Journal of Computational and Applied Mathematics 41 (1992) 237–245
North-Holland

CAM 1175

Distances in random plane-oriented recursive trees

Hosam M. Mahmoud

Department of Statistics / Computer & Information Systems, George Washington University, Washington, DC 20052, United States

Received 25 January 1991
Revised 23 July 1991

Abstract

Mahmoud, H.M., Distances in random plane-oriented recursive trees, Journal of Computational and Applied Mathematics 41 (1992) 237–245.

The average number of nodes in a stratum of random plane-oriented recursive trees is found. The expression is used to determine the exact probability distribution of the depth of the nth node. It is further shown that the limiting distribution of the normalized depth of this node is the standard normal distribution. Via martingales, the normalized external path length is shown to converge almost surely and in L^2 to a limiting random variable.

Keywords: Recursive tree; depth; path length; limit theorem; martingale.

1. Introduction

A tree is a connected graph without cycles (see [1] for basic properties). A tree on n vertices labeled $1, 2, \ldots, n$ is a rooted recursive tree of order n if the node labeled 1 is distinguished as the root, and for each k, $2 \leqslant k \leqslant n$, the labels of the vertices in the unique path joining the root to the vertex labeled k form an increasing sequence. We shall refer to this tree as the "usual" recursive tree. Notice that, by their definition, the children of the usual recursive trees are unordered. In this paper we study the new class of plane-oriented recursive trees when the trees obtained by different orderings of the children of a node of the usual recursive tree are considered as distinct trees. For the rest of this paper the term "tree" without qualification will refer to a plane-oriented recursive tree. The tree grows by adding new nodes at the candidate insertion positions. These are the "gaps" between the edges joining the immediate children of a node to their parent (thinking of the left of the leftmost edge and the right of the rightmost edge as gaps).

It is very convenient in several classes of trees to work with an extension of the trees of the class (see [7] for an example of an extension of binary trees). Such extensions are obtained by

Correspondence to: Prof. H.M. Mahmoud, Department of Statistics/Computer & Information Systems, George Washington University, Washington, DC 20052, United States.

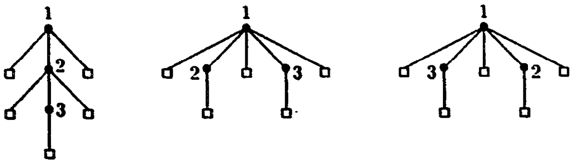

Fig. 1. All extended plane-oriented recursive trees on three vertices.

adding a different type of nodes called *external* at each possible insertion position. We extend a plane-oriented recursive tree by representing each insertion position as a square node joined by an edge to the would-be parent if insertion falls at that position. A node with j children thus has $j+1$ insertion positions, i.e., will have $j+1$ external nodes as immediate children in the extension, one in each gap. Figure 1 illustrates all extended plane-oriented recursive trees of order 3.

We shall analyze plane-oriented recursive trees under the uniform probability model, that is, we assume that all trees of a given order are equally likely. One of the reasons a class of random trees is considered interesting is the existence of a simple growth rule, i.e., a rule by which the addition of a node (to make the transition from a random tree of size $n-1$ to a tree of size n) will appear as if the new tree has been picked at random from its sample space. Such is the situation with the class of random plane-oriented recursive trees. It is not difficult to see that when all the insertion positions in the tree are equally likely, growing the tree by choosing any of the insertion positions with equal probability is equivalent to growing a random tree.

The usual recursive trees under a uniform probability model have been proposed as models for the spread of epidemics [10], for pyramid schemes [4], for the family trees of preserved copies of ancient or medieval texts [11] and for algorithms used to produce convex hulls in higher dimensions [3].

In a pyramid scheme where each entrant competes with those already participating, the experience gained in successful recruiting enhances the prospects of future success as captured by the growth rule of random plane-oriented recursive trees. Similarly, plane-oriented recursive trees under a uniform model will be a better model for any pyramid scheme represented by recursive trees where a nonuniform probability model (on the trees) accounting for the affinity of the nodes by their degrees is more appropriate than a uniform one.

In this paper we study the average number of external nodes in a stratum at distance k from the root of a random plane-oriented recursive tree of order n. This average information is sufficient to establish the exact probability distribution of D_n, the depth of the nth node, i.e., its distance from the root of the tree. This probability distribution involves Stirling numbers of the first kind and the generating function of these combinatorial objects is used to determine the mean and variance of D_n. The generating function will also enable us to prove that D_n is asymptotically normal. The mean of D_n is asymptotically $\frac{1}{2}\ln n$, as $n \to \infty$. The variance of D_n is also asymptotically of logarithmic order guaranteeing the convergence of $D_n/\ln n$ to $\frac{1}{2}$ in probability. Thus the depth of a node with a large index is about half of what it is in usual random recursive trees with high probability [9,10].

As the tree grows by the progressive insertion of nodes, two other cumulative random variables may serve as measures of the overall cost of the construction of the tree, or the cost of later processing of the whole tree if each internal (external) node is to be accessed equally often. The first random variable is the *internal path length*

$$I_n = \sum_{j=1}^{n} D_n,$$

and several of its properties can be derived from information about the different depths of the nodes in a tree. An insertion always increases the number of external nodes by 2, and thus a tree of order n will have $2n-1$ external nodes. Suppose they are indexed by $1, 2, \ldots, 2n-1$, from left to right, say, and that their depths are $x_1, \ldots, x_{2n-1}$. We define the second cumulative random variable X_n by

$$X_n = \sum_{j=1}^{2n-1} x_j.$$

This random variable is called the *external path length.* The strong dependence between the random variables x_j makes it difficult to compute the exact distribution of X_n. In this paper we find a martingale associated with the external path length. The convergence theorem of martingales will then imply that a normalized version of X_n has a limiting distribution, as $n \to \infty$. We show that there exists a random variable X such that $(X_n - n \ln n)/(2n)$ converges to X almost surely and in L^2. As a by-product, the method allows us to compute the mean and variance of the external path length.

To put this work in perspective, it extends several results in the usual recursive trees to plane-oriented recursive trees. The reader is referred to [2,8–10,13] for the counterparts in the usual recursive trees of properties of the plane-oriented recursive trees discussed in this paper.

2. The number of nodes in a stratum

Let Y_{nk} denote the number of external nodes at distance k from the root in a plane-oriented recursive tree of order n. Under our model of randomness, Y_{nk} is a random variable. In this section we determine $E[Y_{nk}]$, the average of this quantity. This average involves $\left[{s \atop j}\right]$, the signless Stirling numbers of the first kind of order s, where $\left[{s \atop j}\right]$ for nonnegative integers s and j is the coefficient of x^j in the product $\langle x \rangle_s = x(x+1)\cdots(x+s-1)$. The average of Y_{nk} also involves the quantity $1 \times 3 \times \cdots \times (2s-1)$ for positive integer s. We shall denote this quantity by $(2s-1)!!$. Another useful notation for this paper appears from the need to differentiate functions of the form $\langle z \rangle_n$ at $z = \frac{1}{2}$. The first two derivatives involve the quantities

$$\alpha_n^{(j)} = 1 + \frac{1}{3^j} + \frac{1}{5^j} + \cdots + \frac{1}{(2n-1)^j}, \quad j = 1, 2.$$

The quantities $\alpha_n^{(j)}$ are related to the harmonic numbers by

$$\alpha_n^{(j)} = H_{2n-1}^{(j)} - \frac{1}{2^j} H_{n-1}^{(j)},$$

where $H_r^{(j)} = 1 + 1/2^j + \cdots + 1/r^j$. (It is customary to drop the superscript when it is 1, and we shall follow this notation in this paper.) We formulate our first result next.

Theorem 1. *The average number of external nodes at distance k from the root in a plane-oriented recursive tree with n nodes is given by*

$$E[Y_{nk}] = \frac{2^{n-k}}{(2n-3)!!}\left[{n \atop k}\right], \quad \textit{for } n \geqslant 2.$$

Proof. The tree T_n evolves from T_{n-1} by an insertion of the nth node at level D_n. An insertion at level j will convert an external node at that level into an internal node and two new sibling external nodes will appear on level j: one to the left and one to the right of the new internal node, and also an external node will appear at level $j+1$ as the only child of the new internal node. The net increase in Y_{nj} is therefore 1. So, an insertion at level $j-1$ will also increase Y_{nj} by 1. Level j is not affected by insertion at any levels other than j and $j-1$. Thus,

$$E\left[Y_{nj} \mid D_n = k\right] = \begin{cases} E[Y_{n-1}, j] + 1, & \text{if } k = j, \\ E[Y_{n-1}, j] + 1, & \text{if } k = j-1, \\ E[Y_{n-1}, j], & \text{otherwise.} \end{cases}$$

It follows from unconditioning the last relation that

$$E\left[Y_{r,j}\right] = E\left[Y_{n-1,j}\right] + \Pr(D_n = j) + \Pr(D_n = j-1). \tag{1}$$

As all $2n-3$ external nodes of T_{n-1} are equally likely,

$$\Pr(D_n = r \mid Y_{n-1,r}) = \frac{Y_{n-1,r}}{2n-3}.$$

That is, unconditionally,

$$\Pr(D_n = r) = \frac{E[Y_{n-1,r}]}{2n-3}. \tag{2}$$

So, (1) can be rewritten as

$$E\left[Y_{nj}\right] = \frac{2n-2}{2n-3} E[Y_{n-1}, j] + \frac{1}{2n-3} E\left[Y_{n-1,j-1}\right]. \tag{3}$$

To solve this recurrence we introduce the leaf polynomials

$$L_n(z) = \sum_{j=0}^{\infty} E\left[Y_{nj}\right] z^j.$$

Multiplying both sides of (3) by z^j and summing over j (noting that $E[Y_{n0}] = 0$), we obtain the following recurrence on the leaf polynomials:

$$L_n(z) = \frac{2n-2+z}{2n-3} L_{n-1}(z),$$

with $L_1(z) \equiv z$. This recurrence has the solution

$$L_n(z) = \frac{2^n}{(2n-3)!!} \tfrac{1}{2}z\left(\tfrac{1}{2}z + 1\right) \cdots \left(\tfrac{1}{2}z + n - 1\right).$$

But the generating function $\langle \frac{1}{2}z \rangle_n$ generates the sequence $2^{-k}\left[{n \atop k}\right]$, $k = 0, 1, \ldots, n$, and the theorem follows. □

3. Exact and limiting distributions for the depth

According to (2), the exact probability distribution of D_n is linked to the average of Y_{nk}. So, the average developed in Theorem 1 provides an immediate proof for the next theorem.

Theorem 2.

$$\Pr(D_n = k) = \frac{2^{n-1-k}}{(2n-3)!!}\left[{n-1 \atop k}\right].$$

The mean value for D_n follows from this exact distribution. It is given by

$$E[D_n] = \frac{2^{n-1}}{(2n-3)!!}\sum_{k=0}^{n-1}\frac{k}{2^k}\left[{n-1 \atop k}\right],$$

and the remaining sum can be handled by differentiating $\langle z\rangle_{n-1}$ once at $z=\frac{1}{2}$. This yields the simple expression

$$E[D_n] = \alpha_{n-1}^{(1)} = H_{2n-3} - \tfrac{1}{2}H_{n-2}.$$

The well-known asymptotics of the harmonic numbers (see [5], for example) admit an asymptotic development with high accuracy:

$$E[D_n] = \tfrac{1}{2}\ln n + \ln 2 + \tfrac{1}{2}\gamma + O\left(\frac{1}{n}\right).$$

Similarly, the second factorial moment of D_n is obtained by first finding an expression for $\sum_{k=0}^{n-1}\left[{n-1 \atop k}\right]k(k-1)/2^k$ from the second derivative of $\langle z\rangle_{n-1}$ at $z=\frac{1}{2}$. The variance follows and is given by

$$\operatorname{Var}[D_n] = \alpha_{n-1}^{(1)} - \alpha_{n-1}^{(2)} = \tfrac{1}{2}\ln n + \ln 2 + \tfrac{1}{2}\gamma - \tfrac{1}{8}\pi^2 + O\left(\frac{1}{n}\right).$$

From an application of Chebychev's inequality we can conclude that

$$\frac{D_n}{\ln n} \to \tfrac{1}{2}, \quad \text{in probability.}$$

The average internal path length I_n can be calculated from

$$E[I_n] = \sum_{k=1}^{n} E[D_k],$$

and from the averages for the depths, the expression

$$E[I_n] = \left(H_{2n-3} - \tfrac{1}{2}H_{n-2}\right)\left(n - \tfrac{1}{2}\right) - \tfrac{1}{2}(n-1)$$

follows. For large n, $E[I_n] \sim \frac{1}{2}n\ln n$.

We next use the exact probability distribution of D_n to prove the asymptotic normality of a normalized version of D_n.

Theorem 3. *The normalized random variable D_n^* defined by*

$$D_n^* = \frac{D_n - \frac{1}{2} \ln n}{\sqrt{\frac{1}{2} \ln n}}$$

has the limiting distribution $\mathcal{N}(0, 1)$, the standard normal distribution with mean zero and variance one.

Proof. Introduce $M_n(t)$, the moment generating function of D_n^*, i.e.,

$$M_n(t) = E\left[e^{D_n^* t}\right],$$

for any fixed real number t. For notational convenience denote $\frac{1}{2}$ ln n by a_n. From the exact probability distribution of Theorem 2,

$$\begin{aligned} M_n(t) &= \sum_{k=0}^{\infty} \exp\left\{\frac{k-a_n}{\sqrt{a_n}} t\right\} \Pr(D_n = k) \\ &= \frac{2^{n-1} e^{-\sqrt{a_n} t}}{(2n-3)!!} \sum_{k=0}^{n-1} \left(\tfrac{1}{2} e^{t/\sqrt{a_n}}\right)^k \left[{n-1 \atop k}\right] \\ &= \frac{2^{n-1} e^{-\sqrt{a_n} t}}{(2n-3)!!} \left\langle \tfrac{1}{2} e^{t/\sqrt{a_n}} \right\rangle_{n-1} \\ &= \frac{2^{n-1} e^{-\sqrt{a_n} t}}{(2n-3)!!} \Gamma^{-1}\left(\tfrac{1}{2} e^{t/\sqrt{a_n}}\right) \Gamma\left(n + \tfrac{1}{2} e^{t/\sqrt{a_n}} - 1\right). \end{aligned}$$

But $(2n-3)!! \sim (2n)!/(2^{n+1} n(n!))$ and the first gamma function approaches $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, as $n \to \infty$. And thus by the Stirling approximation for the second gamma function, for large n,

$$M_n(t) \sim \frac{4^n (n!) n \, e^{-\sqrt{a_n} t}}{(2n)! \sqrt{\pi}} \Gamma(n) n^{\exp(t/\sqrt{a_n})/2-1}.$$

The terms $n! n \Gamma(n)/(2n)!$ can be combined as $\binom{2n}{n}^{-1}$, which is known to have the asymptotic equivalent $\sqrt{\pi n}/4^n$ (see [5]). Hence,

$$M_n(t) \sim \sqrt{n} \exp\left\{-\sqrt{a_n} t + \left(\tfrac{1}{2} e^{t/\sqrt{a_n}} - 1\right) \ln n\right\}.$$

By the Taylor expansion,

$$M_n(t) \sim \sqrt{n} \exp\left\{-\sqrt{a_n} t + \left[\left(\tfrac{1}{2} + \frac{t}{2\sqrt{a_n}} + \frac{t^2}{4a_n} + O\left(a_n^{-3/2}\right)\right) - 1\right] \ln n\right\} \sim e^{t^2/2},$$

the right side of the latter relation being the moment generating function of $\mathcal{N}(0, 1)$. □

4. The external path length

Martingales have been used in the context of path lengths of some classes of trees [8,12]. In this section we find a martingale associated with the external path length X_n, then use it to prove the existence of a limiting random variable almost surely and in L^2 for a properly normalized version of the external path length. The method admits a way of calculating the mean and variance of X_n. Observe that algorithmically a tree T_n of order n is obtained from a tree T_{n-1} of order $n-1$ by inserting the nth node at level D_n. The nth node may replace any of the $2n-3$ external nodes of T_{n-1} with probability $1/(2n-3)$. The new node gives the tree three new external nodes: one to its left, one to its right (which are children of the parent of the new internal node) and its own child, but one of the external nodes of T_{n-1} is lost in the process. The net gain in the external path length is therefore $2D_n + (D_n + 1) - D_n = 2D_n + 1$. Let $\mathcal{F}_n$ denote the sigma field generated by the tree T_n. When the shape of the tree T_{n-1} is available, the levels $x_1, \ldots, x_{2n-3}$ of the external nodes are completely determined. Thus D_n may assume any of the values $x_1, x_2, \ldots, x_{2n-3}$ with equal probability $1/(2n-3)$. We can now formulate a conditional expectation:

$$E[X_n \mid \mathcal{F}_{n-1}] = \frac{1}{2n-3} \sum_{j=1}^{2n-3} (X_{n-1} + 2x_j + 1) = X_{n-1} + 1 + \frac{2}{2n-3} \sum_{j=1}^{2n-3} x_j.$$

But the remaining sum is the external path length of T_{n-1}, i.e.,

$$E[X_n \mid \mathcal{F}_{n-1}] = \frac{2n-1}{2n-3} X_{n-1} + 1. \tag{4}$$

Taking expectations of the last relation we get the following recurrence on expected external path length:

$$E[X_n] = \frac{2n-1}{2n-3} E[X_{n-1}] + 1, \tag{5}$$

which can be easily solved under the initial condition $E[X_1] = 1$ to yield

$$E[X_n] = (2n-1)\alpha_n^{(1)}.$$

The average external path length is asymptotically equivalent to $n \ln n$, twice as much as the asymptotic average internal path length.

Theorem 4. *There exists a limiting random variable X such that*

$$\frac{X_n - n \ln n}{2n} \to X,$$

almost surely and in L^2.

Proof. We prove this theorem by showing that

$$Z_n = \frac{X_n - E[X_n]}{2n-1}$$

is a martingale over the sequence of fields $\{\mathcal{F}_n\}_{n=1}^{\infty}$ with uniformly bounded second moments. Absolute integrability of the sequence $\{Z_n\}_{n=1}^{\infty}$ is guaranteed by the existence of the mean of X_n for each n. Furthermore, by (4) and (5),

$$E[Z_n \mid \mathcal{F}_{n-1}] = \frac{X_{n-1}}{2n-3} - \frac{E[X_n] - 1}{2n-1} = Z_{n-1}.$$

We conclude that the sequence $\{Z_n\}_{n=1}^{\infty}$ is a zero-mean martingale.

To compute the second moment of Z_n we formulate a recurrence for it as follows. Replace X_n by $X_{n-1}+2D_n+1$ in the definition of Z_n and write

$$\begin{aligned} Z_n &= \frac{X_{n-1}+2D_n+1-E[X_{n-1}+2D_n+1]}{2n-1} \\ &= \frac{2n-3}{2n-1}Z_{n-1}+\frac{2}{2n-1}(D_n-E[D_n]). \end{aligned}$$

Squaring the latter relation, then taking expectations yields

$$\begin{aligned} E[Z_n^2] &= \left(\frac{2n-3}{2n-1}\right)^2 E[Z_{n-1}^2]+\frac{4}{(2n-1)^2}\operatorname{Var}[D_n] \\ &\quad+\frac{4(2n-3)}{(2n-1)^2}E[Z_{n-1}(D_n-E[D_n])]. \end{aligned} \tag{6}$$

In the last term we need only to find $E[Z_{n-1}D_n]$ since the component $E[Z_{n-1}E[D_n]]$ is zero. For the required term we compute

$$E[Z_{n-1}D_n]=E[E[Z_{n-1}D_n\,|\,\mathcal{F}_{n-1}]]=E[Z_{n-1}E[D_n\,|\,\mathcal{F}_{n-1}]].$$

But according to the algorithmic development,

$$E[D_n\,|\,\mathcal{F}_{n-1}]=\sum_{j=1}^{2n-3}\frac{x_j}{2n-3}=\frac{X_{n-1}}{2n-3}.$$

So,

$$E[Z_{n-1}D_n]=E[Z_{n-1}^2].$$

Plugging this relation into (6) we arrive at the recurrence

$$E[Z_n^2]=\frac{(2n-3)(2n+1)}{(2n-1)^2}E[Z_{n-1}^2]+\frac{4}{(2n-1)^2}\operatorname{Var}[D_n].$$

The substitution $Q_n=(2n-1)E[Z_n^2]/(2n+1)$ linearizes this recurrence into the simple recurrence

$$Q_n=Q_{n-1}+\frac{4}{(2n-1)(2n+1)}\operatorname{Var}[D_n].$$

By the relation for the variance of D_j, the solution to the last recurrence gives $E[Z_n^2]$ as

$$E[Z_n^2]=\frac{4(2n+1)}{2n-1}\sum_{j=2}^{n}\frac{\alpha_{j-1}^{(1)}-\alpha_{j-1}^{(2)}}{(2j-1)(2j+1)}.$$

The sum in $E[Z_n^2]$ clearly converges; the variance of Z_n is O(1), i.e., $E[Z_n^2]$ is bounded uniformly in n. Convergence almost surely and in L^2 follows from the martingale convergence theorem [6]. □

The standard deviation of the external path length of plane-oriented recursive trees is relatively small compared to the mean value, since the variance of X_n is $E[(2n-1)^2Z_n^2]$, i.e., the standard deviation is $O(n)$ while $E[X_n] \sim n \ln n$, as $n \to \infty$. (This is to be compared with [3] and [8] where the authors independently obtained an $O(n^2)$ bound on the variance of the internal path length in the usual recursive trees.) A standard argument based on Chebychev's inequality shows that $X_n/(n \ln n) \to 1$ in probability. The external path length is asymptotically twice as much as the internal path length with high probability.

Acknowledgements

The author is grateful for the hospitality of the Data Structuring Group at the University of Waterloo who sponsored his visit on sabbatical leave. Special thanks are due to Professor Ian Munro.

References

[1] C. Berge, *Graphs and Hypergraphs* (North-Holland, Amsterdam, 1973).

[2] L. Devroye, Applications of the theory of records in the study of random trees, *Acta Inform.* **26** (1988) 123–130.

[3] R. Dwyer, Las Vegas gift-wrapping is twice as fast, Private communications, 1990.

[4] J. Gastwirth, A probability model of a pyramid scheme, *Amer. Statist.* **31** (1977) 79–82.

[5] R. Graham, D. Knuth and O. Patashnik, *Concrete Mathematics: a Foundation for Computer Science* (Addison-Wesley, Reading, MA, 1989).

[6] P. Hall and C. Heyde, *Martingale Limit Theory and Applications* (Academic Press, New York, 1980).

[7] D. Knuth, *The Art of Computer Programming, Vol. 3: Sorting and Searching* (Addison-Wesley, Reading, MA, 1973).

[8] H. Mahmoud, Limiting distributions for path lengths in recursive trees, *Probab. Engrg. Inform. Sci.* **5** (1991) 53–59.

[9] A. Meir and J. Moon, On the altitude of nodes in random trees, *Canad. J. Math.* **XXX** (5) (1978) 997–1015.

[10] J. Moon, The distance between nodes in recursive trees, in: London Math. Soc. Lecture Note Ser. **13** (Cambridge Univ. Press, London, 1974) 125–132.

[11] D. Najock and C. Heyde, On the number of terminal vertices in certain random trees with an application to stemma construction in philology, *J. Appl. Probab.* **19** (1982) 675–680.

[12] M. Régnier, A limiting distribution for quicksort, *Theoret. Inform. Appl.* **23** (1989) 335–343.

[13] J. Szymański, On the complexity of algorithms on recursive trees, *Theoret. Comput. Sci.* **74** (3) (1990) 355–361.