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 $a_0a_1\cdots a_n$ of n+1 nonnegative integers summing to n, such that $\sum_{j=0}^{i} a_j > i$ for i=0,1,...,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).

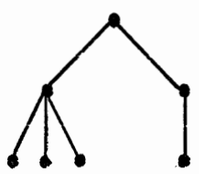
The correspondences between these five sets are illustrated in Figure 1. In general, from a tree t in T_n one gets a sequence p(t) in P_n by traversing t in preorder, writing "(" for each edge passed on the way down and ")" for each edge passed on the way up. (See Figure 1.2.) From p(t) one gets a lattice path $\ell(t)$ 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 $\ell(t)$ 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 <u>legal</u> if in each prefix the number of open parentheses is greater than the number of close parentheses. We use the following lemma:

<u>Cycle Lemma</u> (Dvoretzky and Motzkin [1947]): For any sequence $p_1p_2\cdots p_{m+n}$ of m open and n close parentheses, m>n, there are exactly m-n cyclic permutations

$$p_jp_{j+1}\cdots p_{m+n}p_1\cdots p_{j-1}$$

that are legal.

t = 

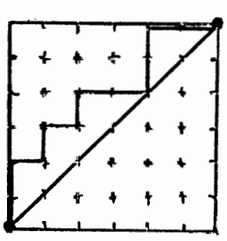
1.1 An ordered tree $t \in T_6$

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

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

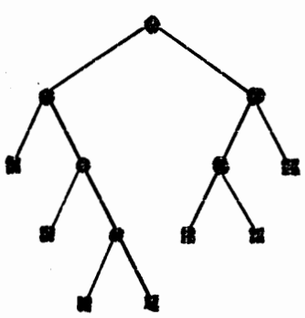
$i(t) = 2300010$

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

$\ell(t)$ = 

(0,0)

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

b(t) = 

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 $L_n(n_0,n_1,\cdots,n_d)$ of trees in T_n with n_j nodes of degree j, $0\leq j\leq d$, and no restrictions on nodes of degree greater than d is

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

where $m=\Sigma n_j$ (the total number of restricted nodes) and $e=\Sigma j n_j$ (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, $0\leq i\leq d$. To count the number of cycles with these restrictions, note that there are

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

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

$$\binom{n-e-d(n+1-m)-1}{n-m}. \qquad \square$$

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

$$\begin{aligned}&(i_0,i_0+\ell-1)\dashrightarrow(i_0,i_0+\ell)\rightarrow\cdots\rightarrow(i_1-1,i_1+\ell)\rightarrow(i_1,i_1+\ell)\rightarrow\cdots\\&\rightarrow\cdots\rightarrow(i_j-1,i_j+\ell)\rightarrow(i_j,i_j+\ell)\rightarrow\cdots\rightarrow(i_d,i_d+\ell)\rightarrow(i_d+1,i_d+\ell)\end{aligned}\qquad\text{(a)}$$

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

$$(0,0)\rightarrow\cdots\rightarrow(i_0,i_0+\ell-1)\qquad\text{(b)}$$

and is followed by a segment

$$(i_d+1,i_d+\ell)\rightarrow\cdots\rightarrow(n,n).\qquad\text{(c)}$$

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 $(i_j-1,i_j+\ell)\rightarrow(i_j,i_j+\ell)$, $1\leq j\leq d$, 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, $0\leq j\leq d$, 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.) □

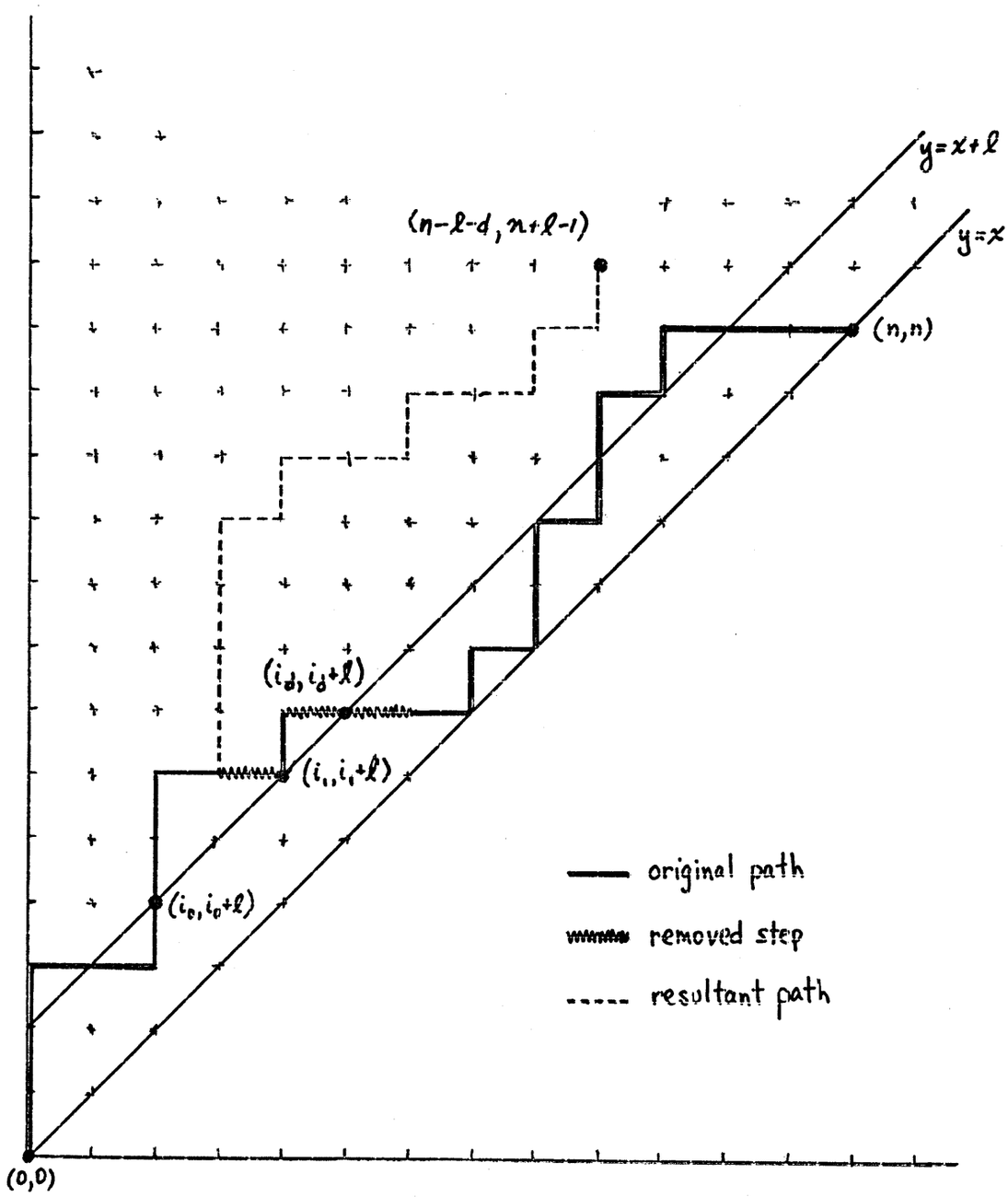


Figure 2. Proof of Theorem 2.

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

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

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

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

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

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

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

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

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

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

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

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

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

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