Discussiones Mathematicae
Graph Theory 16 (1996) 181–195

A PARTITION OF THE CATALAN NUMBERS AND ENUMERATION OF GENEALOGICAL TREES

Rainer Schimming

Institut für Mathematik und Informatik
Ernst-Moritz-Arndt-Universität
D–17487 Greifswald, Germany

Abstract

A special relational structure, called *genealogical tree*, is introduced; its social interpretation and geometrical realizations are discussed. The numbers $C_{n,k}$ of all abstract genealogical trees with exactly $n+1$ nodes and k leaves is found by means of enumeration of code words. For each n, the $C_{n,k}$ form a partition of the n-th Catalan numer C_n, that means $C_{n,1} + C_{n,2} + \cdots + C_{n,n} = C_n$.

Keywords: genealogical tree, Catalan number, generating function.

1991 Mathematics Subject Classification: 05C30, 05A15.

1. Introduction

A *tree* is, by definition, a connected graph which contains no circuits. A *rooted tree* has a specially designated node, called the *root*. All nodes with the same graph-theoretical distance from the root form a *level set* or, shortly, a *level*.

We introduce a *genealogical tree* as a rooted tree with linearly ordered levels such that the following *hereditariness* condition holds. Let nodes x, y, called *parents*, be in a level l, and let nodes x', y', called *children*, be in the next higher level l' and adjacent to x', y' respectively. Then $x < y$, if and only if $x' < y'$.

Some social model is at hand: the progeny of one distinguished person forms a rooted tree by the following interpretation :

node $\cong$ *person*,
root $\cong$ *progenitor*,
adjacency $\cong$ *parent-child relation*,
internal node $\cong$ *person with offspring*,
leaf $\cong$ *person without offspring*,

level $\cong$ *generation.*

We further assume that there is a *social order* in each generation. It might be given by the temporal order of births or by a system of privileges. It shall be hereditary, that means pass over from the parents to their children: if x is socially higher then y, then every child x' of x is socially higher than every child y' of y.

For a geometrical realization in the Euclidean plane it is convenient to represent the parent-child relation nearly vertically, pointing downward, and the social order in a generation horizontally, pointing from left to right. The root or the progenitor is at the top of the graph. We will also present an alternative geometrical realization of the same relational structure by means of non-intersecting circles the centres of which lie on a fixed straight line.

The main aim of the present paper is to enumerate the aforesaid relational structures. We find the number of all *abstract genealogical trees*, that means isomorphism classes of concrete genealogical trees, with exactly $n+1$ nodes and k leaves to be equal to

$$C_{n,k} = \frac{1}{n}\binom{n}{k}\binom{n}{k-1} \equiv \frac{1}{k}\binom{n-1}{k-1}\binom{n}{k-1}.$$

Our method of proof is to encode every abstract genealogical tree by a word built from the alphabet $x, f, (,)$, where x is interpreted as a variable, f as a function, and the parentheses are technical symbols. The words under consideration show a natural recursive structure with respect to the concatenation and this induces a functional equation for the generating function of the double sequence $C_{n,k}$ which is not difficult to solve.

Here are the numbers $C_{n,k}$ for $1 \le k \le n \le 8$:

	k=1	2	3	4	5	6	7	8
n=1	1							
2	1	1						
3	1	3	1					
4	1	6	6	1				
5	1	10	20	10	1			
6	1	15	50	50	15	1		
7	1	21	105	175	105	21	1	
8	1	28	196	490	490	196	28	1

The n-th row of the scheme of the $C_{n,k}$ is a partition of the n-th Catalan

number C_n :

$$\sum_{k=1}^{n} C_{n,k} = C_n :\equiv \frac{1}{n+1}\binom{2n}{n}.$$

The "*Catalan triangle*" built from the numbers $C_{n,k}$ shares some properties with the Pascal triangle of the binomial coefficients : numbers 1 form the boundary, it is naturally extended by values 0 beyond the boundary, and it is symmetric in an obvious sense.

The enumeration method applied here also works for *coloured trees*, where the nodes are partitioned into classes, called *colours* in the graph-theoretical setting or *social groups* in the above social model. If we distinguish N colours of the leaves and M colours of the internal nodes (different from the root), then the number of abstract genealogical trees becomes

$$C_{n,k} \cdot N^k \cdot M^{n-k}.$$

We can also enumerate the trees where each number of the nodes which carry a specific colour is prescribed. Let us present here the formula for $N = M = 2$, i.e. two "colours", called female and male for simplicity. The number of abstract genealogical trees with k_1 female leaves, k_2 male leaves, l_1 female internal nodes ($\neq$ root), and l_2 male internal nodes ($\neq$ root) equals

$$C_{n,k}\binom{k}{k_1}\binom{l}{l_1}$$

where $k = k_1 + k_2$, $l = l_1 + l_2$, $n = k + l$.

The enumeration of abstract genealogical trees by the above numbers $C_{n,k}$ can be found in two sketchy papers [11,12]. More precisely, Wang Zhenyu [11] considers the number $C_{n,k}$ of structurally different "*ordered trees*" with k leaves and l internal nodes. He finds that the generating function

$$C = C(x,y) = \sum_{k,l=1}^{\infty} C_{kl} x^{k-1} y^l$$

is given by

$$xyC^2 + (x+y-1)C + 1 = 0 \ , \qquad C(0,0) = 1.$$

Eight years later in [12] the numbers C_{kl} were explicitely determined from the generating function. The present paper extends the sketchy arguments of [11,12] to full proofs and connects the graph-theoretical problem with some

geometrical and social interpretations. Moreover, the algebraic enumeration method employed here could be applied to other problems too.

A special aspect of the enumeration problem, namely the symmetry $C(x,y) = C(y,x)$ or the symmetry of the "*Catalan triangle*", posed as a problem on code-words, appeared in [9,1,8,10]. The method in [1,8] is more elementary, less general, than the method applied here. The solution [10] explains why $C(x,y) = C(y,x)$, while in [1,8] the symmetry is a mere conclusion from the recursion formula for the $C_{n,k}$ or from the functional equation for $C(x,y)$.

2. Genealogical Trees

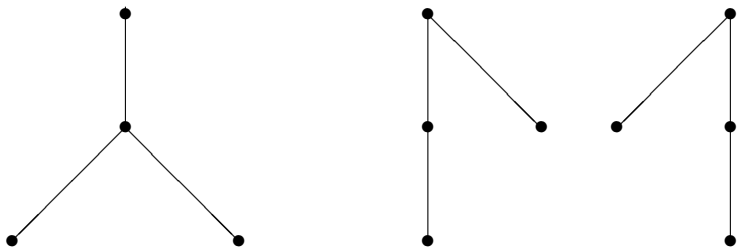
Generally speaking, a *relational structure* $(X, R_1, R_2, \ldots)$ consists of a non-empty set X and relations $R_1, R_2, \ldots$ in X which satisfy given axioms. An *isomorphism* between relational structures $(X, R_1, R_2, \ldots)$, $(X', R'_1, R'_2, \ldots)$ of the same type and with the same axioms is a bijection $f : X \to X'$ which respects the relations, that means xR_iy implies $x'R'_iy'$, where $x' = f(x), y' = f(y)$ $(i = 1, 2, \ldots)$. Let us introduce here a genealogical tree as a special relational structure.

Definition 1. A *genealogical tree* is a triple $G = (X, \wedge, <)$ consisting of a non-empty finite set X and two binary relations $\wedge, <$ in X such that
1. both, the *vertical relation* $\wedge$ and the *horizontal relation*, $<$ are irreflexive partial orders.
2. G with the neighbouring relation N to $\wedge$ is a rooted tree. Here xNy means : $x \wedge y$ and there is no z such that $x \wedge z$ and $z \wedge y$.
3. The restriction of $<$ to a level set of (G, N) is an irreflexive linear order.
4. $<$ is *hereditary* with respect to $\wedge$, that means $x < y$ and $x \wedge x'$, $y \wedge y'$ imply $x' < y'$.

Note that the root of (G, N) is the (unique) minimum with respect to $\wedge$, while the leaves are the maximal elements with respect to $\wedge$.

An isomorphism class of genealogical trees is called an *abstract genealogical tree*. The canonical map which replaces a genealogical tree by its isomorphism class just forgets about names, labels, etc. An abstract tree has nodes, leaves, levels, ... like a concrete tree, but these are merely markers for relations.

Example 1. The abstract genealogical trees with 4 nodes and 2 leaves are given by the pictures :



Note that just the horizontal relation causes the difference between the last two trees. Let $|G|$ denote the number of nodes of G different from the root and $\|G\|$ the number of leaves of G, where G is a concrete or abstract genealogical tree. We want to find the number $C_{n,k}$ of abstract G's with prescribed values $n = |G|$ and $k = \|G\|$. For $k = 1, 2, n, n-1$ elementary combinatorics suffices to solve the problem.

Example 2. There holds

$$C_{n,1} = C_{n,n} = 1, \quad C_{n,2} = C_{n,n-1} = \binom{n}{2}.$$

Proof. A rooted tree with only one leaf looks like Figure 1, hence $C_{n,1} = 1$. Analogously, Figure 2 exhibits $C_{n,n} = 1$. A rooted tree with exactly two leaves bifurcates after $n_0 \geq 0$ nodes into two branches with $n_1 \geq 1$ and $n_2 \geq 1$ nodes; cf. Figure 3. These numbers form a partition $n = n_0 + n_1 + n_2$ of n and there are exactly $\binom{n}{2} = C_{n,2}$ such partitions. Analogous arguments apply to the case $k = n - 1$; cf. Figure 4.

Figure 1

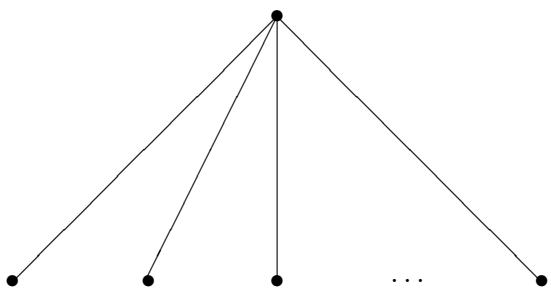


Figure 2

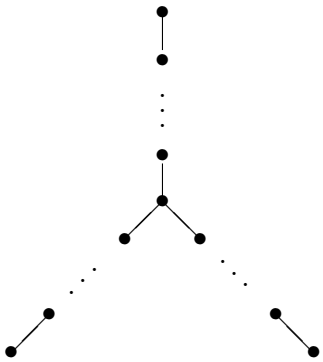


Figure 3

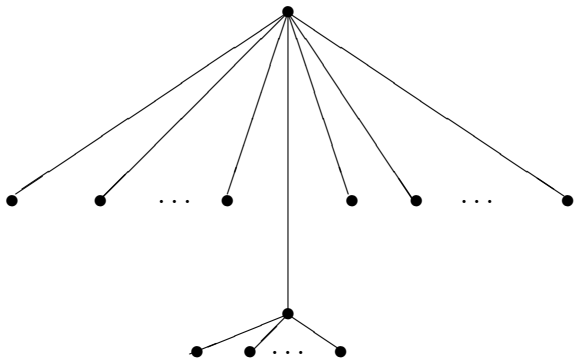


Figure 4

A genealogical tree, taken as a relational structure, can be geometrically realized as a plane graph, like above. There is an alternative realization by circles in the plane, following an idea in the book of Rouse-Ball and Coxeter [7].

Definition 2. A *linear circle configuration*, abbreviated LCC, is a set of circles which have their centres on a fixed straight line and do not intersect each other. Write $x \wedge y$ for circles x, y if y lies in the interior of x, and write $x < y$ if neither $x \wedge y$ nor $y \wedge x$ and if the centre of x is, on the fixed line, left from the centre of y. Further, let $|L|$ denote the number of circles in an LCC L and $\|L\|$ the number of maximal elements with respect to $\wedge$, that means of the most inner circles.

Proposition 1. *Let a given linear circle configuration L be completed by a circle r such that $r \wedge x$ for every $x \in L$. The completed configuration $(L \cup \{r\}, \wedge, <)$ is a genealogical tree and r is its root. Conversely, every genealogical tree $(G, \wedge, <)$ is, as a relational structure, isomorphic to some completed LCC $(L \cup \{r\}, \wedge, <)$.*

Proof. Given L, it is easy to verify the axioms 1. – 4. of a genealogical tree for $(L \cup \{r\}, \wedge, <)$. Given G, the isomorphy condition leads to some obvious geometrical sketch of $L \cup \{r\}$, beginning with the root r as the most outer circle and ending at the leaves as the most inner circles.

Proposition 1 says that abstract LCC's, that means isomorphism classes of LCC's, and abstract genealogical trees are essentially the same objects. Figure 5 shows the abstract LCC's such that $k = 1, 2, n-1, n$, where $n = |L|$, $k = \|L\|$.

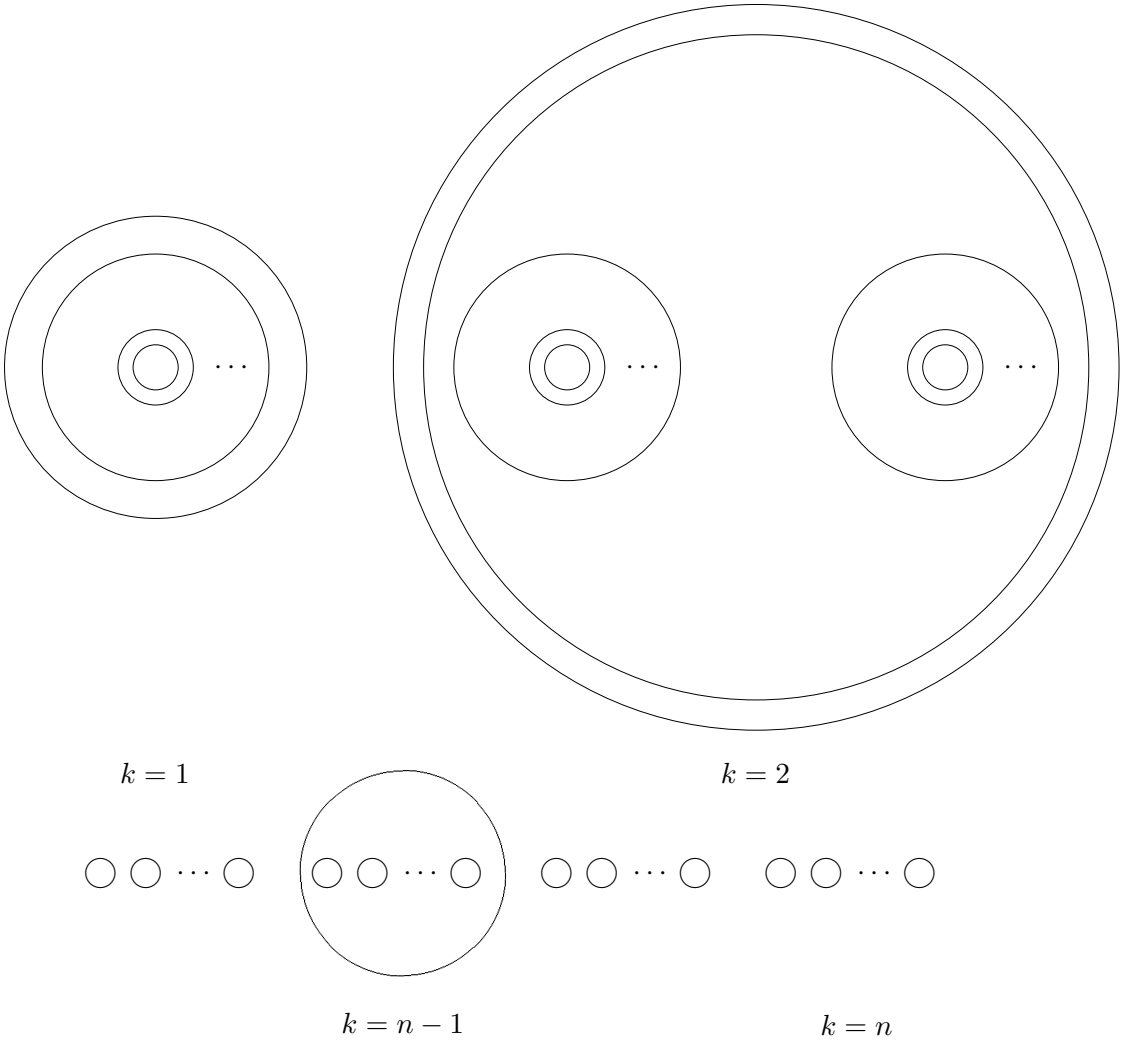


Figure 5

3. Enumeration Through Code-Words

A *semigroup* is, by definition, a non-empty set W together with an associative binary operation $W \times W \to W$, $(v, w) \to v \cdot w$. An *operator* on W is, by definition, an additional unary operation $f : W \to W$, $w \to f(w)$. Let from now on the quadruple $(W, \cdot, x, f)$ denote the free semigroup with exactly one generator x and one operator f. That means, W is the smallest set such that

0. $x \in W$,
1. $w \in W \to f(w) \in W$,
2. $w_1, w_2 \in W \to w_1 \cdot w_2 \in W$.

From a formal point of view, the elements of W are *words* built from the alphabet $x, f, (, \ )$ subject to the above rules. The semigroup operation is represented by the *concatenation* of words. The *length* $|w|$ of a word $w \in W$ is, by definition, the total number of letters x and f in it, while the *degree* $\|w\|$ of w is the number of letters x only. It is convenient to complete W by the empty word 1 which has the properties

$$1w = w1 = w, \ |1| = 0, \ \|1\| = 0.$$

Thus $(W, .)$ becomes a *monoid*, that means a semigroup with a unit element 1. Let us abbreviate :

$$x^n := xx \ldots x, \quad f^n(w) := f(f(\ldots f(w) \ldots)).$$

That means, a word consisting of n copies of one letter is formally written like a power $(n = 1, 2, \ldots)$. Moreover, set $x^0 := 1, \ f^0(w) := w$.

Example 1. The words of degree $1, 2, n, n-1$ have the form

$$f^{n-1}(x),$$

$$f^{k_0}(f^{k_1}(x) f^{k_2}(x)), \text{ where } k_0 + k_1 + k_2 = n - 2, \qquad k_0, k_1, k_2 \geq 0,$$

$$x^n,$$

$$x^{k_0} f(x^{k_1}) x^{k_2}, \text{ where } k_0 + k_1 + k_2 = n - 1, \quad k_0 \geq 0, \quad k_1 \geq 1, \quad k_2 \geq 0,$$

respectively.

It is an essential fact that every word $w \in W$ can be visualized by a genealogical tree and, conversely, every genealogical tree admits a code-word $w \in W$.

Theorem 1. *There is a natural one-to-one map $G \to w$ between abstract genealogical trees G and words w from the monoid $(W, ., x, f,)$ such that $|G| = |w|$, $\|G\| = \|w\|$.*

Proof. We establish some natural one-to-one map $L \to w$ between abstract LCC's L and words w such that $|L| = |w|$, $\|L\| = \|w\|$. Given L, we distinguish three kinds of "*traces*" of the circles of L on the fixed straight line: *left* or *convex traces* of internal circles, *right* or *concave traces* of internal

circles, and *leaves* shrunk to a point, for simplicity. (A *leaf* is maximal with respect to $\wedge$, while an *internal circle* is non-maximal.) That means a *trace* is a point on the fixed straight line together with the additional information left, right, or leaf. Now we construct a code-word w to G successively from left to right along the straight line as follows. A convex trace is encoded by $f($, a concave trace by $)$, and a leaf by x. The construction represents the wanted map. Since LCC's and genealogical trees represent the same relational structures, the theorem is proved.

Example 2. Here are all words of length ≤ 3 together with sketches of their abstract genealogical trees (Figure 6).

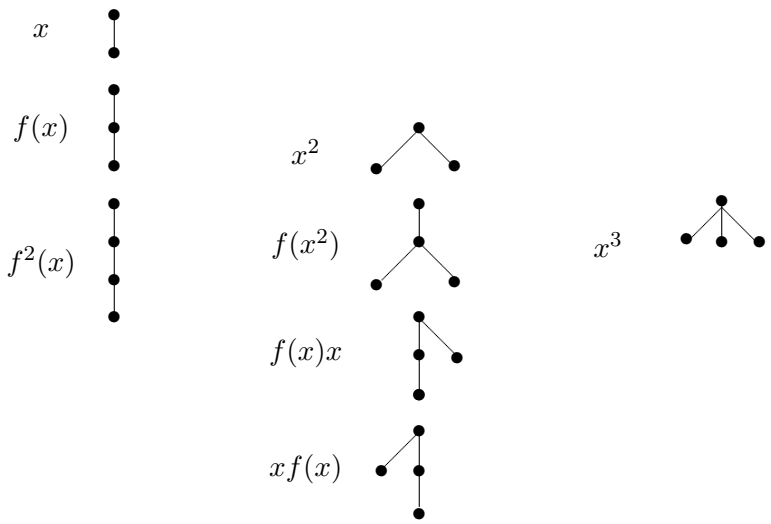


Figure 6

Theorem 1 reduces the enumeration of trees to the enumeration of words: the number $C_{n,k}$ is also equal to the number of words $w \in W$ of length $n = |w|$, and degree $k = \|w\|$. We take into consideration the empty word by setting

$$C_{0,0} = 1,\ C_{0,k} = 0 \text{ for } k \geq 1.$$

Let us introduce a "*forgetful*" *map* π, applied to words $w \in W$ as follows. Parentheses are omitted, a symbol x remains unchanged, and a symbol f is mapped to y, where the images x, y denote commutative variables. That means, π forgets the parentheses and the non-commutativity. Any word w

of length n and degree k is mapped by π to the same commutative word $x^k y^{n-k}$. As a consequence, we have

$$C_{n,k} = |\pi^{-1}(x^k y^{n-k})|$$

where the symbol $|\ |$ now means the cardinality (i.e. number of elements) of a set. In the following more precise setting, π emerges as a homomorphism between rings. We use the standard notations $\mathbf{N}$ = set of nonnegative integers, $\mathbf{Z}$ = set of integers. Let $\mathbf{N}W$ denote the set of all formal sums

$$w_1 + w_2 + \ldots + w_N$$

of words $w_1, \ldots, w_N \in W$. The positive integers $N \in \mathbf{N}$ naturally act on $\mathbf{N}W$ by

$$Nw := w + w + \ldots + w \quad (N\ times).$$

Standard algebraic constructions — namely the bilinear extension of the semigroup product and the introduction of negative elements — extend the semiring $\mathbf{N}W$ to a ring and $\mathbf{Z}$-module $\mathbf{Z}W$. Let further $\mathbf{Z}[x, y]$ denote the ring of polynomials in the commutative indeterminates x, y with integer coefficients.

Proposition 2. *There is a unique ring homomorphism* $\pi : \mathbf{Z}W \to \mathbf{Z}[x, y]$ *such that*

$$\pi(x) = x, \quad \pi(f(w)) = y\pi(w), \quad \pi(1) = 1.$$

Proof. For every $w \in W$ the above rules for π together with $\pi(w_1 w_2) = \pi(w_1)\pi(w_2)$ admit a stepwise calculation of $\pi(w) \in \mathbf{Z}[x, y]$. Obviously, π is the forgetful map introduced earlier.

To our special purpose, we need certain closures of the rings $\mathbf{Z}W$ and $\mathbf{Z}[x, y]$. Let $\overline{\mathbf{Z}W}$ denote the ring of formal series

$$N_1 w_1 + N_2 w_2 + \ldots$$

of words $w_1, w_2, \ldots \in W$ with integer coefficients $N_1, N_2, \ldots$, and $\mathbf{Z}[[x, y]]$ the ring of formal power series in x, y with integer coefficients. Clearly, π can be naturally extended to a homomorphism $\overline{\mathbf{Z}W} \to \mathbf{Z}[[x, y]]$. Note that formal series which differ only by a rearrangement of the summands are identified with each other. There is then a natural identification of a subset $M \subseteq W$ with the sum or series $\sum_{w \in M} w$. In particular, the denumerable whole set W is identified with $\sum_{w \in W} w$.

Proposition 3. *The generating function*

$$C \equiv C(x,y) = \sum_{n=1}^{\infty}\sum_{k=1}^{n} C_{n,k} x^{k-1} y^{n-k} \tag{1}$$

of the double sequence $C_{n,k} := |\pi^{-1}(x^k y^{n-k})|$ *satisfies the quadratic equation*

$$xyC^2 + (x+y-1)C + 1 = 0 \tag{2}$$

and the initial condition

$$C(0,0) = 1. \tag{3}$$

Proof. A word $w \in W$ is either empty, or begins with the letter x, or begins with a factor $f(w'')$:

$$w = 1, \quad or \quad w = xw', \quad or \quad w = f(w'')w',$$

where w' can be empty, while w'' is meant to be non-empty. This classification of words is complete. It defines a partition of W which can be expressed by some equation in the ring $\overline{\mathbf{Z}W}$:

$$W = 1 + xW + f(W-1)W.$$

Let us apply the forgetful homomorphism π; the image $\pi(W) := F$ satisfies

$$F = 1 + xF + y(F-1)F \tag{4}$$

By construction, F is composed of monomials $x^k y^{n-k}$. More precisely, each $x^k y^{n-k}$ appears multiplied with the numerical factor $C_{n,k} = |\pi^{-1}(x^k y^{n-k})|$, that means

$$F = F(x,y) = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} C_{n,k} x^k y^{n-k} = 1 + xC(x,y), \tag{5}$$

where $C = C(x,y)$ is our generating function. Insertion of $F = 1 + xC$ into (4) gives (2). Note that equation (2) is symmetric in x, y, while (4) is not.

Corollary. *There hold*

$$C_{n,k} = C_{n,n+1-k}, \tag{6}$$

$$\sum_{k=1}^{n} C_{n,k} = C_n \equiv \frac{1}{n+1}\binom{2n}{n}. \tag{7}$$

A **proof** can be given without explicitly solving (2). Namely, the problem (2), (3) is symmetric in x, y, hence $C(x,y) = C(y,x)$ which gives (6). Further, by restriction of (4) to the diagonal $x = y$ we obtain

$$xF(x,x)^2 - F(x,x) + 1 = 0, \quad F(0,0) = 1.$$

By this

$$F(x,x) = \sum_{n=0}^{\infty} (\sum_{k=1}^{n} C_{n,k}) x^n$$

is recognized to be the generating function of the Catalan numbers C_n.
We now arrive at the main result of the paper.

Theorem 2. *There holds*

$$C_{n,k} = \frac{1}{n} \binom{n}{k} \binom{n}{k-1}. \tag{8}$$

Proof. Let us remind that the Catalan numbers have the generating function in one variable z

$$\sum_{n=0}^{\infty} C_n z^{n+1} = \frac{1 - (1-4z)^{\frac{1}{2}}}{2}.$$

Our generating function $C = C(x,y)$ in the two variables x, y follows from (2), (3); it is given by

$$2xyC = (1-x-y)[1 - (1-4z)^{\frac{1}{2}}],$$

where now

$$z = xy(1-x-y)^{-2}.$$

We find the expansion

$$C = \sum_{m=0}^{\infty} C_m x^m y^m (1-x-y)^{-2m-1}.$$

Here we insert the well-known binomial series

$$(1-x-y)^{-2m-1} = \sum_{p=0}^{\infty} \binom{2m+p}{2m} (x+y)^p$$

and then the binomial formula

$$(x+y)^p = \sum_{q=0}^{p} \binom{p}{q} x^q y^{p-q}.$$

New summation indices n, k are introduced through

$$p = n-1-2m, \qquad q = k-m-1,$$
$$m+q = k-1, \qquad m+p-q = n-k.$$

After all this, the numerical coefficient of $x^{k-1}y^{n-k}$ in $C(x,y)$ becomes

$$\begin{aligned} C_{n,k} &= \sum_{m=0}^{k-1} C_m \binom{n-1}{2m} \binom{n-1-2m}{k-m-1} \\ &= \frac{1}{n}\binom{n}{k} \sum_{m=0}^{k-1} \binom{n-k}{m} \binom{k}{m+1} \\ &= \frac{1}{n}\binom{n}{k}\binom{n}{k-1}. \end{aligned}$$

Equation (6) expresses a symmetry of the triangular scheme of the numbers $C_{n,k}$. One feels that there ought to be some duality property of the monoid W behind it. We need some new notations in order to actually present this duality. Let the map π on W, introduced above, for the moment only forget the parentheses, but preserve non-commutativity. Let further the map $m \to m^{-1}$ on non-commutative monomials in x, y be defined by

$$x^{-1} = y, \quad y^{-1} = x, \quad (m_1 m_2)^{-1} = m_2^{-1} m_1^{-1}.$$

Let finally, the *duality map* $m \to m^*$ on monomials which end on x be defined by

$$m^* := (m/x)^{-1} x.$$

where m/x originates from m by omission of the letter x at the end.

Proposition 4. *For every non-commutative monomial m in x, y which ends on x the preimages $\pi^{-1}(m) \subset W$ and $\pi^{-1}(m^*) \subset W$ have the same cardinality :*

$$|\pi^{-1}(m)| = |\pi^{-1}(m^*)|.$$

A proof has been given by P. Schreiber [10].
Now $C_{n,k} = C_{n,n+1-k}$ becomes a conclusion from Proposition 4.

4. Discussion

Enumeration of trees has been one of the historical origins of graph theory, cf. [3,2,6,4,5]. Just in this context Arthur Cayley coined the name "tree" for a circuit-free connected graph [3]. At those times, chemical structure formulas were an essential motive for the enumeration of graphs. Later, more motives and more theory evolved. To make a long story short, let us mention Polya's great paper [6] of 1937.

The enumeration problem attacked here belongs to the class of problems which go back to Cayley et al. But it turns out that it is remarkably easy to solve, because of the "horizontal relation" additional to the graph-theoretical adjacency or "vertical relation". The result leads to some partition of the Catalan numbers into positive integers. One may wonder whether others of the many appearances of the Catalan numbers C_n admit a natural partition $C_n = \sum_{k=1}^{n} C_{n,k}$ where the index k has a meaning intrinsically defined by some refinement of the original combinatorial problem. Cf. [4,5] and the literature cited there for enumeration problems leading to the Catalan numbers.

The author gratefully acknowledges stimulating discussions on the topic of the paper with H.-D. Gronau, P. Schreiber and B. Fiedler.

References

[1] H. Bickel and E.H.A. Gerbracht, *Lösung I zu Problem 73*, Math. Semesterber. **42** (1995) 185–187.

[2] H.L. Biggs et al., Graph Theory 1736–1936 (Clarendon Press, Oxford 1976).

[3] A. Cayley, *On the theory of the analytical forms called trees* I, II, Phil. Mag. **13** (1857) 172–176; **18** (1859) 374–378.

[4] R.B. Eggleton and R.K. Guy, *Catalan Strikes Again*, Math. Magazine **61** (1988) 211–219.

[5] P. Hilton and J. Pedersen, *Catalan Numbers, Their Generalizations, and Their Uses*, Math. Intelligencer **13** (1991) 64–75.

[6] G. Polya, *Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen*, Acta Math. **68** (1937) 145–254.

[7] W.W. Rouse Ball and H.S.M. Coxeter, Mathematical Recreations and Essays (12th Edition, Univ. of Toronto Press 1974).

[8] R. Schimming, *Lösung II zu Problem 73*, Math. Semesterber. **42** (1995) 188–189.

[9] P. Schreiber, *Problem 73, Anzahl von Termen.* Math. Semesterber. **41** (1994) 207.

[10] P. Schreiber, *Lösung III zu Problem 73*, Math. Semesterber. **42** (1995) 189–190.

[11] Wang Zhenyu, *Some properties of ordered trees*, Acta Math. Sinica **2** (1982) 81–83.

[12] Wang Zhenyu and Sun Chaoyi, *More on additive enumeration problems over trees*, Acta Math. Sinica **10** (1990) 396–401.

Received 5 July 1996
Revised 14 November 1996