The average height of planted plane trees

Reprinted from:
GRAPH THEORY AND COMPUTING

Academic Press, Inc., New York and London

106

THE AVERAGE HEIGHT OF PLANTED PLANE TREES

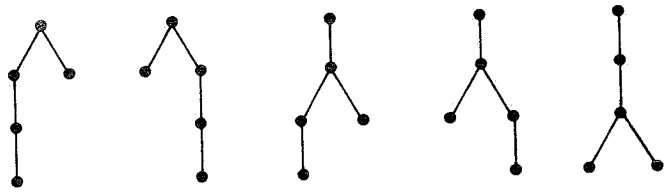
N. G. de Bruijn
Technological University
Eindhoven, The Netherlands

D. E. Knuth†
Stanford University
Stanford, California

S. O. Rice
Bell Telephone Laboratories, Inc.
Murray Hill, New Jersey

A *planted plane tree*, sometimes called an ordered tree, is a rooted tree which has been embedded in the plane so that the relative order of subtrees at each branch is part of its structure. In this paper we shall say simply *tree* instead of *planted plane tree*, following the custom of computer scientists.

The *height* of a tree is the number of nodes on a maximal simple path starting at the root. For example, there are exactly five trees with five nodes and height 4, namely



† This research was supported in part by the National Science Foundation, under grant number GJ-992, and the Office of Naval Research under grant number N-00014-67-A-0112-0057 NR 044-402. Reproduction in whole or in part is permitted for any purpose of the United States Government.

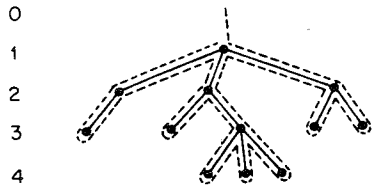


Fig. 1. A tree as a random walk.

The height of a tree is of interest in computing because it represents the maximum size of a stack used in algorithms that traverse the tree [3, pp. 317–318]. Our goal in this paper is to study the average height of a tree with n nodes, assuming that all n-node trees are equally likely. The corresponding problem for oriented, that is, rooted, unordered, trees has been solved by Rényi and Szekeres [6]. Our principal results are stated in Eqs. (32) and (34).

Trees appear in many disguises, and in particular there is a natural correspondence between trees of height less than or equal to h and discrete random walks in a straight line, with absorbing barriers at 0 and $h+1$. If we wander around a tree with n nodes, as shown by the dotted lines in Fig. 1, the vertical component of successive positions describes a path of length $2n-1$ from 1 to 0. For example, the path in Fig. 1 is

$$1,2,3,2,1,2,3,2,3,4,3,4,3,4,3,2,1,2,3,2,3,2,1,0.$$

This is one way a gambler can lose \$1 before winning \$5. This construction, suggested by Harris [2] in 1952, is clearly reversible.

The height of trees plays a similar role in the classical ballot problem. How many ways are there to arrange n ballots for candidate A and n for candidate B in such a way that the number of votes for A never lags behind the number for B, as the ballots are counted, but A is never more than h votes ahead? The answer is the number of trees with $n+1$ nodes and height less than or equal to $h+1$, again by the construction indicated in Fig. 1. The ballot sequence corresponding to that tree is AABBAABAABABABBBAABABB.

We shall begin our study of the asymptotic properties of height by reviewing some known results. Let A_{nn} be the number of trees with n nodes and height less than or equal to h, and let

$$A_h(z) = \sum A_{nh} z^n \tag{1}$$

be the corresponding generating function. We obtain all trees with height less than or equal to $h+1$ by taking a root node and attaching zero or more subtrees each of which has height less than or equal to h. Therefore,

$$\begin{aligned} A_{h+1}(z) &= z(1 + A_h(z) + A_h(z)^2 + A_h(z)^3 + \cdots) \\ &= z/(1-A_h(z)), \qquad h \geqslant 0. \end{aligned} \tag{2}$$

Clearly $A_0(z) = 0$. This relation yields a simple recurrence for the numbers A_{nh},

$$A_{n,h+1} = A_{n-1,h+1}A_{1,h} + A_{n-2,h+1}A_{2,h} + \cdots + A_{1,h+1}A_{n-1,h}, \qquad n \geqslant 2, \quad h \geqslant 0, \tag{3}$$

from which the first few values are easily calculated, as shown in Table I.

TABLE I

	$n=1$	2	3	4	5	6	7	8
$h=1$	1	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1
3	1	1	2	4	8	16	32	64
4	1	1	2	5	13	34	89	233
5	1	1	2	5	14	41	122	365
6	1	1	2	5	14	42	131	417

Since no tree with n nodes can have a height greater than n, we have

$$A_{nh} = A_{nn} = \binom{2n-2}{n-1}\frac{1}{n}, \qquad h \geqslant n, \tag{4}$$

which is the well-known formula for the total number of trees with n nodes [3, p. 389].

Iteration of (2) yields a continued fraction representation of $A_h(z)$. For example,

$$A_4(z) = \cfrac{z}{1-\cfrac{z}{1-\cfrac{z}{1-z}}}. \tag{5}$$

This suggests expressing the generating function as a quotient of polynomials

$$A_h(z) = zp_h(z)/p_{h+1}(z), \tag{6}$$

where

$$p_0(z) = 0, \qquad p_1(z) = 1, \qquad p_{h+1}(z) = p_h(z) - zp_{h-1}(z). \tag{7}$$

The solution to this recurrence is

$$p_h(z) = (1-4z)^{-1/2}\left(\left(\frac{1+(1-4z)^{1/2}}{2}\right)^h - \left(\frac{1-(1-4z)^{1/2}}{2}\right)^h\right), \tag{8}$$

and the form of this solution suggests setting $z = 1/(4\cos^2\theta)$. We obtain

$$p_h(4\cos^2\theta)^{-1} = \sin h\theta/(\sin\theta(2\cos\theta)^{h-1}), \tag{9}$$

$$A_h(4\cos^2\theta)^{-1} = \sin h\theta/(2\cos\theta\sin(h+1)\theta).$$

Incidentally it is easy to verify that $p_h(-1)$ is the Fibonacci number F_h, and that

$$p_h(z) = \sum_{0\leqslant k<h}\binom{h-1-k}{k}(-z)^k, \qquad h \geqslant 1. \tag{10}$$

This leads to another recurrence for the A_{nh}.

Since $p_h(z)^2 - p_{h+1}(z)p_{h-1}(z) = z^{h-1}$, there is a simple generating function for the number of trees with n nodes and height exactly h,

$$A_h(z) - A_{h-1}(z) = z^h/p_{h+1}(z)p_h(z). \tag{11}$$

This formula was recently derived by Kreweras [4, p. 37].

Since p_h is a polynomial of degree $\lfloor (h-1)/2\rfloor$, the roots of $p_h(z) = 0$ are $(4\cos^2(j\pi/h))^{-1}$, for $1 \leqslant j < h/2$. We obtain a partial fraction expansion of the generating function

$$A_h(z) = \sum_{1\leqslant j\leqslant h/2}\frac{\tan^2\theta_{jh}}{(h+1)(1-(4\cos^2\theta_{jh})z)} + a_h + b_h z, \tag{12}$$

where

$$\theta_{jh} = j\pi/(h+1),$$

and

$$a_{2m} = -m, \qquad b_{2m} = 0, \tag{13}$$

$$a_{2m+1} = -m(2m+1)/6(m+1), \qquad b_{2m+1} = (m+1)^{-1}, \qquad m \geqslant 1.$$

This leads immediately to the explicit formula

$$A_{nh} = (h+1)^{-1}\sum_{1\leqslant j\leqslant h/2} 4^n\sin^2(j\pi(h+1))\cos^{2n-2}(j\pi(h+1)), \qquad n \geqslant 2. \tag{14}$$

It is rather remarkable that this formula gives a constant value for fixed n and all $h \geqslant n$. It is perhaps even more remarkable that Lagrange derived a formula in 1775 which essentially includes this as a special case, see Lagrange [5, p. 247]. Feller [1, p. 322] observes that the formula has been rediscovered many times, although it appears in many texts on probability in connection with the equivalent gambler's ruin problem. As a special case of (14) we have the asymptotic formula

$$A_{nh} \sim (4^n/(h+1))\tan^2(\pi/(h+1))\cos^{2n}(\pi/(h+1)), \qquad \text{fixed } h, \quad n \to \infty. \tag{15}$$

Another interesting expression for A_{nh} can be derived by applying complex variable theory. We have

$$A_{nh} = (2\pi i)^{-1} \int^{(0+)} \frac{dz}{z^{n+1}} A_h(z) \tag{16}$$

$$= (2\pi i)^{-1} \int^{(0+)} \frac{dz}{z^n} (1+u) \frac{1-u^h}{1-u^{h+1}},$$

where

$$u = \frac{1-(1-4z)^{1/2}}{1+(1-4z)^{1/2}}, \tag{17}$$

by (6) and (8). Since

$$z = u/(1+u)^2, \tag{18}$$

we have $u \approx z$ when $|z| \ll 1$. Hence, we may change variables in (16) to obtain

$$A_{nh} = (2\pi i)^{-1} \int^{(0+)} \frac{du}{u^n} (1-u)(1+u)^{2n-2} \frac{1-u^h}{1-u^{h+1}}. \tag{19}$$

In other words A_{nh} is the coefficient of u^{n-1} in $(1-u)(1+u)^{2n-2}(1-u^h)/(1-u^{h+1})$. Some simplification now occurs when we consider the number of trees with height *greater* than h,

$$B_{nh} = A_{nn} - A_{nh} \tag{20}$$

$$= (2\pi i)^{-1} \int^{(0+)} \frac{du}{u^{n+1}} (1-u)^2 (1+u)^{2n-2} \frac{u^{h+1}}{1-u^{h+1}}.$$

It follows that

$$B_{n+1,h-1} = \sum_{k \geq 1} \left(\binom{2n}{n+1-kh} - 2\binom{2n}{n-kh} + \binom{2n}{n-1-kh} \right). \tag{21}$$

The *average height* of a tree with n nodes is S_n/A_{nn}, where S_n is the finite sum

$$S_n = \sum_{h \geq 1} h(A_{nh} - A_{n,h-1}) \tag{22}$$

$$= \sum_{h \geq 1} h(B_{n,h-1} - B_{nh})$$

$$= \sum_{h \geq 0} B_{nh}$$

$$= (2\pi i)^{-1} \int^{(0+)} \frac{du}{u^{n+1}} (1-u)^2 (1+u)^{2n-2} \sum_{h \geq 1} \frac{u^h}{1-u^h}$$

$$= (2\pi i^{-1}) \int^{(0+)} \frac{du}{u^{n+1}} (1-u)^2 (1+u)^{2n-2} \sum_{k \geq 1} d(k) u^k.$$

As usual, $d(k)$ denotes the number of positive divisors of k. Therefore,

$$S_{n+1} = \sum_{k\geqslant 1} d(k)\left(\binom{2n}{n+1-k} - 2\binom{2n}{n-k} + \binom{2n}{n-1-k}\right). \tag{23}$$

We shall now proceed to obtain an asymptotic series for the sum

$$f_a(n) = \sum_{k\geqslant 1}\left(\binom{2n}{n+a-k}\Big/\binom{2n}{n}\right) d(k), \qquad \text{fixed } a, \quad n\to\infty, \tag{24}$$

and this will lead to an asymptotic series for S_n.

Let $x = (k-a)/n$. By Stirling's approximation we have

$$\binom{2n}{n+a-k}\Big/\binom{2n}{n} = \exp\left(-2n\left(\frac{x^2}{1\cdot 2}+\frac{x^4}{3\cdot 4}+\cdots\right)+\left(\frac{x^2}{2}+\frac{x^4}{4}+\cdots\right)\right.$$
$$\left.-\frac{1}{6n}(x^2+x^4+\cdots)+O(x^2n^{-3})\right), \tag{25}$$

when $-\frac{1}{2} < x < \frac{1}{2}$, and

$$\binom{2n}{n+a-k}\Big/\binom{2n}{n} = O(\exp(-n^{2\varepsilon}))$$

when $k \geqslant n^{1/2+\varepsilon}+a$, for all fixed $\varepsilon > 0$. Therefore the sum of all terms for $k \geqslant n^{1/2+\varepsilon}+a$ in (24) is negligible, being $O(n^{-m})$ for all $m > 0$, and we may take $x = O(n^{-1/2+\varepsilon})$ in (25).

We now turn to the asymptotic behavior of the function

$$g_b(n) = \sum_{k\geqslant 1} k^b d(k)\exp(-k^2/n), \qquad \text{fixed } b, \quad n\to\infty. \tag{26}$$

Again the terms for $k \geqslant n^{1/2+\varepsilon}$ are negligible, so we can use (25) to express f in terms of g:

$$f_a(n) = g_0(n) + \frac{2a}{n}g_1(n) - \frac{a^2}{n}g_0(n) + \frac{4a^2+1}{2n^2}g_2(n) - \frac{1}{6n^3}g_4(n)$$
$$-\frac{2a^3+a}{n^2}g_1(n) + \frac{4a^3+5a}{3n^3}g_3(n) - \frac{a}{3n^4}g_5(n) + O(n^{-2+\varepsilon}g_0(n)). \tag{27}$$

In principle such an expansion could be carried out as far as we like. Hence, the problem of obtaining an asymptotic expansion for $f_a(n)$ reduces to the analogous problem for $g_b(n)$.

The behavior of $g_b(n)$ can be derived by starting with the well-known formula

$$e^{-x} = (2\pi i)^{-1} \int_{c-i\infty}^{c+i\infty} \Gamma(z) x^{-z}\, dz, \qquad c > 0, \quad x > 1, \tag{28}$$

obtained, for example, by Fourier inversion of $\Gamma(c+2\pi it)$. Then since $\zeta(z)^2 = \sum_{k\geq 1} d(k)/k^z$, we find

$$\begin{aligned} g_b(n) &= \sum_{k\geq 1} (2\pi i)^{-1} \int_{c-i\infty}^{c+i\infty} n^z \Gamma(z) k^{b-2z} d(k)\, dz \\ &= (2\pi i)^{-1} \int_{c-i\infty}^{c+i\infty} n^z \Gamma(z) \zeta(2z-b)^2\, dz, \end{aligned} \tag{29}$$

where now $c > \frac{1}{2}(b+1)$. Let q be a fixed positive number. When $\mathrm{Re}(s) \geq -q$, $\zeta(s) = O(|s|^{q+\frac{1}{2}})$ as $s \to \infty$. Since $n^z\Gamma(z)$ gets small on vertical lines we can shift the line of integration to the left as far as we please if we only take the residues into account. There is a double pole at $z = \frac{1}{2}(b+1)$, and possibly some simple poles at $z = 0, -1, -2, \ldots$. Let $w = z - \frac{1}{2}(b+1)$, we have

$$\begin{aligned} n^z \Gamma(z) \zeta(2z-b)^2 &= n^{(b+1)/2} \Gamma(\tfrac{1}{2}(b+1))(1 + w \ln n + O(w^2)) \\ &\quad \times (1 + w\psi(\tfrac{1}{2}(b+1)) + O(w^2))((2w)^{-2} + \gamma/w + O(1)), \end{aligned}$$

where $\psi(z) = \Gamma'(z)/\Gamma(z)$, hence the residue at the double pole is

$$n^{\frac{1}{2}(b+1)} \Gamma(\tfrac{1}{2}(b+1)) (\tfrac{1}{4} \ln n + \tfrac{1}{4}\psi(\tfrac{1}{2}(b+1)) + \gamma). \tag{30}$$

The residue at $z = -k$ is

$$n^{-k}(-1)^k \zeta(-2k-b)^2/k! = n^{-k}(-1)^k B_{2k+b+1}^2/(2k+b+1)^2 k! \tag{31}$$

which is almost always zero when b is even. The sum of (30) and (31) for all $k \geq 0$ gives an asymptotic series for $g_b(n)$. Hence, we have, for all $m > 0$,

$$\begin{aligned} g_0(n) &= \tfrac{1}{4}(\pi n)^{\frac{1}{2}} \ln n + (\tfrac{3}{4}\gamma - \tfrac{1}{2}\ln 2)(\pi n)^{\frac{1}{2}} + \tfrac{1}{4} + O(n^{-m}); \\ g_1(n) &= \tfrac{1}{4} n \ln n + \tfrac{3}{4}\gamma n + (\tfrac{1}{144}) - (\tfrac{1}{14400}) n^{-1} + O(n^{-2}); \\ g_2(n) &= (n/8)(\pi n)^{\frac{1}{2}} \ln n + (\tfrac{1}{4} + \tfrac{3}{8}\gamma - \tfrac{1}{4}\ln 2) n (\pi n)^{\frac{1}{2}} + O(n^{-m}); \end{aligned} \tag{32}$$

etc. These formulas have been verified by computer calculation. For example, when $n = 10, g_0(n) = 3.96042$ and $\frac{1}{4}(\pi n)^{\frac{1}{2}} \ln n + (\frac{3}{4}\gamma - \frac{1}{2}\ln 2)(\pi n)^{\frac{1}{2}} + \frac{1}{4} = 3.96041$.

Returning to our original problem about trees, we have

$$\begin{aligned} S_{n+1}/(n+1) A_{n+1,n+1} &= f_1(n) - 2f_0(n) + f_{-1}(n) \\ &= (-2/n) g_0(n) + (4/n^2) g_2(n) + O(n^{-\frac{3}{2}} \log n) \end{aligned} \tag{33}$$

by (4), (23), (24), and (27), and this equals $(\pi n)^{-\frac{1}{2}} - \frac{1}{2} n^{-1} + O(n^{-\frac{3}{2}} \log n)$. We have proved the following result.

THEOREM. The average height of a planted plane tree with n nodes, considering all such trees to be equally likely, is

$$(\pi n)^{1/2} - \tfrac{1}{2} + O(n^{-1/2}\log n). \tag{34}$$

The same method can be used to obtain as many further terms of the expansion as desired. The factor $\log n$ in the error term turns out to be unnecessary.

References†

1. Feller, W., "An Introduction to Probability Theory and its Applications," Vol. 1, 2nd ed. Wiley, New York, 1957.
2. Harris, T. E., First passage and recurrence distributions, *Trans. Amer. Math. Soc.* **73**, 471–486 (1952).
3. Knuth, D. E., "The Art of Computer Programming," Vol. 1. Addison-Wesley, Reading, Massachusetts, 1968.
4. Kreweras, G., Sur les eventails de segments, *Cahiers du Bureau Universitaire de Recherche Operationelle* **15**, 1–41 (1970).
5. Lagrange, J. L., Recherches sur les suites récurrentes, *in* "Oeuvres de Lagrange," Vol. 4, pp. 149–251. Paris, 1869.
6. Rényi, A., and Szekeres, G., On the height of trees, *Austral. J. Math.* **7**, 497–507 (1967).
7. Riordan, J., The Enumeration of Trees By Height and Diameter, *IBM J. Res. Develop.* **4**, 473–478 (1960).
8. Riordan, J., Ballots and trees, *J. Combinatorial Theory* **6**, 408–411 (1969).

† We wish to thank Prof. John Riordan for pointing out references [2] and [4].