Half of the nodes of Catalan trees are leaves

Cedric Chauve

LaBRI, Université de Bordeaux I

Unpublished notes, 1999.

HALF OF THE NODES OF CATALAN TREES ARE LEAVES

CEDRIC CHAUVE

Abstract. We give four solutions for a problem about Catalan trees given in [4], and we conclude by an extension of one of our solutions which gives the number of left leaves in complete m-ary trees.

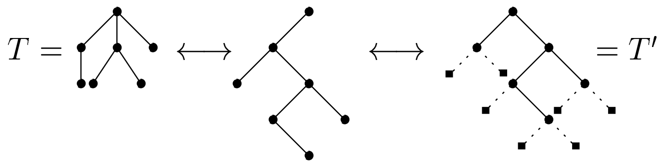
1. Introduction

In [4], Louis Shapiro gives the following problem about Catalan numbers and ordered (rooted) trees (here called Catalan trees). It is well known that the number of Catalan trees with n edges is given by the nth Catalan number $C_n = \binom{2n}{n}/(n+1)$; prove that half of the nodes of all the Catalan trees on n edges are leaves (i.e. there are $(n+1)C_n/2$ leaves among the $(n+1)C_n$ nodes of these trees).

Here we give four different solutions for this problem. The first, which is the more direct and elegant, uses a classical bijection from trees to binary trees and an involution on these trees. The second relies on the encoding of Catalan trees by Dyck words, and is also fully bijective. The third involves the Narayana numbers and the fourth, the easiest to find but the less elegant, is based on a recurrence verified by Catalan trees having a pointed leaf. We conclude by showing that the second solution, which allows to interpret the initial problem in terms of counting left leaves in complete binary trees, can be extended to count left leaves in complete m-ary trees (trees such that each internal node has m sons).

2. A solution involving complete binary trees

There is a classical bijection from Catalan trees on n edges to complete binary trees having n internal nodes and $(n+1)$ leaves (see [3] for example), which verifies that, if T is a Catalan tree on n edges and T' the corresponding complete binary tree having n internal nodes, then the number of leaves of T is the number of left leaves of T' (a leaf is a left leaf if it is the left son of its father).



It follows that the total number of leaves in Catalan trees on n edges is the total number of left leaves in complete binary trees on n internal nodes. Now, for each complete binary tree T on n internal nodes and $n+1$ leaves, consider the complete binary tree T' obtained by exchanging, for each internal node, its left subtree and its right subtree. It is immediate to see that the number of leaves in the two trees T and T' is $2(n+1)$ and the number of left leaves is $(n+1)$. Hence, if we apply this transformation, which is in fact an involution, on all the complete binary trees on n internal nodes, we will obtain each tree of this family twice. It follows that the number of left leaves of this family of trees is half the number of leaves (left and right leaves), i.e. $(n+1)C_n/2$.

3. A solution involving Dyck words

A word $w \in \{a,b\}^{2n}$ is a Dyck word of length $2n$ if and only if w has n letters a and n letters b and there is no prefix v of w such that $|v|_b$ (the number of b in v) is greater than $|v|_a$. A *peak* in a Dyck word is a subword ab (there is at least one peak in a Dyck word).

There is a well known bijection from Catalan trees on n edges to Dyck words of length $2n$. Let T be such a tree and w_T an empty word: during a depth first search in T, for each edge, the first time it is encountered put a letter a at the end of w_T and the last time it is encountered put a letter b; the inverse of this map is immediate. For example the Dyck word corresponding to the tree T displayed in the previous section is $aabbaababbab$.

Moreover, it is immediate to see that the number of leaves of T is exactly the number of peaks of w_T. It follows that the number of leaves in the Catalan trees on n edges is the number of peaks in the Dyck words of length $2n$, or equivalently the number of Dyck words of length $2n$ having a pointed peak (called *pointed Dyck words*).

Now, we will give a bijection from pointed Dyck words of length $2n$ to words of $\{a,b\}^{2n}$ having n letters a and n letters b and such that the last letter is a a (there are clearly $(n+1)C_n/2$ such words).

First let w be a pointed Dyck word of length $2n$: $w = w'\underline{ab}w''$. Let w'_1 be the longest prefix of w' having as much letters a as letters b (w'_1 can be empty). Then we have $w = w'_1w'_2\underline{ab}w''$ and we define the word $W = w'_1bw''w'_2a$. For example

$$\begin{aligned} w &= abaababa\underline{ab}babbab = (ab)(aababa)\underline{ab}(babbab) \\ \Longrightarrow W &= (ab)\underline{b}(babbab)(aababa)\underline{a} = abbbabbabaababa. \end{aligned}$$

Now, let W be a word of $w \in \{a,b\}^{2n}$ having n letters a and n letters b and such that the last letter is a a. Let W' the longest prefix of W such that $|W|_b - |W|_a$ is maximal and W'_1 the longest prefix of W' which is a concatenation of Dyck words. Then we have $W = W'_1W'_2W''$, where W'_1 ends with a b and W'' ends with a a. We define $w = W'_1W''W'_2$ and we point the peak located at the junction of W'_1 and W''. For example

$$\begin{aligned} W &= abbbabbabaababa = (ab)(bbabbab)(aababaa) \\ \Longrightarrow w &= (ab)(aababa)(bbabbab) = abaababa\underline{ab}babbab. \end{aligned}$$

4. A solution involving Narayana numbers

The number of Catalan trees on n edges having k leaves is given by the Narayana number $u(n,k)$ (see [1, Corollary 4.3] for example), given by:

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

It follows that the number of leaves in the Catalan trees on n edges, denoted by $l(n)$, is

$$l(n) = \sum_{k=1}^{n} ku(n,k) = \sum_{k=1}^{n} \frac{k}{n}\binom{n}{k}\binom{n}{k-1} = \sum_{k=1}^{n}\binom{n-1}{k-1}\binom{n}{k-1}.$$

Now, to conclude, we show that

$$\sum_{k=1}^{n}\binom{n-1}{k-1}\binom{n}{k-1}$$

counts the number of words of $\{a,b\}^{2n}$ having n letters a and n letters b and a letter a for last letter (there are $(n+1)C_n/2$ such words):

- first choose among the n last "places" of such a word, excepted the last place which receives a a, $k-1$ places for letters b and fill the $(n-k)$ remaining places with letters a,
- then, choose $k-1$ among the n first "places" for letters a and fill the $(n-k+1)$ remaining with letters b.

5. An inductive solution

We call *pointed Catalan trees* the Catalan trees where a leaf has been pointed. Let $s(n,k)$ be the number of Catalan trees having n edges such that the root has exactly k sons and $r(n,k)$ the number of pointed Catalan trees such that the root has exactly k sons. Here are the first rows of the triangles of these numbers (first $s(n,k)$, then $r(n,k)$).

$n\backslash k$	1	2	3	4
1	1			
2	1	1		
3	2	2	1	
4	5	5	3	1

$n\backslash k$	1	2	3	4
1	1			
2	1	2		
3	3	4	3	
4	10	12	9	4

We want to show that

$$R_n = \sum_{k=1}^{n} r(n,k) = \frac{(n+1)}{2}\, C_n.$$

In order to show this result, we give a recurrence formula for the numbers $r(n,k)$ involving Catalan numbers and the numbers $s(n,k)$, which is based on a decomposition of a pointed Catalan tree obtained by cutting the edge linking the root and its last son. First, it is immediate to see that

$$r(n,1) = R_{n-1}.$$

Next we have that the pointed Catalan trees on n edges such that the root has k sons (for $k>1$) are in bijection with the union of the following sets:

- the pointed Catalan trees on $(n-1)$ edges such that the root has $(k-1)$ sons,
- the Catalan trees on $(n-1)$ edges such that the root has $(k-1)$ sons,
- the pairs of trees (T_1,T_2) where T_1 is a pointed Catalan tree on i edges $(i=2,\ldots,n-k)$ such that the root has $k-1$ sons, and T_2 a Catalan tree on i edges,
- the pairs of trees (T_1',T_2') where T_1' is a Catalan tree on i edges $(i=2,\ldots,n-k)$ such that the root has $k-1$ sons, and T_2' a pointed Catalan tree on i edges.

It follows that

$$r(n,k) = R_{n-1} + C_{n-1} + \sum_{i=2}^{n-k} r(n-i,k-i)C_{i-1} + s(n-i,k-i)R_{i-1},\ \ k>1$$

and then

$$R_n = R_{n-1} + \sum_{k=2}^{n}\left(\sum_{i=2}^{n-k} r(n-i,k-i)C_{i-1} + s(n-i,k-i)R_{i-1}\right),$$

which is equivalent to

$$R_n = 2R_{n-1} + C_{n-1} + \sum_{i=2}^{n-1}\left(R_{n-i}C_{i-1} + C_{n-i}R_{i-1}\right).$$

Now we verify that $R_1 = C_1 = 1$, and we suppose, as an hypothesis of induction that, for $i < n$ $R_i = (i+1)C_i/2$. It follows, by induction, that

$$R_n = \frac{(n+1)}{2}C_{n-1} + \sum_{i=2}^{n-1}\left(\frac{(n-i+1)}{2}C_{n-i}C_{i-1} + \frac{i}{2}C_{n-i}C_{i-1}\right)$$

$$R_n = \frac{(n+1)}{2}C_{n-1} + \frac{(n+1)}{2}\sum_{i=2}^{n-1}(C_{n-i}C_{i-1}).$$

If we recall that the Catalan numbers verify the following relation

$$C_0 = 1,\ C_n = \sum_{i=1}^{n} C_{i-1}C_{n-i},$$

we have

$$R_n = \frac{(n+1)}{2}C_{n-1} + \frac{(n+1)}{2}\left(-2C_{n-1} + \sum_{i=1}^{n}(C_{n-i}C_{i-1})\right)$$

which leads to the final result

$$R_n = \frac{(n+1)}{2}\sum_{i=1}^{n}(C_{n-i}C_{i-1}) = \frac{(n+1)}{2}C_n.$$

6. An extension of the second solution

In the second solution, we use a bijection from Catalan trees on n edges to Dyck words of length $2n$. There is another classical bijection from complete binary trees on n internal nodes to Dyck words of length $2n$. Let T be such a tree and w_T an empty word: during a depth first search in T, for each encountered internal node put a letter a at the end of w_T and for each encountered leaf, excepted the last, put a letter b; the inverse of this map is immediate. Moreover, we can see that the number of left leaves in such a tree is the number of peaks in the corresponding Dyck word. For example the Dyck word corresponding to the tree T' displayed in the first section is $aabbaababbab$. Hence we can say that the number of left leaves in the complete binary trees having n internal nodes is $\binom{2n}{n}/2$.

Now we denote by $D(m,n)$ the words of $\{a,b\}^{mn}$ having n letters a and $(m-1)n$ letters b and such that no prefix v verifies $(m-1)|v|_a < |v|_b$ (the Dyck words of length $2n$ are $D(2,n)$). In the same way than in the previous paragraph, if we consider the complete m-ary trees having n internal nodes (and so $((m-1)n+1)$ leaves), there is a well known bijection between m-ary trees having n internal nodes and words of $D(m,n)$ (the bijection is based on the same principle to associate, during a depth first search in a tree, letters a to the internal nodes and letters b to the leaves, excepted the last, and the number of left leaves in a tree is the number of peaks in the corresponding word). It follows, by the principle of conjugation of words (see [2]), that the number of such trees is $\binom{mn}{n}/((m-1)n+1)$.

Now we call pointed m-ary tree a complete m-ary tree having a pointed leaf, and we denote by $D'(m,n)$ the set of words of $\{a,b\}^{mn}$ having n letters a, $(m-1)n$ letters b and such that the last letter is a. Then, using exactly the same technic than in the second solution (where we replace Dyck words by words from $D(m,n)$, and all the expressions $|w|_a$ by $(m-1)|w|_a$), we can say that there is a bijection between the pointed m-ary trees having n internal nodes and the words of $D'(m,n)$. For example, to the following word w, which is a pointed word from $D(3,4)$

$$w = abba\underline{ab}bbabbb = (abb)(a)\underline{ab}(bbabbb)$$

corresponds the following word W from $D'(3,4)$

$$W = (abb)\underline{b}(bbabbb)(a)\underline{a} = abbbbbabbbaa.$$

Then it follows from the definition of $D'(m,n)$ that the number of left leaves in the family of m-ary trees having n internal nodes is

$$\frac{1}{m}\binom{mn}{n}.$$

Now if we denote respectively by $N(m,n)$ and $L(m,n)$ the number of nodes and the number of left leaves in complete m-ary trees, as a consequence of the previous result, we can say that

$$\lim_{n\to\infty} \frac{N(n,m)}{L(n,m)} = \frac{m^2}{(m-1)}.$$

References

1. W. Y. C. Chen, *A general bijective algorithm for trees*, Proc. Nat. Acad. Sci. U.S.A. **87** (1990), no. 24, 9635–9639.
2. L. Chottin, *Énumération d'arbres et formules d'inversion de séries formelles*, J. Combin. Theory Ser. B **31** (1981), no. 1, 23–45.
3. D. E. Knuth, *The art of computer programming*, Addison-Wesley Publishing Company, Reading, MA, 1973, Volume 3: Sorting and searching.
4. L. Shapiro, *Problem 10753*, Amer. Math. Monthly **106** (1999), no. 8, 777.

LaBRI, Université Bordeaux I, 351 Cours de la Libération, 33405 Talence, France