International Journal of Foundations of Computer Science

DEPTH AND PATH LENGTH OF HEAP ORDERED TREES

HELMUT PRODINGER
Institut für Algebra und Diskrete Mathematik, Technical University of Vienna, Wiedner Hauptstrasse 8–10, A-1040 Vienna, Austria,
Email: Helmut.Prodinger@tuwien.ac.at

Received 17 October 1995
Revised 24 April 1996
Communicated by R. L. Graham

ABSTRACT

A heap ordered tree of size n is a planted plane tree together with a bijection from the nodes to the set $\{1,\dots,n\}$ which is monotonically increasing when going from the root to the leaves. In a recent paper by Chen and Ni, the expectation and the variance of the depth of a random node in a random heap ordered tree of size n was considered. Here, we give additional results concerning level polynomials. From a historical point of view, we trace these trees back to an earlier paper by Prodinger and Urbanek and find formulæ that are proved in the paper by Chen and Ni by ad hoc computations already in a classic book by Greene and Knuth. Also, we mention that a paper by Bergeron, Flajolet and Salvy develops a more general theory of increasing trees, based on simply generated families of trees. Furthermore we consider the path length which is a natural concept when dealing with the depth. Expectation and variance are obtained, both explicitly and asymptotically.

Keywords: Increasing trees, depth, path length, Catalan numbers.

1. Introduction

A *heap ordered tree* with n nodes ("size n") might be described as a *planted plane tree* together with a bijection from the nodes to the set $\{1,\dots,n\}$ which is *monotonically increasing* when going from the root to the leaves.

In Ref. [2], the depth of a random node in a random heap ordered tree was considered. Here, we want to give a streamlined derivation of these results, together with new ones about the coefficients of the *level polynomials*. Then, we turn to the *path length*, which is the sum of the depths of all nodes in the heap ordered tree. Path length is the natural concept when analyzing search costs etc. (Compare Ref. [5]).

Finally we would like to mention, that heap ordered trees appear already in Ref. [7] where monotone labelings of trees where studied in general.

Things that are not explained here in detail are standard and can be found for instance in the textbooks.[5,6]

We would like to emphasize that the present note has a strong tutorial flavor as it emphasizes how heap ordered trees and related parameters should be analyzed in a most efficient, direct, and natural way.

Especially we want to draw the reader's attention to the fact that the important paper[1] develops a general theory of *increasing trees*, based on *simply generated families of trees*, where planted plane trees are just a special case. Although it emphasizes the asymptotic point of view, it has all the generating functions of interest, and getting explicit expressions (involving harmonic numbers etc.) is a complete routine task.

The enumeration of the numbers a_n of heap ordered trees of size n is easy and appears already in Ref. [7]:

The recursion is

$$a_{n+1} = \sum_{k\ge 1} \sum_{j_1+\ldots+j_k=n} \binom{n}{j_1,\ldots,j_k} a_{j_1}\ldots a_{j_k} \quad \text{for } n \ge 1, \quad a_1 = 1 \ ;$$

hence the exponential generating function $A(z) := \sum_{n\ge 0} a_n \frac{z^n}{n!}$ fulfills the differential equation

$$A'(z) = \frac{1}{1-A(z)} \qquad \text{with} \quad A(0) = 0 \ ,$$

with the solution

$$A(z) = 1 - \sqrt{1-2z} \ ,$$

so that

$$a_n = n!\, 2^{1-n}\, \mathcal{C}_n \ .$$

Here and later we use the notion of a (shifted) *Catalan number*

$$\mathcal{C}_n = \frac{1}{n}\binom{2n-2}{n-1} \ .$$

Considering any variety of trees, the concept of *splitting probabilities* is a good idea. We define them by

$$\pi_{n;j_1,\ldots,j_k} := \binom{n}{j_1,\ldots,j_k} \frac{a_{j_1}\ldots a_{j_k}}{a_{n+1}} = \frac{2^k}{\binom{2n}{n}} \mathcal{C}_{j_1}\ldots\mathcal{C}_{j_k} \ .$$

It is a simple exercise to show analytically that they indeed sum to 1, which is the same as to show that

$$\sum_{k\ge 0} \sum_{j_1+\ldots+j_k=n} 2^k\, \mathcal{C}_{j_1}\ldots\mathcal{C}_{j_k} = \binom{2n}{n} \ .$$

In terms of generating functions this means

$$\sum_{k\ge 0} 2^k \left(\frac{1-\sqrt{1-4z}}{2}\right)^k = \frac{1}{\sqrt{1-4z}} \ ,$$

which is obvious.

2. Level Polynomials of Heap Ordered Trees

When studying statistics on the depth of any variety of trees, it is customary to consider *level polynomials* $L_n(u)$. The coefficient of u^k herein should be the expected number of nodes on level k, the root being on level 0, so that $L_n(1) = n$. Compare Refs. [5] and [6].

Here, the (self-explanatory) recursion is

$$L_{n+1}(u) = 1 + u \sum_{k\ge 1} \sum_{j_1+\ldots+j_k=n} \pi_{n;j_1,\ldots,j_k} \Big(L_{j_1}(u) + \ldots + L_{j_k}(u) \Big) .$$

By symmetry, we rewrite the recursion as

$$L_{n+1}(u)\binom{2n}{n} = \binom{2n}{n} + u \sum_{k\ge 1} \sum_{j_1+\ldots+j_k=n} k\, 2^k\, \mathcal{C}_{j_1} \ldots \mathcal{C}_{j_k}\, L_{j_1}(u) .$$

Now we introduce the bivariate generating function

$$L(z,u) := \sum_{n\ge 0} L_n(u)\, \mathcal{C}_n\, z^n \ ;$$

then we get

$$\begin{aligned} \frac{\partial L(z,u)}{\partial z} &= \frac{1}{\sqrt{1-4z}} + u \sum_{k\ge 1} k\, 2^k\, L(z,u) \left(\frac{1-\sqrt{1-4z}}{2} \right)^{k-1} \\ &= \frac{1}{\sqrt{1-4z}} + \frac{2u\, L(z,u)}{1-4z} . \end{aligned}$$

The solution of this differential equation, subject to the condition $L(0,u) = 0$, is

$$L(z,u) = \frac{1}{2(1+u)} \Big((1-4z)^{-u/2} - (1-4z)^{1/2} \Big) .$$

Hence

$$[z^n]L(z,u) = \frac{1}{2(1+u)} \left(\binom{-u/2}{n} (-4)^n + 2\, \mathcal{C}_n \right) .$$

Since (with signless Stirling numbers of the first kind, compare Ref. [3])

$$\binom{-u/2}{n} = \frac{1}{n!} \sum_{j\ge 0} \left[{n \atop j} \right] (-1)^{n-j} \left(-\tfrac{u}{2}\right)^j \qquad \text{and} \qquad \frac{1}{1+u} = \sum_{j\ge 0} (-u)^j ,$$

we find

$$[u^k z^n]L(z,u) = \frac{4^n (-1)^k}{2n!} \sum_{j=0}^{k} \left[{n \atop j} \right] \left(-\tfrac{1}{2}\right)^j + \mathcal{C}_n (-1)^k .$$

Dividing this by $\mathcal{C}_n$, we find the following theorem.

Theorem 1 *The expected number of nodes on level k in random heap ordered trees of size n is given by*

$$\frac{4^n(-1)^k(n-1)!}{2(2n-2)!}\sum_{j=0}^{k}\left[{n \atop j}\right](-\tfrac{1}{2})^j+(-1)^k\ .$$

$\diamondsuit$

Since $L_n(1)=n$, the quantity $L_n(u)/n$ is a probability generating function, and $L_n'(1)/n$ is just the expected depth of a random heap ordered tree of size n. We can perform the differentation with respect to u, followed by the substitution $u=1$, on the function $L(z,u)$, with the result

$$M_1(z):=-\frac{z}{2\sqrt{1-4z}}+\frac{1}{8\sqrt{1-4z}}\log\frac{1}{1-4z}\ .$$

In order to get $[z^n]M_1(z)$, we have to know the formula

$$[z^n]\frac{1}{\sqrt{1-z}}\log\frac{1}{1-z}=4^{-n}\binom{2n}{n}\left(2H_{2n}-H_n\right)\ ,$$

which was proved in Ref. [2] by an *ad hoc method*. However in Greene/Knuth[4] we find the general result ($m\in\mathbb{C}$)

$$[z^n]\frac{1}{(1-z)^{m+1}}\log\frac{1}{1-z}=\binom{n+m}{n}\left(H_{n+m}-H_m\right)\ ;$$

the interpretation of $H_{n+m}-H_m$ is

$$H_{n+m}-H_m=\frac{1}{m+1}+\frac{1}{m+2}+\ldots+\frac{1}{m+n}\ .$$

The formula follows now by a simple computation;

$$\begin{aligned}
H_{n-\frac{1}{2}}-H_{-\frac{1}{2}} &= \frac{1}{\frac{1}{2}}+\frac{1}{\frac{3}{2}}+\frac{1}{\frac{5}{2}}+\ldots+\frac{1}{n-\frac{1}{2}}\\
&= 2\Big(1+\frac{1}{3}+\frac{1}{5}+\ldots+\frac{1}{2n-1}\Big)\\
&= 2H_{2n}-2\Big(\frac{1}{2}+\frac{1}{4}+\ldots+\frac{1}{2n}\Big)\\
&= 2H_{2n}-H_n\ .
\end{aligned}$$

Hence

$$[z^n]M_1(z)=-\frac{1}{2}\binom{2n-2}{n-1}+\frac{1}{8}\binom{2n}{n}\left(2H_{2n}-H_n\right)\ .$$

We must divide this quantity by $n\mathcal{C}_n=\binom{2n-2}{n-1}$ and get a result of Ref. [2] as a corollary.

Theorem 2 *The expected depth d_n of a random heap ordered tree of size n is given by*

$$d_n=\Big(1-\frac{1}{2n}\Big)\left(H_{2n}-\frac{1}{2}H_n\right)-\frac{1}{2}\ .$$

$\diamondsuit$

In order to compute the variance, we differentiate $L(z,u)$ twice with respect to u, followed by $u=1$, which gives

$$M_2(z) := \frac{z}{2\sqrt{1-4z}} - \frac{1}{8\sqrt{1-4z}}\log\frac{1}{1-4z} + \frac{1}{16\sqrt{1-4z}}\log^2\frac{1}{1-4z} \ .$$

For that we need another formula from Greene/Knuth.[4]

$$[z^n]\frac{1}{(1-z)^{m+1}}\log^2\frac{1}{1-z} = \binom{n+m}{n}\Big(\big(H_{n+m}-H_m\big)^2 - \big(H^{(2)}_{n+m} - H^{(2)}_m\big)\Big) \ .$$

The special instance $m=-\frac{1}{2}$ becomes

$$[z^n]\frac{1}{\sqrt{1-z}}\log^2\frac{1}{1-z} = 4^{-n}\binom{2n}{n}\Big(\big(2H_{2n}-H_n\big)^2 - \big(4H^{(2)}_{2n} - H^{(2)}_n\big)\Big) \ .$$

The variance is obtained via $[z^n]M_2(z)/n\mathcal{C}_n + d_n - d_n^2$, which yields another result from Ref. [2].

Theorem 3 *The variance* Var_n *of the depth of a random heap ordered tree of size* n *is given by*

$$\mathrm{Var}_n = \Big(1-\frac{1}{2n}\Big)\Big[\big(H_{2n}-\frac{1}{2}H_n\big)^2 - \big(H^{(2)}_{2n} - \frac{1}{4}H^{(2)}_n\big)\Big] - d_n^2 \ .$$

$\diamondsuit$

Observe that the formula[2] contains an extra square somewhere, which is erroneous. As a consequence, the asymptotic formula is also in error; the correct version is

$$\mathrm{Var}_n = \frac{\log n}{2} + \log 2 + \frac{\gamma}{2} - \frac{1}{4} - \frac{\pi^2}{8} + \mathcal{O}\Big(\frac{\log n}{n}\Big) \ .$$

3. Path Length of Heap Ordered Trees

As indicated earlier, we now engage into the path length, i.e. the sum of the distances of all nodes to the root. Concerning the expected value, it must be the expected depth times the number of nodes. However, for the variance, such a relation does not hold; henceforth independent computations are necessary.

We can set up a sequence of probability generating functions for the path length, viz.

$$F_{n+1}(u) = u^n \sum_{k\ge 0}\sum_{j_1+\ldots+j_k=n} \pi_{n;j_1,\ldots,j_k} F_{j_1}(u)\ldots F_{j_k}(u), \qquad F_0(u)=0 \ ,$$

or

$$F_{n+1}(u)\binom{2n}{n} = u^n \sum_{k\ge 0}\sum_{j_1+\ldots+j_k=n} 2^k\,\mathcal{C}_{j_1}\ldots\mathcal{C}_{j_k}\,F_{j_1}(u)\ldots F_{j_k}(u), \qquad F_0(u)=0 \ .$$

Again, let us define the bivariate generating function

$$F(z,u) := \sum_{n\ge 0} F_n(u)\,\mathcal{C}_n\,z^n \ ;$$

then

$$\frac{\partial F(z,u)}{\partial z} = \frac{1}{1-2F(zu,u)} \ .$$

This functional differential equation is harder than the corresponding equation in the instance of the depth; hence we cannot expect to solve it explicitly. However, it contains enough information to get all moments, at least in principle.

The generating function $f(z)$ of the expected values is given by

$$f(z) = \frac{\partial}{\partial u} F(z,u)\bigg|_{u=1} \ ,$$

which leads to the differential equation

$$f'(z) = \frac{2\,f(z)}{1-4z} + \frac{2z}{(1-4z)^{3/2}}, \qquad f(0)=0 \ .$$

The solution is

$$f(z) = -\frac{z}{2\sqrt{1-4z}} + \frac{1}{8\sqrt{1-4z}} \log\frac{1}{1-4z} \ ,$$

which is the function $M_1(z)$ from before, whence we obtain

Theorem 4 *The expected path length l_n of a random heap ordered tree of size n is given by*

$$l_n = \Big(n-\frac{1}{2}\Big)\big(H_{2n}-\frac{1}{2}H_n\big) - \frac{n}{2} \ .$$

$\diamondsuit$

For the second factorial moment, we have to consider

$$g(z) = \frac{\partial^2}{\partial u^2} F(z,u)\bigg|_{u=1} \ .$$

The differential equation for $g(z)$ is

$$g'(z) = \frac{8}{(1-4z)^{3/2}}\Big(\frac{z}{\sqrt{1-4z}}+f(z)\Big)^2 + \frac{2}{1-4z}\Big(\frac{2z^2}{(1-4z)^{3/2}} + 2z\,f'(z) + g(z)\Big) \ .$$

Maple solves it (subject to $g(0)=0$),

$$\begin{aligned} g(z) &= \frac{1}{32}\frac{1}{(1-4z)^{3/2}}\log^2\frac{1}{1-4z} - \frac{1}{16}\frac{1}{\sqrt{1-4z}}\log^2\frac{1}{1-4z} \\ &+ \frac{1}{16}\frac{1}{(1-4z)^{3/2}}\log\frac{1}{1-4z} - \frac{5}{16}\frac{1}{\sqrt{1-4z}}\log\frac{1}{1-4z} \\ &+ \frac{3}{32}\frac{1}{(1-4z)^{3/2}} + \frac{1}{16\sqrt{1-4z}} - \frac{5}{32}\sqrt{1-4z} \ . \end{aligned}$$

Now we need the formulæ

$$[z^n]\frac{1}{(1-z)^{3/2}}\log\frac{1}{1-z} = 4^{-n}(2n+1)\binom{2n}{n}\big(H_{n+\frac12}-H_{\frac12}\big)$$

and

$$[z^n]\frac{1}{(1-z)^{3/2}}\log^2\frac{1}{1-z} = 4^{-n}(2n+1)\binom{2n}{n}\Big(\big(H_{n+\frac12}-H_{\frac12}\big)^2 - \big(H^{(2)}_{n+\frac12}-H^{(2)}_{\frac12}\big)\Big)\ .$$

We observe

$$H_{n+\frac12}-H_{\frac12} = \frac{2}{2n+1} - 2 + 2H_{2n} - H_n$$

and

$$H^{(2)}_{n+\frac12}-H^{(2)}_{\frac12} = \frac{4}{(2n+1)^2} - 4 + 4H^{(2)}_{2n} - H^{(2)}_n\ .$$

Now we have enough formulæ at hand to get $[z^n]g(z)$. The variance is obtained via $[z^n]g(z)+l_n-l_n^2$. With the help of Maple we find the following theorem.

Theorem 5 *The variance V_n of the path length of a random heap ordered tree of size n is given by*

$$\begin{aligned} V_n &= n^2\Big(\frac14 H^{(2)}_n - H^{(2)}_{2n} + \frac32\Big) \\ &+ n\Big(\frac12 H_n - H_{2n} - \frac14 H^{(2)}_n + H^{(2)}_{2n} - \frac34\Big) \\ &- \frac14 H_n + \frac12 H_{2n} + \frac{1}{16} H^{(2)}_n - \frac14 H^{(2)}_{2n}\ . \end{aligned}$$

Asymptotically, we get

$$V_n = n^2\Big(\frac32 - \frac{\pi^2}{8}\Big) - \frac{n\log n}{2} + n\Big(-\frac12 - \log 2 + \frac{\pi^2}{8} - \frac{\gamma}{2}\Big) + \mathcal{O}(\log n)\ .$$

The constant $\frac32 - \frac{\pi^2}{8}$ is 0.266299449.

$\diamondsuit$

It is interesting to note that the asymptotic orders of depth (expectation and variance) are resp. $\log n$ and $\log n$, whereas for the path length we get $n\log n$ and n^2.

References

1. F. Bergeron, P. Flajolet, and B. Salvy. Varieties of increasing trees. *Lecture Notes in Computer Science*, 581:24–48, 1992.
2. W.-C. Chen and W.-C. Ni. On the average altitude of heap–ordered trees. *International Journal of Foundations of Computer Science*, 15:99–109, 1994.
3. R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics.* Addison Wesley, 1989.
4. D. H. Greene and D. E. Knuth. *Mathematics for the analysis of algorithms.* Birkhäuser, Boston, second edition, 1982.
5. D. E. Knuth. *The Art of Computer Programming*, volume 3: Sorting and Searching. Addison-Wesley, 1973.
6. H. M. Mahmoud. *Evolution of Random Search Trees.* John Wiley, New York, 1992.
7. H. Prodinger and F.J. Urbanek. On monotone functions of tree structures. *Discrete Applied Mathematics*, 5:223–239, 1983.