THE AVERAGE HEIGHT OF PLANTED PLANE TREES

BY

N. G. DE BRUIJN

D. E. KNUTH

S. O. RICE

STAN-CS-218-71

MAY, 1971

COMPUTER SCIENCE DEPARTMENT

School of Humanities and Sciences

STANFORD UNIVERSITY

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Stanford University	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

THE AVERAGE HEIGHT OF PLANTED PLANE TREES

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report, May 1971

5. AUTHOR(S) (First name, middle initial, last name)

N. G. de Bruijn, D. E. Knuth, and S. O. Rice

6. REPORT DATE May 1971	7a. TOTAL NO. OF PAGES 7	7b. NO. OF REFS 2
8a. CONTRACT OR GRANT NO. ONR N-00014-67-A-0112-0057 NR 044-402. b. PROJECT NO. NSF GJ-992 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) STAN-CS-71-218 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None	

10. DISTRIBUTION STATEMENT

Releasable without limitations on dissemination.

11. SUPPLEMENTARY NOTES Details of illustrations in this document may be better studied on microfiche	12. SPONSORING MILITARY ACTIVITY Mathematics Program Office of Naval Research Arlington, Virginia 22217

13. ABSTRACT

An asymptotic expression for the average height of a planted plane tree is derived, based on an asymptotic series for sums such as $\sum_{k\geq 1}\binom{2n}{n+k}d(k)$ and $\sum_{k\geq 1} e^{-k^2/n}\,d(k)$, where $d(n)$ is the number of divisors of n.

DD FORM 1 NOV 68 1473 (PAGE 1)

S/N 0101-807-6801

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Planted plane tree, height of tree, ballot problem, random walk, asymptotic series, divisor function, gamma function, zeta function.						

The Average Height of Planted Plane Trees

by N. G. de Bruijn, D. E. Knuth,* and S. O. Rice

Abstract

An asymptotic expression for the average height of a planted plane tree is derived, based on an asymptotic series for sums such as $\sum_{k\geq 1}\binom{2n}{n+k}d(k)$ and $\sum_{k\geq 1} e^{-k^2/n}\, d(k)$, where $d(n)$ is the number of divisors of n.

AMS 1970 subject classifications: Primary 05C05, 10H25; 95A15, 10A40

Keywords: Planted plane tree, height of tree, ballot problem, random walk, asymptotic series, divisor function, gamma function, zeta function.

* 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.

The Average Height of Planted Plane Trees

by N. G. de Bruijn, D. E. Knuth, and S. O. Rice

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 5 trees with five nodes and height 4, namely

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, p. 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 (i.e., rooted, unordered) trees has been solved by Renyi and Szekeres [6]. Our principal results are stated in equations (32) and (34) below.

Trees appear in many disguises, and in particular there is a natural correspondence between trees of height $\le h$ 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 Figure 1, the vertical component of successive positions describes a path of length $2n-1$ from 1 to 0; for example, the path in Figure 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 T. E. Harris in 1952 [2], 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 $\le h+1$, again by the construction indicated in Figure 1. The ballot sequence corresponding to that tree is AABBAABAABABABBBAABABB.

Figure 1. A tree as a random walk.

We shall begin our study of the asymptotic properties of height by reviewing some known results. Let A_{nh} be the number of trees with n nodes and height $\leq h$, and let

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

be the corresponding generating function. We obtain all trees with height $\leq h+1$ by taking a root node and attaching zero or more subtrees each of which has height $\leq h$; therefore

$$A_{h+1}(z) = z(1 + A_h(z) + A_h(z)^2 + A_h(z)^3 + \ldots)$$

$$= z/(1 - A_h(z)) \quad , \quad \text{for } h \geq 0 \ . \tag{2}$$

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

$$A_{n,h+1} = A_{n-1,h+1}A_{1,h} + A_{n-2,h+1}A_{2,h} + \cdots + A_{1,h+1}A_{n-1,h} \quad , \quad \text{for } n \geq 2 \ , \ h \geq 0 \ , \tag{3}$$

from which it is easy to prepare a table of the first few values:

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

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

$$A_{nn} = A_{nn} = \binom{2n-2}{n-1} \frac{1}{n} \quad , \quad n \geq n \ , \tag{4}$$

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

Iteration of (2) yields a continued fraction representation of $A_h(z)$, e.g.

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

This suggests expressing the generating function as a quotient of polynomials,

$$A_h(z) = \frac{zP_h(z)}{P_{h+1}(z)} \ , \tag{6}$$

where

$$P_0(z) = 0 \ , \ P_1(z) = 1 \ , \ P_{h+1}(z) = P_h(z) - zP_{h-1}(z) \ . \tag{7}$$

The solution to this recurrence is

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

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

$$P_h\left(\frac{1}{4\cos^2\theta}\right) = \frac{\sin h\theta}{\sin\theta(2\cos\theta)^{h-1}} ,$$

$$A_h\left(\frac{1}{4\cos^2\theta}\right) = \frac{\sin h\theta}{2\cos\theta\sin(h+1)\theta} . \tag{9}$$

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

$$P_h(z) = \sum_{0\le k<h} \binom{h-1-k}{k}(-z)^k , \qquad \text{for } h \ge 1 ; \tag{10}$$

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,

$$A_h(z) - A_{h-1}(z) = \frac{z^h}{P_{h+1}(z)P_h(z)} ; \tag{11}$$

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 $1/(4\cos^2(j\pi/h))$, for $1 \le j < h/2$; and we obtain a partial fraction expansion of the generating function,

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

where

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

$$a_{2m} = -m , \quad b_{2m} = 0 ; \quad a_{2m+1} = \frac{-m(2m+1)}{6(m+1)} , \quad b_{2m+1} = \frac{1}{m+1} , \text{ for } m \ge 1 . \tag{13}$$

This leads immediately to the "explicit" formula

$$A_{nh} = \frac{1}{h+1} \sum_{1\le j\le h/2} 4^n \sin^2\frac{j\pi}{h+1} \cos^{2n-2}\frac{j\pi}{h+1} , \text{ for } n \ge 2 . \tag{14}$$

(It is rather remarkable that this formula gives a constant value for fixed n and all $h \ge n$. It is perhaps even more remarkable that Lagrange derived a formula in 1775 which essentially includes this as a special case! See [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

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

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

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

$$= \frac{1}{2\pi i}\int^{(0+)} \frac{dz}{z^n}(1+u)\,\frac{1-u^h}{1-u^{h+1}} \qquad (16)$$

where

$$u = \frac{1-\sqrt{1-4z}}{1+\sqrt{1-4z}} \quad , \qquad (17)$$

by (6) and (8). Since

$$z = \frac{u}{(1+u)^2} \quad , \qquad (18)$$

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

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

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 <u>greater</u> than h :

$$B_{nh} = A_{nn} - A_{nh}$$

$$= \frac{1}{2\pi i}\int^{(0+)} \frac{du}{u^{n+1}}(1-u)^2(1+u)^{2n-2}\,\frac{u^{h+1}}{1-u^{h+1}} \quad . \qquad (20)$$

It follows that

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

The <u>average height</u> of a tree with n nodes is S_n/A_{nn} , where S_n is the (finite) sum

$$S_n = \sum_{h\ge 1} h(A_{nh} - A_{n,h-1}) = \sum_{h\ge 1} h(B_{n,h-1} - B_{nh})$$

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

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

$$= \frac{1}{2\pi i}\int^{(0+)} \frac{du}{u^{n+1}}(1-u)^2(1+u)^{2n-2}\sum_{k\ge 1} d(k)u^k \quad . \qquad (22)$$

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

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

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

$$f_a(n) = \sum_{k\ge 1} \frac{\binom{2n}{n+a-k}}{\binom{2n}{n}} d(k) , \quad \text{fixed } a , \ n\to\infty ; \tag{24}$$

and this will lead to an asymptotic series for S_n.

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

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

when $-\frac{1}{4} < x < \frac{1}{4}$, and

$$\frac{\binom{2n}{n+a-k}}{\binom{2n}{n}} = O(e^{-n^{2\epsilon}}) \qquad \text{when } k \ge n^{\frac{1}{2}+\epsilon} + a ,$$

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

We now turn to the asymptotic behavior of the function

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

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

$$f_a(n) = g_0(n) + \frac{2a}{n} g_1(n) - \frac{a^2}{n} g_0(n) + \frac{4a^2+1}{2n^2} g_2(n) - \frac{1}{6n^3} g_4(n)$$

$$- \frac{2a^3+a}{n^2} g_1(n) + \frac{4a^3+5a}{3n^3} g_3(n) - \frac{a}{3n^4} g_5(n) + O(n^{-2+\epsilon}\, g_0(n)) . \tag{27}$$

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

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

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

$$\varepsilon_b(n) = \sum_{k\ge 1} \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} n^z\,\Gamma(z)k^{b-2z}\,d(k)dz$$

$$= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} n^z\,\Gamma(z)\zeta(2z-b)^2\,dz \qquad (29)$$

where now $c > \frac{1}{2}(b+1)$. Let q be a fixed positive number. When $\mathrm{Re}(s) \ge -q$, $\zeta(s) = O(|s|^{q+\frac{1}{2}})$ as $s \to \infty$; and 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

$$n^z\Gamma(z)\zeta(2z-b)^2 = n^{\frac{1}{2}(b+1)}\,\Gamma(\tfrac{1}{2}(b+1))(1 + w(\ln n) + O(w^2))\cdot(1 + w\psi(\tfrac{1}{2}(b+1)) + O(w^2))\cdot(\frac{1}{w^2} + \frac{\gamma}{w} + O(1))$$

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

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

The residue at $z = -k$ is

$$\frac{n^{-k}(-1)^k}{k!}\zeta(-2k-b)^2 = \frac{n^{-k}(-1)^k}{(2k+b+1)^2\,k!}B_{2k+b+1}^2 \qquad (31)$$

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

$$\varepsilon_0(n) = \tfrac{1}{4}\sqrt{\pi n}\ln n + (\tfrac{\pi}{4}\gamma - \tfrac{1}{2}\ln 2)\sqrt{\pi n} + \tfrac{1}{4} + O(n^{-m}) \quad ;$$

$$\varepsilon_1(n) = \tfrac{1}{4}n\ln n + \tfrac{1}{4}\gamma n + \frac{1}{144} - \frac{1}{14400}n^{-1} + O(n^{-2}) \quad ;$$

$$\varepsilon_2(n) = \tfrac{n}{8}\sqrt{\pi n}\ln n + (\tfrac{1}{8} + \tfrac{\gamma}{8} - \tfrac{1}{4}\ln 2)n\sqrt{\pi n} + O(n^{-m}) \quad ; \qquad (32)$$

etc. These formulas have been verified by computer calculation; for example, when $n = 10$, $\varepsilon_0(n) = 3.960642$ and $\frac{1}{4}\sqrt{\pi n}\ln n + (\frac{\pi}{4}\gamma - \frac{1}{2}\ln 2)\sqrt{\pi n} + \frac{1}{4} = 3.960641$.

Returning to our original problem about trees, we have

$$\frac{g_{n+1}}{(n+1)A_{n+1,n+1}} = f_1(n) - 2f_0(n) + f_{-1}(n)$$

$$= \frac{-2}{n}\varepsilon_0(n) + \frac{4}{n^2}\varepsilon_2(n) + O(n^{-3/2}\log n) \qquad (33)$$

by (4), (25), (34), and (37), and this equals $\sqrt{\pi}\,n^{-\frac{1}{2}} - \frac{1}{2}n^{-1} + O(n^{-3/2}\log n)$. We have proved, in particular, the following result:

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

$$\sqrt{\pi n} - \frac{1}{2} + O(n^{-\frac{1}{2}} \log n) \quad . \tag{34}$$

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.

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

[1] Feller, William, An Introduction to Probability Theory and its Applications, 1, 2nd edition (Wiley, 1957).

[2] Harris, T. E., "First passage and recurrence distributions," Trans. Amer. Math. Soc. 73 (1952), 471-486.

[3] Knuth, Donald E., The Art of Computer Programming, 1 (Addison-Wesley, 1968).

[4] Kreweras, G., "Sur les eventails de segments," Cahiers du Bureau Universitaire de Recherche Operationelle 15 (1970), 1-41.

[5] Lagrange, J. L., "Recherches sur les suites récurrentes," Oeuvres de Lagrange 4 (Paris, 1869), 151-251.

[6] Rényi, A., and Szekeres, G., "On the height of trees," Australian J. Math. 7 (1967), 497-507.

[7] Riordan, John, "The enumeration of trees by height and diameter," IBM J. Res. Devel. 4 (1960), 473-478.

[8] Riordan, John, "Ballots and Trees," J. Comb. Theory 6 (1969), 408-411.