i in a tree (the *stream number* of order i). N_1 is equal to the number of leaves. If I is the order of the whole tree (i.e. of its root) then $N_I = 1$. In figure 1, the first four trees have four branches of order 1 and one of order 2, while the last tree has four branches of order 1, two of order 2, and one of order 3.

The bifurcation ratio B_i of branches of order i and $i+1$ is defined by

$$B_i = N_i / N_{i+1} \,. \tag{1}$$

Studies of river networks show that in many cases, the stream numbers $\{N_i\}$ are very well approximated by a geometric series, such that

$$B_i \approx B \qquad \text{for all } i \,. \tag{2}$$

The observed values of B for different river networks vary between 3 and 5. As shown in [3], a constant bifurcation ratio is a typical property of self-similar tree structures.

Shreve [4] was first to use random binary trees as a model for river networks. Studying the Strahler ordering in this model, he found empirically that the bifurcation ratios tend asymptotically to 4. He also noted that the typical order of a tree of magnitude n was very close to

$$I_{\text{typ}} = \lfloor \log_4 n \rfloor \,. \tag{3}$$

These findings were later substantiated by analytical studies of random binary trees by Kemp [5], Flajolet *et al* [6] and Meir *et al* [7]. They found that the expected order of a tree $I(n)$ is equal to $\log_4 n$ plus a periodic function of $\log_4 n$. Moon [8] had shown that the ratio of the expectation values of successive stream numbers tended to 4. Wang and Waymire [9] have demonstrated recently that the first bifurcation ratio B_1 tends to 4 in a stronger sense.

This paper presents some new results concerning the Strahler ordering of random binary trees. Section 2 extends [9], by evaluating the distributions of bifurcation ratios for higher orders. We show that B_2 also tends strongly to the value 4. While it seems that the bifurcation ratio B_i does tend to 4 asymptotically as long as $i \ll I_{\text{typ}}$, this is not the case when the order i is close to the order of the tree. In section 3 we specifically calculate the expected bifurcation ratio at the root $B_I (= N_{I-1})$. It is found to be less than 4, becoming asymptotically a periodic function of $\log_4 n$.

2. Bifurcation ratios near the top of the tree

2.1. Preliminaries

Denote by $s_i(n,k)$ the number of trees of n leaves with k branches of order i. This number may be calculated recursively by decomposing a binary tree into its two main sub-branches. In general, the number of branches of order i in the whole tree is the sum of the numbers of such branches in its two sub-branches.

For $n = 1$ one has

$$s_i(1,k) = \begin{cases} \delta_{k,1} & i = 1 \\ \delta_{k,0} & i \geqslant 2 . \end{cases} \tag{4}$$

For $n \geqslant 2$

$$s_i(n,k) = \sum_{m=1}^{n-1}\sum_{j=0}^{k} s_i(m,j)s_i(n-m,k-j) + \theta_{i,2}\sum_{m=1}^{n-1}(\delta_{k,1}-\delta_{k,0})s_{i-1}(m,1)s_{i-1}(n-m,1) \tag{5}$$

where $\theta_{i,2} = 1$ if $i \geqslant 2$, otherwise it is equal to 0. When $i \geqslant 2$, one has to take into account the occurence of two sub-branches of order $i-1$ which results in a tree of order i.

Let us introduce the generating functions

$$S_i(y,x) = \sum_{n=1}^{\infty}\sum_{k=0}^{n} s_i(n,k)y^n x^k \tag{6}$$

$$Q_i(y) = \sum_{n=1}^{\infty} s_i(n,0)y^n \tag{7}$$

$$R_i(y) = \sum_{n=1}^{\infty} s_i(n,1)y^n \tag{8}$$

$$C(y) = \sum_{n=1}^{\infty} C_n y^n = S_i(y,1) . \tag{9}$$

For $i = 1$ one calculates, from (4)–(6),

$$S_1(y,x) = xy + \left[S_1(y,x)\right]^2 \tag{10}$$

the solution of which is

$$S_1(y,x) = \frac{1-(1-4xy)^{1/2}}{2} = C(xy) . \tag{11}$$

This yields C_n, the total number of trees of magnitude n

$$C_n = \frac{1}{2n-1}\binom{2n-1}{n} \tag{12}$$

the well known result of Cayley [10], and the distribution of branches of order 1

$$s_1(n,k) = C_n \delta_{n,k} . \tag{13}$$

For $i \geqslant 2$, one finds

$$S_i(y,x) = y + \left[S_i(y,x)\right]^2 + (x-1)\left[R_{i-1}(y)\right]^2 . \tag{14}$$

The two supplementary generating functions $Q_i(y)$ and $R_i(y)$ obey the following recursion relations:

$$Q_i(y) = y + \left[Q_i(y)\right]^2 - \left[R_{i-1}(y)\right]^2 \tag{15}$$

$$R_i(y) = 2Q_i(y)R_i(y) + \left[R_{i-1}(y)\right]^2 . \tag{16}$$

A tree with no branches of order i has either 0 or 1 streams of order $i-1$, therefore

$$s_i(n,0) = s_{i-1}(n,0) + s_{i-1}(n,1) . \tag{17}$$

This implies a relation for the generating functions of order i

$$Q_i(y) = Q_{i-1}(y) + R_{i-1}(y) \tag{18}$$

in terms of the generating functions of the previous order.

2.2. *Calculation of B_1*

For $i = 2$ we calculate

$$S_2(y,x) = \tfrac{1}{2} - \tfrac{1}{2}\left[(1-2y)^2 - 4y^2x\right]^{1/2} . \tag{19}$$

A Taylor expansion in x and y of (19) gives the distribution $s_2(n,k)$. This expansion is convergent as long as $y < \frac{1}{4}$. This does not hamper us, as we are interested in the limit $y \to 0$ when extracting the distributions $s_i(n,k)$ from the generating functions.

For $1 \leqslant k \leqslant \lfloor n/2 \rfloor$, we find

$$s_2(n,k) = \frac{1}{2k-1}\binom{2k-1}{k}\binom{n-2}{n-2k}2^{n-2k} . \tag{20}$$

There are no trees with more than $\lfloor n/2 \rfloor$ second-order branches. This is because a pair of first-order branches is needed to create a second-order branch.

We define the probability $p_i(n,k)$ of finding a tree of n leaves with k branches of order i by $p_i(n,k) = s_i(n,k)/C_n$. For $i = 2$,

$$p_2(n,k) = \frac{n!(n-1)!(n-2)!2^{n-2k}}{k!(k-1)!(n-2k)!(2n-2)!} . \tag{21}$$

For large n, and k of the order of n, we simplify (21) using the Stirling formula

$$p_2(n,k) \simeq \sqrt{\frac{8}{\pi n}} \exp\left\{-\frac{8}{n}\left(k - \frac{n}{4}\right)^2\right\} . \tag{22}$$

For large n, the distribution $p_2(n,k)$ tends to a normal distribution, with mean $n/4$ and variance $n/16$. This shows that the bifurcation ratio B_1 tends asymptotically to 4. This result may also be found directly by calculating the moments of the distribution $s_2(n,k)$ directly from the generating function (19), as was done in [9].

2.3. Calculation of B_2

The generating function for $n = 3$ is

$$S_3(y, x) = \frac{1}{2} - \frac{1}{2(1-2y)}\left[(1-4y+2y^2)^2 - 4y^4x\right]^{1/2}. \tag{23}$$

The distribution $s_3(n, k)$, for $1 \leqslant k \leqslant \lfloor n/4 \rfloor$, is given by

$$s_3(n, k) = \frac{1}{2k-1}\binom{2k-1}{k}\sum_{l=0}^{n-4k} 2^{n-4k-l} \sum_{m=\lfloor (l+1)/2 \rfloor}^{l} \binom{2k+m-2}{m}\binom{m}{l-m} 2^{3m-l}(-1)^{l-m}. \tag{24}$$

Again, there are no trees with more than $\lfloor n/4 \rfloor$ third-order branches, as a pair of second-order branches is needed to form a third-order branch.

We have also calculated the first two moments with respect to k of this distribution

$$\mu_3^{[1]}(n) = \sum_{m=0}^{n-4}\binom{2m}{m}(n-m-3)2^{n-m-4} \tag{25}$$

$$\mu_3^{[2]}(n) = \sum_{m=0}^{n-8}\binom{2m}{m}\frac{(n-m-5)(n-m-6)(n-m-7)(2m+1)}{3}2^{n-m-8}. \tag{26}$$

A numerical study of (25), (26) shows that, asymptotically,

$$\mu_3^{[1]}(n) \sim \frac{n}{16} \tag{27}$$

$$\mu_3^{[2]}(n) - \left(\mu_3^{[1]}(n)\right)^2 \sim \frac{5n}{256} \tag{28}$$

so that the bifurcation ratio B_2 also tends strongly to 4 for large n.

2.4. Calculation of higher orders

For higher orders, the direct calculation of the distributions $s_i(n, k)$, as well as their moments, becomes impossible practically. Yet it is still possible to find the asymptotic behaviour of the first moments, for large n, from (14).

For this we use the solution for $R_i(y)$ found in [5–7]

$$R_i(y) = \frac{y^{1/2}\sin(\theta)}{\sin(2^i\theta)} \tag{29}$$

where

$$\theta(y) = -\mathrm{i}\ln\frac{C(y)}{y^{1/2}}. \tag{30}$$

It is seen to be the solution by rewriting (16) using (18) and the trigonometric identity $\cos(2\phi) = 2\cos^2(\phi) - 1$.

With this result at hand, we can now solve for $S_i(y, x)$ from (14)

$$S_i(y, x) = \frac{1}{2} - \frac{y^{1/2}\sin(\theta)}{\sin(2^{i-1}\theta)}\left[\cos^2(2^{i-1}\theta) - x\right]^2. \tag{31}$$

The first moment with respect to k of this distribution is

$$M_i(y) = \frac{\partial S_i(y,x)}{\partial x}\Big|_{x=1}$$
$$= y(1-4y)^{-1/2}\left[\frac{\sin(\theta)}{\sin(2^{i-1}\theta)}\right]^2 . \tag{32}$$

Let us now inspect the expression for $M_2(y)$, which is easily found from (19)

$$M_2(y) = y^2(1-4y)^{-1/2} . \tag{33}$$

This function has a singularity at $y = \frac{1}{4}$. The expansion of this function around $y = 0$ is

$$M_2(y) = \sum_{n=2}^{\infty} \binom{2n-4}{n-2} y^n$$
$$= \sum_{n=2}^{\infty} \mu_2^{[1]}(n) y^n . \tag{34}$$

For large n the terms of the expansion behave as

$$\mu_2^{[1]}(n) \simeq \frac{1}{16\sqrt{\pi}} n^{-1/2}\, 4^n . \tag{35}$$

The radius of convergence of this series is $\frac{1}{4}$, as expected.

It is possible to extract the asymptotic behaviour of $\mu_i^{[1]}(n)$ by studying the singular behaviour of the functions $M_i(y)$ in the vicinity of $y = \frac{1}{4}$ (as is done, for example, in [11]). The generating functions $M_i(y)$ given in (32) all display the same singularity at $y = \frac{1}{4}$. We set

$$y = \frac{1-\delta}{4} .$$

To find the behaviour of $M_i(y)$ in the vicinity of this singularity, we note that for $\delta \ll 1$

$$\theta = -\mathrm{i} \ln \frac{1-\delta^{1/2}}{(1-\delta)^{1/2}} \simeq \mathrm{i}\delta^{1/2} . \tag{36}$$

Expanding the sine functions, we can immediately write

$$M_i \simeq \tfrac{1}{2} 4^{-i} \delta^{-1/2} . \tag{37}$$

The ratios between the singular parts of the moments $M_i(y)$ in the vicinity of $y = \frac{1}{4}$ will yield the ratios of the expected stream numbers of successive Strahler orders. It is seen directly that

$$B_i = \lim_{n\to\infty} \frac{\mu_i^{[1]}(n)}{\mu_{i+1}^{[1]}(n)} = 4 . \tag{38}$$

Note that for $\delta \ll 1$,

$$M_2(y) \simeq \tfrac{1}{16}\delta^{-1/2}.$$

On the other hand, the series expansion (34) may be approximated as

$$\begin{aligned} M_2(y) &\simeq \frac{1}{16\sqrt{\pi}} \sum_{n=2}^{\infty} n^{-1/2}\,(1-\delta)^n \\ &\simeq \frac{\delta^{-1/2}}{16\sqrt{\pi}} \sum_{n=2}^{\infty} \delta(n\delta)^{-1/2}\,\exp\{-n\delta\} \\ &\simeq \frac{\delta^{-1/2}}{16\sqrt{\pi}}\Gamma\left(\frac{1}{2}\right) = \frac{1}{16}\delta^{-1/2}. \end{aligned}$$

The moral of this short exercise is that for $\delta \ll 1$ and $n \gg 1$

$$n \sim \delta^{-1}. \tag{39}$$

The previous analysis holds only for values of i such that $\left|2^{i-1}\theta\right| \ll 1$, in order to be able to expand the term $\sin(2^{i-1}\theta)$ in (32). By (36) and (39), this condition is equivalent to

$$i \ll 1 + \log_4 n. \tag{40}$$

Not surprisingly, $I_{\mathrm{typ}} = 1 + \lfloor \log_4 n \rfloor$ is the typical value of the order of a tree of n leaves, in the limit of large n [5–7].

Although we have only shown that for any fixed order i the ratios of *expected* stream numbers tend to the value 4, it reasonable to assume that these ratios tend to this value in a stricter sense, as is the case for $i = 2, 3$. This will be true for any value of i sufficiently small, complying with (40). This property will break down near the root of a tree, and we expect the bifurcation ratio at the root of a random binary tree to be different from 4.

3. Bifurcation ratio at the root

Let us denote by $t_i(n,k)$ the number of trees with n leaves with a root of order i having k branches of order $i-1$. Evidently, summing $t_i(n,k)$ over k gives $s_i(n,1)$. In analogy to (5) one can write a recursion relation for $t_i(n,k)$ by decomposing a tree into its two main sub-branches. We find the following relation

$$\begin{aligned} t_i(n,k) = 2\sum_{m=1}^{n-1} \{t_i(m,k)s_{i-1}(n-m,0) + t_i(m,k-1)s_{i-1}(n-m,1)\} \\ + \sum_{m=1}^{n-1} s_{i-1}(m,1)s_{i-1}(n-m,1)\delta_{k,2} \end{aligned} \tag{41}$$

with boundary conditions

$$t_i(n,0) = t_i(n,1) = 0 \qquad \text{for all } n \geqslant 2 \tag{42}$$

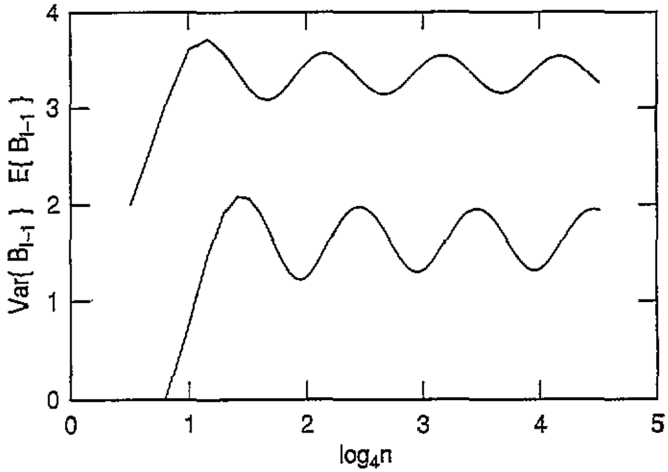

Figure 2. The expectation value and variance of the bifurcation ratio at the root of a random binary tree, as a function of n, calculated numerically from (41)–(43).

and initial conditions, for $n = 1$,

$$t_i(1, k) = \delta_{i,1}\delta_{k,0} \,. \tag{43}$$

For any given value of n, k is bounded between 2 (the minimal number of branches of order $I - 1$ needed to form a branch of order I) and $n - 2$ (for a tree of order 2 with the form of a comb).

One can now proceed in the manner used earlier, and define the generating function

$$T_i(y, x) = \sum_{n=1}^{\infty}\sum_{k=1}^{n} t_i(n, k) y^n x^k \,. \tag{44}$$

This generating function obeys the equation

$$T_i(y, x) = \frac{x^2[R_i(y)]^2}{1 - 2Q_i(y) - 2xR_i(y)} \tag{45}$$

where $R_i(y)$ and $Q_i(y)$ are defined in (7), (8). In principle, one may now use the solution for $R_i(y)$ in order to obtain an expression for $T_i(y, x)$. We have not been able to extract information from this solution. Instead, we solved the recursion relation (41) numerically. Using this solution, we calculated the expected value of k, the bifurcation ratio at the root

$$B_{I-1}(n) = C_n{}^{-1} \sum_{i=1}^{n}\sum_{k=1}^{n} k t_i(n, k) \tag{46}$$

and its variance, both of which are found to be of order 1. Figure 2 shows the mean and variance of the bifurcation ratio at the root, as a function of $\log_4 n$. $B_{I-1}(n)$ tends asymptotically to a periodic function with a mean value of 3.34 and of amplitude 0.19. The bifurcation ratio at the root is therefore always less than 4. The variance is also a periodic function of $\log_4 n$, with an average value of 1.54. Such a periodicity is not surprising, as it is also seen in the expected order of a tree of magnitude n.

4. Conclusion

Random binary trees are an interesting example of a random branching structure, as it is possible to calculate analytically many of their statistical properties. This paper focused on the Horton–Strahler ordering of these trees. We have shown that in order to observe the asyptotically self-similar behaviour of the topological structure of these trees, one has to satisfy conditions similar to those required with respect to the metric properties of self-similar objects. One has to go to large system sizes (the magnitude of the trees tending to infinity), and stay far from the cut-off scales (the root of the tree).

References

[1] Horton R E 1945 *Bul. Geol. Soc. Am.* **56** 275
[2] Strahler A N 1957 *Trans. Am. Geophys. Union* **38** 913
[3] Yekutieli I, Mandelbrot B B and Kaufman H 1993 *Yale University preprint*
[4] Shreve R L 1966 *J. Geol.* **74** 17; 1967 *J. Geol.* **75** 178
[5] Kemp R 1979 *Acta Informatica* **11** 363
[6] Flajolet P, Raoult J C and Vuillemin J 1979 *Theor. Comput. Sci.* **9** 99
[7] Meir A, Moon J W and Pounder J R 1980 *SIAM J. Alg. Disc. Meth.* **1** 25
[8] Moon J W 1978–79 *Math. Colloq. Univ. Cape Town* **XII** 1
[9] Wang S X and Waymire E C *SIAM J. Disc. Math.* **4** 1
[10] Cayley A 1858 *Phil. Mag.* **28** 374
[11] Penaud J G 1991 *Discrete Appl. Math.* **31** 1