

APPLIED TREE ENUMERATIONS

Nachum Dershowitz*
Department of Computer Science
University of Illinois
Urbana, Illinois 61801
U.S.A.

Shmuel Zaks
Department of Computer Science
Technion
Haifa 32000
Israel

I. INTRODUCTION

In this paper we consider the class T_n of ordered trees with n edges and give combinatorial proofs to several enumeration formulae concerning T_n . In particular, closed-form expressions are given for (1) the number of trees in T_n with n_0 leaves, n_1 unary nodes, ..., n_d nodes with d children, and no restrictions on nodes with more than d children, and for (2) the number of nodes in T_n on level l with d children. Several statistical results are derived from these.

The combinatorial tools we use to prove our results include one-to-one correspondences between ordered trees and other combinatorial objects, the Cycle Lemma, and lattice-path techniques. Many of these results could, alternatively, have been obtained using generating functions and the Lagrange inversion formula.

We demonstrate the use of these enumerations in analyzing the following applications: (1) a sorting problem, (2) the average height of a stack during tree-traversal, (3) algorithms for threaded binary trees, and (4) a pattern-matching problem.

II. CORRESPONDENCES

We consider ordered trees (see Knuth [1968] for definitions). Each node has a degree (the number of its children). A node of degree 0 is termed a leaf; otherwise it is called an internal node. The level of a node is its distance from the root (the root is on level 0). The number of trees in the set T_n of ordered trees with n edges is the well-known Catalan number

*Research supported in part by the National Science Foundation under Grant MCS 79-04897.

$$|T_n| = C_n = \frac{1}{n+1} \binom{2n}{n}$$

(see, for example, Gardner [1976]).

There are numerous one-to-one correspondences between elements of these sets of ordered trees and other combinatorial objects (see, for example, Kuchinski [1977]). Among them, the correspondences between the following sets help in our enumerations:

T_n : the set of ordered trees with n edges.

B_n : the set of binary trees with n internal nodes, each having exactly two children.

P_n : the set of sequences of n open parentheses and n close parentheses, where each open parenthesis has a matching close parenthesis.

I_n : the set of sequences $a_0 a_1 \dots a_n$ of $n+1$ nonnegative integers summing to n , such that $\sum_{j=0}^i a_j > i$ for $i=0,1,\dots,n-1$.

L_n : the set of shortest lattice-paths from $(0,0)$ to (n,n) that do not go below the diagonal $y=x$ (all steps are either up or to the right).

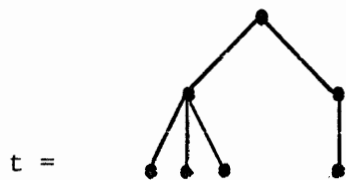
The correspondences between these five sets are illustrated in Figure 1. In general, from a tree t in T_n one gets a sequence $p(t)$ in P_n by traversing t in preorder, writing "(" for each edge passed on the way down and ")" for each edge passed on the way up. (See Figure 1.2.) From $p(t)$ one gets a lattice path $\lambda(t)$ in L_n by starting at $(0,0)$ and going up one coordinate for each open parenthesis and going right one coordinate for each close parenthesis. (See Figure 1.3.) From t one gets a sequence $i(t)$ in I_n by reading the degrees of all the nodes of t in preorder. (See Figure 1.4.) From $\lambda(t)$ a binary tree $b(t)$ in B_n is built in preorder, each step up on the path corresponding to an internal node and each step to the right corresponding to a leaf (a final leaf is also added). (See Figure 1.5.)

A sequence of open and close parentheses is called legal if in each prefix the number of open parentheses is greater than the number of close parentheses. We use the following lemma:

Cycle Lemma (Dvoretzky and Motzkin [1947]): For any sequence $p_1 p_2 \dots p_{m+n}$ of m open and n close parentheses, $m > n$, there are exactly $m-n$ cyclic permutations

$$p_j p_{j+1} \dots p_{m+n} p_1 \dots p_{j-1}$$

that are legal.



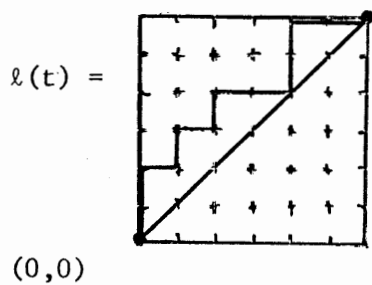
1.1 An ordered tree $t \in T_6$

$p(t) = (())()()()$

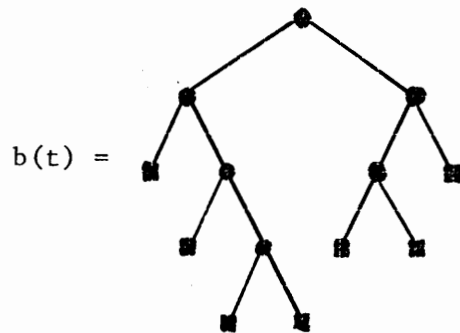
$i(t) = 2300010$

1.2 A sequence $p(t) \in P_6$

1.4 A sequence $i(t) \in I_6$



1.3 A lattice path $\ell(t) \in L_6$



1.5 A binary tree $b(t) \in B_6$

Figure 1. Correspondences between T_n , B_n , P_n , I_n , and L_n .

From this lemma it follows that there is a one-to-one correspondence between ordered trees in T_n and cycles with $n+1$ open parentheses and n close parentheses, there being only one legal permutation of $p(t)$ with an extra open parenthesis prepended. Furthermore, each such cycle of parentheses corresponds to a cycle of $n+1$ nonnegative integers summing to n (representing the number of close parentheses between pairs of open parentheses). These correspondences are the basis for our use of the Cycle Lemma in the enumerations.

III. ENUMERATIONS

In this section we present the main enumeration results and discuss some of their consequences. All trees are in T_n .

Theorem 1: The number $L_n(n_0, n_1, \dots, n_d)$ of trees in T_n with n_j nodes of degree j , $0 < j < d$, and no restrictions on nodes of degree greater than d is

$$\frac{1}{n+1} \binom{n-e-d(n-m+1)-1}{n-m} \binom{n+1}{n_0, n_1, \dots, n_d, n-m+1}$$

where $m = \sum n_j$ (the total number of restricted nodes) and $e = \sum j n_j$ (the total number of edges accounted for).

This generalizes a result of Narayana [1959] for $d=0$ (proved in Dershowitz and Zaks [1980] using the Cycle Lemma) and the multinomial formula for the case $d=n$ (Erdelyi and Etherington [1940], proved in Raney [1960] using the Cycle Lemma).

Proof: By the Cycle Lemma we have to count the cycles with $n+1$ nonnegative integers summing to n , of which n_i are i , $0 < i < d$. To count the number of cycles with these restrictions, note that there are

$$\frac{1}{n+1} \binom{n+1}{n_0, \dots, n_d, n+1-m}$$

ways of placing the degrees of the restricted nodes on a cycle with $n+1$ positions. The remaining $n+1-m$ unrestricted nodes must have degrees ranging from $d+1$ to n and summing to $n-e$. The number of ways to place these degrees is the same as the number of ways of decomposing the integer $n-e-(d+1)(n+1-m)$ into $n+1-m$ integers ranging from 0 to $n-(d+1)$, which is

$$\binom{n-e-d(n+1-m)-1}{n-m}, \quad \square$$

From this theorem, we can derive the following

Consequences:

1.1) The number $L_n(k)$ of trees with k leaves is

$$L_n(k) = \frac{1}{n+1} \binom{n-1}{n-k} \binom{n+1}{k} = \frac{1}{k} \binom{n-1}{k-1} \binom{n}{k-1}.$$

This is Narayana's [1959] result (see also Mohanty [1979]).

1.2) The number $L_n(k)$ of trees with k leaves is equal to the number $L_n(n+1-k)$ of trees with $n+1-k$ leaves.

1.3) The expected number of leaves (or internal nodes) is $\frac{n+1}{2}$. (Here, and in the sequel, all trees in T_n are assumed equiprobable). This result has also been given by Dasarathy and Yang [1980].

1.4) The expected degree of an internal node is

$$\frac{n C_n}{\frac{n+1}{2} C_n} = \frac{2n}{n+1} \approx 2.$$

Theorem 2: The total number $N_n(\ell, d)$ of nodes in T_n of degree d on level ℓ is

$$N_n(\ell, d) = \binom{2n-d-1}{n+\ell-1} - \binom{2n-d-1}{n+\ell} = \frac{2\ell+d}{2n-d} \binom{2n-d}{n+\ell}.$$

This result was first proved in Dershowitz and Zaks [1980]. We give here a lattice-path proof.

Proof: The number of lattice paths from $(0,0)$ to $(n-d-\ell, n+\ell-1)$ that do not go below the diagonal $y=x$ is

$$\frac{2\ell+d}{2n-d} \binom{2n-d}{n+\ell}$$

(see, for example, Mohanty [1979]). We give a correspondence between each such path and each node in T_n of degree d on level ℓ . (The correspondence applies to the case $\ell > 1$; for $\ell=0$ a simpler correspondence works.)

Consider a tree t in T_n and the corresponding lattice path $\ell(t)$ from $(0,0)$ to (n,n) . A node x of degree d on level ℓ of t corresponds to a path segment

$$\begin{aligned} & (i_0, i_0 + \ell - 1) \rightarrow (i_0, i_0 + \ell) \rightarrow \dots \rightarrow (i_1 - 1, i_1 + \ell) \rightarrow (i_1, i_1 + \ell) \rightarrow \dots \\ & \rightarrow \dots \rightarrow (i_j - 1, i_j + \ell) \rightarrow (i_j, i_j + \ell) \rightarrow \dots \rightarrow (i_d, i_d + \ell) \rightarrow (i_d + 1, i_d + \ell) \end{aligned} \quad (a)$$

that does not go below the diagonal $y=x+\ell$ (except at the two ends). This segment is preceded by another segment

$$(0,0) \rightarrow \dots \rightarrow (i_0, i_0 + \ell - 1) \quad (b)$$

and is followed by a segment

$$(i_d + 1, i_d + \ell) \rightarrow \dots \rightarrow (n, n). \quad (c)$$

To get the desired lattice path from $(0,0)$ to $(n-d-\ell, n-\ell-1)$, we do the following (see Figure 2):

- (a) The lattice path steps $(i_j - 1, i_j + \ell) \rightarrow (i_j, i_j + \ell)$, $1 \leq j < d$, as well as $(i_d, i_d + \ell) \rightarrow (i_d + 1, i_d + \ell)$, are removed from segment (a), yielding a path segment from $(i_0, i_0 + \ell - 1)$ to $(i_d - d, i_d + \ell)$.
- (b) Segment (b) is left intact.
- (c) Segment (c) is inverted by reversing both the order and the direction of its steps, i.e. every step right becomes a step down and every step up becomes a step left, yielding a path segment that ends at $(n-d-\ell, n+\ell-1)$ and begins at $(i_d - d, i_d + \ell)$.

Since this correspondence between nodes and paths is one-to-one, the desired result is proved. (To get the corresponding node, given a path, note that ℓ and d are determined by the endpoint $(n-d-\ell, n+\ell-1)$ and that for all j , $0 \leq j < d$, the path segment from $(i_j, i_j + \ell)$ to the endpoint does not return below the diagonal $y=x+\ell+j$. Both the deleted steps and the inverted path segment can be easily restored.) \square

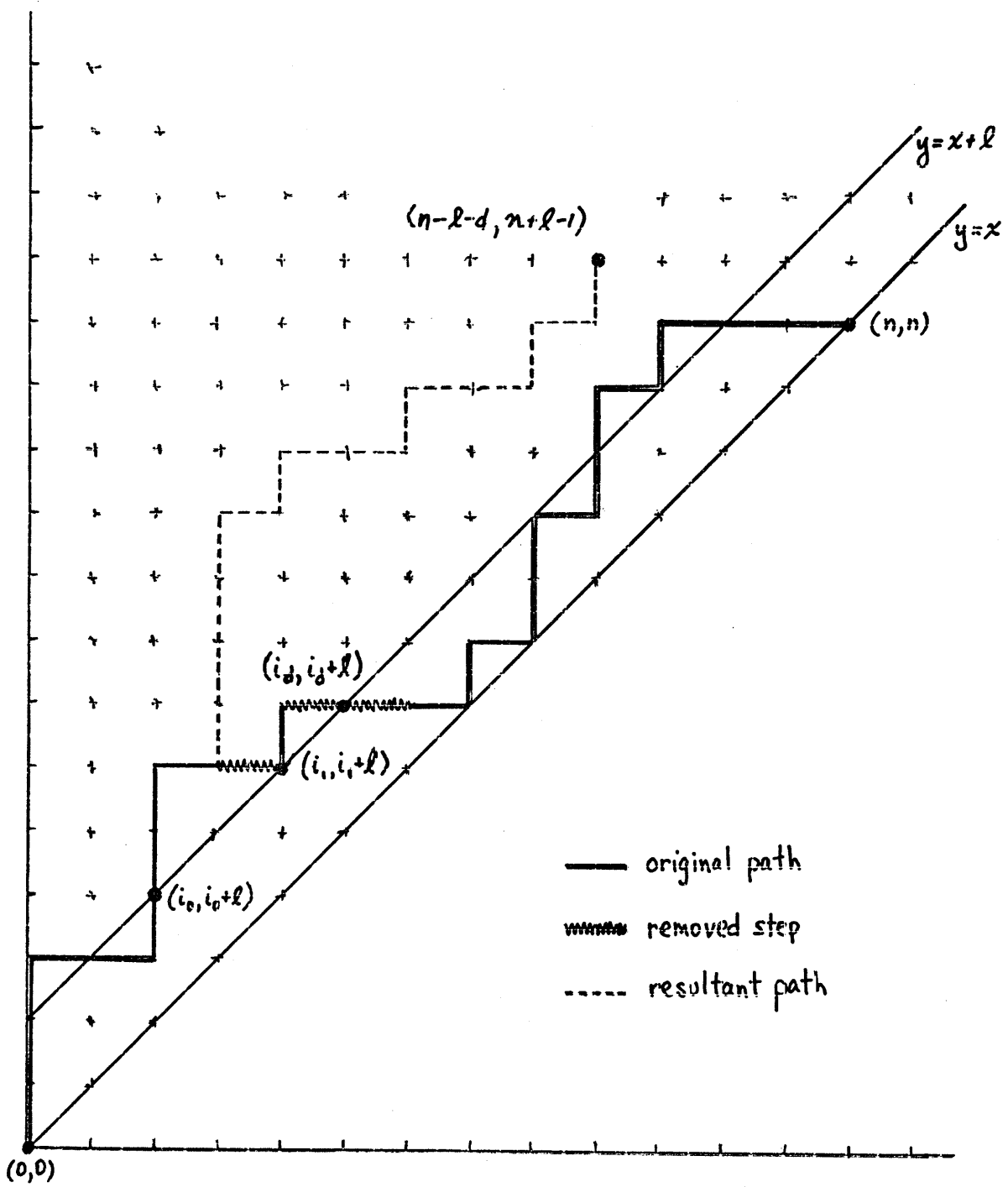


Figure 2. Proof of Theorem 2.

From this theorem, we can derive the following

Consequences:

2.1) The total number of nodes of degrees i through j on level ℓ is

$$\sum_{d=i}^j N_n(\ell, d) = \sum_{d=i}^j \frac{2\ell+d}{2^{n-d}} \binom{2n-d}{n+\ell} = \frac{2\ell+i+1}{2^{n-i+1}} \binom{2n-i+1}{n+\ell+1} - \frac{2\ell+j+2}{2^{n-j}} \binom{2n-j}{n+\ell+1}.$$

In particular, the total number of nodes on level ℓ is

$$\frac{2\ell+1}{2^{n+1}} \binom{2n+1}{n-\ell}.$$

2.2) The expected number of leaves on level ℓ of a tree in T_n is

$$\frac{N_n(\ell, 0)}{C_n} = \frac{\frac{\ell}{n} \binom{2n}{n+\ell}}{\frac{1}{n+1} \binom{2n}{n}} = \frac{\ell \binom{2n}{n-\ell}}{\binom{2n}{n-1}} < \ell.$$

For small ℓ ,

$$\frac{N_n(\ell, 0)}{C_n} \approx \ell.$$

2.3) The expected level of a leaf is

$$\frac{\sum \ell N_n(\ell, 0)}{\frac{n+1}{2} C_n} = \frac{2^{2n-1}}{\binom{2n}{n}} \approx \frac{\sqrt{\pi n}}{2}.$$

(This sum and subsequent ones may be evaluated using the identities in, for example, Riordan [1968].) Thus, the expected external path length of a tree (as defined in Knuth [1968]) is

$$\frac{2^{2n-1}}{\binom{2n}{n}} \frac{n+1}{2} \approx \frac{\sqrt{\pi n}(n+1)}{4}.$$

2.5) The expected number of internal nodes on level ℓ is

$$\frac{\sum_{d=1}^n N_n(\ell, d)}{C_n} = \frac{(\ell+1) \binom{2n}{n-\ell-1}}{\binom{2n}{n-1}} < \ell+1.$$

For small ℓ ,

$$\frac{\sum_{d=1}^n N_n(\ell, d)}{C_n} \approx \ell + 1.$$

2.5) The expected level of an internal node is

$$\frac{\sum \ell \frac{\ell+1}{n} \binom{2n}{n-\ell-1}}{\frac{n+1}{2} C_n} = \frac{2^{2n-1}}{\binom{2n}{n}} - 1 \approx \frac{\sqrt{\pi n}}{2} - 1.$$

Thus, the expected internal path length of a tree is

$$\left[\frac{2^{2n-1}}{\binom{2n}{n}} - 1 \right] \frac{n+1}{2} \approx \frac{\sqrt{\pi n}(n+1)}{4} - \frac{n+1}{2}.$$

2.6) The expected level of a node is

$$\frac{\sum \ell \frac{2\ell+1}{2n+1} \binom{2n+1}{n-\ell}}{\binom{2n}{n}} = \frac{2^{2n-1}}{\binom{2n}{n}} - \frac{1}{2} \approx \frac{\sqrt{\pi n}}{2} - \frac{1}{2}.$$

This result has been given by Volosin [1974], Meir and Moon [1978], and Dasarathy and Yang [1980]. Higher moments of the node level can be calculated in the same way.

2.7) The total number of nodes of degree d on levels i through j is

$$\sum_{\ell=i}^j N_n(\ell, d) = \sum_{\ell=i}^j \left[\binom{2n-d-1}{n+\ell-1} - \binom{2n-d-1}{n+\ell} \right] = \binom{2n-d-1}{n+i-1} - \binom{2n-d-1}{n+j}.$$

In particular, the total number $D_n(d)$ of nodes of degree d is

$$D_n(d) = \binom{2n-d-1}{n-1}.$$

It follows that the total number of nodes on levels i through j is

$$\binom{2n}{n-i} - \binom{2n}{n-j-1}$$

and the total number of nodes of degrees i through j is

$$\binom{2n-1}{n} - \binom{2n-j-1}{n}.$$

2.8) The expected number of nodes of degree d in a tree in T_n is

$$\frac{D_n(d)}{C_n} = \frac{\binom{2n-d-1}{n-1}}{\frac{1}{n+1} \binom{2n}{n}} < \frac{n+1}{2^d}.$$

For small d ,

$$\frac{D_n(d)}{C_n} \approx \frac{n+1}{2^{d+1}}.$$

2.9) The number $R_n(r)$ of trees with root degree r is

$$R_n(r) = N_n(0, r) = \frac{r}{n} \binom{2n-r-1}{n-1}.$$

The expected root degree is

$$\frac{\sum r R_n(r)}{C_n} = \frac{\sum \frac{r^2}{n} \binom{2n-r-1}{n-1}}{\frac{1}{n+1} \binom{2n}{n}} = \frac{\frac{3}{n+2} \binom{2n}{n+1}}{\frac{1}{n} \binom{2n}{n-1}} = \frac{3n}{n+2} \approx 3.$$

These results are also given in Ruskey and Hu [1977]. In a similar manner higher moments can be calculated. For example, the variance of the root degree is

$$\frac{\sum r^2 R_n(r)}{C_n} - \left[\frac{3n}{n+2} \right]^2 = \frac{2 \binom{2n}{n+3} + 3 \binom{2n+1}{n+3} + \binom{2n+2}{n+3}}{\frac{n}{n+1} \binom{2n}{n}} - \left[\frac{3n}{n+2} \right]^2 \approx \frac{1}{4}.$$

IV. APPLICATIONS

In this section we illustrate how the formulae of the previous section can help in the analysis of various algorithms.

1. A sorting problem: Given an ordered tree with n edges, with a number stored at each node, we want to order the edges so that the children of each node will be in increasing order. (The need to sort trees in this manner arose in an algorithm for computing the recursive path ordering on terms, as defined in Dershowitz [1981]). A node with d children needs no more than $d^2/3$ comparisons to sort them, hence the total number of comparisons is bounded by $\sum i^2 n_i / 3$ for a tree with n_i nodes of degree i . To find the average \bar{s} of this function over all trees in T_n , we make use of the closed form for the number $D_n(d)$ of nodes of degree d in T_n (Consequence 2.7):

$$\bar{s} = \frac{\frac{1}{3} \sum d^2 D_n(d)}{C_n} = \frac{\frac{1}{3} \sum d^2 \binom{2n-d-1}{n-1}}{\frac{1}{n+1} \binom{2n}{n}} = \frac{n}{n+2}.$$

Thus, on the average, less than one comparison per node is required. Considering that sorting d children actually requires only on the order of $d \cdot \log_2 d$ comparisons, a slightly better result should be possible.

2. Average height of a stack: The height of an ordered tree is the worst stack size that is formed while traversing the tree in, say, preorder. deBruijn, Knuth, and Rice [1972] and others have shown that the expected value of this height, over all trees in T_n , is about $\sqrt{\pi n}$. If, instead of studying the worst stack size, we look at the average while traversing the tree, we need the expected level of a node, which we have already seen (Consequence 2.6) is

$$\frac{2^{2n-1}}{\binom{2n}{n}} - \frac{1}{2} \approx \frac{\sqrt{\pi n}}{2} - \frac{1}{2}.$$

If the average is weighted by the number of times during the traversal that the stack is of that height, i.e. by the degree of the node plus one, then we get

$$\frac{\sum_{\lambda, d} (d+1) \lambda N_n(\lambda, d)}{(2n+1) C_n} = \frac{2^{2n} - \binom{2n+1}{n}}{\binom{2n+1}{n}} \approx \frac{\sqrt{\pi n}}{2} - 1.$$

3. Threaded binary trees: In a recent paper by Brinck and Foo [1981], algorithms on threaded binary trees are investigated. The analyses are based on certain enumeration lemmas, proved using recurrence relations. Our enumeration techniques enable one to prove those lemmas using a direct combinatorial approach, which also sheds light on the structure of these trees.

For example, the expected number of left leaves (i.e. backward threads) in a binary tree in B_n , and the expected number of right leaves (i.e. forward threads), are equal to the expected number of leaves in an ordered tree (by the correspondence between T_n and B_n), which we have seen (Consequence 1.3) is $\frac{n+1}{2}$. Also, the expected distance from the root of a binary tree to the leftmost (or rightmost) leaf is equal to the expected root degree of an ordered tree (letting each digit in a sequence in I_n correspond to the number of internal leaves that precede a leaf in the preorder construction of a tree in B_n), which we have seen (Consequence 2.9) is $\frac{3n}{n+2}$.

4. Pattern matching: Flajolet and Steyaert [1980] have recently investigated tree-matching algorithms. Using our techniques one can deal with similar problems. For example, a pattern is an ordered tree some of whose leaves are designated open. A pattern p is said to occur in a tree x if x contains a subtree of the same form as p with arbitrary trees substituted for the open nodes. In order to know how many times a certain pattern p -- with e edges and d open leaves -- occurs as a subtree in some tree in T_n , we simply replace p by a node of degree d , and ask how many nodes of degree d are in T_{n-e} ; this number is given (Consequence 2.7) by

$$D_{n-e+d}(d) = \binom{2n-2e+d-1}{n-e}.$$

Similarly, the number of occurrences of p on level ℓ is given by

$$N_{n-e+d}(\ell, d) = \frac{2\ell+d}{2n-2e+d} \binom{2n-2e+d}{n+\ell}.$$

More generally, the number of occurrences of the combination of m distinct nonoverlapping patterns p_1, p_2, \dots, p_m within T_n is

$$(n-e+d)^m D_{n-e+d}(d),$$

where e is the total number of edges in the patterns, d is the total number of open leaves, and $x^m = x(x-1)\dots(x-m+1)$. To prove this, note that the number of occurrences of the patterns p_1 in trees with n nodes is equal to the number of

trees with m labelled nodes having degrees d_1, \dots, d_m (where d_i is the number of open leaves in p_i) and $n-e+d$ edges. Each such tree corresponds (by the Cycle Lemma) to a cycle of $n-e+d+1$ degrees -- m of them specified and labelled -- summing to $n-e+d$. There are

$$(n-e+d)^m$$

ways of placing the labelled nodes on the cycle and there are

$$D_{n-e+d}(d)$$

ways of specifying the degrees of the remaining nodes.

In a similar manner it can be shown that

$$(n-e+d)^m \binom{2n-2e+d}{n-e}$$

is the number of occurrences of m patterns in the set B_n of binary trees, in which case e denotes the total number of internal nodes in the patterns. For example, the number of left nodes in B_n that have forward threads (i.e. the number of internal nodes that are left children and have right leaves) is

$$(n-2+2)^1 \binom{2n-4+2}{n-2} = \binom{2n-2}{n-2}$$

(cf. Brinck and Foo [1981]).

REFERENCES

- [1] K. Brinck and N. Y. Foo [1981], Analysis of algorithms on threaded trees, to appear in Comp. J.
- [2] N. deBruijn, D. E. Knuth, and O. Rice [1972], The average height of planted plane trees, in Graph Theory and Computing, (R. C. Read, ed.), Academic Press, New York, 15-22.
- [3] B. Dasarathy and C. Yang [1980], A transformation on ordered trees, Comp. J. 23 (2), 161-164.
- [4] N. Dershowitz [1981], Orderings for term-rewriting systems, to appear in J. Theoretical Computer Science.
- [5] N. Dershowitz and S. Zaks [1980], Enumerations of ordered trees, Discrete Math 31 (1), 9-28.
- [6] A. Dvoretzky and Th. Motzkin [1947], A problem of arrangements, Duke Math. J. 14, 305-313.
- [7] A. Erdelyi and I. M. H. Etherington [1940], Some problems of non-associative combinations (2), Edin. Math. Notes 32, 7-12.
- [8] Ph. Flajolet and J. M. Steyaert [1980], On the analysis of tree-matching algorithms, Proc. 7th Intl. Conf. Automata, Languages and Programming, Amsterdam, 208-220.
- [9] M. Gardner [1976], Mathematical games: Catalan numbers, Scientific American 234 (6), 120-125.
- [10] D. E. Knuth [1968], The Art of Computer Programming, Vol. 1: Fundamental algorithms, Addison-Wesley, Reading, MA.
- [11] M. J. Kuchinski [1977], Catalan structures and correspondences, M.S. Thesis, Dept. of Mathematics, West Virginia Univ., Morgantown, WV.
- [12] A. Meir and J. Moon [1978], On the altitude of nodes in random trees, Can. J. Math. 30 (5), 997-1015.
- [13] S. G. Mohanty [1979], Lattice Path Counting and Applications, Academic Press, New York.
- [14] T. V. Narayana [1959], A partial order and its applications to probability, Sankhya 21, 91-98.
- [15] G. H. Raney [1960], Functional composition patterns and power series reversion, Trans. AMS 94, 441-451.
- [16] J. Riordan [1968], Combinatorial Identities, Wiley, New York.
- [17] R. Ruskey and T. C. Hu [1977], Generating binary trees lexicographically, SIAM J. Computing 6 (4), 745-758.
- [18] Ju. M. Volosin [1974], Enumeration of the terms of object domains according to the depth of embedding, Sov. Math. Dokl. 15, 1777-1782.