MIT/LCS/TM-271

# PATTERNS IN TREES

NACHUM DERSHOWITZ

SHMUEL ZAKS

January 1985

# PATTERNS IN TREES[1]

Nachum Dershowitz[2] *and* Shmuel Zaks[3]

June 1984

[1] A preliminary version of this paper was presented at the 9th Colloquium on Trees in Algebra and Programming, Bordeaux, France, March 1984.

[2] Department of Mathematics and Computer Science, Bar-Ilan University, Ramat-Gan, Israel; on leave from the Department of Computer Science, University of Illinois, Urbana, IL 61801, supported in part by NSF grant MCS-79-04897.

[3] Laboratory for Computer Science, M.I.T., Cambridge, MA 02139, supported by NSF grant MCS-83-02391; on leave from the Department of Computer Science, Technion, Haifa, Israel.

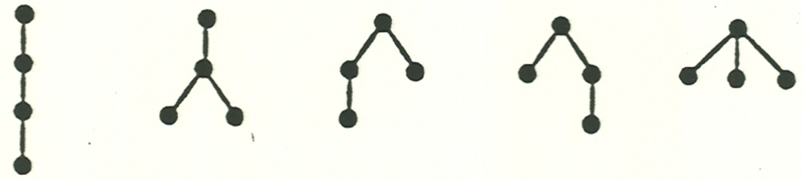
# ABSTRACT

A very general enumeration formula for occurrences of a pattern, or set of patterns, in the class of ordered trees with a given number of edges is presented, and its wide usefulness is demonstrated.

## 1 INTRODUCTION

An ordered or (plane-planted) tree is a tree in which the order of the outgoing edges of each node is significant. We denote by $T_n$ the class of trees of n edges. For example, there are 5 trees in $T_3$:

$T_3$:



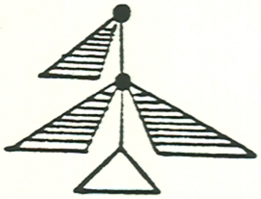
The number of trees in $T_n$ is the well-known Catalan number $\frac{1}{n+1}\binom{2n}{n}$.

Our main result is a closed-form expression for the number of occurrences of multisets of patterns in classes of ordered trees. Its proof is based on an extension of the Cycle Lemma. This enumeration formula is widely applicable, and many known results fit within its framework.

We first discuss the notion of patterns and their occurrences in trees (Section 2), and then present the Cycle Lemma and its extension to trees (Section 3). This is followed by the enumeration formula (Section 4) and its applications (Section 5).

## 2 PATTERNS AND THEIR OCCURRENCES

A pattern is like a tree except that some of its nodes are designated open slots and some of its edges are designated closed slots and do not end in nodes. For example, the pattern



occurs wherever a node has a grandchild through its youngest child. Slots in a pattern represent arbitrary trees and are depicted as triangles; closed slots as shaded triangles hanging off a node; open slots as unshaded triangles hanging off an edge. An open slot 'matches' any subtree, including the empty tree (consisting of a single node), while a closed slot acts like a variable number of open slots and 'matches' any number (including zero) of edges along with the subtrees (= open slots) below them. (Thus, there is no reason in having adjacent closed slots in a pattern.) For example, the above pattern occurs five times in the class $T_3$ (see figure in Section 1): twice in the first tree (once at the root and once at its child), twice in the second (once for each of the two grandchildren), and once in the fourth.

Precise definitions of patterns and their occurrences follow : A pattern is said to occur at a subtree of an ordered tree according to the following rules:

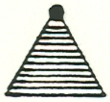
1) The leaf pattern

●

occurs at the empty subtree

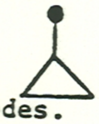
●

2) The closed slot pattern



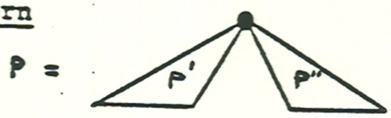
occurs at all subtrees.

3) The open slot pattern

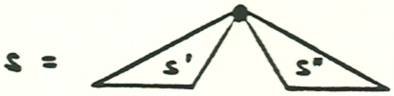


occurs at all unary nodes.

4) If $p'$ and $p''$ are patterns occurring at subtrees $s'$ and $s''$, respectively, then the adjoined pattern

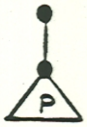


occurs at the subtree

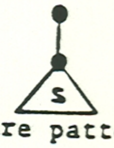


If there is more than one way of partitioning s into s' and s'' so that p' occurs at s' and p'' at s'', then they give rise to multiple occurrences of p at s. Thus, if s can be divided in k ways into $s'_1$ and $s''_1$, $s'_2$ and $s''_2$, ..., $s'_k$ and $s''_k$, and p' occurs $m_i$ times at $s'_i$ and p'' $n_i$ times at $s''_i$, then p occurs $m_1n_1 + m_2n_2 + \ldots + m_k n_k$ times at s.

5) If the pattern p occurs n times at a subtree s, then the extended pattern

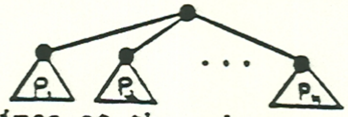


occurs n times at the tree

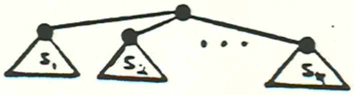


.

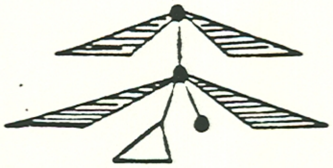
6) It follows that if $p_1, p_2, \ldots, p_k$ are patterns occurring $n_1, n_2, \ldots, n_k$ times at subtrees $s_1$, $s_2$, ..., $s_k$, respectively, then the composite pattern



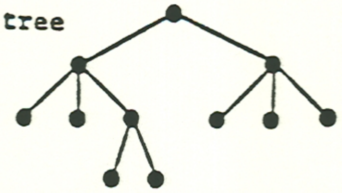
occurs $n_1 \cdot n_2 \cdot \ldots \cdot n_k$ times at the subtree



For example, the pattern



occurs at any node that has a childless grandchild with an older sibling. It occurs four times in the tree

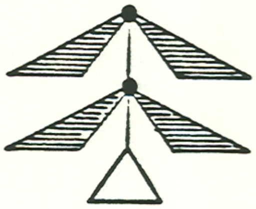


,

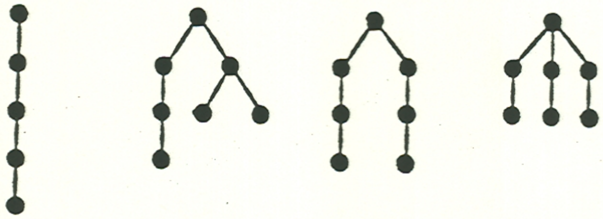
three times at the root and once at its oldest child.

Let $p_1, p_2, \ldots, p_k$ be distinct patterns. The multiset containing $n_1$ copies of $p_1$, $n_2$ copies of $p_2, \ldots, n_k$ copies of $p_k$, is denoted by $\{n_1 * p_1, n_2 * p_2, \ldots, n_k * p_k\}$.

A multiset of patterns is said to *occur* in a tree if each of its individual patterns occurs, and the (non-slot) nodes of the occurrences are disjoint. For example, the pattern



matches any grandparent-grandchild relation between two nodes. The multiset of *two* such patterns occurs six times among the four trees



,

once in the first tree, twice in the second, and three times in the third. It does not occur in the fourth tree at all, since any two such relations share the grandparent node.

# 3 THE CYCLE LEMMA FOR TREES

A sequence $p_1, p_2, \ldots, p_\ell$ of boxes and circles is called k-dominating if for every position $i$, $1 < i < \ell$, the number of boxes in $p_1, p_2 \cdots p_i$ is more than $k$ times the number of circles ($k$ is a positive integer). For example, the sequence □ □ □ □ □ ○ ○ □ □ □ ○ is 2-dominating; the sequence □ □ □ □ ○ ○ □ □ ○ is 1-dominating (or just dominating) but not 2-dominating; the sequences

○ □ □ □ □ ○ ○ □ □ and □ □ ○ ○ □ □ ○ □ □ are not even 1-dominating.

The following lemma has been rediscovered many times. Though it is not difficult to prove, it is a powerful tool in enumeration arguments. Cycle Lemma (Dvoretzky and Motzkin [1947]): For any sequence $p_1 p_2 \cdots p_{m+n}$ of $m$ boxes and $n$ circles, $m > kn$, there exist exactly $m-kn$ cyclic permutations $p_j p_{j+1} \cdots p_{m+n}\, p_1 \cdots p_{j-1}$, $1 < j < m+n$, that are k-dominating.

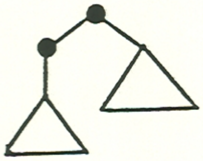
For example, of the nine cyclic permutations of the sequence ○ □ □ □ □ ○ ○ □ □ of six boxes and three circles, only three are dominating: □ □ ○ □ □ □ □ ○ ○ , □ □ □ □ ○ ○ □ □ ○ and □ □ □ ○ ○ □ □ ○ □ . None are 2-dominating. As a special case of this lemma, if $m = n+1$, then there is a unique dominating permutation. See, for example, Raney [1960] or Dershowitz and Zaks [1983].

As a corollary we get the Cycle Lemma for Trees, giving the number of cyclic permutations of patterns of trees that can be grafted together. It applies to any "simply generated" family of trees (see Flajolet and Steyaert [1980]), like ordered trees or binary trees.

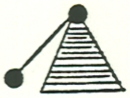
We first define the grafting operation. If $p$ and $p'$ are two patterns, then the graft of $p$ and $p'$ is the pattern $p \circ p'$ obtained by replacing the rightmost (open or closed) slot of $p'$ with $p$, and is the sequence $(p, p')$ when $p'$ contains no slots.

For example, if $p$ and $p'$ are as follows:

$p =$

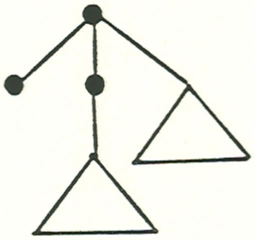


$p' =$

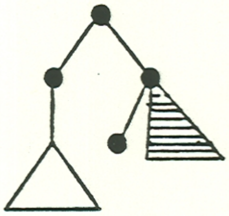


then

$p \circ p' =$



$p' \circ p =$

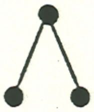


.

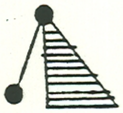
This grafting operation can be extended to sequences of patterns. The graft of a sequence of patterns is the sequence obtained by performing all possible grafting operations between adjacent patterns.

For example, if

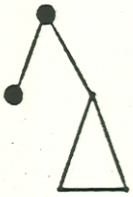
$p_1 = p_2 =$



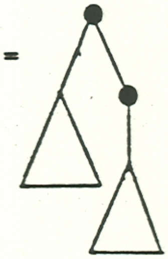
$p_3 =$



$p_4 =$

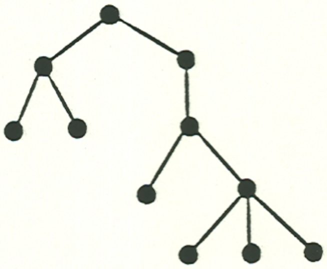


$p_5 =$



then

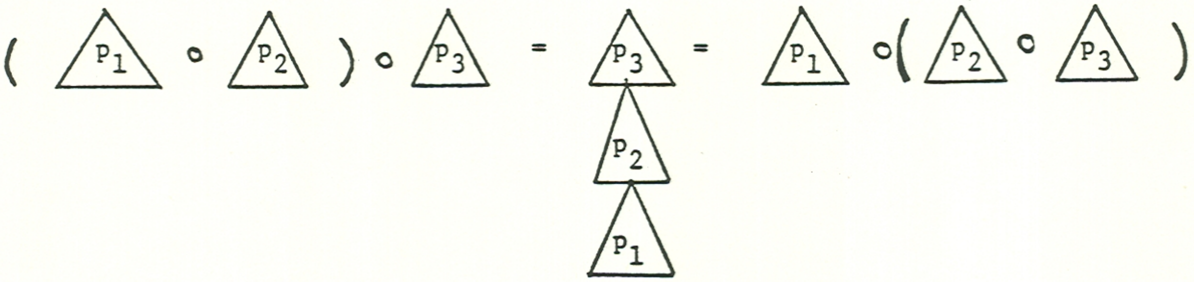
$p_1 \circ p_2 \circ p_3 \circ p_4 \circ p_5 =$



This operation is well defined due to the following property:

Lemma: The grafting operation is associative.

Proof: (by diagram): Let $p_1$, $p_2$, $p_3$ be three patterns. There are four cases to be considered, according as $p_2$ and $p_3$ either do or do not contain slots. The interesting case is when $p_2$ and $p_3$ each have at least one slot. Then



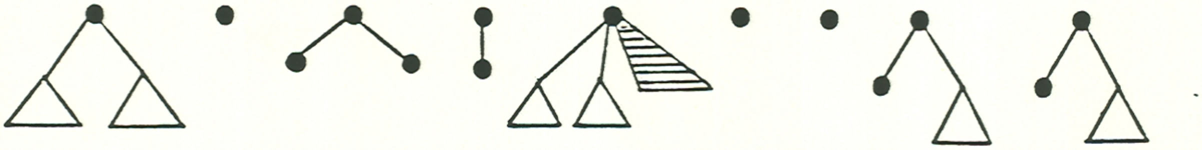
with only the rightmost slot pictured. The other three cases are simpler. ■

The corollary now follows:

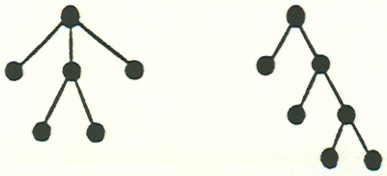
Cycle Lemma for Trees: For any sequence $p_1$, $p_2$, ..., $p_m$ of m patterns, containing n (open or closed) clots, $m > n$, there exist exactly $m - n$ cyclic permutations $p_jp_{j+1}, \cdots p_m\ p_1 \cdots p_{j-1}$, $1 \leq j \leq m$, that can be grafted together to form a forest of $(m - n)$ slotless trees.

Proof: To see this, we adapt the Cycle Lemma with each pattern acting as a box followed by as many circles as the pattern has slots. Arrange the m patterns on a cycle. Since there are more patterns than slots, there must be at least one pattern p without slots that is followed by a pattern p' with slots. Graft p to p'. Continue grafting in this manner until no slots are left. Clearly $m - n$ trees remain. Any of the $m - n$ patterns at the roots of the remaining trees could be at the end of a cyclic permutation whose graft yields a forest of $m - n$ slotless trees. ■

For example, the sequence



of nine patterns can be grafted into a forest beginning with the second or sixth one and continuing around. In either case, the forest obtained contains the two trees



# 4 MAIN RESULT

Our main result is the following enumeration formula for occurrences of a multiset of patterns in the class of ordered trees with a given number of edges:

Theorem: The total number of occurrences of a multiset $\{n_0*p_0, n_1*p_1, \dots, n_k*p_k\}$ of patterns among all ordered trees with n edges is

$$\frac{1}{n-e+d+1}\binom{n-e+d+1}{n_0,\ n_1,\dots,n_k}\binom{2n-m-2e+s}{n-e},$$

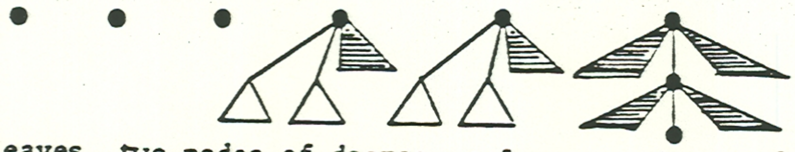
where e is the total number of edges in the patterns (excluding slots), s is the total number of (open and closed) slots, d is the number of open slots, and $m = \sum n_i \leqslant n-e+d+1$ is the total number of patterns.

By the second factor in the formula, we intend the multinomial coefficient

$$\frac{(n-e+d+1)!}{n_0!n_1!\dots n_k!(n-e+d+1-m)!}.$$

which is taken to be 0 when $m > n-e+d+1$.

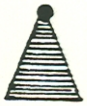
For example, the number of occurrences of the multiset of patterns



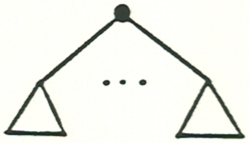
(three leaves, two nodes of degree at least two, and one leaf below level one) in the class $T_8$ of 1430 ordered trees with eight edges is ( $n=8$, $k=2$, $n_0=3$, $n_1=2$, $n_2=1$, $m=6$, $e=6$, $s=10$, $d=4$)

$$\frac{1}{7}\binom{7}{3,2,1}\binom{8}{2} = 1680.$$

Proof: The $m$ given patterns include $m + e-d$ nodes. That leaves $(n+1)-(m+e-d)$ of the nodes in a tree unrestricted, for each of which we add single closed slot patterns



The $(s-d) + (n+1) - (m + e-d) = n + 1-m-e + s$ closed slots we now have can each be replaced with an open slot pattern of the form



containing among themselves the $n-e$ edges unaccounted for in the given patterns, in $\binom{(n + 1-m-e + s) + (n-e)-1}{n-e} = \binom{2n-m-2e + s}{n-e}$ ways. The $m + (n + 1)-(m + e-d)=n-e + d + 1$ patterns can be placed on a cycle in

$$\frac{1}{n-e + d + 1} \binom{n-e + d + 1}{n_0, n_1, \ldots, n_k} \qquad \text{ways.}$$

By the Cycle Lemma for Trees, such a cycle of patterns - containing $n - e + d$ slots and $n + 1 - e + d$ patterns - corresponds to a unique tree. (We use here the Cycle Lemma for Trees with only open slots.)

Each possible arrangement of patterns on the cycle yields a different occurrence. Since only open slots are filled, the patterns do not share nodes. ∎

The above theorem generalizes

- (Harary, Prins, and Tutte [1964]) the Catalan number

$$\frac{1}{n+1}\binom{2n}{n}$$

for unrestricted ordered trees with $n$ edges ($k = -1$, $m=e=s=d=0$):

- (Cayley [1859]) the Catalan number

$$\frac{1}{2r+1} \binom{2r+1}{r}$$

for unrestricted binary trees with r binary nodes and hence r+1 leaves ($n=2r$, $k=1$, $n_0=r$, $n_1= r+1$, $m=1\,r+1$, $e=d=s=2r$);

- (Tutte [1964]) the multinomial formula

$$\frac{1}{n+1} \binom{n+1}{n_0n_1,\ldots,n_n}$$

for enumerating trees with $n_i$ nodes of degree i and a total of n edges ($k=n$, $m=n+1$, $e=s=d=n$);

- (Flajolet and Steyaert [1980]) the binomial formula

$$\binom{2n-2e+s-1}{n-e}$$

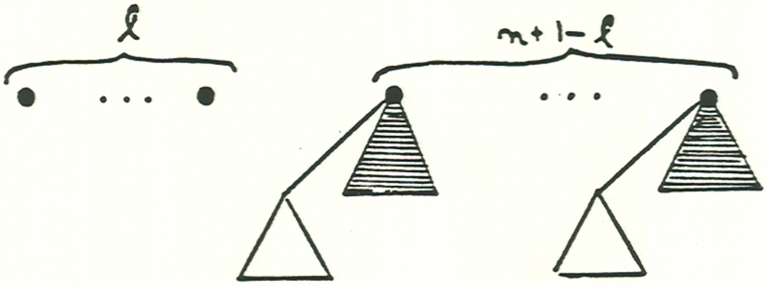
for occurrences of a single pattern (with no closed slots and no leaves) among all ordered trees with n edges ($k=0$, $m=1$).

Such a theorem can also be proved using generating functions (Steyaert [1983]).

# 5 APPLICATIONS

The theorem of the previous section has wide applicability. We give here some representative illustrations.

<u>Example 1.</u> The number of ordered trees with n edges and exactly $\ell$ leaves is equal to the number of occurrences of the patterns

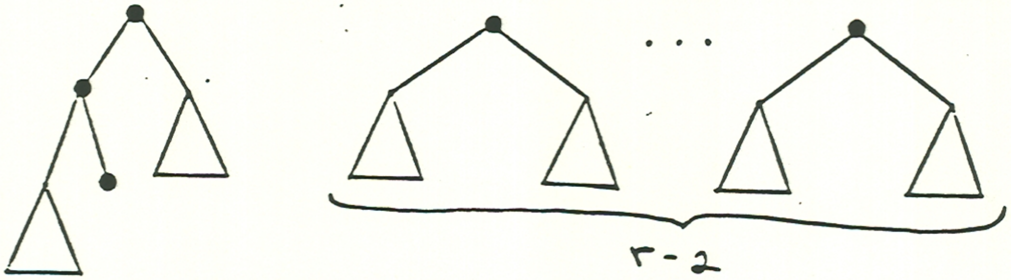


($\ell$ leaves and $n+1-\ell$ nodes of degree at least one), since there can be only one occurrence per tree. By letting $k=1$, $n_0=\ell$, $n_1=n+1-\ell$, $m=n+1$, $e=n+1-\ell$, $d=n+1-\ell$, and $s=2(n+1-\ell)$ in the theorem, we get

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

These well-known numbers appear in Narayana [1959] in the context of ballots and in Riordan [1968] in reference to a communication problem. See also Dershowitz and Zaks [1980], where these numbers are derived using the Cycle Lemma.

Example 2. The number of left children having right leaves among all binary trees with r internal nodes is equal to the number of occurrences of the patterns

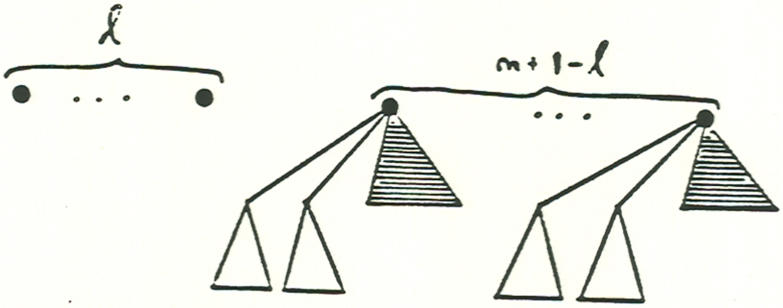


(a left child with a right leaf and the remaining r-2 nodes). By letting n=2r, k=1, $n_0$=1, $n_1$=r-2, m=r-1, e=2r, d=2r-2, and s=2r-2 in the theorem, we get

$$\frac{1}{2r-1}\binom{2r-1}{1,r-2}\binom{r-1}{0}=\binom{2r-2}{r}.$$

This formula is derived in Brinck and Foo [1981] using generating functions. See also Gouyou-Beauchamps [1975].

Example 3. The number of ordered trees with n edges, ℓ leaves, and no unary nodes is equal to the number of occurrences of the patterns



(ℓ leaves and n+1-ℓ nodes with at least two edges). By letting k=1, $n_0$=ℓ, $n_1$=n+1-ℓ, m=n+1, e=2(n+1-ℓ), d=2(n+1-ℓ), and s=3(n+1-ℓ) in the theorem, we get

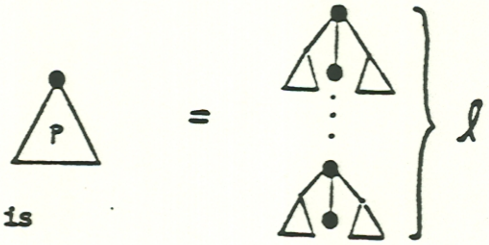
$$\frac{1}{n+1}\binom{n+1}{\ell}\binom{\ell-2}{n-\ell}.$$

Summing this for all n, we get

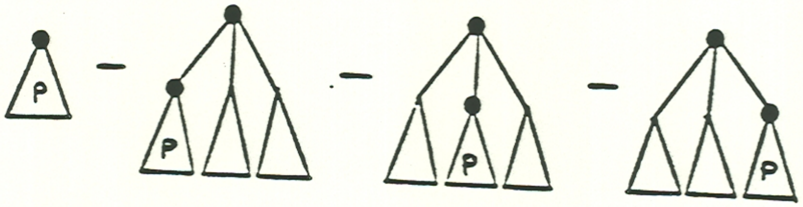
$$\frac{1}{\ell}\sum_{n}\binom{\ell-2}{n-\ell}\binom{n}{\ell-1}$$

for any given number of leaves $\ell$. These numbers were investigated by Schröder [1870]; their relation to polygon partitions is discussed in Motzkin [1948]; their relation to trees appears in Knuth [1968]. See also Donaghey [1977,1980], Donaghey and Shapiro [1977], and Rogers and Shapiro [1977].

Example 4. The number of ternary trees with r internal nodes and a leaf $\ell$ levels directly below the root may be determined in the following manner: Let the pattern p be



The desired number is



(a leaf directly beneath a node, but not beneath a non-root node), where each pattern denotes the number of times it occurs in the class of r-node ternary trees. By letting $n=3r$, $k=1$, $n_0=1$, $n_1=r-\ell$, $m=r-\ell+1$, $e=3r$, and $d=s=3r-\ell$ in the theorem, we get

$$\frac{1}{3r-\ell+1}\binom{3r-\ell+1}{1,r-\ell}\binom{3r-1}{0} = \binom{3r-\ell}{r-\ell}$$

for the pattern p; by letting $n=3r$, $k=1$, $n_0=1$, $n_1=r-\ell-1$, $m=r-\ell$, $e=3r$, and $d=s=3r-\ell-1$ in the theorem, we get

$$\frac{1}{3r-\ell}\binom{3r-\ell}{1,r-\ell-1}\binom{3r-2}{0} = \binom{3r-\ell-1}{r-\ell-1}$$

for each of the excluded patterns. Thus, the desired number of occurrences is

$$\binom{3r-\ell}{r-\ell}-3\binom{3r-\ell-1}{r-\ell-1}= \frac{2\ell}{3r-\ell}\binom{3r-\ell}{r-\ell}.$$

This is the same as the number of ordered forests of $2\ell$ ternary trees with a total of $r-\ell$ internal nodes. Similarly, it can be shown that the number of $\ell$ tree forests of $t$-ary trees with a total of $r$ internal nodes (of degree $t$) is

$$\frac{\ell}{tr+\ell}\binom{tr+\ell}{r} .$$

## REFERENCES

1. K. Brinck and N. Y. Foo [1981], "Analysis of algorithms on threaded trees," *Comp. J.*, Vol. 24, no. 2, pp. 148-155.
2. A. Cayley [1859], "On analytical forms called trees," *Phil. Mag.*, Vol. 20, pp. 374-378. Also in *Collected Math. Papers*, Vol. 4, pp. 112-115.
3. N. Dershowitz and S. Zaks [1980], "Enumerations of ordered trees," *Discrete Math.*, Vol. 31, no. 1, pp. 9-28.
4. N. Dershowitz and S. Zaks [1982], "The Cycle Lemma and some applications," Technical Report 238, Department of Computer Science, Technion, Haifa, Israel.
5. R. Donaghey [1977], "Restricted plane tree representations of four Motzkin-Catalan equations," *J. Combinatorial Theory* (B), Vol. 22, pp. 114-121.
6. R. Donaghey [1980], "Automorphisms on Catalan trees and bracketings," *J. Combinatorial Theory* (B), Vol. 29, pp. 75-90.

7. R. Donaghey and L. W. Shapiro [1977], "Motzkin numbers," J. Combinatorial Theory (A), Vol. 23, pp. 291-301.

8. A. Dvoretzky and Th. Motzkin [1947], "A problem of arrangements," Duke Math. J., Vol. 14, pp. 305-313.

9. Ph. Flajolet and J. M. Steyaert [1980], "On the analysis of tree-matching algorithms," Proc. 7th Intl. Conf. Automata, Languages and Programming, Amsterdam, pp. 208-220.

10. D. Gouyou-Beauchamps [1975], "Deux propriétés combinatoires du langage de Łukasiewicz," R.A.I.R.O., Vol. 3, pp. 13-24.

11. F. Harary, G. Prins, and W. T. Tutte [1964], "The number of plane trees," Indagationes Math., Vol. 26, pp. 319-329.

12. D. E. Knuth [1968], The Art of Computer Programming, Vol. 1: "Fundamental algorithms," Addison-Wesley, Reading, MA.

13. Th. Motzkin [1948], "Relations between hypersurface cross ratios and a combinatorial formula for partititons of a polygon, for permanent preponderance, and for non-associative products," Bull. Amer. Math. Soc., Vol. 54, pp. 352-360.

14. T. V. Narayana [1959], "A partial order and its applications to probability," Sankhya 21, pp. 91-98.

15. G. M. Raney [1960], "Functional composition patterns and power series reversion," Transactions of the AMS, Vol. 94, pp. 441-451.

16. J. Riordan [1968], Combinatorial Identities, Wiley, New York.

17. D. G. Rogers and L. W. Shapiro [1977], "Some correspondences involving the Schröder numbers and relations," Lecture Notes in Mathematics 686, pp. 267-274.

18. E. Schröder [1870], "Vier combinatorische Probleme," Zeit. fur Math. und Physik, Vol. 15, pp. 361-376.

19. J. M. Steyaert [1983], personal communication.

20. W. T. Tutte [1964], "The number of planted plane trees with a given partition," Amer. Math. Monthly, Vol. 71, pp. 272-277.