Compact Textbooks in Mathematics

Steven Roman

An Introduction to Catalan Numbers

Birkhäuser

Compact Textbooks in Mathematics

Birkhäuser

Compact Textbooks in Mathematics

This textbook series presents concise introductions to current topics in mathematics and mainly addresses advanced undergraduates and master students. The concept is to offer small books covering subject matter equivalent to 2- or 3-hour lectures or seminars which are also suitable for self-study. The books provide students and teachers with new perspectives and novel approaches. They feature examples and exercises to illustrate key concepts and applications of the theoretical contents. The series also includes textbooks specifically speaking to the needs of students from other disciplines such as physics, computer science, engineering, life sciences, finance.

More information about this series at http://www.springer.com/series/11225

Steven Roman

An Introduction to Catalan Numbers

Birkhäuser

Steven Roman
Professor Emeritus
California State University, Fullerton
Irvine, CA, USA

ISSN 2296-4568 ISSN 2296-455X (electronic)
Compact Textbooks in Mathematics
ISBN 978-3-319-22143-4 ISBN 978-3-319-22144-1 (eBook)
DOI 10.1007/978-3-319-22144-1

Library of Congress Control Number: 2015952313

Mathematics Subject Classification (2010): 05A05, 05A15, 05A18, 05C05, 05C30, 05C38, 05C75

Springer Cham Heidelberg New York Dordrecht London

Cover design: deblik, Berlin

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

To Scott and to Greg
Réquiem ætérnam dona eis,
requiéscant in pace.

Foreword

There have been many recent expositions of the fascinating sequence

$$1, \ 1, \ 2, \ 5, \ 14, \ 42, \ \ldots$$

of Catalan numbers. Indeed, I myself have just completed a monograph on this topic. Steven Roman does an admirable job of providing an introduction to Catalan numbers of a different nature from the previous ones. In particular, he gives complete details of the background information necessary to understand various aspects of Catalan numbers. He has made an excellent choice of topics in order to convey the flavor of Catalan combinatorics. For instance, the discussion of interval structures, stack-sortability, and semiorders provides introductions to these topics not readily available elsewhere in such a straightforward form. The reader who has successfully absorbed the material here may want to try next my own book, which proceeds at a less leisurely pace. Even for those who wish to go no further, they will acquire a good feeling for why so many mathematicians are enthralled by the remarkable ubiquity and elegance of Catalan numbers.

March 12, 2015 Richard Stanley

Preface

On page 219 of Richard Stanley's book *Enumerative Combinatorics*, Volume II (Cambridge University Press), there is an exercise with 66 parts (surely a student's nightmare). Each part defines a finite set of mathematical objects that is counted by the Catalan numbers. Moreover, Stanley has recently completed a monograph called simply *Catalan Numbers* that describes 214 objects counted by the Catalan numbers, along with an additional 68 in the problem sets. The monograph (also Cambridge University Press) will appear sometime in the early part of 2015.

The purpose of the present little book is to provide an introduction to these remarkable numbers. We will look at a smorgasbord of the more prominent combinatorial objects that are counted by the Catalan numbers, after we have discussed the numbers themselves.

The organization of this book is by topic, as a glance at the table of contents will immediately reveal. For example, one chapter is devoted to Catalan numbers and trees and another chapter is devoted to Catalan numbers and permutations. I have endeavored to place the more accessible topics in the earlier part of the book to provide a gradual acclivity in mathematical sophistication.

For those wishing to test their grasp of the subject matter, I have included some exercises at the end of the book. These exercises come primarily from Richard Stanley's books *Enumerative Combinatorics* (Volume II) and *Catalan Numbers*. The relevant citation from these books is given with each problem. I also provide hints or solutions to these exercises.

Irvine, CA, USA — Steven Roman

Contents

Introduction

1

The Binary Decomposition Model

Speaking very generally, many finitary combinatorial objects are *stitched together* in some orderly manner from smaller objects of the same type. For instance, the Cartesian product $A \times B$ of two finite sets A and B is a simple example of stitching together smaller objects to make a larger object of the same type.

Viewed from the opposite direction, many combinatorial objects can be *decomposed* into smaller objects of the same type. For example, a full binary tree with $n > 1$ vertices can be decomposed (with some loss) into its left subtree and its right subtree, as shown in Figure 1.1.

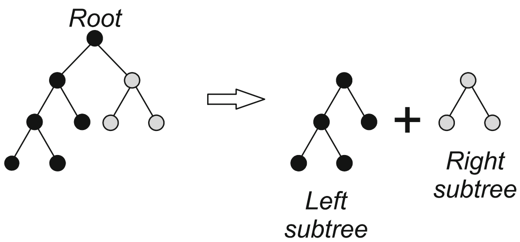


Figure 1.1 Decomposition of a binary tree

(A **full binary tree** is a binary tree in which each vertex has either zero or two children. For more details, please visit the Appendix.) Again speaking very generally, suppose that *widgits* come in nonnegative integral sizes and let $\mathcal{W}_n$ be the finite family of widgits of size n. We wish to count the number of widgits in $\mathcal{W}_n$.

Suppose further that the family $\mathcal{W}_n$ can be *partitioned* into blocks:

$$\mathcal{P}_n = \left\{ \mathcal{W}_{n,k} \middle| \ell \leq k \leq u \right\}$$

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1_1

Thus, by definition, each block is nonempty, the blocks are pairwise disjoint, and their union is $\mathcal{W}_n$. Suppose further that for each k, there is a *bijection*:

$$\theta_{n,k} : \mathcal{W}_{n,k} \to \mathcal{W}_{f(n,k)} \times \mathcal{W}_{g(n,k)}$$

where $f(n,k) < n$ and $g(n,k) < n$. Thus, $\theta_{n,k}$ *decomposes* each widgit $W \in \mathcal{W}_{n,k}$ into a pair of smaller widgits:

$$\theta_{n,k}(W) = (W_1, W_2)$$

where $W_1 \in \mathcal{W}_{f(n,k)}$ and $W_2 \in \mathcal{W}_{g(n,k)}$. It follows that

$$|\mathcal{W}_{n,k}| = \left|\mathcal{W}_{f(n,k)}\right| \ \left|\mathcal{W}_{g(n,k)}\right|$$

and the numbers $D_n = |\mathcal{W}_n|$ satisfy the recurrence relation

$$D_n = \sum_{k=\ell}^{u} D_{f(n,k)} D_{g(n,k)} \qquad [1.1]$$

It is often the case that

$$f(n,k) + g(n,k) = n + \alpha$$

where $\alpha = 0,\ 1,\ -1$, but this is not essential. Let us refer to the partition $\mathcal{P}_n$ and the family of bijections $\theta_{n,k}$ as a **binary decomposition model** for the family $\mathcal{W}_n$. We will also refer to equation [1.1] as the **recurrence relation associated to the decomposition model**.

Defining a binary decomposition model first involves determining the partition $\mathcal{P}_n$ and this usually involves identifying a specific property of widgits that is indexed by the integers k. For example, for full binary trees with n vertices, we can index on the number of vertices in the left subtree, which may be any number k between 0 and $n-1$.

The next step is to define the bijections $\theta_{n,k}$ and this usually involves finding a link or **nexus** within a widgit's structure at which the widgit can be broken into smaller widgits. For example, the root vertex (and its edges) form a nexus for decomposing a full binary tree into its left and right subtrees. Notice that the decomposition process causes a loss of one vertex (the root) and so, in the symbolism above, we have

$$f(n,k) = k \quad \text{and} \quad g(n,k) = n - k - 1$$

whence

$$f(n,k) + g(n,k) = n - 1$$

Note also that we must assign a root to each of the subtrees because the smaller objects must be of the same type as the larger object from whence they came. In this case, it is natural to declare the two vertices that were adjacent to the original root as the roots of the subtrees.

In general, a decomposition map $\theta_{n,k}$ will be injective if the decomposition is *reversible* and surjective if the recombination process can be applied to *any* pair of smaller widgits of size $f(n, k)$ and $g(n, k)$ to produce a widgit of size n. For example, the decomposition of a full binary tree into its left and right subtrees is clearly reversible: We can simply reintroduce a root vertex and connect it to the root vertices of each subtree (in the proper left/right orientation, of course). Also, the recombination process applies to *any* pair of appropriately sized full binary trees.

As we will see, a binary decomposition model applies to many classes of combinatorial objects, such as certain types of trees, permutations, partitions, integer sequences, geometric objects, algebraic objects, partially ordered sets, families of intervals, and more. Perhaps the most well-behaved binary decomposition model occurs when

$$\ell = 0, \quad u = n-1, \quad f(n,k) = k \quad \text{and} \quad g(n,k) = n-1-k$$

in which case the associated recurrence relation is the elegant

$$C_n = \sum_{k=0}^{n-1} C_k C_{n-1-k} \qquad [1.2]$$

This happens to be the recurrence relation that defines the Catalan numbers (along with the initial condition $C_0 = 1$) and perhaps this is why the Catalan numbers count such a large variety of elegant combinatorial objects. Equation [1.2] is called the **Catalan recurrence relation**.

Thus, one way to show that a sequence D_n of integers is the Catalan sequence C_n is to show that $D_0 = 1$ and the underlying widgits have a binary decomposition model whose associated recurrence relation is the Catalan recurrence. This is one of our two primary counting techniques. We describe the other technique next.

Counting by Characterization

The most direct, and one could argue the *only* way to count, is to find a bijection between the set A whose size we wish to determine and a set B whose size is already known.

Put a bit more colorfully, if $f : A \to B$ is a bijection, then we can say that the elements of A are **characterized** or **represented** by the elements of B. To be sure, in simple cases, this language is a bit *too* colorful. For example, even though there is a bijection $\{a, b, c\} \to \{1, 2, 3\}$ one would generally not expect to hear the phrase "the letters a, b and c are characterized by the numbers 1, 2, and 3."

On the other hand, it is not uncommon to say that the subsets of a set of size n are characterized by binary strings of length n, that real polynomials of degree n or less are characterized by ordered $(n+1)$-tuples of real numbers, that graphs (of the combinatorial variety) are characterized by their adjacency matrices and that finite cyclic groups are characterized by the positive integers, to name but a few.

We will encounter several examples of sets of objects that can be characterized as sets of other objects whose quantity we have already determined.

Words Over an Alphabet

Before starting our journey into the Catalan numbers, we need to establish a few preliminaries. There are some additional "preliminaries" in the Appendix for those who might wish to partake.

A **word** over a set $S = \{a_1, \ldots, a_k\}$ is a sequence of members of A, written without punctuation. For example, if $S = \{A, B\}$, then the following is a word over S,

$$w = BAABA$$

The **length** $\text{len}(w)$ of w is the number of symbols in w. For example, the word $w = BAABA$ has length 5. The set S is called the **alphabet**. The set of all words over S is denoted by S^* and the set of all words over S of length k is denoted by S^k. Concatenation (juxtaposition) is a binary operation on S^*. It is associative but not commutative. The **empty word** ϵ with no elements is the identity under concatenation.

Our interest centers around alphabets of size two, often taken to be $\{A, B\}$. However, we must attend to an important technical detail. Depending on the application, we will use symbols other than A and B for the two letters in the alphabet and since the letters do *not* play a symmetric role in the upcoming discussions, we need a way to distinguish them. Accordingly, we will order the alphabet by writing it as an ordered pair (A, B) and refer to the first letter A as the **dominant** letter in the alphabet. Thus, for example, when we speak of a word over the alphabet ((,)), we mean a word whose letters are the two parentheses (and), such as

$$())\,((()) $$

with the left parenthesis being dominant. (This is the only case where the notation ((,)) is rather dreadful.)

Let $\mathcal{W}_{a,b}$ be the set of all words over (A, B) that contain a dominant letters and b nondominant letters. We will need the following concepts. Let $\omega \in (A, B)^*$.

1) The **complement** w' of w is the word formed by replacing A with B and B with A throughout the word. For example,

$$w = AABA \quad \Rightarrow \quad w' = BBAB$$

2) The **prefix** or **initial segment** of w of length k is denoted by $[w]_k$. Thus,

$$w = a_1 \cdots a_n \quad \Rightarrow \quad [w]_k = a_1 \cdots a_k$$

for $1 \leq k \leq n$.

3) The number of A's in w is denoted by $N_A(w)$ and the number of B's by $N_B(w)$. We refer to $N_A([w]_k)$ as the ***A*-count** of w at position k and similarly for the ***B*-count**. For example, if

$$w = AABBBA$$

then the A-count at position 3 is two and the B-count is one. The **count difference** is the function

$$\Delta_{A,B}\left([w]_k\right) = N_A\left([w]_k\right) - N_B\left([w]_k\right)$$

Some Notation

We will write $[n]$ for the set $\{1, 2, \ldots, n\}$. If $(S, \leq)$ is a totally ordered set, then a **closed interval** in S is a set of the form

$$[a, b] = \left\{x \in S \middle| a \leq x \leq b\right\}$$

and an **open interval** is a set of the form

$$(a, b) = \left\{x \in S \middle| a < x < b\right\}$$

The **interior** I° of an interval I is the largest open interval contained in I and so $[a, b]^\circ = (a, b)$. We write $\mathrm{Int}(S)$ for the poset of all *closed* intervals in S, ordered by set inclusion, that is,

$$\mathrm{Int}(S) = \left\{[a, b] \middle| a, b \in S\right\}$$

Whenever readability is not affected, we will write sequences as words $a_1 a_2 \cdots a_n$, rather than in the more traditional notation $a_1, a_2, \ldots, a_n$.

2 Dyck Words

The German mathematician Walther Franz Anton von Dyck (1856–1934) studied words in $\mathcal{W}_{a,b}$ for $a = b$ with the property that the A-count is at all times greater than or equal to the B-count, that is, for which

$$\Delta_{A,B}\left([w]_k\right) \geq 0$$

for all k. Those words have since become known as **Dyck words**.

Incidentally, von Dyck has the distinction of being the first person to define (in 1882) the notion of a mathematical group in the modern *fully axiomatic* sense, by making explicit the notion of inverse. Prior to von Dyck's work on groups, both Cauchy and Cayley had contributed to the axiomatic definition of a group, but had not specifically axiomatized the notion of an inverse. (To Galois, a group was simply a table of ordered arrangements.)

Since we do not want to require initially that $a = b$ and for other reasons as well, we make the following definition, which is not standard in the literature.

Definition 2.1

1) A **strong Dyck word** $w \in \mathcal{W}_{a,b}$ is a word whose A-count is always greater than its B-count, that is,

$$\Delta_{A,B}\left([w]_k\right) > 0$$

for all k. We will assume that $a > b$, lest the set be empty.

2) A **weak Dyck word** $w \in \mathcal{W}_{a,b}$ is a word whose A-count is always greater than or equal to its B-count, that is,

$$\Delta_{A,B}\left([w]_k\right) \geq 0$$

for all k. We will assume that $a \geq b$. □

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1_2

Now it should be clear why we order the alphabet and define a dominant letter—the phrase "strong Dyck word over $\{X, Y\}$" must give the same result as the phrase "strong Dyck word over $\{Y, X\}$" and so we do not know which letter is intended to be more copious than the other.

To count the number of *strong* Dyck words, as often happens, it is easier to count what we *do not* want and then subtract that from the total. Now, since all Dyck words start with an A, we take our universe to be the set $\mathcal{W}_{a,b}(A)$ of all words in $\mathcal{W}_{a,b}$ that start with an A. This particular universe has size

$$|\mathcal{W}_{a,b}(A)| = \binom{a+b-1}{b}$$

The set that we do not want is the set T of all words in $w \in \mathcal{W}_{a,b}(A)$ for which

$$\Delta_{A,B}\left([w]_k\right) \le 0$$

for *some* k. But since $\Delta_{A,B}(w) = a - b > 0$ and since $\Delta_{A,B}([w]_k)$ changes by only one as k increases, there must be a value k for which

$$\Delta_{A,B}\left([w]_k\right) = 0$$

that is, there must be a *tie* in the letter counts at some point. Thus, T is the set of all words in $\mathcal{W}_{a,b}(A)$ that have a tie in the letter counts at some point.

Let $w \in \mathcal{W}_{a,b}(A)$ and let $k = 2m$ be the position of the *last* tie in w, where $2 \le k < a + b$. Then w has the form

$$w = [w]_k \alpha$$

where $\alpha \neq \epsilon$, $N_A(\alpha) = a - m$, and $N_B(\alpha) = b - m$. If we replace α by its complement to get

$$w_1 = [w]_k \alpha'$$

then

$$N_A(w_1) = m + (b - m) = b \quad \text{and} \quad N_B(w_1) = m + (a - m) = a$$

and so $w_1 \in \mathcal{W}_{b,a}(A)$. Clearly, the map $\theta : T \mapsto \mathcal{W}_{b,a}(A)$ sending w to w_1 is injective.

To see that θ is surjective, note that if $u \in \mathcal{W}_{b,a}(A)$, then since u starts with an A but has more Bs than As, it must also have a last tie, say at position $k = 2m$, in which case

$$u = [u]_k \alpha$$

and so θ sends $[u]_k \alpha'$ to u. Thus, θ is a bijection from T to $\mathcal{W}_{b,a}(A)$ and so the number of strong Dyck words is

$$
\begin{aligned}
|\mathcal{W}_{a,b}(A)| - |T| &= |\mathcal{W}_{a,b}(A)| - |\mathcal{W}_{b,a}(A)| \\
&= \binom{a+b-1}{b} - \binom{a+b-1}{a} = \frac{a-b}{a+b}\binom{a+b}{a}
\end{aligned}
\tag{2.1}
$$

To find the number of weak Dyck words, we simply observe that $w \in \mathcal{W}_{a,b}$ is a weak Dyck word if and only if $Aw \in \mathcal{W}_{a+1,b}$ is a strong Dyck word. Therefore, since the prefix map $w \mapsto Aw$ is a bijection, Equation [2.1] with a replaced by $a+1$ gives

$$
\frac{a+1-b}{a+1+b}\binom{a+1+b}{a+1} = \frac{a+1-b}{a+1}\binom{a+b}{a}
$$

Theorem 2.1

1) *The number of strong Dyck words with a* A s *and b* B s *is*

$$
D^{st}_{a,b} = \frac{a-b}{a+b}\binom{a+b}{a}
$$

2) *The number of weak Dyck words with a* A s *and b* B s *is*

$$
D^{wk}_{a,b} = \frac{a+1-b}{a+1}\binom{a+b}{a} \qquad \square
$$

Bertrand's Ballot Problem

Dyck words (which generalize Catalan words) can be applied in a variety of situations but since our main interest is in Catalan numbers, we will be brief. **Bertrand's ballot problem** (from 1887) can be described as follows. Consider an election with two candidates A and B. There are n voters in the election and their votes are tallied one at a time. Hence, a **voting history** is just a word w of length n over the alphabet $\{A, B\}$. For instance, the voting history

$$
w = ABAABBBA
$$

means Voter 1 voted for Candidate A, Voter 2 voted for Candidate B, and so on.

Bertrand's ballot problem is to determine the *probability* (assuming that all voting histories are equally likely) that Candidate A is always ahead of Candidate B in the election. A voting history in which this happens is precisely a strong Dyck word. We can also inquire about the probability that Candidate A is never behind Candidate B. These possibilities correlate bijectively to weak Dyck words.

Thus, the answer to Bertrand's ballot problem is an immediate consequence of Theorem 2.1.

Corollary 2.1
Consider an election with two candidates A and B where A receives *a votes and B receives b* votes, *with* $a \geq b$. *Assume that all voting histories are equally likely.*

1) *The probability that Candidate A is always ahead of Candidate B is*

$$\frac{a-b}{a+b}$$

2) *The probability that Candidate A is never behind Candidate B is*

$$\frac{a+1-b}{a+1}$$

□

Counting Paths

Dyck words can be characterized as certain types of paths in the plane. Here are two examples.

Monotonic Paths

Figure 2.1 shows an 8×6 grid. A path in this grid consists of a sequence of edges starting at the lower left corner (the **origin**) that move up (U) or to the right (R) one square at a time, ending at the upper right corner.

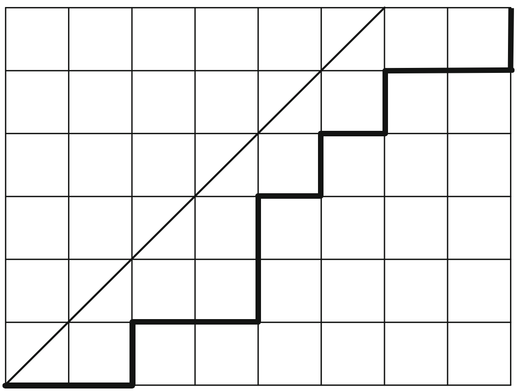


Figure 2.1 A path that does not intersect the diagonal

A path can therefore be described as a word over the alphabet (R, U). For example, the path in Figure 2.1 is

$$w = RRU\ RRUU\ RU\ RU\ RRU$$

Suppose the grid has size $n \times k$ where $n \geq k$. Then a path that does not cross (although may intersect) the diagonal line emanating from the origin corresponds to a weak Dyck word with n Rs and k Us and there are $D^{wk}_{n,k}$ such words. A path that does not intersect the diagonal corresponds to a strong Dyck word with n Rs and k Us and there are $D^{st}_{n,k}$ such words.

Dyck Paths

Consider the path shown in Figure 2.2.

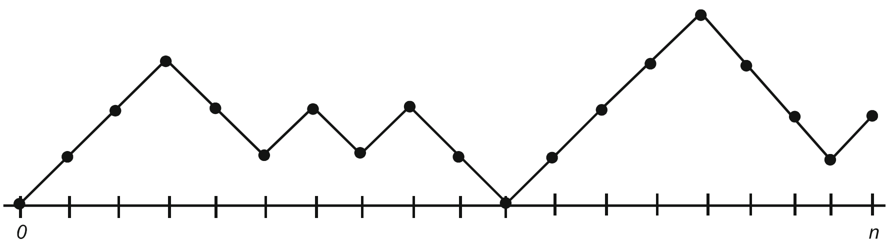


Figure 2.2 A Dyck path

This path starts at the origin $(0, 0)$ and moves either one unit up (U) diagonally or one unit down (D) diagonally for each tick mark on the x-axis. The paths that never drop below the x-axis correspond bijectively to weak Dyck words over the alphabet (U, D). These paths (especially those that end on the x-axis) are called **Dyck paths**. Hence, if there are $n + 1$ tick marks, the Dyck words have length n and if the ending point is (n, k), then there are k more ups than downs and so there are a total of $D^{wk}_{(n+k)/2,\,(n-k)/2}$ Dyck paths.

The paths that never touch the x-axis (except at the origin) correspond bijectively to strong Dyck words over (U, D) and so there are a total of $D^{st}_{(n+k)/2,\,(n-k)/2}$ of these so-called **elevated Dyck paths**.

The Catalan Numbers

3

The most important special case of Dyck words comes when $a = b$ so that the words have the same number of dominant and nondominant letters. A slight variation on these Dyck words is *ballot sequences*, which we also define here for completeness.

Definition 3.1

a) We will refer to a weak Dyck word in $\mathcal{W}_{n,n}$ as a **Catalan word** (no confondre amb paraules en l'idioma català). For $n = 3$, the five Catalan words over (A, B) are

$$AAABBB, \quad AABABB, \quad AABBAB, \quad ABAABB, \quad ABABAB$$

Let $\mathcal{C}_n(A, B)$ denote the set of Catalan words over (A, B) of length $2n$. The numbers

$$C_n = |\mathcal{C}_n(A, B)|$$

are called the **Catalan numbers** (although there are many other equivalent definitions of these numbers). Note that $C_0 = 1$, since the empty word is a Catalan word.

b) A **ballot sequence** is a sequence of n ones and n negative ones such that every partial sum is nonnegative. For $n = 3$, writing a plus sign for $+1$ and a minus sign for -1, the ballot sequences are

$$+++---, \quad ++-+--, \quad ++--+-, \quad +-++--, \quad +-+-+-$$

(This notation suggests that a more appropriate term for these sequences might be **charge sequence**: a sequence of positive and negative electric charges for which the total charge up to any point in the sequence is nonnegative.) □

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1_3

Catalan numbers were named by the mathematician John Riordan in 1948 after the Belgian mathematician **Eugène Charles Catalan** (1814–1894), who worked in a variety of areas of mathematics, including continued fractions, number theory, and combinatorics. Catalan showed (in 1838) that the Catalan number

$$C_n = \frac{(2n)!}{n!(n+1)!}$$

counts the number of ways to fully parenthesize a string of $n + 1$ letters. We will revisit this application later in the book.

Actually, the Catalan numbers were known to the Mongolian mathematician **Minggantu** (1692–1763). He used these numbers to express $sin(2x)$ and $sin(4x)$ in terms of $sin(x)$, to wit,

$$sin\,(2x) = 2\,sin\,(x) - \sum_{k=1}^{\infty} \frac{C_{k-1}}{4^{k-1}} \sin^{2k+1}(x)$$

As an aside, Eugène Catalan is also known for his **Catalan conjecture**, namely that $2^3 = 8$ and $3^2 = 9$ are the only pairs of powers of positive integers that differed by 1, that is, the equation $x^j - y^k = 1$ has only one solution in positive integers. This conjecture lay unproven for 58 years, until it was finally justified by Preda Mihailescu, using some very sophisticated tools from number theory!

Our discussion of Dyck words immediately yields the following.

Corollary 3.1 *The number of Catalan words of length* $2n$ *is*

$$C_n = \frac{1}{n+1}\binom{2n}{n} \qquad \square$$

The first few Catalan numbers (starting with C_0) are

1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, 208012, 742900, 2674440, 9694845

and as you can see, the Catalan numbers grow large *very* quickly.

Basic Properties of the Catalan Numbers

Before looking at what the Catalan numbers can count, we wish to establish some properties of these numbers. The following formulas are a more or less direct consequence of Corollary 3.1.

Theorem 3.1 *For the Catalan numbers C_n, we have the following*:

1) $C_n = \binom{2n}{n} - \binom{2n}{n-1}$
2) $C_n = \frac{1}{n+1}\sum_{k=0}^{n}\binom{n}{k}^2$
3) *Nonlinear recurrence relation*

$$C_{n+1} = \frac{2(2n+1)}{n+2}C_n, \quad C_0 = 1$$

4) *Asymptotic expression*

$$C_n \Big/ \left(\frac{4^n}{\sqrt{\pi}n^{3/2}}\right) \to 1 \quad as \quad n \to \infty$$

Proof. For 2), we know that

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

and the Vandermonde convolution formula gives

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \frac{1}{n+1}\sum_{k=0}^{n}\binom{n}{k}^2$$

For 4), Stirling's approximation to $n!$ is

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

and so

$$C_n = \frac{1}{n+1}\frac{(2n)!}{n!n!} \sim \frac{1}{n+1}\frac{2\sqrt{\pi n}\left(\frac{2n}{e}\right)^{2n}}{2\pi n\left(\frac{n}{e}\right)^{2n}} = \frac{1}{n+1}\frac{1}{\sqrt{\pi n}}4^n \sim \frac{1}{n}\frac{1}{\sqrt{\pi n}}4^n = \frac{4^n}{\sqrt{\pi}n^{3/2}} \quad \square$$

Integral Representation

The Catalan numbers also have an integral representation. A useful approach when a recurrence formula such as

$$C_{n+1} = \frac{2(2n+1)}{n+2} C_n, \quad C_0 = 1$$

is available is to start with an integral of the form

$$I_n = \int_{\ell}^{u} x^{n+\alpha}(a+bx)^{\beta} dx$$

for $n \geq 0$ and integrate by parts for $n \geq 1$. Let

$$du = (a+bx)^{\beta} dx \quad \text{and} \quad v = x^{n+\alpha}$$

Then

$$u = \frac{1}{b(\beta+1)}(a+bx)^{\beta+1} \quad \text{and} \quad dv = (n+\alpha)x^{n-1+\alpha}$$

Hence,

$$I_n = \frac{1}{b(\beta+1)} x^{n+\alpha}(a+bx)^{\beta+1}\Big|_{\ell}^{u} - \frac{n+\alpha}{b(\beta+1)} \int_{\ell}^{u} x^{n-1+\alpha}(a+bx)^{\beta+1} dx$$

At this point, we set the limits of integration so that the first term on the right is zero,

$$\ell = 0 \quad \text{and} \quad u = -\frac{a}{b}$$

The integral on the right can be expanded

$$\begin{aligned} \int x^{n-1+\alpha}(a+bx)^{\beta+1} dx &= a\int x^{n-1+\alpha}(a+bx)^{\beta} dx + b\int x^{n+\alpha}(a+bx)^{\beta} dx \\ &= aI_{n-1} + bI_n \end{aligned}$$

and so we get

$$I_n = -\frac{n+\alpha}{b(\beta+1)}(aI_{n-1} + bI_n)$$

or finally,

$$I_n = -\frac{a(n+\alpha)}{b(\beta+1+n+\alpha)} I_{n-1}$$

for $n \geq 1$. Now we need to select the parameters so that the desired recurrence and initial conditions

$$C_n = \frac{2(2n+1)}{n+2} C_n, \quad C_0 = 1$$

are obtained. Taking the naive view, we set

$$a(n+\alpha) = 2(2n-1) \quad \text{and} \quad -b(\beta + 1 + n + \alpha) = n + 1$$

Thus,

$$a = 4, \ \ \alpha = -1/2, \ \ b = -1, \ \ \beta = 1/2$$

and so

$$I_n = \frac{2(2n-1)}{(n+1)} I_{n-1}$$

for $n \geq 1$. To deal with the initial conditions, we multiply I_n by an appropriate constant. Since

$$I_0 = \int_0^4 x^{-1/2}(4-x)^{1/2} dx = 2\pi$$

the appropriate constant is $1/2\pi$, giving the following integral representation.

Theorem 3.2 *The Catalan numbers have the integral representation*

$$C_n = \frac{1}{2\pi} \int_0^4 x^{n-1/2}(4-x)^{1/2} dx$$

for $n \geq 0$. □

Recurrence Relation and Generating Function

The key to finding a binary decomposition model for Catalan words is that all Catalan words must have a tie at some point, even if that point comes only at the very end of the word. To save ink, let us write $\mathcal{C}_n$ for $\mathcal{C}_n(A, B)$. If the *first* tie occurring in a Catalan word $w \in \mathcal{C}_n$ is at position $2k$, then we can write

$$w = [w]_{2k} v$$

where $v \in \mathcal{C}_{n-k}$, since we are in a sense starting fresh at position $2k+1$. Moreover, $[w]_{2k}$ must start with an A and end with a B and so

$$[w]_{2k} = AxB$$

where $x \in \mathcal{C}_{k-1}$. Thus, we conclude that all Catalan words have the form

$$w = AxBv$$

where $x \in \mathcal{C}_{k-1}$ and $v \in \mathcal{C}_{n-k}$. The two letters A and B above provide the nexus for a decomposition of w into the pair $(x, v) \in \mathcal{C}_{k-1} \times \mathcal{C}_{n-k}$. Note that we lose both A and B in the decomposition. Specifically, if $\mathcal{C}_{n,k}$ is the family of Catalan words whose first tie is at position k, then we define a map

$$\theta_{n,k} : \mathcal{C}_{n,k} \to \mathcal{C}_{k-1} \times \mathcal{C}_{n-k}$$

by

$$\theta_{n,k}(AxBv) = (x, v)$$

This map is injective since the decomposition is clearly reversible and surjective since *every* pair $(x, v) \in \mathcal{C}_{k-1} \times \mathcal{C}_{n-k}$ can be recomposed into a Catalan word $AxBv$ in $\mathcal{C}_n$.

Hence, the associated recurrence relation is

$$C_n = \sum_{k=1}^{n} C_{k-1}C_{n-k} = \sum_{k=0}^{n-1} C_k C_{n-1-k}$$

with $C_0 = 1$, which is the Catalan recurrence. In other words, the Catalan numbers satisfy the Catalan recurrence relation.

Generating Function

The right side of the Catalan recurrence relation might look familiar: It is the coefficient of x^{n-1} in the product

$$\left(\sum_{k=0}^{\infty} C_k x^k\right)^2$$

So if the generating function of the Catalan numbers is

$$C(x) = \sum_{k=0}^{\infty} C_k x^k$$

then for all $n \geq 0$,

$$\operatorname{coef}_{x^{n-1}}\left(C^2(x)\right) = C_n$$

or equivalently,

$$\operatorname{coef}_{x^n}\left(xC^2(x)\right) = C_n$$

for $n \geq 1$. For $n = 0$, the left side is 0 and the right side is 1 and so

$$xC^2(x) = C(x) - 1$$

This quadratic can be formally solved for $C(x)$ to get

$$C(x) = \frac{1 \pm \sqrt{1-4x}}{2x}$$

Expanding $\sqrt{1-4x}$ in a Taylor series shows that we must take the negative sign in order to get positive terms in the expansion and so we arrive at the generating function for the Catalan numbers.

Theorem 3.3

1) *The Catalan numbers C_n satisfy the Catalan recurrence relation*

$$C_n = \sum_{k=0}^{n-1} C_k C_{n-k-1}$$

for $n \geq 1$, with $C_0 = 1$.

2) *The generating function of the Catalan numbers is*

$$C(x) = \sum_{k=0}^{\infty} C_k x^k = \frac{1 - \sqrt{1-4x}}{2x} \qquad \square$$

We will frequently encounter a **shifted Catalan recurrence relation** of the form

$$D_n = \sum_{k=0}^{n-1-a} D_{k+a} D_{n-k-1}, \quad D_a = 1$$

for $n \geq a+1$. Setting $F_{n-a} = D_n$ gives $F_0 = D_a = 1$ and

$$F_{n-a} = \sum_{k=0}^{n-1-a} F_k F_{n-k-1-a}$$

Then replacing n by $n + a$ gives the Catalan recurrence

$$F_n = \sum_{k=0}^{n-1} F_k F_{n-k-1}$$

for $n \geq 1$, whence $D_{n+a} = F_n = C_n$.

Theorem 3.4 *If a sequence D_n of integers satisfies the recurrence relation*

$$D_n = \sum_{k=0}^{n-1-a} D_{k+a} D_{n-k-1}, \quad D_a = 1$$

for $n \geq a + 1$, then

$$D_n = C_{n-a}$$

for all $n \geq a$. □

Catalan Numbers and Paths 4

We begin our examination of the types of objects that are counted by the Catalan numbers with paths, because the work has already been done.

Monotone Paths

We have seen that a Catalan word of length $2n$ over (U, R) can be described as a monotonic path in an $n \times n$ grid that does not rise above the diagonal, as shown in Figure 4.1.

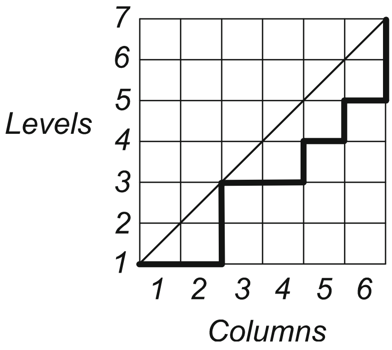


Figure 4.1 A monotonic path

Theorem 4.1
C_n counts the number of monotonic paths in an $n \times n$ grid that do not rise above the diagonal. □

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1_4

Dyck Paths

We have also seen that a Catalan word over (U, D) of length $2n$ can be characterized as a Dyck path of length $2n$ that ends on the horizontal axis, as shown in Figure 4.2. These are usually referred to simply as **Dyck paths**.

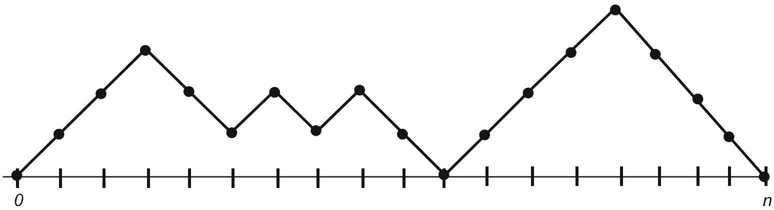


Figure 4.2 A Dyck path ending on the horizontal axis

Theorem 4.2
C_n counts the number of Dyck paths of length $2n$ that end on the horizontal axis. □

Path Summary

Theorem 4.3
C_n counts the number of

1) *monotonic paths in an $n \times n$ grid that do not rise above the diagonal,*
2) *Dyck paths of length $2n$ that end on the horizontal axis.* □

5 Catalan Numbers and Trees

The Catalan numbers excel at counting a variety of types of trees. (The Appendix of this book contains a brief introduction to the subject of trees for those who are interested.)

Ordered Trees

Let $\mathcal{R}_n$ be the family of all ordered trees with n vertices. If $n \geq 2$ and if $T \in \mathcal{R}_n$ is an ordered tree whose root has degree $d > 0$, then each of the d edges incident with the root will also be incident with a distinct subtree of the root (perhaps consisting of a single vertex only). Figure 5.1 shows a tree whose root has degree 3.

Since $d > 0$, the root vertex will have a leftmost subtree T_ℓ and we can use the edge connecting the root with that subtree as the nexus with which to decompose T into two disjoint subtrees T_ℓ and T_r, as shown in Figure 5.1. Here T_r is the *complement* of T_ℓ, excluding the nexus edge but including the original root. However, in order that the left subtree be of the same type as the original tree, we need to specify its root, which we take to be the vertex incident with the nexus.

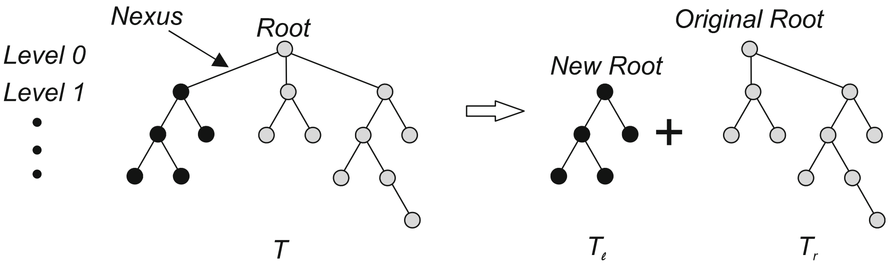


Figure 5.1 A decomposition of a rooted tree

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1_5

Thus, if $\mathcal{R}_{n,k}$ is the family of all members of $\mathcal{R}_n$ whose left subtree has size k, then we can define a map

$$\theta_{n,k} : \mathcal{R}_{n,k} \to \mathcal{R}_k \times \mathcal{R}_{n-k}$$

by

$$\theta_{n,k}(T) = (T_\ell, T_r)$$

where T_ℓ has size k, and T_r has size $n-k$ and $1 \le k \le n-1$ (for $n \ge 2$). It is clear that $\theta_{n,k}$ is a bijection. Specifically, $\theta_{n,k}$ is injective because the decomposition is clearly reversible. Also, $\theta_{n,k}$ is surjective because we can recombine *any* two ordered trees $T_1 \in \mathcal{R}_k$ and $T_2 \in \mathcal{R}_{n-k}$ by reversing the decomposition process, specifically, by placing the root of T_1 at level 1 and the root of T_2 at level 0 and then connecting the two roots with a new edge, declaring the root of T_2 to be the root of the new tree.

Thus, if D_n is the number of ordered trees on n vertices, then since $D_1 = 1$, the decomposition model yields the recurrence

$$D_n = \sum_{k=1}^{n-1} D_k D_{n-k} = \sum_{k=0}^{n-2} D_{k+1} D_{n-k-1}$$

for $n \ge 2$, which is a shifted Catalan recurrence (Theorem 3.4 with $a = 1$) and so

$$D_{n+1} = C_n$$

for all $n \ge 0$.

Theorem 5.1
C_n counts the number of ordered trees with $n+1$ vertices. □

Binary Trees

Let $\mathcal{B}_n$ be the family of binary trees with n vertices for $n \ge 1$. If $T \in \mathcal{B}_n$ then we can use the root vertex together with its incident edges as our nexus, decomposing T into a (possibly empty) left subtree T_ℓ with k vertices and a (possibly empty) right subtree T_r with $n-k-1$ vertices, thereby losing the original root, as shown in Figure 5.2. The roots for the subtrees are the vertices adjacent to the original root.

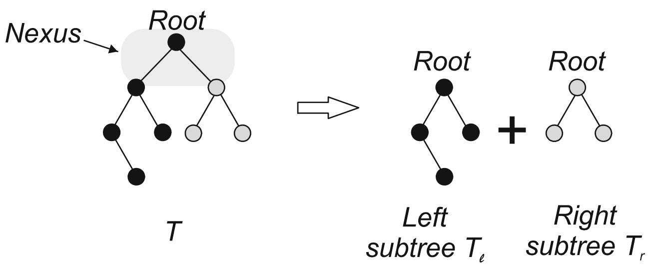


Figure 5.2 Decomposition of a binary tree

If $\mathcal{B}_{n,k}$ is the family of all members of $\mathcal{B}_n$ that have a left subtree of size k, for $0 \le k \le n-1$, then this decomposition into a left and right subtrees defines a family of bijections

$$\theta_{n,k} : \mathcal{B}_{n,k} \longrightarrow \mathcal{B}_k \times \mathcal{B}_{n-k-1}$$

by

$$\theta_{n,k}(T) = (T_\ell, T_r)$$

Hence, the number D_n of binary trees with n vertices satisfies the recurrence

$$D_n = \sum_{k=0}^{n-1} D_k D_{n-k-1}$$

for $n \ge 1$. Since $D_0 = 1$, this is the Catalan recurrence and so $D_n = C_n$ for all $n \ge 0$.

By the way, lest it should seem that we are cheating by arbitrarily taking $D_0 = 1$ simply to make things work out in a Catalan fashion, note that $D_0 = 1$ is *required* for the count. For example, if there is no right subtree, we still want to count the decomposition into a nonempty left subtree and an empty right subtree and so D_0 *must* equal 1.

Theorem 5.2
C_n counts the number of binary trees with n vertices. □

Full Binary Trees

A full binary tree T consists of a root vertex together with both a left and a right subtree, both empty or both nonempty. Note that T must have an odd number of vertices, a fact easily proved by induction. Let $\mathcal{F}_n$ be the family of full binary trees with $2n+1$ vertices (for $n \ge 0$) and let $\mathcal{F}_{n,k}$ be the members of $\mathcal{F}_n$ whose left subtree has size $2k+1$, for $0 \le k \le n-1$ (and $n \ge 1$).

If $T \in \mathcal{F}_n$, then we can again use the root vertex and its incident edges as our nexus, decomposing T into a full left subtree T_ℓ with $2k+1$ vertices and a full right subtree T_r with

$$(2n+1) - (2k+1) - 1 = 2n - 2k - 1$$

vertices, thereby losing the original root vertex, as shown in Figure 5.3.

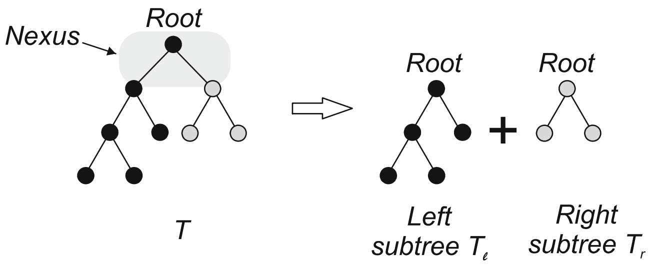


Figure 5.3 Decomposition of a full binary tree

This decomposition defines a bijection

$$\theta_{n,k} : \mathcal{F}_{n,k} \longrightarrow \mathcal{F}_k \times \mathcal{F}_{n-k-1}$$

by

$$\theta_{n,k}(T) = (T_\ell, T_r)$$

for $0 \le k \le n-1$ and so the number D_n of full binary trees with $2n+1$ vertices satisfies the recurrence

$$D_n = \sum_{k=0}^{n-1} D_k D_{n-k-1}, \quad D_0 = 1$$

for $n \ge 1$, which is the Catalan recurrence.

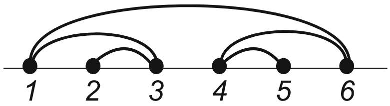
Theorem 5.3
C_n counts the number of full binary trees with $2n+1$ vertices. □

Noncrossing, Alternating Trees

Here is another, somewhat more unusual example of counting trees. These trees can be characterized in other interesting ways, as we will see in the sequel. As shown in Figure 5.4, suppose we draw n points on a line, label them $1, \ldots, n$, and connect two integers i and j by an edge that does not drop below the line and with the following properties.

1) **Noncrossing property**: No two edges intersect except possibly at vertices of the graph.
2) **Alternating property**: At each vertex, the edges all exit in one direction (all exit to the left or all exit to the right).

Figure 5.4 A noncrossing alternating tree



Let $\mathcal{G}$ be the resulting graph. Figure 5.5 shows the five noncrossing, alternating graphs with four vertices.

Figure 5.5 The five noncrossing, alternating trees with four vertices

Now, it is clear that $\mathcal{G}$ is acyclic, since any vertex in a cycle would have edges exiting in both directions. Therefore, the maximum number of edges in $\mathcal{G}$ is $n-1$. Also, $\mathcal{G}$ is connected and is therefore a tree if and only if it has precisely $n-1$ edges.

Let us assume that $\mathcal{G}$ is a tree. Such trees are called **noncrossing, alternating trees**. The term "alternating" comes from the fact that if we follow any path in the tree, the vertex labels alternate in relative size: larger, smaller, larger, smaller, ... (or smaller, larger, smaller, larger, ...).

To save trees, we will refer to noncrossing, alternating trees as **NA-trees**. We wish to determine the size of the family $\mathcal{T}_n$ of all NA-trees with n vertices.

Note first that any $T \in \mathcal{T}_n$ must contain the edge $\{1, n\}$. For if not, then we may suppose that the largest vertex adjacent to vertex 1 is vertex $k < n$. But then the noncrossing property implies that vertices $1, \ldots, k$ are isolated from vertices $k+1, \ldots, n$, since any edge $\{i, j\}$ with $i \le k$ and $j \ge k+1$ would intersect the edge $\{1, k\}$ at a nonvertex. This implies that the graph is not connected, a definite falsehood.

If we remove edge $\{1, n\}$ from the NA-tree T, the resulting graph consists of exactly two connected components. (This is true for any tree.) Let T_1 be the component that contains the vertex 1 and let T_n be the component that contains the vertex n. Since the properties that define an NA-tree pass to subtrees, both T_1 and T_n are NA-trees and so $\{1, k\}$ is an edge in T_1 and $\{k+1, n\}$ is an edge in T_n. Hence, the vertex set for T_1 is $[1, k]$ and the vertex set for T_n is $[k+1, n]$.

Thus, if $\mathcal{T}_{n,k}$ is the family of all NA-trees of size n with largest vertex k, then we have a family of evidently injective maps

$$\theta_{n,k} : \mathcal{T}_{n,k} \to \mathcal{T}_k \times \mathcal{T}_{n-k}$$

defined by

$$\theta_{n,k}(T) = (T_1, T_n)$$

To see that $\theta_{n,k}$ is also surjective, observe that if T_1 and T_n are any NA-trees with vertex set $[1,k]$ and $[k+1,n]$, respectively, then we can draw them as described above and stitch them together by adding the edge $\{1,n\}$ to get an NA-tree with n vertices.

Hence, if D_n denotes the number of NA-tree on n vertices, we have

$$D_n = \sum_{k=1}^{n-1} D_k D_{n-k} = \sum_{k=0}^{n-2} D_{k+1} D_{n-k-1}$$

which is a shifted Catalan recurrence (Theorem 3.4 with $a = 1$) and so $D_{n+1} = C_n$.

Theorem 5.4
C_n counts the number of noncrossing, alternating trees with $n+1$ vertices. □

Tree Summary

Theorem 5.5
C_n counts the number of

1) *ordered trees with $n+1$ vertices,*
2) *binary trees with n vertices,*
3) *full binary trees with $2n+1$ vertices,*
4) *noncrossing, alternating trees with $n+1$ vertices.* □

Catalan Numbers and Geometric Widgits

6

Catalan numbers can also count a variety of geometric objects.

Nonintersecting Chords

Consider the vertices of a convex $2n$-gon, as shown on the left in Figure 6.1. There are many ways to draw chords connecting pairs of vertices in such a way that all vertices are incident with a chord and that no two chords intersect (even at the vertices). We call this a **nonintersecting chording** of the polygon.

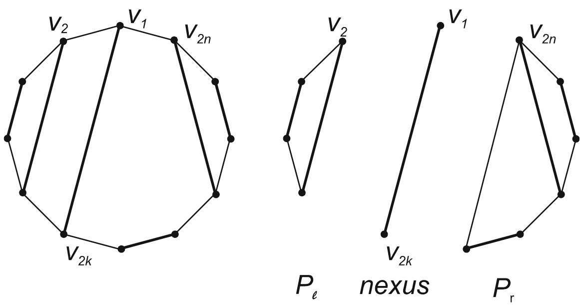


Figure 6.1 Chording a convex polygon

Figure 6.2 shows the five possible nonintersecting chordings of a hexagon $(n = 3)$.

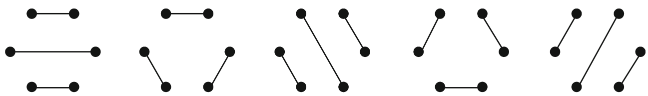


Figure 6.2 Nonintersecting chordings of a hexagon

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1_6

To count the number D_n of ways to chord a convex $2n$-gon P, we fix a **root vertex** and label it v_1. Then we label the remaining vertices in counterclockwise order v_2 through v_{2n}.

In any nonintersecting chording, the root vertex v_1 is incident with a chord that we call the **root chord**. The root chord must be incident with a vertex v_{2k} of even index, so that there are an even number of vertices on each side of the root chord. Let $\mathcal{P}_n$ be the family of all rooted chorded $2n$-gons and let $\mathcal{P}_{n,k}$ be the members of $\mathcal{P}_n$ whose root vertex is adjacent to the vertex v_{2k}, for $1 \le k \le n$.

If $P \in \mathcal{P}_{n,k}$, then the root chord and its incident vertices define the nexus for a decomposition of P into two smaller convex polygons P_ℓ of size $2k-2$ and P_r of size $2n-2k$, as shown in Figure 6.1. Note that if the root vertex is adjacent to either v_2 or v_{2n}, then either P_ℓ or P_r will be empty. In any case, if P_ℓ or P_r is nonempty, then it is properly chorded and we declare the root of P_ℓ to be the vertex v_2 and the root of P_r to be the vertex v_{2n}.

Thus, for each $1 \le k \le n$, we have an injective map

$$\theta_{n,k} : \mathcal{P}_{n,k} \to \mathcal{P}_{k-1} \times \mathcal{P}_{n-k}$$

defined by

$$\theta_{n,k}(P) = (P_\ell, P_r)$$

The map $\theta_{n,k}$ is also surjective, since any two chorded rooted polygons can be recombined by the addition of a nexus chord (and concomitant vertices) to produce a new larger rooted chorded polygon. Note that we must connect the roots of P_1 and P_2 to the new root of the larger polygon.

Thus, if $D_n = |\mathcal{P}_n|$, then the associated recurrence relation is precisely the Catalan recurrence

$$D_n = \sum_{k=1}^{n} D_{k-1} D_{n-k} = \sum_{k=0}^{n-1} D_k D_{n-1-k}$$

and since $D_0 = 1$, we have $D_n = C_n$ for all $n \ge 0$.

Theorem 6.1
C_n counts the number of ways to chord a convex $2n$-gon with nonintersecting chords. $\square$

For example, since $C_6 = 132$, there are 131 other ways to partition the dodecagon in Figure 6.1. Just in case you think this sort of problem is frivolous, consider that we have just proved that if you are sitting at a round table with 11 mutual strangers, there are *exactly* 132 ways in which every stranger can shake hands with another stranger *at the same time* without interference from the arms of the others seated at the table! Enough said.

Tilings of a Staircase

The far left portion of Figure 6.3 shows an **inverted staircase**. This staircase has size 3×3 because it fits over a 3×3 grid of equal-sized squares.

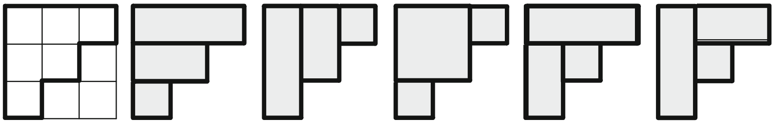


Figure 6.3 Tiling with n tiles

There are several ways in which an inverted staircase of size $n \times n$ can be tiled with exactly n rectangular tiles. The five ways that this can be done for the case $n = 3$ are also shown in Figure 6.3. Let us count the ways.

Note first that each tile can contain at most one diagonal square and since there are n diagonal squares and n tiles, it follows that each tile must contain *exactly one* diagonal square.

The upper left tile, that is, the tile that contains the upper left corner of the staircase, can be used as a nexus to decompose the tiling into two smaller tilings as follows. With reference to Figure 6.4, if the upper left tile has size $r \times c$ (r rows and c columns of squares), then it is easy to see that

$$c = n + 1 - r$$

and so we can use r to uniquely indentify the upper left tile, where $1 \leq r \leq n$.

Let $\mathcal{T}_n$ be the family of tilings of an $n \times n$ inverted staircase and let $\mathcal{T}_{n,r}$ be the members of $\mathcal{T}_n$ for which the upper left tile has size $r \times (n + 1 - r)$. By removing this upper left tile, we decompose the original $n \times n$ tiling T into an $(r - 1) \times (r - 1)$ tiling T_1 and an $(n - r) \times (n - r)$ tiling T_2, as shown in Figure 6.4.

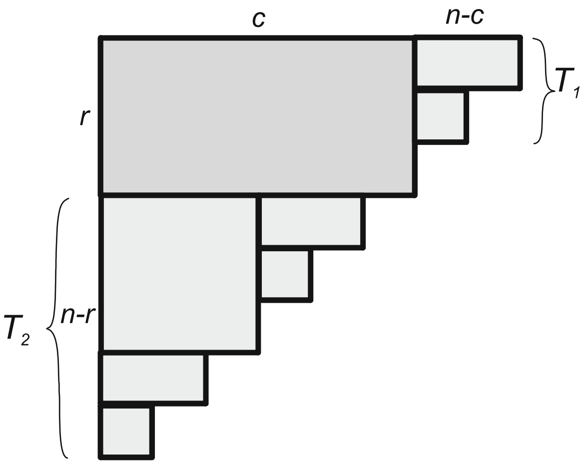


Figure 6.4 The decomposition

If $r = 1$ or $r = n$, then one of these tilings is empty. Now, it is clear that this decomposition is reversible, that is, the map

$$\theta_{n,r} : \mathcal{T}_{n,k} \mapsto \mathcal{T}_{r-1} \times \mathcal{T}_{n-r}$$

defined by

$$\theta_{n,r}(T) = (T_1, T_2)$$

is injective. Moreover, $\theta_{n,r}$ is surjective since any pair of tilings of sizes $(r-1) \times (r-1)$ and $(n-r) \times (n-r)$ can be recombined using a tiling of size $r \times (n+1-r)$. Hence, if D_n is the number of $n \times n$ tilings, then taking $D_0 = 1$, we have

$$D_n = \sum_{r=1}^{n} D_{r-1} D_{n-r} = \sum_{r=0}^{n-1} D_r D_{n-r-1}$$

and so $D_n = C_n$.

Theorem 6.2
C_n counts the number of staircase tilings of an $n \times n$ grid using n tiles. □

Noncrossing, Alternating Chords

If we start with a noncrossing, alternating tree, such as the one on the far left in Figure 6.5 and *carefully* move the vertices as shown in the figure, the result is a chorded convex polygon P whose vertices are the vertices of the NA-tree.

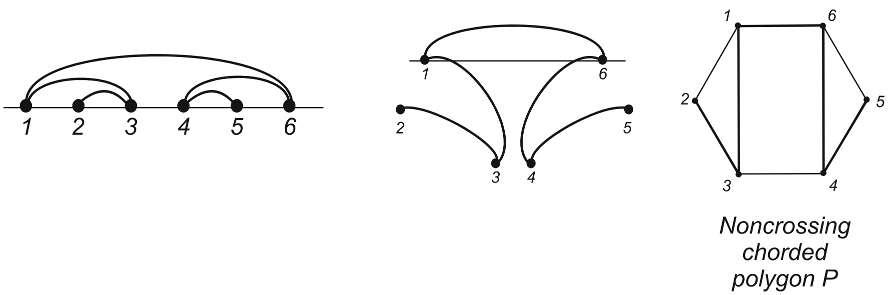


Figure 6.5 Morphing an NA-tree into an NA-chorded polygon

Moreover, the noncrossing and alternating properties carry over to the polygon, that is,

1) **Noncrossing property**: The chords do not intersect except perhaps at a vertex of P.
2) **Alternating property**: No integer in $[n]$ is **adjacent** (that is, connected by a chord) to both a smaller vertex and a larger vertex.

For instance, in the example in Figure 6.5, vertex 2 is adjacent to a larger vertex only and vertex 3 is adjacent to smaller vertices only.

Theorem 6.3
C_n counts the number of chorded convex $(n+1)$-gons with n chords and with the following properties.

1) ***Noncrossing property**: The chords do not intersect except perhaps at a vertex of P.*
2) ***Alternating property**: No integer in $[n+1]$ is adjacent to both a smaller vertex and a larger vertex.* □

Triangulations of a Convex Polygon

A **triangulation** of a convex polygon is a division of the interior of the polygon into triangles using noncrossing chords that connect vertices of the polygon. Figure 6.6 shows the triangulations of a hexagon.

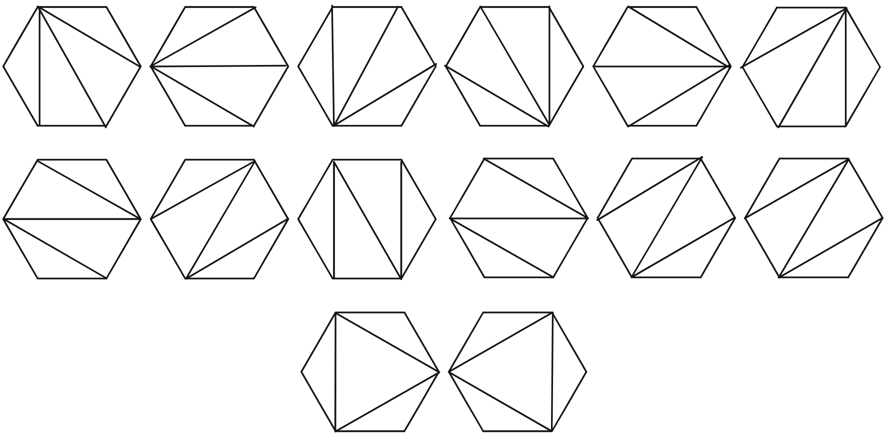


Figure 6.6 Triangulations of a hexagon

In the 18th century, the famous mathematician **Leonhard Euler** took an interest in determining the number of triangulations of a convex n-gon. We do as well.

Note that the first two triangulations in Figure 6.6 are considered distinct even though each can be obtained from the other by a mere rotation. This is because the vertices (and edges) of the polygon are considered distinguishable. To make this plain, referring to the left-hand side of Figure 6.7, let us assume that the vertices are labeled v_1 through v_n in clockwise order. We refer to the edge connecting v_1 and v_n as the **root edge**, with **right root vertex** v_1 and **left root vertex** v_n. The triangle containing the root edge is the **root triangle** and the third vertex of the root triangle is the **opposite vertex**.

The following terminology will also come in handy. As on the left in Figure 6.7, when the vertices are equally spaced around a circle, with the root edge placed horizontally at the top, we say that the polygon is in **center standard position**.

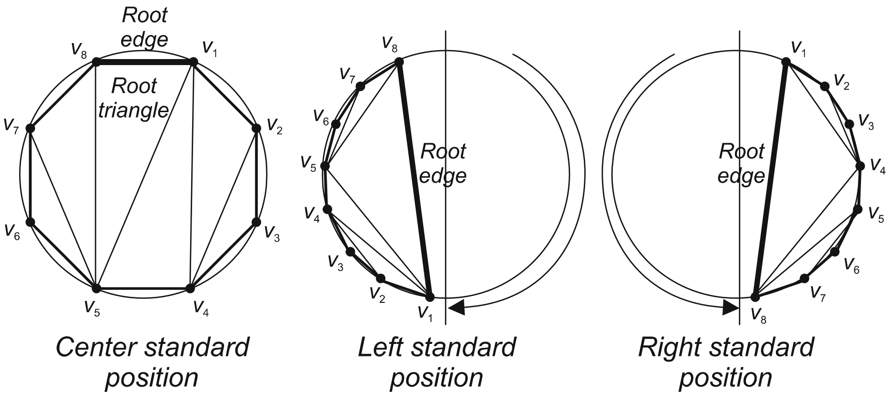


Figure 6.7 Vertex positions

Now, it is clear that we may nudge the vertices of the polygon around the circle without affecting the triangulation, as long as we do not change the order of the vertices. If we cram the vertices into the left half of the circle as shown in the middle of Figure 6.7, with the root edge having a large negative slope, we say that the triangulated polygon is in **left standard position**. Similarly, **right standard position** is shown on the right in Figure 6.7.

Let $\mathcal{P}_n$ be the family of all rooted triangulated n-gons. As shown in Figure 6.8, each $P \in \mathcal{P}_n$ can be decomposed into a rooted triangulated polygon P_ℓ of size k that is in left standard position and a rooted triangulated polygon P_r of size j that is in right standard position as follows.

1) Put the polygon in center standard position.
2) Delete the root edge.
3) Split the opposite vertex v_k into two vertices v_k' and v_k''.
4) Nudge the vertices $v_{n-1}, \ldots, v_{k+1}$ and v_k' so that P_ℓ is in left standard position. Nudge the vertices $v_1, \ldots, v_{k-1}$ and v_k'' so that P_r is in right standard position.
5) The remaining edges of the original root triangle become the roots of P_ℓ and P_r.

Figure 6.8 shows two examples of the decomposition process for a hexagon. The root triangles are shaded.

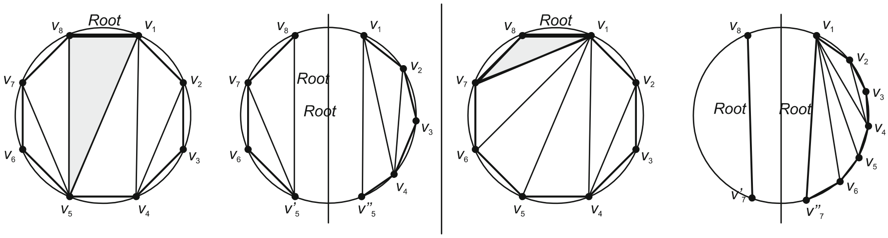


Figure 6.8 Decomposing a triangulation at the root edge

For the two extreme cases where the root triangle has *two* edges on the polygon, as on the right in Figure 6.8, the decomposition is into a line segment (a 1-gon) and an $(n-1)$-gon, with no net change in edge count.

For the other cases, the decomposition results in a k-gon P_ℓ and an $(n-k+1)$-gon P_r, where $3 \le k \le n-2$, for a net *gain* of one edge. Hence, the possibilities for the sizes $\left(|P_\ell|, |P_r|\right)$ of the constituent polygons are

$$(1, n-1), \quad (n-1, 1) \quad \text{and} \quad (k, n-k+1)_{3 \le k \le n-2}$$

If $\mathcal{P}_{n,k}$ is the set of triangulated n-gons for which P_ℓ has size k, for $3 \le k \le n-2$, the map

$$\theta_{n,k} : \mathcal{P}_{n,k} \to \mathcal{P}_k \times \mathcal{P}_{n-k+1}$$

defined by

$$\theta_{n,k}(P) = (P_\ell, P_r)$$

is injective, since the decomposition is reversible.

This map is also surjective. To see this, suppose that P_1 is a rooted triangulated k-gon and P_2 is a rooted triangulated $(n-k+1)$-gon. Place P_1 in left standard position and place P_2 in right standard position on the same circle. Then combine the two polygons by connecting the left root vertex of P_1 to the right root vertex of P_2 and combining the right root vertex of P_ℓ with the left root vertex of P_r.

As to the extreme cases, $k = 1$ and $k = n-1$, it is easy to see that both families $\mathcal{P}_{n,1}$ and $\mathcal{P}_{n,n-1}$ are in bijection with $\mathcal{P}_{n-1}$.

Therefore, if $D_n = |\mathcal{P}_n|$, then setting $D_2 = 1$ and including all of the possibilities for k, we get the recurrence relation

$$D_n = 2D_{n-1} + \sum_{k=3}^{n-2} D_k D_{n-k+1} = \sum_{k=2}^{n-1} D_k D_{n-k+1} = \sum_{k=0}^{n-3} D_{k+2} D_{n-k-1}$$

for $n \ge 3$. Since $D_3 = 1$, this is the recurrence for a shifted Catalan recurrence (Theorem 3.4 with $a = 2$) and so $D_{n+2} = C_n$.

Theorem 6.4
C_n counts the number of triangulations of a convex polygon with $n + 2$ sides. □

Disk Stacking

Sometimes it is easier to find a characterization of one type of object in terms of another type of object whose count we already know than to directly count the original objects. Here is an example.

Figure 6.9 shows one way to stack equal-sized disks in the plane, a task that we often find ourselves wishing to do. Let D_n be the number of possible disk stackings, where the bottom row has n disks.

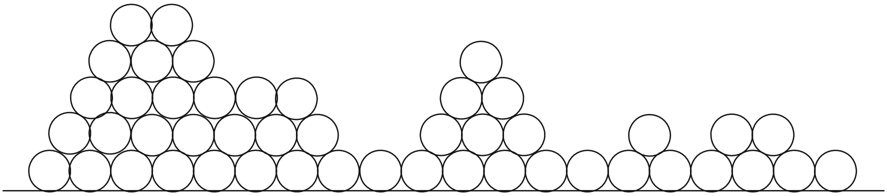


Figure 6.9 A disk stacking

At first glance, it looks like we might be able to characterize a stack of disks by a Dyck path. To this end, call a disk **exposed** if a perfectly vertical rainstorm falling from above bedashes the disk. As shown in Figure 6.10, a path connecting the centers of the exposed disks is almost a Dyck path: the problem is the flats in the path.

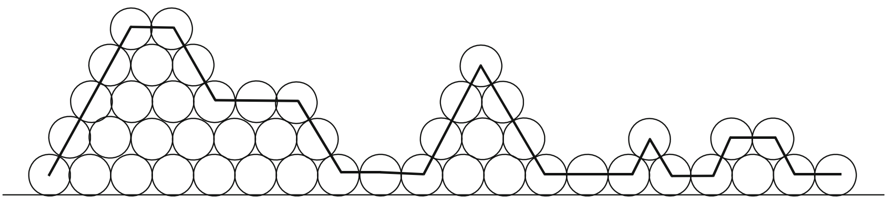


Figure 6.10 Flats problem

The key to dealing with this is the fact that this type of path does not have any V-formations because, as shown in Figure 6.11, a V-formation implies that the path goes through a nonexposed disk.

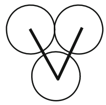


Figure 6.11 No V's in the path

Referring to Figure 6.12, let us add a new row of $n + 1$ disks to the bottom of the stack.

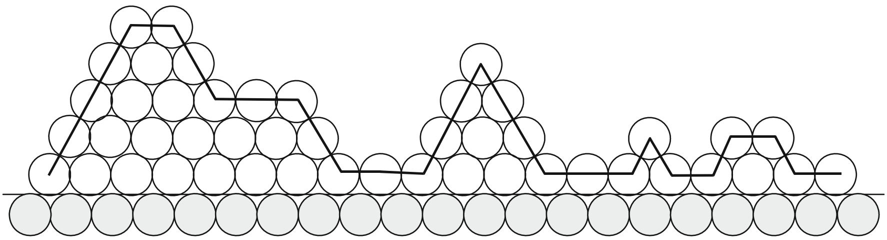
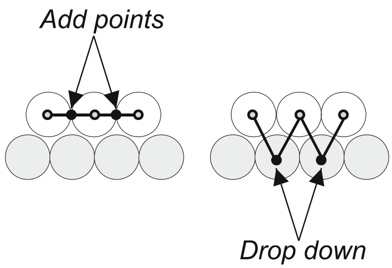


Figure 6.12 Add a bottom row

Now we can eliminate the flats with the two-step process shown in Figure 6.13. In particular, between each pair of adjacent disks involved in a flat, add an additional vertex at the point of tangency and then drop the vertex down one level to create a V. Because of the additional row, this will always be possible. Note that this step is reversible precisely because there were no Vs in the original path.

Figure 6.13 Removing flats



The result of this manipulation is a Dyck path, as shown in Figure 6.14.

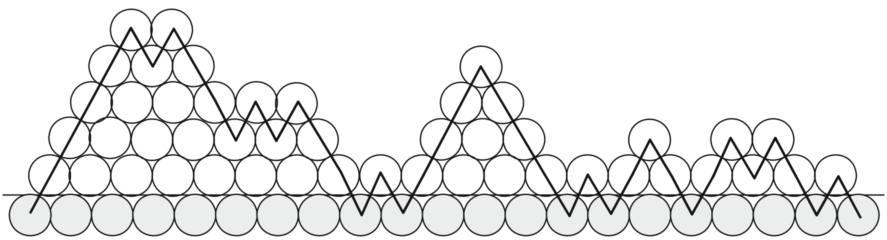


Figure 6.14 Redraw path with no flats

Moreover, since the manipulations are reversible, the map from stacks of disks with n disks on the bottom row to Dyck paths of width $n+1$ is injective.

To see that it is also surjective, let us remove all disks whose centers do not contact the path, as shown in Figure 6.15. This step is also reversible and it seems pretty clear that there is essentially no difference between the path and the collection of disks.

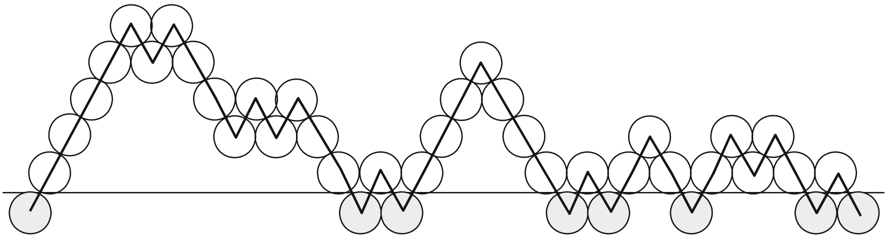


Figure 6.15 The final result

Finally, since there are $n+1$ disks on the bottom row, the Dyck path, which advances one disk radius horizontally for each step, has length $2n$.

Theorem 6.5
C_n counts the number of ways to stack equal-sized disks with n disks on the bottom row. □

Geometric Widgit Summary

Theorem 6.6
C_n counts the number of

1) *ways to chord a convex $2n$-gon with nonintersecting chords,*
2) *staircase tilings of an $n \times n$ grid using n tiles,*
3) *noncrossing, alternating, chorded convex $(n+1)$-gons with n chords,*
4) *triangulations of a convex polygon with $n+2$ sides,*
5) *ways to stack equal-sized disks with n disks on the bottom row.* □

7 Catalan Numbers and Algebraic Widgits

Correct Parenthesizing Under a Nonassociative Binary Operation

Consider a binary operation defined on a set $\mathcal{A}$ that is represented by juxtaposition. If the operation is not associative, then a word of the form *abcd* is not well defined unless we insert parentheses. For the word *abcd*, there are five ways this can be done:

$$((ab)c)d, \quad (ab)(cd), \quad (a(bc))d, \quad a((bc)d), \quad a(b(cd))$$

We would like to count the number of ways there are to **fully parenthesize** a word

$$w = a_1 a_2 \cdots a_n$$

of length n. We will assume that a full parenthesizing does not include parentheses surrounding the entire word, as in $(a(bc))$ or parentheses surrounding a single letter, as in $a((b)c)$. A full parenthesization contains just enough parentheses to disambiguate the expression.

Let $\mathcal{F}_n$ be the family of fully parenthesized words of length n over $\mathcal{A}$. Assume that $n \geq 3$. We can use the first pair of matching parentheses as the nexus to decompose a fully parenthesized word w into smaller fully parenthesized words. In particular, w has the form

$$w = \alpha(\beta)\gamma \tag{7.1}$$

where α, β, and γ are fully parenthesized words, and $\operatorname{len}(\beta) > 1$ and the parentheses shown are the first matching pair, that is, the open parenthesis is the first open parenthesis in w and the closing parenthesis matches the open parenthesis.

Note first that one of α or γ must be the empty word, since otherwise w is not fully parenthesized, for it is not possible to further parenthesize this expression in

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1_7

order to group α and β together or β and γ together without introducing an open parenthesis somewhere to the left of the first open parenthesis.

Moreover, α must have length 0 or 1, since otherwise w would have the form

$$w = a_1 a_2 \cdots (\beta)\gamma$$

with no open parenthesis to the left of the one showing and so w would not be fully parenthesized.

If α is the empty word, then $w = (\beta)\gamma$ and there are two possibilities for γ:

$$w = \begin{cases} (\beta)g & \text{if } \gamma = g \in \mathcal{A} \\ (\beta)(\gamma') & \text{if } \gamma \text{ has length } > 1 \end{cases}$$

where β and γ' are fully parenthesized. On the other hand, if $\mathrm{len}(\alpha) = 1$ then γ must be the empty word and w has the form

$$w = a(\beta)$$

where $a \in \mathcal{A}$ and β is fully parenthesized.

In summary, w has one of the following forms

$$w = a(\beta) \quad \text{or} \quad w = (\beta)g \quad \text{or} \quad w = (\beta)(\gamma')$$

where $a, g \in \mathcal{A}$, $\mathrm{len}(\beta) > 1$, $\mathrm{len}(\gamma') > 1$, and both β and γ' are fully parenthesized. The first two cases are easy to enumerate, since there are evident bijections between the fully parenthesized words of length n and of type $a(\beta)$ (or type $(\beta)g$) and the fully parenthesized words of length $n-1$.

For $2 \le k \le n-2$, let $\mathcal{F}_{n,k}$ be the subfamily of $\mathcal{F}_n$ consisting of those fully parenthesized words of type

$$w = (\beta)(\gamma')$$

where

$$\mathrm{len}(\beta) = k \quad \text{and} \quad \mathrm{len}(\gamma') = n - k$$

Then the map

$$\theta_{n,k} : \mathcal{F}_{n,k} \longrightarrow \mathcal{F}_k \times \mathcal{F}_{n-k}$$

defined by

$$\theta_{n,k}(w) = (\beta, \gamma')$$

is bijective, its inverse map being $(\beta, \gamma) \mapsto (\beta)(\gamma)$.

Letting $D_n = \left|\mathcal{F}_n\right|$ and $D_1 = 1$, we have the associated recurrence relation

$$D_n = 2D_{n-1} + \sum_{k=2}^{n-2} D_k D_{n-k} = \sum_{k=1}^{n-1} D_k D_{n-k} = \sum_{k=0}^{n-2} D_{k+1} D_{n-k-1}$$

for $n \geq 2$. This is the recurrence for a shifted Catalan recurrence (Theorem 3.4 with $a = 1$) and so

$$D_{n+1} = C_n$$

for all $n \geq 0$.

Theorem 7.1
C_n counts the number of ways to fully parenthesize a word of length $n+1$ under a nonassociative binary operation. □

Balanced Parentheses

An inductive characterization of Catalan words can sometimes be useful in solving a counting problem. The key to the inductive definition is that every Catalan word starts with A and so it must contain the subword AB. Moreover, the removal of any such subword simply produces a shorter *Catalan* word.

Theorem 7.2
A word w of length $2n$ over the alphabet (A, B) is a Catalan word if and only if it comes from the empty word ϵ or from a Catalan word of length $2n-2$ by the insertion of the subword AB at any location within the word. □

Consider a string consisting of n opening parentheses and n closing parentheses, such as

$$((())) \quad \text{or} \quad)()()(\quad \text{or} \quad)))(((\,$$

In the first string, the parentheses are correctly matched, or *balanced* but in the second string, they are not. What does it mean to say that a string of parentheses is balanced?

In examining a string w of parentheses, perhaps the most natural way to determine whether or not it is balanced is to repeatedly remove *adjacent* matched pairs of parentheses, that is, substrings of the form () until we reach a point where the string is either empty, in which case w is balanced or is nonempty but we can reduce no further, in which case w is not balanced.

This provides an inductive definition of balanced: A word w of length $2n$ over the alphabet ((,)) is **balanced** if and only if it comes from the empty word ϵ or from a balanced string of length $2n-2$ by the insertion of the subword (). But this is precisely the inductive characterization of Catalan words given earlier.

Theorem 7.3
C_n counts the number of balanced strings of parentheses of length $2n$. □

Null Sums in $\mathbb{Z}_{n+1}$

Here is another algebraic arena in which the Catalan numbers appear. Let $\mathbb{Z}_{n+1} = \{0, 1, \ldots, n\}$ be the integers modulo $n+1$. A multiset M over $\mathbb{Z}_{n+1}$ has **null sum** if the sum of its elements is zero modulo $n+1$. For example, if $n=5$, the multiset

$$S = \{1, 1, 2, 4, 4\}$$

has sum $12 \equiv 0 \bmod 6$ and so S has null sum. Let $\mathcal{Z}_n$ be the family of all multisets over $\mathbb{Z}_{n+1}$ of size n (counting multiplicities of the elements) that have null sum. Thus, $S \in \mathcal{Z}_5$.

Here is an elegant method for determining the size of $\mathcal{Z}_n$ due to Richard Stanley. Define a binary relation on $\mathcal{Z}_n$ by saying that $S = \{a_1, \ldots, a_n\}$ is related to each of its *translations*,

$$S + k = \{a_1 + k, \ldots, a_n + k\}$$

for $k \in \mathbb{Z}_{n+1}$. This is easily seen to be an equivalence relation on $\mathcal{Z}_n$.

Now, the $n+1$ translates

$$S, S+1, \ldots, S+n$$

are all distinct because n and $n+1$ are relatively prime. To see this, suppose that $S+k = S+j$, where $0 \le j \le k \le n$. Using the notation $\sum X$ for the sum of the elements in the set X we have

$$\sum (S+k) \equiv nk \mod(n+1)$$

and

$$\sum (S+j) \equiv nj \mod(n+1)$$

Then $S+k = S+j$ implies that $nk \equiv nj \bmod(n+1)$, that is,

$$n+1 \big| n(k-j)$$

and since n and $n+1$ are relatively prime, we deduce that $k=j$. Therefore, each equivalence class has size $n+1$.

Now, the total number of multisets is equal to the multiset coefficient

$$\left(\binom{n+1}{n}\right) = \binom{2n}{n}$$

and so there are precisely

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

equivalence classes in $\mathcal{M}_n$. However, because n and $n+1$ are relatively prime, the sums for each multiset in a given equivalence class are all different and so exactly one of the multisets in each equivalence class has null sum.

Theorem 7.4
C_n counts the number of multisets of size n in $\mathbb{Z}_{n+1}$ with null sum. □

Algebraic Widgit Summary

Theorem 7.5
C_n counts the number of

1) *ways to fully parenthesize a word of length $n+1$ under a nonassociative binary operation,*
2) *balanced strings of parentheses of length $2n$,*
3) *multisets of size n in $\mathbb{Z}_{n+1}$ with null sum.* □

Catalan Numbers and Interval Structures 8

There are many interesting families of intervals counted by the Catalan numbers. First, however, let us make a remark about antichains in the poset $\mathrm{Int}([n])$ of intervals on $[n]$. If $\mathcal{A}$ is such an antichain, then no two intervals in $\mathcal{A}$ can share a common left endpoint or a common right endpoint. Moreover, if we order the intervals so that the left endpoints are strictly increasing, then the right endpoints must also be strictly increasing. In fact, a family

$$\mathcal{F} = \left\{[a_i, b_i] \middle| 1 \le i \le n\right\}$$

of intervals in $\mathrm{Int}([n])$ is an antichain if and only if both the left endpoint sequence $a_1 \cdots a_n$ and the right endpoint sequence $b_1 \cdots b_n$ can be ordered *at the same time* in strictly increasing order, in symbols, $\mathcal{F}$ is an antichain if and only if (after *simultaneous* reindexing of the a's and b's if necessary)

$$a_1 < \cdots < a_n, \quad b_1 < \cdots < b_n, \quad a_i \le b_i$$

for all i.

Separated Families of Intervals

The following concept will prove useful later.

Definition 8.1

Let $[j, n] \subseteq \mathbb{Z}$. A family $\mathcal{F}$ of intervals in $Int([j, n])$ is **separated** if any two intervals in $\mathcal{F}$ are either disjoint or else one interval is contained in the interior of the other. Let $\mathcal{S}_{[j,n]}$ be the collection of all separated families in $Int([j, n])$. □

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1_8

To count the number D_n of separated families in $\mathcal{S}_{[1,n]}$, we partition the families into groups.

1) For $1 \le k \le n$, let $\mathcal{S}^k_{[1,n]}$ be the set of separated families that contain the interval $[k, n]$.
2) Let $\mathcal{S}^{\neg n}_{[1,n]}$ be the set of separated families for which n is not in any interval of the family.

Since $\mathcal{S}^{\neg n}_{[1,n]} = \mathcal{S}_{[1,n-1]}$, we have

$$\left|\mathcal{S}^{\neg n}_{[1,n]}\right| = D_{n-1}$$

The members of $\mathcal{S}^1_{[1,n]}$ contain the interval $[1, n]$ and so correspond bijectively to the members of $\mathcal{S}_{[2,n-1]}$ and the members of $\mathcal{S}^n_{[1,n]}$ contain the interval $[n, n]$ and so correspond bijectively to the members of $\mathcal{S}_{[1,n-1]}$.

For $2 \le k \le n-1$, a family $\mathcal{F} \in \mathcal{S}^k_{[1,n]}$ can be decomposed into two smaller families using the nexus interval $[k, n]$, because any interval $[i, j] \in \mathcal{F}$ other than $[k, n]$ must be either disjoint from $[k, n]$, that is, contained in $[1, k-1]$ or else contained in $[k, n]^\circ = [k+1, n-1]$. Hence, $\mathcal{F}$ can be written as a *disjoint union*

$$\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 \cup \{[k, n]\}$$

where $\mathcal{F}_1 \in \mathcal{S}_{[1,k-1]}$ and $\mathcal{F}_2 \in \mathcal{S}_{[k+1,n-1]}$. This defines a family of injective maps

$$\theta_{n,k} : \mathcal{S}^k_{[1,n]} \to \mathcal{S}_{[1,k-1]} \times \mathcal{S}_{[k+1,n-1]}$$

by

$$\theta_{n,k}(\mathcal{F}_1 \cup \mathcal{F}_2 \cup \{[k, n]\}) = (\mathcal{F}_1, \mathcal{F}_2)$$

Moreover, since *any* pair $(\mathcal{F}_1, \mathcal{F}_2) \in \mathcal{S}_{[1,k-1]} \times \mathcal{S}_{[k+1,n-1]}$ can be combined into a family $\mathcal{F}_1 \cup \mathcal{F}_2 \cup \{[k, n]\}$ in $\mathcal{S}^k_{[1,n]}$ the map $\theta_{n,k}$ is a bijection.

Set $D_{-1} = D_0 = 1$. Also, $D_1 = 2$ since both the empty family and the family $\{[1]\}$ are separated. The recurrence relation that arises from the decomposition model is then

$$D_n = D_{n-1} + D_{n-2} + D_{n-1} + \sum_{k=2}^{n-1} D_{k-1} D_{n-k-1} = \sum_{k=0}^{n} D_{k-1} D_{n-k-1}$$

This is the recurrence for a shifted Catalan recurrence (Theorem 3.4 with $a = -1$) and so

$$D_n = C_{n+1}$$

Theorem 8.1
C_n counts the number of separated families of intervals in Int$([n-1])$. □

Covering Antichains in Int([n])

The Catalan numbers count the number of **covering antichains** in the interval poset Int ([n]), that is, antichains in Int([n]) with the property that every element of [n] is contained in some interval of the antichain. Figure 8.1 contains an example for $n = 7$.

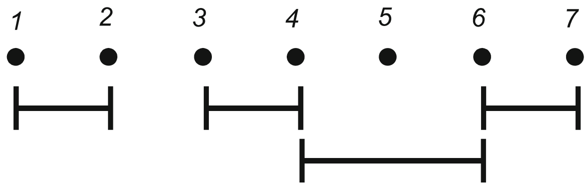


Figure 8.1 A covering antichain in *Int*([7])

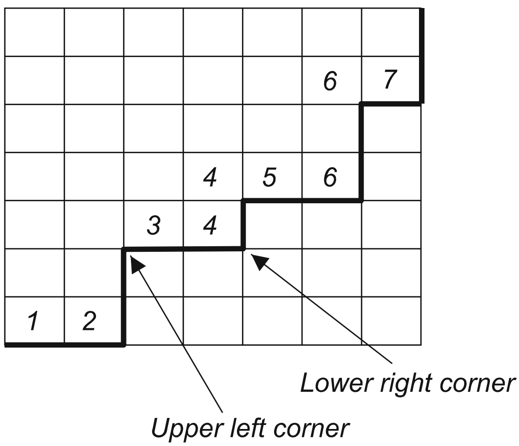
The intervals in this case are

$$[1,2],[3,4],[4,6],[6,7]$$

We can characterize covering antichains as monotonic paths as follows. First, draw an $n \times n$ grid of squares, as shown in Figure 8.2. Denote the cell at row r and column c by $C(r,c)$.

Place the interval $[k,j]$ in row k, starting in the diagonal cell $C(k,k)$ and ending in the cell $C(k,j)$. Then construct a path in the grid as follows. First include the bottom border of the *lowest* occupied cells in each column. Since the intervals cover [n], there is an integer in every column of the table and so there is a lower border in every column. Then connect these line segments with vertical lines to complete the path.

Figure 8.2 Path created from Figure 8.1



The resulting path will not cross the diagonal because if it did cross the diagonal, its next move to the right would be under an empty cell, which it does not do. Note that because the path does not cross the diagonal, it has zero or more **upper left corners** and one or more **lower right corners**, also shown in Figure 8.2.

We thus have a map θ from covering antichains $\mathcal{A}$ in Int([n]) to monotonic paths in an $n \times n$ grid that do not cross the diagonal. To recover the antichain $\mathcal{A}$ from the path $\theta(\mathcal{A})$, for each row r that has a lower right corner at cell $C(r, c)$, include the interval $[r, c]$. Thus, θ is injective.

To see that θ is also surjective, we must show that *any* monotone path P that does not cross the diagonal comes from a covering antichain $\mathcal{A}$ of intervals, that is, $\theta(\mathcal{A}) = P$. Let

$$\mathcal{A} = \left\{[r, c] \middle| C(r, c) \text{ contains a lower right corner of } P\right\}$$

These intervals form an antichain since both the left endpoints and the right endpoints increase as we examine the intervals from the bottom row up. It is also clear that the intervals cover $[n]$, since the path moves horizontally across *every* column and so every column number is contained in an interval.

Thus, $\mathcal{A}$ is a covering antichain.

Theorem 8.2
C_n counts the number of covering antichains in Int$([n])$. □

Antichains in Int([$n - 1$])

Not only do the Catalan numbers count the number of *covering* antichains in Int([n]), but they also count the number of *all* antichains $\mathcal{A}$ in Int$([n-1])$.

The key is that a Catalan word can be completely described by giving the A-count and B-count at certain points only. Specifically, let w be a word over (A, B). We refer to an instance of A in w as **new** if it occurs *immediately following* a B.

Every Catalan word w over (A, B) of length $2n$ has k new As, where $0 \le k \le n - 1$. Moreover, w can be completely characterized by giving the sequence $\alpha_1\alpha_2\cdots\alpha_k$ of A-counts and the sequence $\beta_1\beta_2\cdots\beta_k$ of B-counts for the *new* As only.

To see this, consider two consecutive new As, with letter counts α_j, β_j, and α_{j+1}, β_{j+1}, respectively. Then the substring u of w that lies exclusively between these new As must consist of consecutive As followed by consecutive Bs, as pictured below:

$$A\,\underbrace{A \cdots AB \cdots B}_{u}\,A$$

Moreover, the positions p_j and p_{j+1} of the two consecutive new As are

$$p_j = \alpha_j + \beta_j \quad \text{and} \quad p_{j+1} = \alpha_{j+1} + \beta_{j+1}$$

and the A and B counts for u are

$$\begin{aligned} N_A(u) &= \alpha_{j+1} - \alpha_j - 1 \\ N_B(u) &= \beta_{j+1} - \beta_j - 1 \end{aligned}$$

Hence, the positions of the new As and the letter counts between each pair of consecutive new As are completely determined by the sequences $\alpha_1\alpha_2\ldots\alpha_k$ and $\beta_1\beta_2\ldots\beta_k$. In other words, w is completely determined by these two sequences.

For example, let $n=6$ and consider the count sequences

$$(\alpha_1,\alpha_2,\alpha_3) = (3,5,6) \quad \text{and} \quad (\beta_1,\beta_2,\beta_3) = (1,3,4)$$

Then the new-A positions are

$$(p_1,p_2,p_3) = (4,8,10)$$

and a fill-in-the-gaps construction will construct the unique word w with these count sequences. First, we insert the new As where they belong:

$$w = \underset{1}{}\,\underset{2}{}\,\underset{3}{}\,\underset{4}{A}\,\underset{5}{}\,\underset{6}{}\,\underset{7}{}\,\underset{8}{A}\,\underset{9}{}\,\underset{10}{A}\,\underset{11}{}\,\underset{12}{}$$

Then we compute the number of As between consecutive new As (and before the first new A and after the last new A):

$$\begin{aligned} \alpha_1 - 1 &= 2 \\ \alpha_2 - \alpha_1 - 1 &= 1 \\ \alpha_3 - \alpha_2 - 1 &= 0 \\ n - \alpha_3 &= 0 \end{aligned}$$

Now we can fill in the gaps to get the Catalan word:

$$w = \underset{1}{A}\,\underset{2}{A}\,\underset{3}{B}\,\underset{4}{\boldsymbol{A}}\,\underset{5}{A}\,\underset{6}{B}\,\underset{7}{B}\,\underset{8}{\boldsymbol{A}}\,\underset{9}{B}\,\underset{10}{\boldsymbol{A}}\,\underset{11}{B}\,\underset{12}{B}$$

It is clear that both count sequences $\alpha_1\cdots\alpha_k$ and $\beta_1\cdots\beta_k$ are strictly increasing because each gap between new As must include at least one B. Moreover, the Catalan property implies that

$$1 \le \beta_i < \alpha_i \le n \tag{8.1}$$

for all i, with strict inequality because $\Delta_{a,b}\left([w]_p\right) > 0$ at each *new-A* position p.

Now let $\mathcal{C}_{n,k}$ be the set of all Catalan words over (A, B) of length $2n$ that have k new As. Define a map $\theta_{n,k}$ by setting

$$\theta_{n,k}(w) = (\alpha_1 \cdots \alpha_k, \beta_1 \cdots \beta_k)$$

This map is an injection from $\mathcal{C}_{n,k}$ to the family of pairs of strictly increasing sequences of length k for which [8.1] holds.

To see that $\theta_{n,k}$ is also surjective, let $\alpha_1 \cdots \alpha_k$ and $\beta_1 \cdots \beta_k$ be two increasing sequences satisfying [8.1]. Construct a word w as follows. Begin w with $\alpha_1 - 1$ As followed by β_1 Bs, followed by a (new) A. Since $\beta_1 \leq \alpha_1 - 1$, the A-count does not fall below the B-count at any point in this construction. Next, append $\alpha_2 - \alpha_1 - 1$ As followed by $\beta_2 - \beta_1$ Bs, followed by a (new) A. Since $\beta_2 \leq \alpha_2 - 1$, the A-count does not fall below the B-count at any point in the word thus far constructed. Continuing in this manner will produce a *Catalan* word w whose A and B counts at each new A are given by the sequences $\alpha_1 \cdots \alpha_k$ and $\beta_1 \cdots \beta_k$. Thus, $\theta_{n,k}$ is a bijection.

It is tempting to consolidate these two sequences into a family of intervals

$$\mathcal{F}_k = \left\{[\beta_1, \alpha_1], \ldots, [\beta_k, \alpha_k] \middle| 1 \leq \beta_i < \alpha_i \leq n\right\}$$

thus producing an antichain $\mathcal{F}_k$ in Int($[n]$). Note that since $\beta_i < \alpha_i$, the family $\mathcal{F}_k$ contains no singleton intervals. But if we reduce each α_i by 1 (a reversible operation), we obtain an antichain

$$\mathcal{A}_k = \left\{[\beta_1, \alpha_1], \ldots, [\beta_k, \alpha_k] \middle| 1 \leq \beta_i < \alpha_i \leq n - 1\right\}$$

in Int$([n-1])$ with no restrictions on interval size.

Theorem 8.3
C_n counts the number of antichains (equivalently, down sets) in the interval poset Int$([n-1])$.

Interval Summary

Theorem 8.4
C_n counts the number of

1) *separated families of intervals in Int($[n]$),*
2) *noncrossing, alternating families of n intervals in $[n+1]$,*
3) *covering antichains in Int($[n]$),*
4) *antichains (equivalently, down sets) in Int$([n-1])$.*

Catalan Numbers and Partitions

9

Noncrossing Partitions

Catalan numbers count the number of *noncrossing partitions* of the set $[n]$.

Definition 9.1

A partition $\mathcal{P} = \{B_1, \ldots, B_m\}$ of $[n]$ is **noncrossing** if whenever

$$1 \leq i < j < k < \ell \leq n$$

we cannot have $i, k \in B_u$ and $j, \ell \in B_v$ for $u \neq v$. □

This concept is most easily visualized by placing n equally spaced points around a circle and using unbroken regions interior to the circle to indicate the blocks of the partition. For example, the partition of $\{1, \ldots, 5\}$ on the left in Figure 9.1 is noncrossing but the partition on the right is crossing.

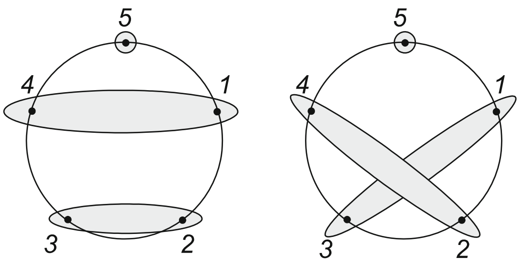


Figure 9.1 Crossing and noncrossing partitions of $\{1, \ldots, 5\}$

(Note that this use of the term "noncrossing" is a bit different than earlier uses. For example, two edges of a noncrossing, alternating tree may intersect at their vertices. However, two blocks of a partition are always disjoint as sets).

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1_9

We want to show that the number of noncrossing partitions of $[n]$ is the Catalan number C_n. To this end, let us make the following definition.

Definition 9.2

Let $\mathcal{P}$ be a noncrossing partition of $[n]$.

1) The **principal block** R of $\mathcal{P}$ is the block containing the integer n. The set of nonprincipal blocks of $\mathcal{P}$ is denoted by $\mathcal{P}'$.
2) The **extent** of a nonprincipal block $B \in \mathcal{P}$ is the interval

$$e(B) = [min\{B\}, max\{B\}]$$

We denote the lower bound of $e(B)$ by $\ell(B)$ and the upper bound by $u(B)$ and so

$$e(B) = [\ell(B), u(B)]$$

3) The **family of extents** of a partition $\mathcal{P}$ is the family of extends of the nonprincipal blocks of $\mathcal{P}$. □

If B is a nonprincipal block of $\mathcal{P}$ and we place the integers 1 through $n-1$ in a line, then B has a form similar to the *blob* (with apologies to Mary Katherine Linaker) shown in Figure 9.2.

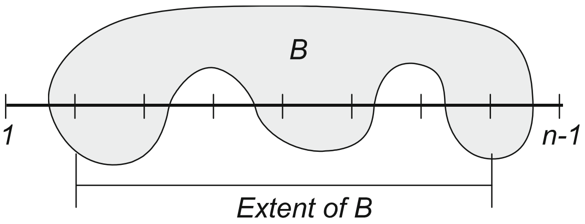


Figure 9.2 The extent of a block

Note that the endpoints of an extent $[\ell(B), u(B)]$ are contained in the defining block B. As a consequence, the endpoints of two distinct extents are disjoint, as are the endpoints of an extent and the nonprincipal block R.

Theorem 9.1
If $\mathcal{P}$ is a noncrossing partition of $[n]$, then the following hold:

1) *The family $\mathcal{I}$ of extents of $\mathcal{P}$ is separated.*
2) *The extents do not intersect the principal block R.*

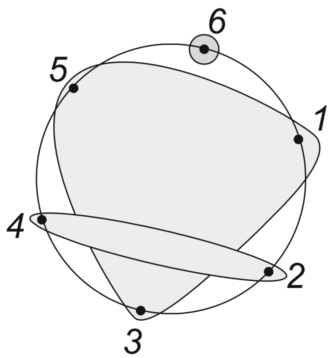
Proof If $\mathcal{F}$ is any family of intervals whose endpoints are distinct, then $\mathcal{F}$ is either crossing or separated. For the second statement, if $x \in [\ell(B), u(B)] \cap R$, then $x \notin B$ and so $n \in R$ implies that

$$\ell(B) < x < u(B) < n$$

violates the noncrossing property. □

Note that the converse of this theorem does not hold. For example, as shown in Figure 9.3, the partition

Figure 9.3 Counterexample



$$\mathcal{P} = \{\{2,4\},\{1,3,5\},\{\{6\}\}$$

of [6] is crossing but the family of extents $\{[2,4],[1,5]\}$ is separated and $R = \{6\}$ does not intersect any other block of the partition.

Next we observe that the blocks of $\mathcal{P}$ can be recovered from the family of extents. Note that if $\mathcal{F}$ is a separated family of intervals, then $I \subseteq J^{\circ}$ is equivalent to $I \subset J$ for $I,\ J \in \mathcal{F}$.

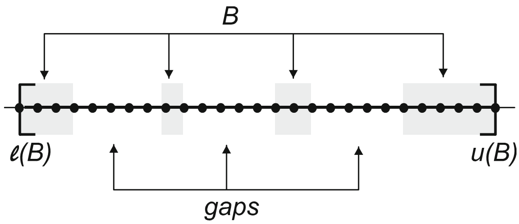
Theorem 9.2
Let $\mathcal{P}$ be a noncrossing partition of $[n]$. Then any nonprincipal block B of $\mathcal{P}$ can be written in the form

$$B = e(B) \backslash \left(\bigcup_{\substack{C \in \mathcal{P}' \\ e(C) \subset e(B)}} e(C) \right)$$

Hence, the partition $\mathcal{P}$ is completely determined by its family of extents.

Proof Figure 9.4 shows an extent $e(B)$ for a nonprincipal block B.

Figure 9.4 An extent $e(B)$



The simple fact is that every nonprincipal block C for which $e(C) \subset e(B)$ is entirely contained within one of the gaps in $e(B)$, lest the noncrossing property be violated. Therefore, since $R \cap e(B) = \emptyset$, if we remove all such extents $e(C)$, what remains must be B itself.

For those who are not convinced by this argument, here are the explicit details. Let S be the set on the right. If $C \in \mathcal{P}'$ is one of the blocks in the union defining S, that is, if $e(C) \subset e(B)$, then $e(C)$ is disjoint from B, for if $b \in B \cap e(C)$ then

$$\ell(B) < \ell(C) < b < u(C)$$

which violates the noncrossing property. It follows that $B \subseteq S$.

For the reverse inclusion, suppose that $x \in S$ but $x \notin B$. Then $x \in e(B) \backslash B$ and since $R \cap e(B) = \emptyset$, it follows that $x \notin R$. Hence, $x \in C$ for some nonprincipal block $C \in \mathcal{P}'$ other than B. Since $e(C)$ and $e(B)$ are not disjoint (both containing x), we must have either $e(C) \subset e(B)$ or $e(B) \subset e(C)$. If $e(C) \subset e(B)$, then $e(C)$ is one of the extents that is removed from S and so $x \notin S$, a contradiction. On the other hand, if $e(B) \subset e(C)$, then $x \in e(B) \backslash B$ implies that

$$\ell(C) < \ell(B) < x < u(B) < u(C)$$

which violates the noncrossing property. Thus, neither case is possible and so $x \in B$. □

Theorem 9.2 implies that the map θ sending any noncrossing partition $\mathcal{P}$ to its family of extents is injective. The following result shows that θ is also surjective.

Theorem 9.3
Any separated family $\mathcal{F}$ of intervals in $Int([n-1])$ is the family of extents for a noncrossing partition $\mathcal{P}$ of $[n]$. The nonprincipal blocks of $\mathcal{P}$ are the sets

$$B_I = I \backslash \left(\bigcup_{J \subset I} J \right)$$

for all $I \in \mathscr{F}$ and the principal block is the rest of $[n]$,

$$R = [n] \backslash \bigcup_{I \in \mathcal{F}} B_I$$

Proof Consider the intersection of two blocks

$$B_I \cap B_K = \left[I \backslash \left(\bigcup_{J \subset I} J \right) \right] \cap \left[K \backslash \left(\bigcup_{J \subset K} J \right) \right]$$

If I and K are disjoint, then clearly so are B_I and B_K. If I and K are not disjoint, then we can assume that $I \subset K$, in which case I is one of the intervals that is removed in defining B_K and so $I \cap B_K = \emptyset$, which implies a fortiori that $B_I \cap B_K = \emptyset$. Thus, the family

$$\mathcal{P} = \left\{ B_I \middle| I \in \mathcal{I} \right\} \cup \{R\}$$

is a partition of $[n]$.

To see that $\mathcal{P}$ is noncrossing, suppose that

$$1 \leq i < j < k < \ell \leq n$$

If $i,\ k \in B_I$ and $j,\ \ell \in B_J$ for $I \neq J$, then $i,\ k \in I$ and $j,\ \ell \in J$ and since I and J are intervals, we have

$$\{j, k\} \subseteq [i, k] \subseteq I \quad \text{and} \quad \{j, k\} \subseteq [j, \ell] \subseteq J$$

whence $j,\ k \in I \cap J$. Hence, one of I or J is contained in the interior of the other. If $J \subset I$, then $B_I \cap J = \emptyset$, contradicting the fact that $k \in B_I \cap J$. Similarly, if $I \subset J$, then $B_J \cap I = \emptyset$, contradicting the fact that $j \in B_J \cap I$. Hence, no two nonprincipal blocks cross.

On the other hand, if $i, k \in B_I$, and $j, \ell \in R$, then $j \in I \cap R$. Since $j \in R$, it is not in any of the blocks B_K. In particular, $j \notin B_I$ and since $j \in I$, it follows that j is in some interval I_1 *properly* contained in I. But $j \notin B_{I_1}$ and since $j \in I_1$, it follows that j is in some interval I_2 *properly* contained in I_1. This argument is clearly headed straight for a contradiction. Hence, no nonprincipal block crosses R and so $\mathcal{P}$ is noncrossing. □

Thus, the map θ sending noncrossing partitions of $[n]$ to a separated family in $\mathrm{Int}([n-1])$ is a bijection. We can now invoke Theorem 8.1.

Theorem 9.4
C_n counts the number of noncrossing partitions of the set $[n]$. □

Noncrossing Partitions and Davenport-Schinzel Sequences

Davenport-Schinzel sequences where first described by H. Davenport and A. Schinzel in their paper "A Combinatorial Problem Connected with Differential Equations," *American Journal of Mathematics*, Vol. 87, No. 3, July, 1965, pp. 684–694. We consider a special case of these sequences.

First a bit of terminology. For any sequence $w = a_1 \cdots a_m$ of integers, the **underlying set** $U(w)$ of w is the set of distinct elements of w. For instance, if $w = 1231234$ then $U(w) = \{1, 2, 3, 4\}$. A sequence of integers is said to be ***abab*-avoiding** if it contains no subsequence of the form *abab*. For instance, the sequence

$$123452674$$

is not *abab*-avoiding because it contains the subsequence 2424. Patterns of the form $abab\cdots ab$ of length $2k$ consisting of k pairs ab are called **alternating patterns** of length $2k$. Finally, if $I, J \subseteq \mathbb{Z}$, we write $I < J$ if every element of I is less than every element of J.

Definition 9.3

Let $I \subseteq \mathbb{Z}$ be a nonempty subset of integers. A **Davenport-Schinzel sequence** (or **DS sequence**) over the set I is a sequence w over I with the following properties:

1) No two consecutive elements of w are the same.
2) w is *abab*-avoiding.

A **normalized Davenport-Schinzel sequence** (or **NDS sequence**) is a DS sequence w with the following additional properties:

3) $U(w) = I$, that is, every element of I appears in w.
4) The first occurrences in w of each integer in I are in increasing order.

A **maximal normalized Davenport-Schinzel sequence** (or **MNDS sequence**) is an NDS sequence over I that is not a proper subsequence of another NDS sequence over I. Let $NDS(I)$ denote the family of NDS sequences over I and let $MNDS(I)$ denote the family of MNDS sequences over I. We use the notation $(M)NDS(I)$ to represent either $NDS(I)$ or $MNDS(I)$. □

In the literature, I is usually the set $[n]$. We will write DS(n) for DS($[n]$) and DS(k, n) for DS($[k, n]$). Also, Davenport and Schinzel's original definition of DS sequences involves another parameter k by specifying that a DS sequence has no alternating pattern $abab\cdots ab$ of length $2k$.

Here are the NDS(4) sequences. The $C_3 = 5$ sequences in MNDS(4) are marked with an asterisk.

1213141*
121341
1213431*
123141
1232141*
123241
1232421*
123421
123431
1234321*
123431

Our interest is in determining the size of the set MNDS(n). We leave it as an exercise (which might be better attempted after finishing this section) to show that the sequences in MNDS(n) have length $2n - 1$.

DS Sequences and Partitions

We can view MNDS sequences in the following light. Let $w = b_1 \cdots b_{2n-1} \in \text{MNDS}(n)$. Define a partition $\mathcal{P}$ of $[2n-1]$ with n blocks by specifying that b_k is the block number of the integer k. The normality condition ensures that the blocks are both nonempty and indistinguishable, as they must be in a partition. Maximality simply ensures that we are partitioning a single integer $2n - 1$.

For example, for $1213431 \in \text{MNDS}(4)$, we write the block numbers below the integers 1 through 7,

$$\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 2 & 1 & 3 & 4 & 3 & 1 \end{array}$$

which shows clearly that the concomitant partition is

$$\mathcal{P} = \{\{1, 3, 7\}, \{2\}, \{4, 6\}, \{5\}\}$$

Now, requirement 1) in the definition of DS sequences is equivalent to the statement that no block of $\mathcal{P}$ contains two *consecutive* integers and the *abab*-pattern avoidance of w is equivalent to the fact that $\mathcal{P}$ is noncrossing. Thus, MNDS(n) sequences characterize noncrossing partitions of $[2n - 1]$ with the additional property that no block contains two consecutive integers.

Counting MNDS Sequences

It is clear that the underlying set I of MNDS(I) is not really relevant to the combinatorial properties of MNDS(I). Put more technically, if $w = a_1 \cdots a_m \in (\mathrm{M})$ NDS(I) and if $\sigma : I \to J \subseteq \mathbb{Z}$ is an order-preserving bijection, then the induced map

$$\sigma w = (\alpha a_1) \cdots (\sigma a_m) \in (\mathrm{M})\mathrm{NDS}(J)$$

is a bijection from (M)NDS(I) to (M)NDS(J). In particular, if $\left|I\right| = n$, then

$$|(\mathrm{M})\mathrm{NDS}(I)| = |(\mathrm{M})\mathrm{NDS}(n)|$$

Theorem 9.5
Suppose that $w \in NDS(n)$.

1) *Then w starts with 1 and if w is also maximal, then it ends with 1.*
2) *If we expose all of the occurrences of 1, that is, if*

$$w = 1x_1 1x_2 \cdots 1x_r(1)$$

where $1 \notin U(x_i)$, then there are integers $p_1, \ldots, p_{r-1}$ for which

$$\begin{aligned} &x_1 \in NDS(2, p_1) \\ &x_2 \in NDS(p_1 + 1, p_2) \\ &x_3 \in NDS(p_2 + 1, p_3) \\ &\quad\vdots \\ &x_r \in NDS(p_{r-1} + 1, n) \end{aligned}$$

Moreover, w is maximal if and only if each x_i is maximal.

Proof For part 1), normality implies that w starts with 1. Moreover, if w is maximal but

$$w = 1 \cdots u$$

where $u \neq 1$, then $w1$ will be in NDS(n) provided that it does not have an $a1a1$ pattern using the trailing 1. But if w has this pattern, then it also has the $1a1a$ pattern using the leading 1, which is not the case. Hence, $w1 \in \mathrm{NDS}(n)$, which contradicts the maximality of w. We conclude that w ends in 1.

For part 2), it is clear that each x_i is a DS sequence satisfying part 4) of the definition. As to the underlying sets, if $a \in U(x_i) \cap U\left(x_j\right)$ for some $i \neq j$, then w contains the proscribed pattern $1a1a$. Hence, the sets $U(x_i)$ are pairwise disjoint and partition the set $[2, n]$. Moreover, if $a \in U(x_i)$ and $b \in U\left(x_j\right)$ where $i < j$ but

$a > b$, then normality implies that b must also appear before a, say $b \in U(x_\ell)$ for $\ell \leq i < j$, which contradicts the disjointness of the underlying sets. Thus,

$$i < j \quad \Rightarrow \quad U(x_i) < U(x_j)$$

and so the sets $U(x_i)$ must have the desired form.

For the final statement of part 2), since the underlying sets are disjoint, insertion of an integer $a \in U(x_i)$ into x_i will create a forbidden pattern in x_i if and only if it creates a forbidden pattern in w. □

We can now decompose a member of MNDS(n) using the first *internal* appearance of 1 (if there is such an appearance). Specifically, Theorem 9.5 implies that any $w \in \mathrm{MNDS}(n)$ has one of the following two forms:

1) $w = 1x1$, where $x \in \mathrm{MNDS}(2, n)$
2) $w = 1x1y1$, where

$$x \in \mathrm{MNDS}(2, k) \quad \text{and} \quad 1y1 \in \mathrm{MNDS}(\{1\} \cup [k+1, n])$$

for some $2 \leq k \leq n - 1$.

Let $\mathrm{MNDS}_1(n)$ be the members of MNDS(n) of type 1 and let $\mathrm{MNDS}_{2,k}(n)$ be the members of MNDS(n) of type 2 with $x \in \mathrm{MNDS}(2, k)$. Let $D_n = \left|\mathrm{MNDS}_{(n)}\right|$.

For case 1), we define the decomposition map

$$\theta_{n,1} : \mathrm{MNDS}_1(n) \to \mathrm{MNDS}(2, n)$$

by

$$\theta_{n,1}(1x1) = x$$

This map is clearly a bijection and so

$$|\mathrm{MNDS}_1(n)| = D_{n-1}$$

For case 2), we define the decomposition map

$$\theta_{n,k} : \mathrm{MNDS}_{2,k}(n) \to \mathrm{MNDS}(2, k) \times \mathrm{MNDS}(\{1\} \cup [k+1, n])$$

by

$$\theta_{n,k}(1x1y1) = (x, 1y1)$$

which is also a bijection and so

$$|\mathrm{MNDS}_{2,k}(n)| = |\mathrm{MNDS}(2,k)| \times |\mathrm{MNDS}(\{1\} \cup [k+1,n])| = D_{k-1}D_{n-k+1}$$

Thus, we are led to the recurrence relation (setting $D_1 = 1$)

$$D_n = D_{n-1} + \sum_{k=2}^{n-1} D_{k-1}D_{n-k+1} = \sum_{k=2}^{n} D_{k-1}D_{n-k+1} = \sum_{k=0}^{n-2} D_{k+1}D_{n-k+1}$$

which is a shifted Catalan recurrence (Theorem 3.4 with $a = 1$) and so

$$C_n = D_{n+1} = |\mathrm{MNDS}(n+1)|$$

for all $n \geq 0$.

Theorem 9.6
C_n counts the number of

1) *maximal normalized Davenport-Schinzel sequences on* $[n+1]$
2) *noncrossing partitions of* $[2n+1]$ *into* $n+1$ *blocks, none of which contain two consecutive integers.* □

Partition Summary

Theorem 9.7
C_n counts the number of

1) *noncrossing partitions of the set* $[n]$.
2) *noncrossing partitions of* $[2n+1]$ *into* $n+1$ *blocks, none of which contain two consecutive integers (equivalently, members of MNDS* $(n+1)$*).* □

Catalan Numbers and Permutations 10

We turn now to a vast subject called **pattern avoiding permutations** that is currently the subject of much research. Let S_n denote the set of all permutations of $[n]$.

A **pattern** of length k is a permutation of the set $[k]$, whose entries are meant to indicate *relative* order. For example, the pattern 312 means "an integer followed by a smaller integer followed by an integer between the first two integers." The pattern 12345 means an increasing sequence of five integers. The numbers in the pattern have no meaning in absolute terms—3 does not mean 3—it only means "something bigger than what is represented by 2 and smaller than what is represented by 4."

A **subsequence** of a permutation $\pi = a_1 a_2 \cdots a_n \in S_n$ of length k is the permutation obtained simply by deleting $n - k$ elements from π. This preserves the order of the elements in the subsequence. Note that the elements of the subsequence need not be consecutive within the original permutation.

If $b_1 b_2 \cdots b_k$ is a pattern of length $k \leq n$, then a permutation $\pi \in S_n$ is $\boldsymbol{b_1 b_2 \cdots b_k}$**-avoiding** if it contains no subsequence with this pattern. For instance, the permutation 2154376 is 312-avoiding, although it may take a moment or two to see this.

As mentioned, the subject of counting pattern avoiding permutations is quite complex and is an active area of current research in combinatorics. We will determine the number of pattern-avoiding permutations in S_n for all patterns of length three. Guess what?

Permutations Obtained from Stacks and Queues

A **stack** is a data structure that stores objects in a linear array and for which insertion and removal occur only at one end of the array, known as the **top** of the stack because the stack is often visualized as a vertical column of objects. The prototypical example of a stack is a stack of dinner plates: a new plate is inserted at the top of the stack only and a plate is removed from the top of the stack only.

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1_10

Adding an object to the top of a stack is referred to as **pushing** and removing an object from the top is called **popping**.

On the other hand, a **queue** is a data structure that stores objects in a linear array and for which insertion is done at the end of the queue and removal is done at the beginning of the queue. The prototypical example of a queue is the long, long, long line at the bank.

Now, consider a game that has an input queue, an output queue, and a stack, as shown on the left in Figure 10.1.

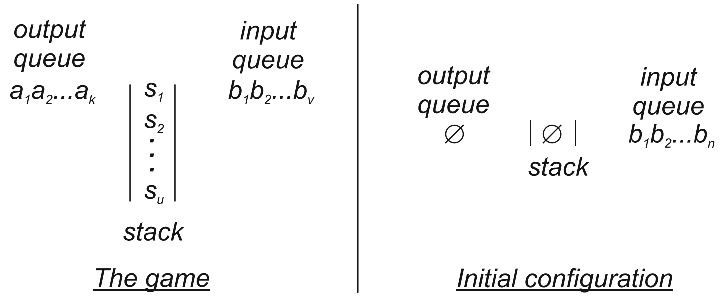


Figure 10.1 The game and its initial configuration

As shown on the right in Figure 10.1, initially the output queue and stack are empty and the input queue contains a permutation $b_1 \cdots b_n$ of $[n]$. At any stage in the game, we can do one of the two things: push the item at the front of the input queue onto the stack or pop an item off the stack onto the end of the output queue. Both operations are shown in Figure 10.2.

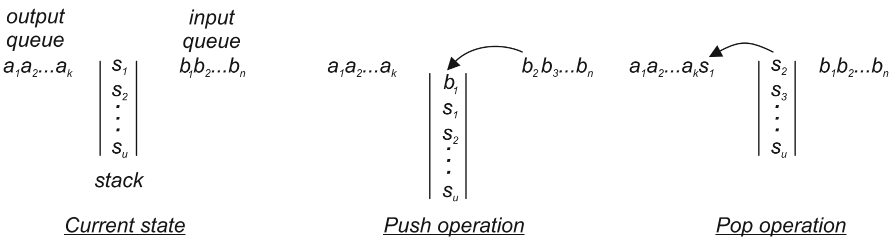


Figure 10.2 The push and pop operations

Let us agree that the **reading order** of the two queues is from left to right and the reading order of the stack is from *bottom* to *top*. The undecorated term *order* will refer to reading order.

We refer to a sequence of pushes and pops that results in the entire input permutation being moved to the output queue as a **legal procedure**. In general, there are many legal procedures, each of which must consist of exactly n pushes and n pops. Once a legal procedure is completed, the permutation in the output queue is called the **output permutation**. Here is an example, where the input queue is 1234.

	Output		Stack		Input
0)	$\varnothing$		$\|\varnothing\|$		1234
1)	$\varnothing$		$\|1\|$	$\leftarrow$	234
2)	$\varnothing$		$\begin{vmatrix} 2 \\ 1 \end{vmatrix}$	$\leftarrow$	34
3)	2	$\leftarrow$	$\|1\|$		34
4)	21	$\leftarrow$	$\|\varnothing\|$		34
5)	21		$\|3\|$	$\leftarrow$	4
6)	213	$\leftarrow$	$\|\varnothing\|$		4
7)	213		$\|4\|$	$\leftarrow$	$\|\varnothing\|$
8)	2134	$\leftarrow$	$\|\varnothing\|$		$\varnothing$

Thus, the output permutation is 2134.

If we denote a push operation by U and a pop operation by O, then a legal procedure can be described by a word w over the alphabet (U, O). However, not all words describe a legal procedure. For example, we cannot start with a pop operation since the stack is initially empty. Also, at any stage, the number of pops cannot exceed the number of pushes, since the stack is not a hot-air popper. (Sorry.) Of course, at the end, the number of pops must equal the number of pushes. Does all this sound familiar?

Thus, a legal procedure is just a Catalan word of length $2n$ over (U, O). Conversely, any Catalan word w over (U, O) of length $2n$ represents a legal procedure. To see this, note that the first letter in w is a U and we may certainly perform this push operation. Suppose that we have performed the first k stack operations specified by w. The issue is whether or not we can perform the $(k+1)$ st stack operation specified by w.

If the $(k+1)$st operation is a push, then we have done at most $n-1$ prior pushes and so the input queue is not empty and we can perform the required push. If the $(k+1)$st operation is a pop, then since $[w]_{k+1}$ is a weak Dyck word, $[w]_k$ is a strong Dyck word, that is,

$$N_O([w_k]) < N_U([w_k])$$

which means that we have not popped off the stack as many integers as we have pushed onto the stack and so there is at least one integer available for popping. Thus, we can always perform the next operation in w and so w is a legal procedure.

Thus, legal procedures are characterized by Catalan words and so for any given input permutation of length n, there are precisely C_n possible output permutations.

Stack Permutations

Now let us suppose that the input queue for the game is the permutation $12\cdots n$. In this case, output permutations are called **stack permutations**. We note the following for future reference:

1) The integers on the stack at any time are in increasing (reading) order.
2) The integers on the stack at any time are smaller than the integer at the end of the output queue.
3) Everything in the input queue is larger than everything in the stack or the output queue.

We can characterize stack permutations using pattern avoidance. In fact, a stack permutation σ is 312-avoiding. To see this, suppose that σ contains the pattern 31, that is,

$$\sigma = \cdots c \cdots a \cdots \quad \text{where} \quad c > a$$

At the moment that a is popped off the stack, everything else on the stack is smaller than a and everything in the input queue is larger than c and so there are no integers between a and c available to add to the output permutation. Hence, σ cannot have the 312 pattern.

Conversely, any 312-avoiding permutation $\sigma = a_1 \cdots a_n$ is a stack permutation. Clearly, we can place a_1 first in the output queue using an appropriate sequence of pushes and pops. Suppose that we have *just popped* a_k onto the output queue, as shown below:

$$a_1 \cdots a_k \qquad \left|\begin{array}{c} s_u \\ \vdots \\ s_1 \end{array}\right| \qquad b_1 \cdots b_v$$

Then

$$s_1 < \cdots < s_u < a_k$$

and we must show that a_{k+1} can be moved to the output queue immediately following a_k.

First note that a_{k+1} cannot be buried in the stack, because then

$$a_{k+1} < s_u < a_k$$

But s_u is not equal to any of $a_1, \ldots, a_k$ or a_{k+1} and so σ has the form

$$\sigma = a_1 \cdots a_k a_{k+1} \cdots s_u \cdots$$

which has the forbidden 312 pattern. This leaves two possibilities. If a_{k+1} is on the top of the stack, we can certainly pop it onto the end of the output queue following a_k. Also, if a_{k+1} is in the input queue, then we can push integers onto the stack until a_{k+1} is at the top of the stack and then pop it onto the output queue. Thus, nothing can prevent us from placing a_{k+1} on the output queue immediately following a_k. Hence, σ is a stack permutation.

Theorem 10.1
1) *A permutation is a stack permutation if and only if it is 312-avoiding.*
2) C_n *counts the number of stack permutations of* $[n]$. □

But there is more that we can say. First, we need a couple of definitions.

The **reverse** π^R of a permutation $\pi \in S_n$ is the permutation formed by reversing the order of the integers in π. For example, if $\pi = 14253$ then $\pi^R = 35241$. The reverse operation is clearly a bijection on S_n. The **complement** of $\pi \in S_n$ is the permutation π' formed by replacing each integer k in π by $n + 1 - k$. For example, the complement of $\pi = 14253$ is $\pi' = 52413$. This operation is also a bijection on S_n.

Here is the reason for these definitions:

1) A permutation $\pi \in S_n$ is 312-avoiding if and only if its **reverse** π^R is 213-avoiding.
2) A permutation $\pi \in S_n$ is 312-avoiding if and only if its complement is 132-avoiding.
3) A permutation $\pi \in S_n$ is 312-avoiding if and only if its reverse complement is 231-avoiding.

Theorem 10.2
C_n *counts the sizes of each of the following sets:*

1) *The set of 312-avoiding permutations,*
2) *The set of 213-avoiding permutations,*
3) *The set of 132-avoiding permutations,*
4) *The set of 231-avoiding permutations.* □

Stack-Sortable Permutations

Let us look again at 231-avoiding permutations. A permutation σ of $[n]$ is **stack-sortable** if there is a legal procedure that produces the output permutation $12\cdots n$ from the *input* σ.

If σ is stack-sortable, then any procedure that produces the output $12\cdots n$ must adhere to the following simple rule:

$$\text{always keep the stack in } \textit{decreasing} \text{ order from bottom to top} \qquad [10.1]$$

For if $k > j$ and k appears higher on the stack than j, then k will appear in the output *before* j and so we can never achieve the required output $12\cdots n$.

It is easy to see that the presence of the 231 pattern will cause a violation of this rule at some stage in the procedure. For suppose that σ has the 231 pattern, say

$$\sigma = \cdots a_i \cdots a_j \cdots a_k \cdots \quad \text{for} \quad a_k < a_i < a_j$$

Then if rule [10.1] has been followed up to the point at which a_j is about to be pushed onto the stack, the smaller a_i must already be in the output queue. But then it is too late to place a_k in the output queue ahead of the larger a_i. Thus, if σ is stack-sortable, it must be 231-avoiding.

For the converse, we show that any 231-avoiding permutation σ is stack-sortable using a procedure that adheres to rule [10.1]. The first step is to show that we can place the integer 1 at the front of the output queue. But since σ is 231-avoiding, the integers that come before 1 in σ must be in decreasing order and so we can push those integers onto the stack without violating rule [10.1] until we reach the integer 1, which can then be moved to the front of the output queue.

Now assume that we have placed the integers $12 \cdots k$ in the output queue in order, all the while adhering to rule [10.1]. Thus, the situation is as follows *just after* popping k onto the output queue,

$$12 \cdots k \quad \left| \begin{array}{c} s_u \\ \vdots \\ s_1 \end{array} \right| \quad b_1 \cdots b_v$$

where

$$s_1 > s_2 > \cdots > s_u$$

We must show that $k+1$ can be placed onto the output queue immediately following k. Note that $k+1$ cannot be anywhere on the stack except at the top, in which case we simply pop it onto the output queue.

On the other hand, suppose that $k+1$ is in the input queue, say

$$12 \cdots k \quad \left| \begin{array}{c} s_u \\ \vdots \\ s_1 \end{array} \right| \quad b_1 \cdots b_j (k+1) \cdots$$

Since the integers $b_1, \ldots, b_j$ (if any) are larger than $k+1$, they must appear in decreasing order from left to right in order to avoid the 231 pattern. Moreover, these integers are smaller than anything on the stack and so we can push them onto the stack without violating rule [10.1]. Then we can move $k+1$ to its rightful position on the output queue.

Theorem 10.3

1) *A permutation σ is stack-sortable if and only if it is 231-pattern avoiding.*
2) *C_n counts the number of stack-sortable permutations of* $[n]$. □

321-Avoiding and 123-Avoiding Permutations

Theorem 10.2 accounts for four of the six three-term patterns. We are left with the patterns 321 and 123. But since the reverse map sends a 321-avoiding permutation to a 123-avoiding permutation, we can restrict attention to 321-avoiding permutations. These are the permutations that do not have a three-term *decreasing* subsequence. Let $\mathcal{F}_n$ denote the set of all 321-avoiding permutations of $[n]$.

We begin with the observation that if we divide the permutation $12\cdots n$ into two subsequences $b_1b_2\cdots b_k$ and $c_1c_2\cdots c_{n-k}$ and then *interlace* these two sequences in any manner (retaining the increasing order of each subsequence), the result is a permutation $\sigma = a_1\cdots a_n$ that is 321-avoiding. This is a simple application of the pigeonhole principle: Given any three integers a_i, a_j, and a_k in σ with $i < j < k$, at least two of them must belong to one of the increasing subsequences and so we cannot have $a_i > a_j > a_k$.

This raises the question of whether all 321-avoiding permutations are constructed in this manner. To address this question, note that if $\sigma = a_1a_2\cdots a_n \in \mathcal{F}_n$, then we can easily find one increasing subsequence simply by taking each integer in σ that is larger than any integer that came before it. For example, in the permutation

$$\sigma = 41237856$$

each of the integers 4, 7, and 8 is larger than any integer that comes before it and so the subsequence 478 is increasing.

Let us pause for a few definitions. If a is an integer in σ, let $p(a)$ denote the **position** of a in σ. For example, if $\sigma = 41237865$, then $p(3) = 4$.

Definition 10.1

Let $\sigma = a_1a_2\cdots a_n$ be in S_n.

1) The integer a_k is **superior** if it is the largest integer in the initial segment $a_1\cdots a_k$. The **superior subsequence** σ_{sup} of σ is the subsequence consisting of the superior integers.
2) The integer a_k is **inferior** if it is not superior. The **inferior subsequence** σ_{inf} of σ is the subsequence consisting of the inferior integers. □

For example, for the permutation

$$\sigma = 41237856$$

we have

$$\sigma_{sup} = 478 \quad \text{and} \quad \sigma_{inf} = 12356$$

As it happens, the 321-avoiding condition is equivalent to the inferior subsequence being increasing.

Theorem 10.4
Let $\sigma = a_1 a_2 \cdots a_n \in S_n$.

1) a) *The superior subsequence* $\sigma_{sup} = c_1 \cdots c_k$ *of* σ *is increasing.*
   b) *For all* i,

$$c_i \geq p(c_i)$$

   c) *The length of the superior sequence satisfies*

$$1 \leq \operatorname{len}(\sigma_{sup}) \leq n$$

   with equality possible.
2) a) *The inferior subsequence* $\sigma_{inf} = b_1 \cdots b_k$ *of* σ is increasing if and only if σ is 321-avoiding.
   b) *If* σ is 321-avoiding, then

$$b_i < p(b_i)$$

   for all i.
   c) *The length of the inferior sequence satisfies*

$$0 \leq \operatorname{len}(\sigma_{inf}) \leq n - 1$$

   with equality possible.

Proof. For part 1), it is clear that σ_{sup} is increasing. Moreover, if $c_i < p(c_i)$ for some i, then there are $p(c_i) - 1$ positions in σ to the left of c_i but only $c_i - 1 < p(c_i) - 1$ positive integers smaller than c_i and so one of the integers occurring before c_i must be larger than c_i, in contradiction to its superiority. Thus, $c_i \geq p(c_i)$ for all i. As to the length of the superior subsequence, n is always superior and so $1 \leq \operatorname{len}(\sigma_{sup}) \leq n$. Finally, all integers in the permutation $12 \cdots n$ are superior and only one integer in the permutation $n(n-1) \cdots 1$ is superior.

For part 2), if σ_{inf} is increasing then as discussed earlier, σ is 321-avoiding. For the converse, suppose that σ is 321-avoiding. Then since any inferior integer is preceded by something larger, it cannot be followed by anything smaller without creating a 321 pattern. Hence, the inferior subsequence of a 321-avoiding permutation must be increasing.

Moreover, if σ is 321-avoiding and $b_i \geq p(b_i)$ for some i, then since b_i is inferior, the *first* integer c_j greater than b_i must occur before b_i and must be superior. The number of positions in σ to the left of b_i, not including c_j, is $p(b_i) - 2$, but there are $b_i - 1 \geq p(b_i) - 1$ positive integers less than b_i and so one of these integers a_k must occur later than b_i, in which case $c_j b_i a_k$ is decreasing, a contradiction.

Finally, the statement about length follows from part 1) and the fact that $\mathrm{len}(\sigma_{inf}) + \mathrm{len}(\sigma_{sup}) = n$. □

Thus, we see that any 321-avoiding permutation σ is indeed composed of two interlaced increasing sequences—its inferior subsequence and its superior subsequence—and that $b_i < p(b_i)$ for all inferior integers b_i. Moreover, given only the inferior subsequence of a permutation, along with the *position* of each inferior integer, we can reconstruct the permutation. All we need to do is place the inferior integers in their proper position and then fill in the gaps with the rest of the integers, in increasing order.

To illustrate, the permutation

$$\sigma = 41237856 \in S_8^3$$

can be described by the inferior/position sequences

$$S = (12356, 23478)$$

To reconstruct σ, we place the inferior integers in their correct positions:

$$/123//56$$

and then insert the remaining (superior) integers 4, 7, and 8 in the empty positions, in increasing order. Let us call this process the **interlacing procedure**.

The interlacing procedure and Theorem 10.4 imply that the map θ taking a 321-avoiding permutation $\sigma \in S_n$ to the pair of increasing sequences

$$\theta(\sigma) = (b_1 \cdots b_k, p(b_1) \cdots p(b_k))$$

where $\sigma_{inf} = b_1 \cdots b_k$ is injective. Let us show that θ is also surjective.

Theorem 10.5
Let

$$(b_1 \cdots b_k, p_1 \cdots p_k)$$

be a pair of increasing sequences in $[n]$ *with* $b_i < p_i$ *for all* i*. Then* $b_1 \cdots b_k$ *is the inferior subsequence of a 321-avoiding permutation* σ *with* $p(b_i) = p_i$ *for all* i*.*

Proof Begin the creation of σ by placing each b_i at position p_i. Then fill in the gaps with the remaining integers

$$\{c_1, \ldots, c_{n-k}\} = [n] \setminus \{b_1, \ldots, b_k\}$$

in increasing order. Since σ is composed of two interleaved increasing sequences, it is 321-avoiding.

To see that each b_i is inferior, note that there are $p(b_i) - 1 > b_i - 1$ positions to the left of b_i and only $b_i - 1$ integers less than b_i and so at least one integer to the left of b_i must be greater than b_i. Hence, b_i is inferior.

To see that each c_i is superior, suppose that c_i is inferior for some $i > 1$ (c_1 is superior). Then c_i must be preceded by a larger integer, which must be one of the b's, say $b_j > c_i$ precedes c_i. But since b_j is inferior, it must be preceded by a $c_k > b_j$ and so we have

$$c_k > b_j > c_i$$

which is a contradiction to the increasing nature of the c's. Thus, $b_1 \cdots b_k$ is the inferior subsequence of σ. □

Thus, the map θ taking the 321-avoiding permutations in $\mathcal{F}_n$ to pairs of increasing sequences

$$(b_1 \cdots b_k, p_1 \cdots p_k), \quad b_i < p_i$$

is bijective. Of course, such pairs of sequences can be characterized as antichains of intervals

$$\mathcal{F} = \{[b_i, p_i] \,|\, 1 \le i \le k\}$$

in $\text{Int}([n])$ with no singletons. By reducing p_i by one, these correspond bijectively to antichains

$$\mathcal{A} = \{[b_i, p_i - 1] \,|\, 1 \le i \le k\}$$

on $\text{Int}([n-1])$ with no restrictions. Theorem 8.3 then gives the following.

Theorem 10.6
C_n counts the number of

1) *321-avoiding permutations of size n*
2) *123-avoiding permutations of size n* □

Permutation Summary

Theorem 10.7
C_n counts the number of abc-avoiding permutations of $[n]$, for any given three-digit pattern abc. □

Catalan Numbers and Semiorders 11

(The Appendix of this book contains a brief introduction to the subject of partial orders for those who are interested.)

The Definition of Semiorder

Semiorders are intended to model **personal preference** and play a large role in decision theory. A lack of preference between two items is referred to as **indifference**. For convenience, we will say that two items are *indifferent* when we really mean that the decision maker is indifferent to the two items.

One aspect of personal preference that seems to be born out by observation is that *indifference is not transitive*. For example, we are surely indifferent to two pieces of identical chocolate whose weights differ by 10^{-10} grams because this difference is totally imperceptible. Therefore, if indifference were transitive, then we would be indifferent to two pieces of chocolate whose weights differed by one ounce. How ridiculous is that?

In 1956, R. Duncan Luce proposed the following definition for a semiorder.

Definition 11.1

Let $\prec$ and $\sim$ be two binary relations on a nonempty finite set S. Read $a \prec b$ as "b is preferred to a" and $a \sim b$ as "b is indifferent to a." Then $(\prec, \sim)$ is a **semiorder** on S if the following hold:

S1) (Indifference corresponds to nonpreference) For each pair $a, b \in S$, exactly one of the following holds:

$$a \prec b, \quad b \prec a \quad or \quad a \sim b$$

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1_11

S2) (Indifference is reflexive) For all $a \in S$,

$$a \sim a$$

S3) (Indifference is compatible with transitivity) For $a, b, c, d \in S$,

$$a \prec b \sim c \prec d \quad \Rightarrow \quad a \prec d$$

S4) (Indifference is not too liberal) If $a \prec b \prec c$, then no $d \in S$ can be indifferent to all three elements a, b, and c.

If $(\prec, \sim)$ is a semiorder on S, we say that S is **semiordered** by $(\prec, \sim)$ or that $(S, \prec, \sim)$ is a **semiordered set**. □

In the theory of partial order, if $(S, \leq)$ is a partially ordered set, we say that a, $b \in S$ are *comparable* if $a \leq b$ or $b \leq a$. Thus, an element is comparable to itself, which certainly reflects the natural meaning of the term *comparable*. However, since preference is a *strict* relation, no element is preferred to itself and so we will shun the term *comparable*.

Instead, for a semiordered set $(S, \prec, \sim)$, we say that a, $b \in S$ are **preferentially related** if $a \prec b$ or $b \prec a$. In view of property S1), we do not need a new symbol for preferential relationships, since

$$(a \prec b \quad \text{or} \quad b \prec a) \quad \text{iff} \quad a \nsim b$$

In this case, we also say that the pair (a, b) is preferentially related. Thus, an element of S is *not* preferentially related to itself.

Here are some direct consequences of the axioms of a semiorder.

1) Preference is transitive and irreflexive and is thus a strict partial order on S.
2) Preference determines indifference and so completely determines the semiorder, since being indifferent is equivalent to being non-preferentially related, in symbols,

$$a \sim b \quad \text{iff} \quad (a \nprec b \quad \text{and} \quad b \nprec a)$$

3) Indifference is symmetric.

In view of these facts, we can adopt an equivalent definition of semiorder that places the concept in the more familiar context of partially ordered sets. Specifically, a semiordered set $(S, \prec)$ is a partially ordered set $(S, \preceq)$ under the "preference or equality" partial order

$$a \preceq b \quad \text{if} \quad (a = b \quad \text{or} \quad a \prec b)$$

Moreover, preference satisfies the additional properties S1)–S4).

Conversely, if $(S, \leq)$ is a partially ordered set, then we can define a relation $\sim$ by

$$a \sim b \quad \text{iff} \quad (a = b \quad \text{or} \quad a \| b)$$

Then properties S1) and S2) are automatically satisfied. If S3) and S4) also hold, then $(S, <, \sim)$ is a semiordered set. This makes it clear that a semiordered set is just a special type of partially ordered set.

Definition 11.2 (Semiorder as a Special Type of Partial Order)

Let $(S, \leq)$ be a partially ordered set. Define a relation $\sim$ on S by

$$a \sim b \quad \text{iff} \quad (a = b \quad \text{or} \quad a \| b)$$

Then $(<, \sim)$ is a **semiorder** on S if the following hold:

S3) For $a, b, c, d \in S$,

$$a < b \sim c < d \quad \Rightarrow \quad a < d$$

S4) If $a < b < c$, then no $d \in S$ can satisfy $d \sim a$, $d \sim b$, and $d \sim c$.

If S3 and S4 hold, then $\sim$ is called **indifference**. □

Characterization by Maximal Completely Indifferent Subsets

Let $(\prec, \sim)$ be a semiorder on S. We say that a subset $C \subseteq S$ is **completely indifferent** (**CI**) if every pair of elements of C are indifferent. We say that a subset $M \subseteq S$ is **maximal completely indifferent** (**MCI**) if it is CI and no superset is also CI. Let $\mathcal{M}$ be the family of all MCI subsets of S. The maximality of each member of $\mathcal{M}$ implies that $\mathcal{M}$ is an *antichain* in the power set $\wp(S)$. In particular, if $M \neq M'$ are in $\mathcal{M}$, then both of the set differences $M \setminus M'$ and $M' \setminus M$ are nonempty.

The family $\mathcal{M}$ has several interesting properties. Let $M, M', M'' \in \mathcal{M}$.

P1) $\mathcal{M}$ completely determines indifference, to wit, $a \sim b$ if and only if a and b lie in the same MCI set (although they may also be in different MCI sets).

If $a \sim b$ then $\{a, b\}$ is CI and so is contained in some MCI set. Conversely, if a and b are in the same MCI set, then $a \sim b$.

P2) Existence of preferentially related pairs I: Let $M \neq M'$.

a) Then

$$a \in M \setminus M' \quad \Rightarrow \quad \exists b \in M' \setminus M \quad \text{that is preferentially related to } a$$

In particular, there exists a preferentially related pair $(a, b) \in M \times M'$.

b) All preferentially related pairs $(a,a')\in M\times M'$ satisfy $a\prec a'$ or all preferentially related pairs $(a,a')\in M\times M'$ satisfy $a'\prec a$.

For part a), if not, then a is indifferent to every element of M' and so $M'\cup\{a\}$ is CI, which contradicts the maximality of M'. For part b), if $(a,a')\in M\times M'$ satisfies $a\prec a'$ and $(b,b')\in M\times M'$ satisfies $b'\prec b$, then

$$a\prec a'\sim b'\prec b$$

and so S3 implies that $a\prec b$, which is false.

P3) $\mathcal{M}$ can be totally ordered.

Property P2 implies that we can define a binary relation $\prec$ on $\mathcal{M}$ by setting

$$M\prec M' \quad \text{if} \quad (a\prec a' \text{ for some } a\in M, a'\in M')$$

Moreover, exactly one of the following holds:

$$M\prec M' \quad \text{or} \quad M'\prec M \quad \text{or} \quad M=M'$$

Hence, the relation is irreflexive. To see that it is transitive, suppose that $M\prec M'\prec M''$. Then there are elements

$$m_1\in M, \quad m_{21},m_{22}\in M', \quad m_3\in M''$$

for which

$$m_1\prec m_{21}\sim m_{22}\prec m_3$$

and so S3 implies that $m_1\prec m_3$, whence $M\prec M''$. Thus, $\prec$ is a strict *total* order on $\mathcal{M}$, say

$$\mathcal{M}=\{M_1\prec M_2\prec\cdots\prec M_m\}$$

P4) $\mathcal{M}$ completely determines the preference order and therefore the semiorder on S.

Let $a,b\in S$. Then $a\sim b$ if and only if a and b are in the same member of $\mathcal{M}$. If indifference is ruled out, then $a\prec b$ if and only if $(a,b)\in M\times M'$ for some $M\prec M'$.

P5) Existence of preferentially related pairs II: If $M\prec M'$, then

$$\begin{aligned} a\in M\setminus M' \quad &\Rightarrow \quad \exists b\in M'\setminus M \text{ with } a\prec b\\ a\in M'\setminus M \quad &\Rightarrow \quad \exists a\in M\setminus M' \text{ with } a\prec b \end{aligned}$$

P6) Interval structure: If $M\prec M'\prec M''$, then

$$M\cap M''\subseteq M'$$

Moreover, if $M \cap M'' \neq \emptyset$, then

$$M \cap M'' \subseteq M' \subseteq M \cup M''$$

and so

$$M' \setminus M \subseteq M''$$

If $x \in M \cap M''$ but $x \notin M'$, then, $x \in M \setminus M'$ and $x \in M'' \setminus M'$ and so P5 implies that there exist $a', b' \in M'$ for which $x \prec a' \sim b' \prec x$, which implies by S3 that $x \prec x$, a blatant falsehood.

Next, suppose that $b \in M'$ but $b \notin M \cup M''$. Then $b \in M' \setminus M$ and $b \in M' \setminus M''$ and so there is an $a \in M$ and a $c \in M''$ for which $a \prec b \prec c$. However, any $x \in M \cap M''$ is also in $M \cap M' \cap M''$ and so is indifferent to all three elements a, b, and c, in contradiction to S4.

Property P6 is the key to showing that there is a total order $<$ on S with the following properties:

O1) The order is **preference compatible**, that is, $a \prec b$ implies that $a < b$.
O2) Each set $M \in \mathcal{M}$ is an *interval* under the total order $<$.

Hence, under this total order, $\mathcal{M}$ is a *covering antichain* in Int(S).

We first note that the elements of $U_2 = M_1 \cup M_2$ can be totally ordered following properties O1 and O2, as shown in Figure 11.1. Specifically, we place the elements of $M_1 \setminus M_2$ first (in any order), then the elements of $M_1 \cap M_2$, and finally the elements of $M_2 \setminus M_1$.

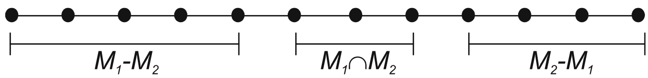


Figure 11.1 Ordering $M_1 \cup M_2$

Suppose that we have totally ordered the elements of $U_{k-1} = M_1 \cup \cdots \cup M_{k-1}$ satisfying properties O1 and O2. We want to include the elements of M_k in this ordering. The elements of M_k that are not yet included, that is, the elements of $M_k \setminus U_{k-1}$ can be attached to the end of the current list in any order. The problem is that some of the elements of M_k may already be in U_{k-1}, that is, $M_k \cap U_{k-1}$ may not be empty and these elements must be "pushed" to the end of the current list so that they can form an interval with the other elements $M_k \setminus U_{k-1}$. Let us refer to the elements of $M_k \cap U_{k-1}$ as **potential offenders**.

Property P6 will come to our aid here, since it implies that if $x \in M_k \cap M_i$ for some $i \leq k-1$, then $x \in M_{k-1}$. Hence, the potential offenders all lie in the largest interval M_{k-1} in the current list U_{k-1}, that is,

$$M_k \cap U_{k-1} \subseteq M_{k-1}$$

Assume that x is the smallest element in $M_k \cap U_{k-1}$ and that $M_i = [a, b]$ is the smallest interval containing x. There are two possibilities to consider.

The first possibility is that $M_i = M_{k-1}$. In this case, $M_k \cap U_{k-1}$ does not intersect any interval comprising U_{k-1} other than the largest one M_{k-1}. This is pictured in Figure 11.2. Now, we may reorder the elements of $M_{k-1} \backslash M_{k-2}$ *arbitrarily* because these elements are pairwise indifferent and so their reordering does not destroy preference compatibility and because they do not cross an interval boundary and so their reordering does not destroy the interval structure created thus far. In particular, we may push all of the potential offenders $M_k \cap U_{k-1}$ to the far right and so, as discussed above, we can adjoin M_k to U_{k-1} as a new interval.

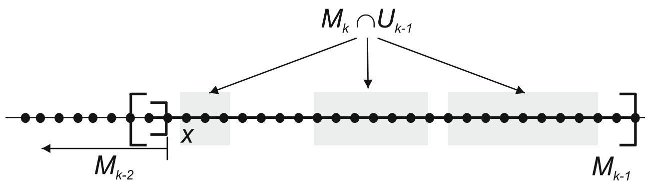


Figure 11.2

The second possibility is that $M_i \prec M_{k-1}$, as shown in Figure 11.3. Then $M_i \prec M_{k-1} \prec M_k$ and so P6 implies that $M_{k-1} \backslash M_i \subseteq M_k$. In other words, *all* of the elements of $M_{k-1} \backslash M_i$ are potential offenders, which is good, because these potential offenders are already “pushed” against the right endpoint of M_{k-1}.

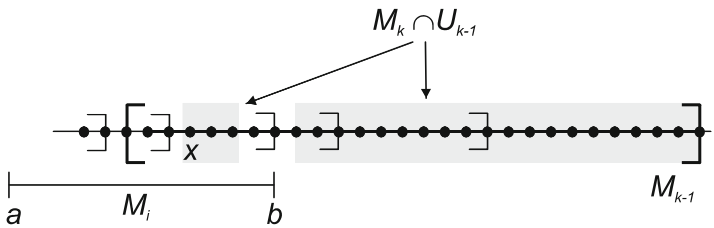


Figure 11.3

Moreover, we can reorder the elements in the interval $[x, b]$ (see Figure 11.3) in any way we desire for the reasons discussed earlier. Hence, we can push all of the elements of $M_i \cap U_{k-1}$ up against the right endpoint of M_i and so, as shown by the shaded portions of Figure 11.4, the elements of $M_k \cap U_{k-1}$ form an interval at the far right end of U_{k-1}.

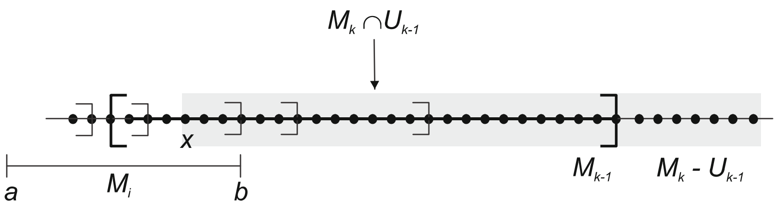


Figure 11.4 M_k is now an interval

Thus, we have shown that there is a preference-compatible total ordering $<$ of S under which $\mathcal{M}$ is a *covering antichain* in Int(S). To be sure, there is some ambiguity in how the total order on S is defined because indifferent elements that do not cross interval borders can be reordered at will. However, we may assume that some method is selected that removes this ambiguity. For example, we may assume that S has a preexisting total order $\ll$ (such as the usual order on the set $S = [n]$) and that whenever the elements $a, b \in S$ with $a \ll b$ can be placed in either order when defining $<$, we always choose the order $a < b$. Simply put, we assure whenever possible that $<$ is compatible with $\ll$.

Again we observe that the semiorder is completely characterized by the covering antichain $\mathcal{M}$ of intervals as follows:

$$\begin{aligned} a \sim b \quad &\text{iff} \quad a, b \in M \text{ for some } M \in \mathcal{M} \\ a \prec b \quad &\text{iff} \quad a \nsim b \text{ and } (a, b) \in M \times M' \text{ where } M \prec M' \end{aligned}$$

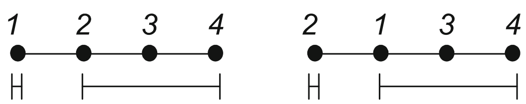
Now suppose we fix a finite set S, which may as well be $S = [n]$. As we scan through all of the semiorders on $[n]$, we can jot them down using the total order to label n equally spaced dots on a horizontal line and then drawing the intervals in $\mathcal{M}$ on that line.

For example, let $n = 4$ and consider the semiorder given by

$$\begin{aligned} &a \sim a \text{ for all } a \in [4] \\ &2 \sim 3 \sim 4 \\ &1 \prec 2, 1 \prec 3, 1 \prec 4 \end{aligned}$$

Under the usual total order on [4], this semiorder is shown on left-hand side of Figure 11.5.

Figure 11.5 Two semiorders



As another example, consider the semiorder given by

$$\begin{aligned}&a \sim a \ \text{ for all } \ a \in [4]\\&1 \sim 3 \sim 4\\&2 \prec 1, 2 \prec 3, 2 \prec 4\end{aligned}$$

Under the total order $2 \ll 1 \ll 3 \ll 4$ on [4], this semiorder is shown on the right-hand side of Figure 11.5.

It should be clear that there is no significant difference between these two semiorders, because their interval structures are the same, to wit, an interval of length one that is less than an interval of length three. More precisely, the bijection

$$\theta(a) = b, \quad \theta(b) = a, \quad \theta(c) = c, \quad \theta(d) = d$$

preserves the semiorder structure (both indifference and preference) and so these two semiorders are order isomorphic.

Therefore, as we scan the semiorders on $[n]$, we are really interested in jotting down a complete list of *nonisomorphic* semiorders. Let us be a bit more precise about this.

Canonical Forms for Semiordered Sets

It is easy to show that order isomorphism is an equivalence relation on the family $\mathcal{S}_n$ of all semiordered sets on $[n]$. This equivalence relation induces a partition on $\mathcal{S}_n$, the blocks of which are called **isomorphism classes**. For example, the isomorphism class that contains the two isomorphic semiorders in Figure 11.5 contains $4! = 24$ semiorders on [4]—just leave the intervals in the left side of Figure 11.5 alone but permute the four integers in [4].

Therefore, to get a complete but nonredundant picture of the semiorder structures on $[n]$, we need a *single* representative from each isomorphism class. Such a representative is called a **canonical form**, a term that is used extensively in linear algebra. A complete set of canonical forms, that is, a set consisting of one canonical form from each isomorphism class, is referred to as a **set of canonical forms**.

Fortunately, we can construct a set of canonical forms for the semiorders on $[n]$ quite easily (at least in theory) by selecting that element of each isomorphism class for which the dots are labeled in the usual order $1, 2, \ldots, n$. (This is the first one in Figure 11.5.)

From the opposite point of view, we can construct "all" semiorders on $[n]$ by writing down n equally spaced dots in a line and numbering them from left to right in the usual order $1, 2, \ldots, n$. We then lay out all possible covering antichains of intervals over these dots. Of course, we must show that *any* covering antichain of $[n]$ belongs to an isomorphism class, that is, defines a semiorder on $[n]$.

Figure 11.6 shows an example of a covering antichain:

$$\mathcal{I} = \{[u_i, v_i] | 1 \le i \le k\}$$

Let us refer to the intervals in $\mathcal{I}$ as **indifference intervals**.

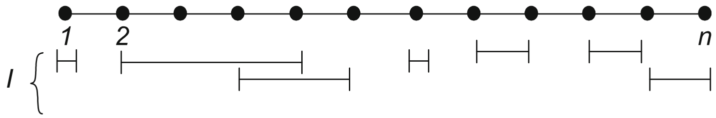


Figure 11.6 A covering antichain

We define a semiorder $(\prec, \sim)$ on $[n]$ by

$$\begin{array}{lll} a \sim b & \text{iff} & a, b \in I \text{ for some } I \in \mathcal{I} \\ a \prec b & \text{iff} & a \nsim b \text{ and } a < b \end{array}$$

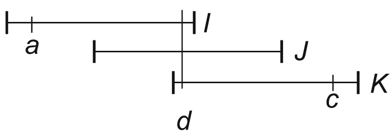
Referring to the definition of semiorder, it is clear that S1 and S2 are satisfied. As to S3), if $a \prec b \sim c \prec d$, then b and c are in the same indifference interval $I = [v_i, v_i]$ and $a < u_i$ and $v_i < d$. Therefore, a and d cannot be in the same indifference interval and so $a \prec d$.

As to S4), if $a \prec b \prec c$ and d is indifferent to a, b, and c, then there are three *distinct* indifference intervals for which

$$a, d \in I = [u_i, v_i], \quad b, d \in J = [u_j, v_j] \quad \text{and} \quad c, d \in K = [u_k, v_k]$$

As to the relative position of these intervals, $a \prec b$ implies that a and b are not in the same interval and $a < b$. Therefore, $u_i < u_j$, that is, I, lies at least partially to the left of J. Similarly, $b \prec c$ implies that J lies at least partially to the left of K. However, since d is in all three intervals, they must overlap, as in Figure 11.7.

Figure 11.7



But where does this leave b? It cannot be in J without being in either I or K, both of which are untrue. This contradiction implies that S4 holds. Thus, we have a semiorder.

Of course, our interest here is in counting the *number* of nonisomorphic semiorders on $[n]$, that is, the number of isomorphism classes. We have proved that this is just the number of covering antichains on $[n]$.

Theorem 11.1
C_n counts the number of semiorders on $[n]$, up to order isomorphism. □

More on Semiorders

As mentioned earlier, there are a variety of ways to characterize semiorders, each of which gives another counting application of the Catalan numbers.

Characterization by Forbidden Subposet

Theorem 11.2
Let $(S, \preceq)$ be a partially ordered set and define a binary relation on S by

$$a \sim b \quad \text{iff} \quad \left(a = b \quad or \quad a||b\right)$$

Then $(S, \prec, \sim)$ is a semiordered set if and only if S has no induced subposets of the form $\mathbf{2} + \mathbf{2}$ or $\mathbf{3} + \mathbf{1}$, as pictured in Figure 11.8.

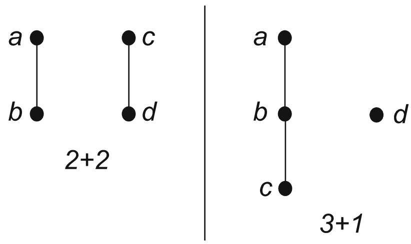


Figure 11.8 Forbidden subposets

Proof Suppose first that S is semiordered by $(\prec, \sim)$. If $a \prec b$ and $c \prec d$, with no other preferences between $a,\ b,\ c$ and d, then

$$a \prec b \sim c \prec d$$

But then S3 implies that so $a \prec d$, a contradiction. Thus, S does not have an induced subposet $\mathbf{2} + \mathbf{2}$. Also, if $a \prec b \prec c$, then S4 implies that for any $d \in S$, there must be a preference between d and at least one of a, b, or c and so the induced subposet $\{a, b, c, d\}$ is not $\mathbf{3} + \mathbf{1}$.

For the converse, suppose that $(S, \preceq)$ has no forbidden subposets. Then S1) and S2) hold. Also, S4) is a direct consequence of the fact that S has no induced subposet of the form $\mathbf{3} + \mathbf{1}$. As to S3), suppose that

$$a \prec b \sim c \prec d$$

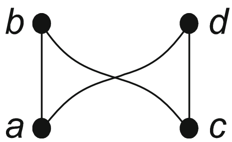
We must rule out the possibilities that $d \prec a$ and $a \sim d$.

1) If $d \prec a$, then

$$c \prec d \prec a \prec b$$

and so transitivity implies that $c \prec b$, which is false.
2) If $a \sim d$, then we have the situation shown in Figure 11.9, where the wavy line indicates indifference.

Figure 11.9



As to the relationship between a and c, if $a \prec c$, then $a \prec c \prec d$ and so $a \prec d$, which is false. If $c \prec a$ then $c \prec a \prec b$ which implies that $c \prec b$, also false. Hence, $a \sim c$. A similar argument shows that $b \sim d$ and so the induced subposet $\{a, b, c, d\}$ is $\mathbf{2} + \mathbf{2}$, contrary to assumption.

Thus, we have ruled out the two undesirable possibilities and are left with $a \prec d$, which is S3. □

Characterization by Unit Interval Order

Let us say that the interval $[i, j]$ is **completely to the left** of the interval $[k, \ell]$ if $j < k$.

Definition 11.3

1) A poset $(P, <)$ is said to have an **interval order** if each $a \in P$ can be assigned an interval $I_a = [\ell_a, r_a]$ on the real line in such a way that $a < b$ if and only if I_a is completely to the left of I_b.
2) A poset $(P, <)$ is said to have a **unit interval order** if each $a \in P$ can be assigned a unit interval $I_a = [\ell_a, \ell_a + 1]$ on the real line in such a way that $a < b$ if and only if I_a is completely to the left of I_b. □

Figure 11.10 shows a unit interval order on [6].

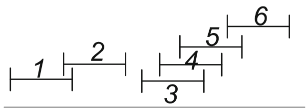


Figure 11.10 A unit interval order on [6]

This order is given by

$$\begin{aligned}1 &< 3,4,5,6\\ 2 &< 3,4,5,6\\ 3 &< 6\\ 4 &< 6\end{aligned}$$

and nothing else.

The Scott-Suppes theorem says that a poset $(P,<)$ has a unit interval order if and only if it is a semiorder, where indifference is the union of equality and noncomparability.

Theorem 11.3 (Scott-Suppes Theorem)
A finite poset $(P,\leq)$ has a unit interval order if and only if $(<,\sim)$ is a semiorder on P, where

$$a \sim b \quad \text{iff} \quad \left(a = b \quad \textit{or} \quad a\middle|\middle|b\right)$$

Proof Suppose first that $(P,\leq)$ has a unit interval order, with unit intervals

$$\mathcal{I} = \left\{I_k = [\ell_k, \ell_k + 1]\middle| k = 1, \ldots, m\right\}$$

Note that $a = b$ or $a\big|\big|b$ if and only if $I_a \cap I_b \neq \emptyset$. We survey the definition of semiorder. It is clear that S1 and S2 hold. As to S3), suppose that

$$a < b\big|\big|c < d$$

If $x \in I_b \cap I_c$, then I_a is completely to the left of x and I_d is completely to the right of x, whence I_a is completely to the left of I_d, that is, $a < b$. As to S4, suppose that

$$a < b < c$$

Then since the intervals I_a, I_b, and I_c all have unit length and do not intersect pairwise, it is clear that no other unit interval can overlap all three of these intervals. Thus, $\left(<, \big|\big|\right)$ is a semiorder on P.

For the converse, it is sufficient to show that a semiorder $(<,\sim)$ on $[n]$ is a unit interval order and this we do by induction on n. If $n = 1$, then the result is clear. Assume that the result holds for all $m < n$.

Let $\mathcal{A}$ be the covering antichain in Int($[n]$) associated with the semiorder. Thus,

$$a \sim b \quad \text{iff} \quad a, b \in I \text{ for some } I \in \mathcal{A}$$
$$a \prec b \quad \text{iff} \quad a \nsim b \text{ and } a < b$$

We wish to identify the maximal elements of $[n]$ under the *semiorder* on $[n]$ (not the usual order on $[n]$). If the interval in $\mathcal{A}$ that contains n is $[k, n]$, then every element $a \in [k, n]$ is a maximal element in $[n]$ because there are no integers greater than n. But if $a \notin [k, n]$, then a and n cannot be in the same interval of the antichain and so $a \prec n$. Hence, $[k, n]$ is the set of maximal elements of S under the semiorder.

Suppose that the second largest interval in $\mathcal{A}$ is $[i, m]$. Then $i < k$ and $m < n$. Let us remove from $\mathcal{A}$ the interval $[k, n]$, along with the integers larger than m, so that the remaining intervals form a covering antichain $\mathcal{A}'$ on $[m]$. We can now identify three groups of elements.

1) $\{m + 1, \ldots, n\}$: The maximal elements of $[n]$ that are lost when we cut back from $[n]$ to $[m]$.
2) $\{k, \ldots, m\}$: The maximal elements of $[n]$ (if any) that survive the cut and are therefore maximal in $[m]$.
3) $\{i, \ldots, k - 1\}$: The *new* maximal elements, that is, elements maximal in $[m]$ but not in $[n]$.

Now, the induction hypothesis implies that the inherited semiorder on $[m]$ is a unit interval order $\mathcal{F}$. As illustrated in Figure 11.11, let $I_i, \ldots, I_k$ be the unit intervals corresponding to the new maximal elements $\{i, \ldots, k\}$ and let $I_k, \ldots, I_m$ be the unit intervals corresponding to the surviving maximal elements $\{k, \ldots, m\}$.

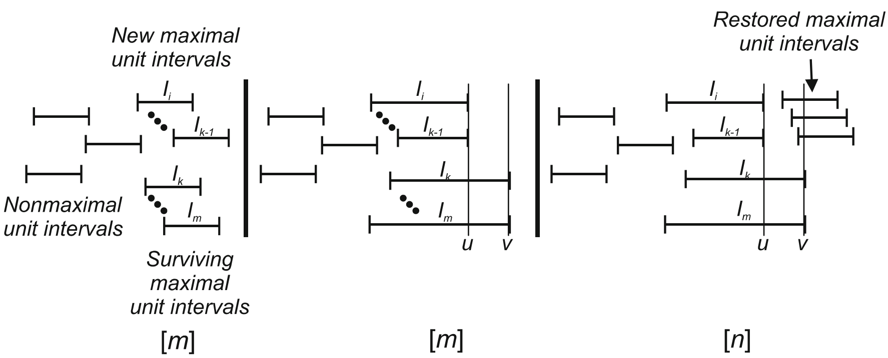


Figure 11.11

Note that the *right* endpoints of the maximal unit intervals do not intersect any nonmaximal unit intervals and so we may extend any of these maximal unit intervals to the right as far as desired without affecting the semiorder in $[m]$.

So as shown in Figure 11.11, let us extend the new maximal unit intervals $I_i, \ldots, I_{k-1}$ to a point u greater than any existing right endpoint and extend the surviving maximal unit intervals to a point $v > u$. Then we can reintroduce the elements $m+1, \ldots, n$ that were cut from $[n]$ by adding new unit intervals $I_{m+1}, \ldots, I_n$ in such a manner that each new interval has its left endpoint in the open interval (u, v). The new unit intervals therefore overlap the surviving unit intervals and each other and no others.

Hence, the indifference relation induced by the totality of unit intervals is that of the original semiorder on $[n]$. Moreover, since the preference order on $[m]$ is induced by the original semiorder on $[n]$ and since the maximality of $m+1, \ldots, n$ is reflected in the unit interval order, the original semiorder is represented by the unit interval order on $[n]$. □

Note that there was nothing special about the fact that all of the intervals in the interval order had unit length. The only salient point is that all intervals have the *same* length.

Semiorder Summary

Theorem 11.4
C_n counts the number of

1) *semiorders on* $[n]$,
2) *partial orders on* $[n]$ *with no induced subposets of the form* $\mathbf{2}+\mathbf{2}$ *or* $\mathbf{3}+\mathbf{1}$,
3) *unit interval orders on* $[n]$. □

Recap

Here is a recap of the objects that we have counted using the Catalan numbers.

Theorem 1
C_n counts the number of

1) *monotonic paths in an $n \times n$ grid that do not rise above the diagonal,*
2) *Dyck paths of length $2n$ that end on the horizontal axis,*
3) *ordered trees with $n+1$ vertices,*
4) *binary trees with n vertices,*
5) *full binary trees with $2n+1$ vertices,*
6) *noncrossing, alternating trees with $n+1$ vertices,*
7) *ways to chord a convex $2n$-gon with nonintersecting chords,*
8) *staircase tilings of an $n \times n$ grid using n tiles,*
9) *noncrossing, alternating, chorded convex $(n+1)$-gons with n chords,*
10) *triangularizations of a convex polygon with $n+2$ sides,*
11) *ways to stack equal-sized disks with n disks on the bottom row,*
12) *ways to fully parenthesize a word of length $n+1$ under a nonassociative binary operation,*
13) *balanced strings of parentheses of length $2n$,*
14) *multisets of size n* in $\mathbb{Z}_{n+1}$ *with null sum,*
15) *separated families of intervals in $Int([n])$,*
16) *covering antichains in $Int([n])$,*
17) *antichains (down sets) in $Int([n-1])$,*
18) *noncrossing partitions of $[n]$,*
19) *noncrossing partitions of $[2n+1]$ into $n+1$ blocks, none of which contain two consecutive integers,*
20) *maximal normalized Davenport-Schinzel sequences on $[n+1]$,*
21) *abc - avoiding permutations of $[n]$, for any given three-digit pattern abc,*
22) *semiorders on $[n]$ (up to isomorphism),*
23) *partial orders on $[n]$ with no induced subposets of the form* $\mathbf{2}+\mathbf{2}$ or $\mathbf{3}+\mathbf{1}$,
24) *unit interval orders on $[n]$.* □

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1

Exercises

As mentioned in the Preface, the following exercises come primarily from Richard Stanley's book *Enumerative Combinatorics*, Volume II, herein denoted by ECII or from his book *Catalan Numbers*, herein denoted by CN. Hints or solutions (that is, longer hints) follow the exercise set.

In each case, show that the Catalan number C_n counts the size of the set described in the exercise.

Paths

1. Figure 1 shows the $C_3 = 5$ Dyck paths of length $2n$ for $n = 3$. Note that there are a total of $14 = C_4$ **bottom points**, that is, points on the horizontal axis. Hmm.

Figure 1 The five Dyck paths of length $2n$ for $n = 3$

2. (CN 32) Dyck paths of length $4n$ such that every **descent** (sequence of consecutive edges with negative slope) has length 2.

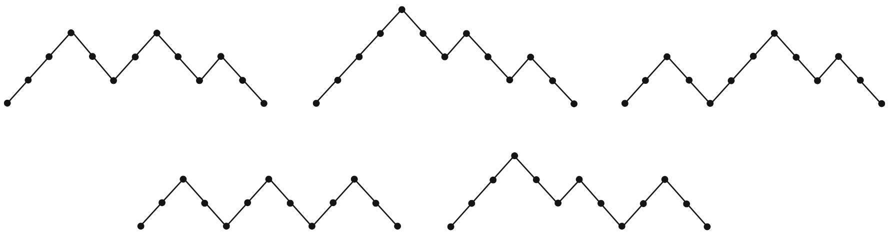


Figure 2 The case $n = 3$

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1

3. (ECII 6.19.k, CN 27) Dyck paths from $(0,0)$ to $(2n+2,0)$ with no peaks (local maxima) at height two.

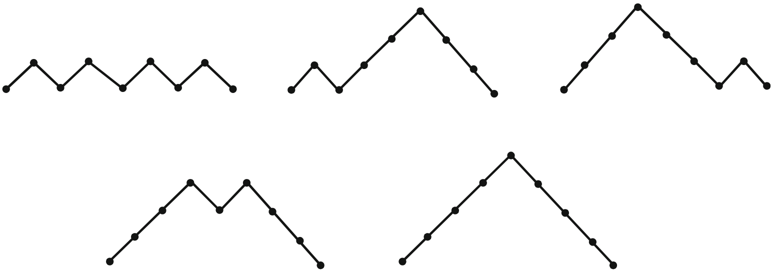


Figure 3 The case $n = 3$

4. The left side of Figure 4 shows the various ways to chord a 4-gon (sometimes also known as a square) with nonintersecting chords where in this case we do not require that all vertices be incident with a chord. The right side of the figure shows the number of paths where we can move diagonally up, diagonally down, or straight ahead. These paths are called **Motzkin paths**.

   Show that the number of ways chordings of a convex n-gon or Motzkin paths of length n is

$$M_n = \sum_{k=0}^{[n/2]} \binom{n}{2k} C_k$$

These are called **Motzkin numbers**.

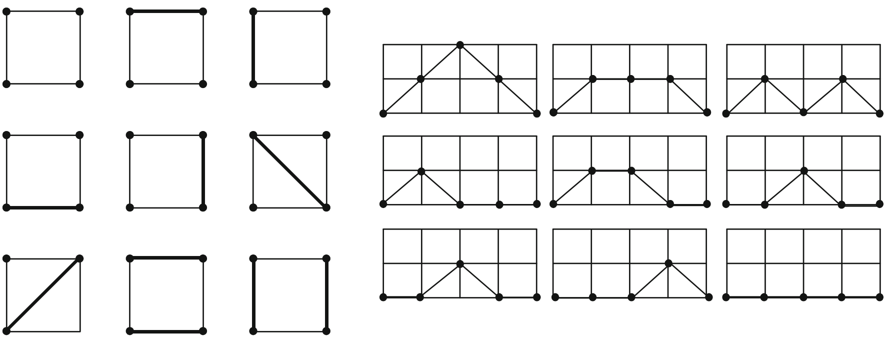


Figure 4

Trees

5. (ECII 6.19f, CN 7) Planted trivalent trees with $2n+2$ vertices. A tree is **planted** if its root has degree one and **trivalent** if all nonleaf vertices (except the root) have degree 3.

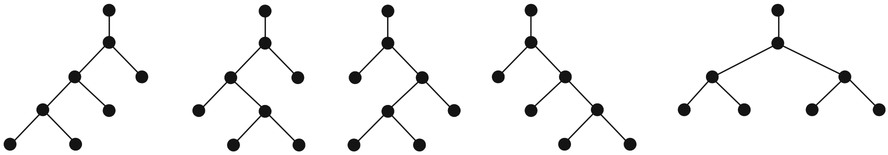


Figure 5 The case $n = 3$

6. (ECII 6.19g, CN 8) Ordered trees with $n+2$ vertices such that the rightmost path of each subtree of the root has an even number of vertices.

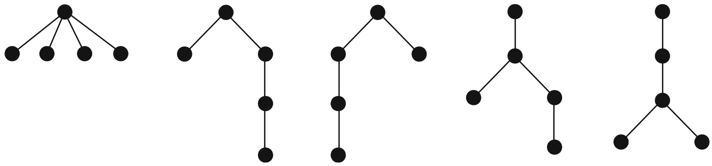


Figure 6 The case $n = 3$

7. (CN 13) Ordered trees for which every vertex has 0, 1, or 3 children and for which the total number of vertices with 0 or 1 children is $n+1$.

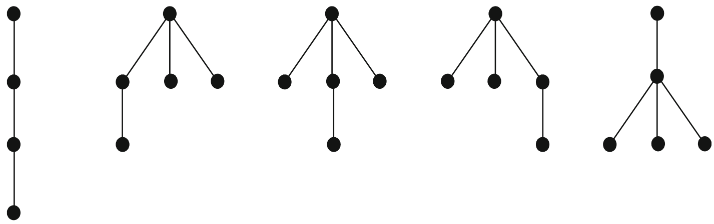


Figure 7 The case $n = 3$

8. (ECII 6.19.o, CN 61) Show that the Catalan number C_n counts the number of ways to connect $2n$ points lying on a line in the plane with n nonintersecting arcs with the property that each arc connects two points and lies above the line. This is called a **noncrossing complete matching** of the vertices.

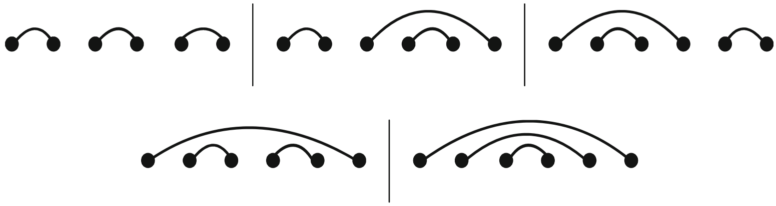


Figure 8 The case $n = 3$

9. (ECII 6.19.q, CN 63) (*Compare with noncrossing, alternating trees.*) Show that the Catalan number C_n counts the number of ways to draw $n + 1$ points lying on a horizontal line L in the plane and n arcs connecting those points with the following properties: (1) the arcs do not pass below L, (2) the graph thus formed is a tree, (3) no arc (including its endpoints) lies strictly below another arc, (4) at every point, all of the arcs exit in the same direction.

Figure 9 The case $n = 3$

Geometry

10. (CN 75) There are a total of $C_3 = 5$ disjoint chordings shown in Figure 6.2, repeated here in Figure 10. There are also a total of five horizontal chords in these chording. Intriguing, isn't it?

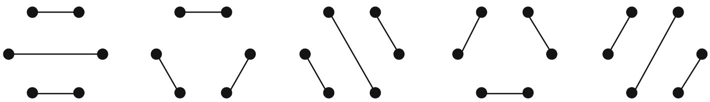


Figure 10 The disjoint chordings of a hexagon

11. (CN 2) Figure 11 shows the 14 triangulations of the hexagon. The triangles marked with a X are the triangles with the property that their vertices consist of vertex 1 together with two other *consecutive* vertices of the hexagon. How many such triangles are there? Is this a coincidence?

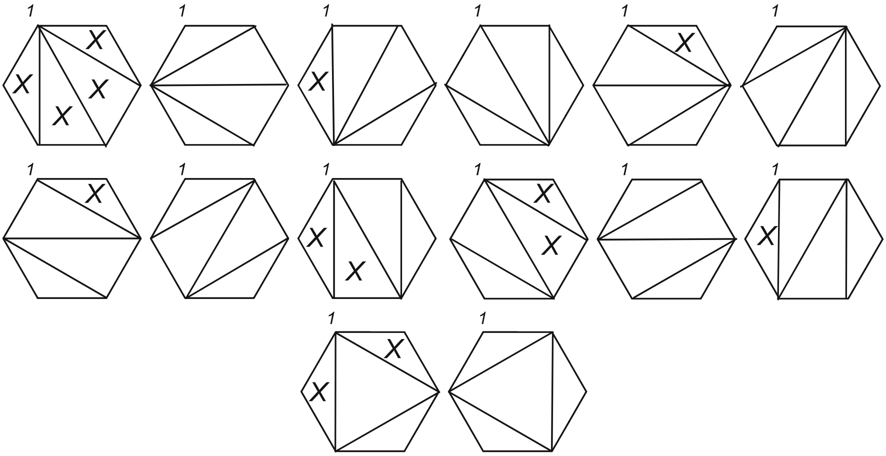


Figure 11

12. (ECII 6.19.l, CN 57) Sets of two monotonic paths on the same square grid with the following properties: (1) both paths start at $(0,0)$ and end at the same point (a,b), (2) each path has $n+1$ steps, and (3) the paths intersect only at their endpoints $(0,0)$ and (a,b). These regions are called **parallelogram polyominoes**

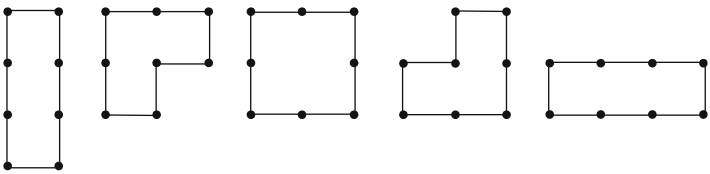


Figure 12 The case $n=3$

13. (ECII 6.19.ggg, CN 60)
 a) Ways to join some of the vertices of a rooted convex $(n-1)$-gon using disjoint line segments and circling a subset of the other vertices.
 b) Prove **Touchard's identity**:

$$C_{n+1} = \sum_{k \geq 0} \binom{n}{2k} C_k 2^{n-2k}$$

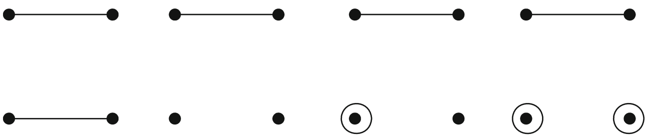


Figure 13 A few of the $C_5 = 42$ possibilities

Integer Sequences

14. (ECII 6.19.s, CN 78) Integer sequences

$$1 \leq a_1 \leq a_2 \leq \cdots \leq a_n$$

for which $a_i \leq i$ for all $i = 1, 2, \ldots, n$.

15. (ECII 6.19.t, CN 79) Strictly increasing positive integer sequences $a_1 a_2 \cdots a_{n-1}$ for which $1 \leq a_i \leq 2i$ for all $i = 1, 2, \ldots, n$.

16. (ECII 6.19.u, CN 80) Positive integer sequences $a_1 a_2 \cdots a_n$ for which $a_1 = 0$ and $0 \leq a_{i+1} \leq a_i + 1$ for all $i = 1, 2, \ldots, n-1$.

Permutations

17. (ECII 6.19.cc, CN 113) Permutations of the multiset

$$\{1, 1, 2, 2, 3, 3, \ldots, n, n\}$$

(there are two copies of each integer in $[n]$) of length $2n$ with the following properties: (1) the first occurrences of each distinct integer in $[n]$ occur in increasing order and (2) there are no *subsequences* of the form *abab*. Here are the permutations for $n = 3$:

$$112233, \quad 112332, \quad 122331, \quad 123321, \quad 122133$$

18. (ECII 6.19.dd, CN 114) Permutations of the set $[2n]$ for which the following hold:
    a) The odd integers appear in increasing order.
    b) The even integers appear in increasing order.
    c) $2k-1$ appears before $2k$ for $1 \leq k \leq n$.

19. (CN 120) Figure 14 shows a structure that has m queues $q_1, \ldots, q_m$ in parallel with an input (source) and an output (sink). The input can be any permutation in S_n and the output is also a permutation in S_n. There are two *legal operations* associated with this structure: (1) move the next integer from the front of the input onto the end of one of the queues and (2) move an integer from the beginning of one of the queues to the end of the output.

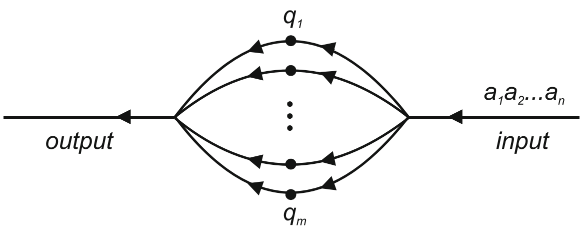


Figure 14 A queue-based sorting structure

A permutation $\sigma \in S_n$ is ***m*-queue sortable** if it is possible to produce the output $\pi = 12 \cdots n$ from the input σ using legal operations. If $\sigma \in S_n$, let $d(\sigma)$ denote the length of the longest decreasing subsequence in σ. Prove that a permutation is m-queue sortable if and only if $d(\sigma) \leq m$, that is, if and only if avoids the pattern $m(m-1)\cdots 1$. In particular, the number of 2-queue sortable permutations is C_n.

Partitions

20. Prove that the Davenport-Schinzel sequences in MNDS(n) have length $2n - 1$.

21. (ECII 6.19.uu, CN 164) **Nonnesting partitions**, that is, partitions of $[n]$ such that if $i < j < k < \ell$ and $i, \ell, \in B$ and $j, k \in C$ where B and C are distinct blocks then there is an $x \in B$ for which $j < x < k$.

Miscellaneous

22. (ECII 6.19.uu, CN 187) Binary relations R on $[n]$ that are reflexive and symmetric and for which if $1 \leq i < j < k \leq n$ and iRk then iRj and jRk. These relations are called **similarity relations**. Here are the similarity relations for $n = 3$:

$$
\begin{aligned}
R_1 &= \{\emptyset\} \\
R_2 &= \{(1,1),(2,2),(3,3),(1,2),(2,1)\} \\
R_3 &= \{(1,1),(2,2),(3,3),(2,3),(3,2)\} \\
R_4 &= R_2 \cup R_3 \\
R_5 &= [3] \times [3]
\end{aligned}
$$

Solutions and Hints

1. *Hint*: To count the number of bottom points in all of the paths, for each path of length n, choose a bottom point, insert a diagonal up-step immediately after that point, and then insert a diagonal down-step at the end of the path. Examine the associated map from bottom points to paths of length $n + 2$.
2. Such a path has exactly $n - 1$ valleys (down-step immediately followed by an up-step). To see this, let u be the number of up-steps, let d be the number of down-steps, and let v be the number of valleys. Then $u = d$ and $u + d = 4n$ and so $d = 2n$. But $v = d/2 - 1 = n - 1$, since the last pair of down-sets does not make a valley. Now, given such a path, delete the first and last steps and every valley (with apologies to Handel). The result is a path of length $4n - 2(n - 1) - 2 = 2n$ and this gives a bijection to the Dyck paths of length $2n$. How do you reverse the deletion of a valley?
4. This result is due to Paul Peart and Wen-Jin Woan and can be found in the *Journal of Integer Sequences* 4 (2001), Article 01.1.3. Here is an outline of their proof using generating functions. They also present a bijection between Dyck paths of length $2n + 2$ with no peaks at height 2 and all Dyck path of length $2n$.
   a) First, show that the generating function $C(x)$ for the Catalan numbers satisfies

$$\frac{1}{1 - x^2C^2(x)} = \frac{C(x)}{1 + xC(x)}$$

   b) Let $D_{n,m}$ be the number of Dyck paths of length $2n$ that have no peaks at height m and let $D_m(x)$ be the generating function for $D_{n,m}$. Show that

$$D_m(x) = \frac{1}{1 - xD_{m-1}(x)}$$

for $m \geq 2$. *Hint*: Any Dyck path of length $2n$ with no peaks at height m starts with an up-step from level 0 to level 1, followed by a Dyck path of length $2k$ starting and ending at level 1 and with no peak at height $m - 1$, followed by a down-step back to level 0, followed by a Dyck path of length $2n - 2 - 2k$ with no peak at height m. Show that the corresponding recurrence for this decomposition is

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1

$$D_{n,m} = \sum_{k=0}^{n-1} D_{k,m-1} D_{n-1-k,m}$$

c) We must now consider D_{n-1}. Show that

$$D_{n,1} = \sum_{k=1}^{n-1} C_k D_{n-1-k,1}$$

Note the lower limit of summation. Show that

$$D_1(x) = \frac{1}{1 - x^2C^2(x)} = \frac{C(x)}{1 + xC(x)}$$

d) Thus,

$$D_2(x) = \frac{1}{1 - xD_1(x)} = 1 + xC(x)$$

and so for $n \geq 1$,

$$D_{n+1,2} = C_n$$

5. *Hint*: Removing the root gives a full binary tree.
6. *Hint*: Stanley says that there is "an elegant bijection" between these trees and ordered trees with $n + 1$ vertices, given to him in a private communication in 1996 by F. Bernhardt.
7. Call a vertex *small* if it has 0 or 1 children. Let $\mathcal{B}_n$ be the family of all such trees with n small vertices and let $B_n = |\mathcal{B}_n|$. Let $T \in \mathcal{B}_n$ have root r. Decompose T by throwing away r. There are two cases (for $n \geq 2$). First, if r is small, then it has degree 1 and so the decomposition yields a single member of $\mathcal{B}_{n-1}$. Second, if r is not small, it has degree 3 and the decomposition yields three trees, say in $\mathcal{B}_i$, $\mathcal{B}_j$, and $\mathcal{B}_k$ where $i + j + k = n$ because the root was (and is) not small. It is also clear that this decomposition is bijective. Hence, for $n \geq 2$,

$$B_n = B_{n-1} + \sum_{i+j+k=n} B_i B_j B_k$$

If the generating function is $b(x) = \sum_{n \geq 0} B_n x^n$, then

$$b(x) = x + xb(x) + b^3(x)$$

since $B_1 = 1$ and $B_0 = 0$. Now, if the generating function of the Catalan numbers is $c(x)$, then let $u(x) = xc(x)$. Since $xc^2(x) - c(x) + 1 = 0$, it follows that $u^2(x) - u(x) + x = 0$ and so $u^3(x) - u^2(x) + xu(x) = 0$, whence $u(x) = x + xu(x) + u^3(x)$ and so $xc(x) = u(x) = b(x)$. Thus, $B_{n+1} = C_n$.

8. *Hint*: Relate these objects to nonintersecting chordings of $2n$ evenly placed points on a circle by cutting the circle at some location on its circumference (not at one of the points) and straightening it into a straight line.
9. *Hint*: This may be a bit hard.
10. Count incidences of chords and slopes. From the point of view of chords, each of the nC_n chords is incident with one slope. From the point of view of slopes, all slopes occur the same number x of times, as can be seen by rotating all of the $2n$-gons so that one slope value is "rotated" into another. Hence, Hence, $xn = nC_n$ and so $x = C_n$.
11. *Hint*: How many triangulations are there that involve a triangle with vertices $1, i, i+1$? Answer: $C_{i-1}C_{n-i+1}$. Then use the recurrence relation.
12. Let $L_1, \ldots, L_k$ be the columns of the polyomino. Let a_i be the number of squares in column L_i and let b_i be the number of squares in common to columns L_i and L_{i+1}. Set $b_0 = b_k = 1$. Define a sequence of charges by

$$\sigma = \left[(+)^{a_1-b_0+1}(-)^{a_1-b_1+1}\right]\ \left[(+)^{a_2-b_1+1}(-)^{a_2-b_2+1}\right]\cdots\left[(+)^{a_k-b_{k-1}+1}(-)^{a_k-b_k+1}\right]$$

where exponentiation means repetition. Then the number of plus signs and minus signs in σ is

$$N_+ = \sum_{i=1}^{k}(a_i - b_{i-1} + 1) = \sum_{i=1}^{k} a_i - \sum_{i=1}^{k} b_{i-1} + k = \sum_{i=1}^{k} a_i - \sum_{i=1}^{k-1} b_i + k - 1$$

and

$$N_- = \sum_{i=1}^{k}(a_i - b_i + 1) = \sum_{i=1}^{k} a_i - \sum_{i=1}^{k} b_i + k = \sum_{i=1}^{k} a_i - \sum_{i=1}^{k-1} b_i + k - 1$$

whence $N_+ = N_-$. Moreover, the sum

$$\sum_{i=1}^{k} a_i - \sum_{i=1}^{k-1} b_i$$

is equal to the number of squares that remain after tossing out all squares that have a neighbor square to its left. Hence, this sum is equal to the number of up-steps in the top path. But k is the number of right steps and so N_+ is equal to one less than the number of steps in each path, that is,

$$N_+ = N_- = n$$

As to the charge at any point, note that the total charge within the ith square bracket defining σ above is

$$(a_i - b_{i-1} + 1) - (a_i - b_i + 1) = b_i - b_{i-1}$$

and so the total charge after m such terms is

$$\sum_{i=1}^{m} (b_i - b_{i-1}) = b_m - b_0 = b_m - 1 \geq 0$$

Since these are the worst-case scenarios, it follows that the charge is nonnegative at all points in σ. Thus, σ is a charge sequence.

13. Let $\mathcal{S}_n$ denote the family of all possible configurations described in part a) and let $s_k = |\mathcal{S}_k|$. Then grouping by the number of chords gives

$$|\mathcal{S}_n| = \sum_{k \geq 0} \binom{n}{2k} C_k 2^{n-2k}$$

To show that $|\mathcal{S}_n| = C_{n+1}$, let $\mathcal{S}_{n,k}$ be the members of $\mathcal{S}_n$ whose root vertex v_1 is connected to vertex v_k by a chord. Then removal of vertices v_1 and v_k and the chord produces a member of $\mathcal{S}_{k-1}$ and a member of $\mathcal{S}_{n-2-k}$. It follows that

$$|\mathcal{S}_{n,k}| = \sum_{k=0}^{n-2} s_k s_{n-2-k}$$

Let $S_n^{'}$ be the members of $\mathcal{S}_n$ for which the root vertex v_1 is not connected to a chord and not circled. Then removal of the root vertex produces a member of $\mathcal{S}_{n-1}^{'}$ and so $\left|\mathcal{S}_n^{'}\right| = |\mathcal{S}_{n-1}|$. Similarly, if $\mathcal{S}_n^{''}$ is the set of members of $\mathcal{S}_n$ for which the root vertex v_1 is circled, then $\left|\mathcal{S}_n^{''}\right| = |\mathcal{S}_{n-1}|$. Hence,

$$s_n = \sum_{k=0}^{n-2} s_k s_{n-2-k} + 2s_{n-1}$$

If the generating function for the sequence $\mathcal{S}_n$ is

$$S(x) = \sum_{n \geq 0} s_n x^n$$

then the previous summation shows that

$$S(x) = 1 + 2xS(x) + x^2 S^2(x)$$

Multiplying both sides by x and letting $U(x) = xS(x)$, we get

$$U(x) = x\big(1 + 2U(x) + U^2(x)\big) = x(1 + U(x))^2$$

and so if $V(x) = U(x) + 1 = xS(x) + 1$, then

$$V(x) - 1 = xV^2(x)$$

whence $V(x) = C(x)$, the Catalan generating function and so

$$C(x) = 1 + xS(x)$$

from which the result follows.

14. *Hint*: Figure 15 shows the five monotonic paths that do not cross the diagonal along with the five integer sequences described in the problem. Feel free to generalize.

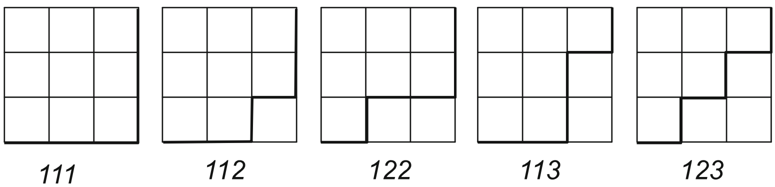


Figure 15

15. This follows fairly easily from Exercise 14, to wit, subtract $i - 1$ from a_i and append a 1 at the beginning of the sequence.
16. Let $b_i = a_i - a_{i+1} + 1$ and replace a_i with a single 1 followed by b_i -1s for $1 \le i \le n$ (where $a_{n+1} = 0$). This produces a ballot sequence.
17. There is a simple bijection between these permutations and the graphs in Exercise 8. Place $2n$ points on a horizontal line and label then with the integers in the permutation in order. Then connect like-labeled points with an arc above the line. Figure 16 shows the process for $n = 3$.

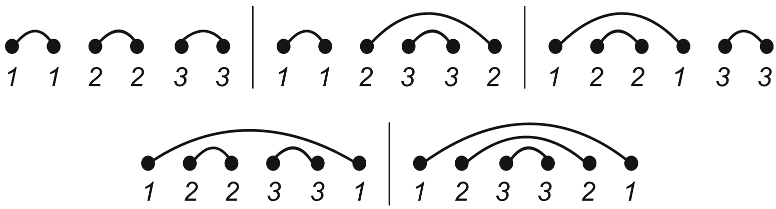


Figure 16

18. *Hint*: Replace each odd number by + and each even number by − to get a charge sequence.

19. First, we observe that if a smaller integer is ever sent through a given queue q_k at some time after a larger integer has gone through q_k, then the output cannot be π. But if $d(\sigma) > m$, then this must happen. Hence, if σ is m - queue sortable, then $d(\sigma) \leq m$. For the converse, suppose that $d(\sigma) \leq m$. Consider the following procedure for loading up the queues with all of the integers in σ.
   a) Place the first integer in queue q_1. Note that the entries at the end of each queue are in decreasing order, since there is only one entry at the end of the queues.
   b) Assume that we have placed $a_1, \ldots, a_{k-1}$ in the queues in such a way that, scanning the last entries in queues q_1 through q_m in that order, the resulting sequence $b_1 \cdots b_m$ is decreasing. Place a_k in the *first* queue q_j (queue with smallest index j) for which $b_j < a_k$. There must be such a queue since otherwise the decreasing subsequence $b_1 \cdots b_m a_k$ of σ would have length greater than m. Moreover, the last queue entries are now

$$b_1 \cdots b_{j-1} a_k b_{j+1} \cdots b_m$$

   where

$$b_m < \cdots < b_{j+1} < b_j < a_k < b_{j-1} < \cdots b_1$$

   and so are still in decreasing order.

   Thus, we may move all of the input integers onto the queues in such a way that each queue contains an increasing sequence of integers (read from the beginning to the end). It is then a simple matter to move the integers to the output to produce the permutation π.

20. Use induction. The result holds for $n = 1$. Assume that it holds for sequences shorter than n. If $w \in \mathrm{MNDS}(n)$, then w has one of the forms $w = 1x1$ where $x \in \mathrm{MNDS}(2, n)$ or $w = 1x1y1$, where

$$x \in \mathrm{MNDS}(2, k) \quad \text{and} \quad 1y\,1 \in \mathrm{MNDS}(\{1\} \cup [k+1, n])$$

   for some $2 \leq k \leq n - 1$. In the former case,

$$\mathrm{len}(w) = 2 + \mathrm{len}(x) = 2 + 2(n-1) - 1 = 2n - 1$$

   In the latter case,

$$\mathrm{len}(w) = 1 + \mathrm{len}(x) + \mathrm{len}(y) = 1 + 2(k-1) - 1 + 2(n-k+1) - 1 = 2n - 1$$

21. Let $\mathcal{P}$ be a nonnesting partition of $[n]$. Write the elements of a block B of $\mathcal{P}$ in increasing order

$$B = \{b_1 < b_2 < \cdots < b_u\}$$

   Let $\mathcal{A}_{\mathcal{P}}$ be the family of all intervals $[b_i, b_{i+1}]$, for all blocks B of $\mathcal{P}$. It is not hard to see that $\mathcal{A}_{\mathcal{P}}$ is an antichain in $[n]$. Clearly, no two intervals in $\mathcal{A}_{\mathcal{P}}$ that come from the same block of $\mathcal{P}$ have a subset relationship. If $[b_i, b_{i+1}]$ comes from block B and $[c_j, c_{j+1}]$ come from a different block C of $\mathcal{P}$ and

$$[b_i, b_{i+1}] \subset [c_j, c_{j+1}]$$

then the endpoints must be distinct since B and C are disjoint. Thus,

$$c_j < b_i < b_{i+1} < c_{j+1}$$

and so the nonnesting property implies that there is a $c \in C$ for which

$$c_j < b_i < c < b_{i+1} < c_{j+1}$$

contradicting the fact that $[c_j, c_{j+1}]$ is in $\mathcal{A}$. Hence, $\mathcal{A}_{\mathcal{P}}$ is an antichain. Note also that all intervals in $\mathcal{A}_{\mathcal{P}}$ have length at least 2.

We can describe the blocks of $\mathcal{P}$ in terms of the antichain $\mathcal{A}_{\mathcal{P}}$ by saying that the blocks of $\mathcal{P}$ are composed of the *maximal* sequences of the form

$$\alpha : a_1 < a_2 < \cdots a_u$$

for which each of the intervals $[a_i, a_{i+1}]$ is in $\mathcal{A}_{\mathcal{P}}$. Moreover, since i and j are in the same block of $\mathcal{P}$ if and only if there is such a maximal sequence containing i and j, the map $\mathcal{P} \mapsto \mathcal{A}_{\mathcal{P}}$ is injective.

Now let $\mathcal{A}$ be an antichain in Int($[n]$) with the property that all intervals have size at least 2. Note that no two distinct intervals in $\mathcal{A}$ have a common left endpoint or a common right endpoint. Consider the family $\mathcal{P}_{\mathcal{A}}$ of all maximal sequences of the form

$$\alpha : a_1 < a_2 < \cdots < a_u$$

for which each of the intervals $[a_i, a_{i+1}]$ is in $\mathcal{A}$. If

$$\beta : b_1 < b_2 < \cdots < b_v$$

is another such maximal sequence and the two sequences have a term $x = a_i = b_j$ in common, then this equality will propagate throughout the sequences as much as possible, that is,

$$a_{i-1} = b_{j-1} \text{ and } a_{i+1} = b_{j+1}$$

assuming that these terms exist. Put another way, all of the integers in the intersection $[a_1, a_u] \cap [b_1, b_v]$ that belong to one of the sequences also belong to the other. Now, if $a_1 < b_1$, then $b_1 \in [a_s, a_{s+1}]$ for some s and so we could extend β by prefixing it with the appropriate choice of $a_1 < \cdots < a_{s-1}$ or $a_1 < \cdots < a_s$, which is impossible. Continuing this argument shows that the two sequences are equal. Hence, the maximal sequences in $\mathcal{P}_{\mathcal{A}}$ are disjoint.

Note that since every interval in $\mathcal{A}$ is part of some maximal sequence, we can recover the antichain $\mathcal{A}$ from the partition $\mathcal{P}$ and so the map $\mathcal{A} \mapsto \mathcal{P}_{\mathcal{A}}$ from antichains to partitions is injective.

We have shown that the number of nonnesting partitions of $[n]$ is the same as the number of antichains in *Int*($[n]$) that have no singleton intervals, which is the same as the number of antichains in $Int([n-1])$, which number is C_n.

22. Since a similarity relation is symmetric, it is completely determined by its *lower half*, that is, by the values iRj for $i \leq j$.

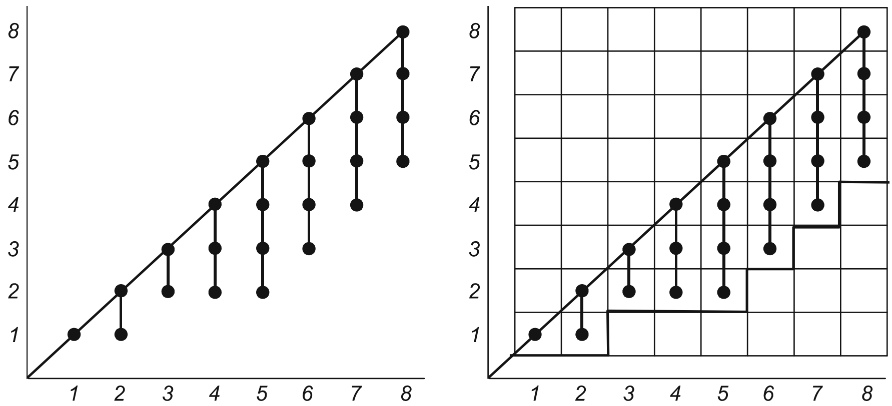


Figure 17

The left side of Figure 17 shows the lower half of the graph of a reflexive and symmetric relation R on $[n]$ for $n = 8$. For each $k \in [n]$, let

$$s_k = \min\{i \in [n] \,|\, iRk\}$$

Then the similarity condition is equivalent to the following two properties:

1) Each $k \in [n]$ is related to every integer between s_k and k (the vertical lines in Figure 17 have no breaks).
2) $s_1 \leq \cdots \leq s_n$

To prove this, first assume that R is a similarity relation. Then $s_k R k$ implies that property 1) holds. As for property 2), if $s_k = k$, then

$$s_{k-1} \leq k - 1 < k = s_k$$

and if $s_k < k$ then $s_k Rk$ implies that $s_k R(k-1)$ and so $s_{k-1} \leq s_k$. Conversely, suppose that R is reflexive and symmetric and satisfies these two properties. If iRk and $i < j < k$, then property 2) implies that

$$s_i \leq s_j \leq s_k \leq i < j < k$$

and so property 1) implies that iRj and jRk.

Now, superimposing an appropriate grid over the left side of Figure 17 gives the right side of Figure 17 and so we glean a bijective correspondence between similarity relations and monotonic paths that do not cross the diagonal.

Appendix

A Brief Introduction to Partially Ordered Sets

For those who are not familiar with partially ordered sets, here is a very brief introduction—just what we need for this book (well, perhaps a tiny bit more). For those who wish a deeper look into this fascinating subject, please allow me to recommend my book *Lattices and Ordered Sets*, published by Springer.

Definition 1

Let A be a nonempty set. A **binary relation** on A is a subset R of the cartesian product $A \times A$. We write $(a,b) \in R$ as $a \sim b$. A binary relation is

1) **reflexive** if for all $a \in A$,

$$a \sim a$$

2) **irreflexive** if for all $a \in A$,

$$a \nsim a$$

3) **symmetric** if for all $a, b \in A$,

$$a \sim b \quad \Rightarrow \quad b \sim a$$

4) **asymmetric** if for all $a, b \in A$,

$$a \sim b \quad \Rightarrow \quad b \nsim a$$

5) **antisymmetric** if for all $a, b \in A$,

$$a \sim b, \quad b \sim a, \quad \Rightarrow \quad a = b$$

6) **transitive** if for all $a, b \in A$,

$$a \sim b, \quad b \sim c, \quad \Rightarrow \quad a \sim c$$

□

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1

Note that asymmetry implies irreflexivity, since asymmetry and $a \sim a$ imply $a \nsim a$, which is absurd. Moreover, if $\sim$ is transitive, then the converse holds. In particular, f $\sim$ is irreflexive and $a \sim b$, then $b \sim a$ implies that $a \sim b \sim a$ and transitivity give $a \sim a$, which is false. Thus, $\sim$ is asymmetric.

Definition 2 A **partial order** (or just an **order**) on a nonempty set P is a binary relation $\leq$ on P that is reflexive, antisymmetric, and transitive, specifically, for all $x, y, z \in P$:

1) (**reflexive**)

$$x \leq x$$

2) (**antisymmetric**)

$$x \leq y, \quad y \leq x \quad \Rightarrow \quad x = y$$

3) (**transitive**)

$$x \leq y, \quad y \leq z \quad \Rightarrow \quad x = z$$

The pair $(P, \leq)$ is called a **partially ordered set** or **poset**, although it is often said that P is a poset, when the order relation is understood. If $x \leq y$, then x is **less than or equal to** y or y is **greater than or equal to** x. If $x \leq y$ but $x \neq y$, we write $x < y$ or $y > x$. If $x \leq y$ or $y \leq x$, then x and y are said to be **comparable**. Otherwise, x and y are **incomparable**, denoted by $a \| b$. □

If S and T are subsets of a poset P, then $S \leq T$ means that $s \leq t$ for all $s \in S, t \in T$. If $T = \{t\}$, then $S \leq \{t\}$ is written $S \leq t$ and similarly for $s \leq T$.

Definition 3 A poset $(P, \leq)$ is **totally ordered** or **linearly ordered** if every $x, y \in P$ are comparable, that is,

$$x \leq y \quad or \quad y \leq x$$

In this case, the order is said to be **total** or **linear**. □

Example 1

1) Let $\mathbb{N} = \{0, 1, \ldots\}$ be the set of natural numbers. Then $(\mathbb{N}, \leq)$ is a poset under the ordinary order. Also, $(\mathbb{N}, |)$ is a poset under division, that is, where $x | y$ means that $y = kx$ for some $k \in \mathbb{N}$, that is, x divides y.
2) If X is a nonempty set, then the **power set** $\wp(X)$ of X is the set of all subsets of X. It is well known that $\wp(X)$ is a poset under set inclusion.
3) The set $\mathcal{P}$ of all partitions of a nonempty set X is a poset, where $\lambda \leq \sigma$ if λ is a **refinement** of σ, that is, if every block of σ is a union of blocks of λ. Put another

way, the blocks of λ are constructed by further partitioning some (or none) of the blocks of σ.

4) The set $\boldsymbol{n} = \{0, 1, \ldots, n-1\}$ is linearly ordered by ordinary order. $\square$

The Product and Sum of Posets

The cartesian product $P \times Q$ of two posets $(P, \leq)$ and $(Q, \preceq)$ can be made into a poset in two natural ways. The **product order** on $P \times Q$ is defined by

$$(p_1, q_1) \leq (p_2, q_2) \quad \text{if} \quad p_1 \leq p_2 \quad \text{and} \quad q_1 \preceq q_2$$

The set $P \times Q$ with this order is called the **product** of P and Q. On the other hand, the **lexicographic order** on the cartesian product $P \times Q$ is defined by

$$(p_1, q_1) \leq (p_2, q_2) \quad \text{if} \quad p_1 < p_2 \quad \text{or} \quad (p_1 = p_2 \quad \text{and} \quad q_1 \preceq q_2)$$

This is also a partial order on $P \times Q$.

If $(P, \leq)$ and $(Q, \preceq)$ are posets, their **sum** $P + Q$ is the poset formed by taking the disjoint union $P \uplus Q$ of P and Q (to avoid any duplicate elements). As to the order on $P \uplus Q$, two elements that both belong to P or both belong to Q have the same relationship that they had in either $(P, \leq)$ or $(Q, \preceq)$ and that $p \| q$ for all $p \in P$ and $q \in Q$.

Induced Subposets

If $(P, \leq)$ is a poset and $S \subseteq P$, then the order relation is also a binary relation on S and so $(S, \leq)$ is a poset in its own right. It is called the **subposet** of P induced by S.

Strict Orders

To every partial order on a set P there corresponds a strict partial order.

Definition 4 A binary relation $<$ on a nonempty set P is called a **strict partial order** (or **strict order**) if it is asymmetric and transitive, or equivalently, irreflexive and transitive. $\square$

Given a partially ordered set $(P, \leq)$, we may define a strict order $<$ by

$$x < y \quad \text{if} \quad x \leq y \ \text{ and } \ x \neq y$$

Conversely, if $<$ is a strict order on a nonempty set P, then the binary relation defined by

$$x \leq y \quad \text{if} \quad x < y \ \text{ or } \ x = y$$

is a partial order and so $(P, \leq)$ is a poset. Thus, there is a one-to-one correspondence between partial orders and strict partial orders on a nonempty set P and so a partially ordered set can be defined as a nonempty set with a strict order relation.

Chains and Antichains

Totally ordered subsets of a poset play an important role in the theory of partial orders.

Definition 5 Let $(P, \leq)$ be a poset.

1) A nonempty subset S of P is a **chain** in P if S is totally ordered by $\leq$. A finite chain with n elements can be written in the form

$$c_1 < c_2 < \cdots < c_n$$

Such a chain is said to have **length** $n - 1$. If $a < b$, then a **chain from *a* to *b*** in P is a chain in P whose smallest element is a and whose largest element is b. A **maximal chain** from a to b is a chain from a to b that is not contained in a larger (in the sense of set inclusion) chain from a to b.
2) A nonempty subset S of P is an **antichain** in P if every two elements of S are incomparable. An antichain with n elements is said to have **width** n. A **maximal antichain** is an antichain that is not contained in a larger (in the sense of set inclusion) antichain. □

Maximal and Minimal Elements

Maximal and maximum elements can be defined in posets.

Definition 6 Let $(P, \leq)$ be a partially ordered set.

1) An element $m \in P$ is **maximal** if no element of P is larger than m, that is,

$$p \in P, \quad m \leq p \quad \Rightarrow \quad m = p$$

An element $m \in P$ is **maximum** (**largest** or **greatest**) if it is greater than every other element of P, that is,

$$p \in P \quad \Rightarrow \quad p \leq m$$

2) An element $n \in P$ is **minimal** if no element is smaller than n, that is,

$$p \in P, \quad p \leq n \quad \Rightarrow \quad p = n$$

An element $n \in P$ is **minimum** (**smallest** or **least**) if it is smaller than all other elements of P, that is,

$$p \in P \quad \Rightarrow \quad n \leq p$$

A partially ordered set is **bounded** if it has both a 0 and a 1. □

Definition 7 If a poset P has a smallest element 0, then any cover of 0 is called an **atom** or **point** of P. The set of all atoms of a poset P is denoted by $\mathcal{A}(P)$. A poset with 0 is **atomic** if every nonzero element contains an atom. If P has a 1, then any element covered by 1 is called a **coatom** or **copoint** of P. □

Upper and Lower Bounds

Upper and lower bounds can be defined in a poset.

Definition 8 Let $(P, \leq)$ be a partially ordered set and let $S \subseteq P$.

1) An **upper bound** for S is an element $x \in P$ for which

$$S \leq x$$

2) A **lower bound** for S is an element $x \in P$ for which

$$x \leq S$$ □

Topological Sorting

It is often useful to be able to write down the elements of a finite poset $(P, \leq)$ one at a time with the property that if $a \in P$ is written down before $b \in P$, then either $a < b$ or $a \| b$. The act of writing down the elements of P in a linear fashion is equivalent to defining a *total* order $\preceq$ on P and the property stated above is that

$$a < b \quad \Rightarrow \quad a \preceq b$$

For a *finite* poset, a compatible total order can be found using a simple algorithm called **topological sorting**, implemented by simply taking a minimal element at each stage. In this way, larger (or smaller) elements are input first.

Down-Sets

Definition 9 A **down-set** (also called an **order ideal** by some authors) in a partially ordered set $(P, \leq)$ is a nonempty subset $I \subseteq P$ with the property that

$$s \in I, \quad p \leq s \quad \Rightarrow \quad p \in I$$

Note that a down-set $I \subseteq P$ is also a poset in its own right under the same binary relation $\leq$. □

Definition 10 If $S \subseteq P$, then the **down-set generated** by S is the set of all elements in P that are less than or equal to at least one element of S, that is,

$$\downarrow S = \{p \in P | p \leq s \ for \ some \ s \in S\}$$ □

A down-set I in a finite poset $(P, \leq)$ is generated by the collection $\mathcal{M}_I$ of all maximal elements of I, where we mean maximal within the poset I, not the poset P. Note that $\mathcal{M}_I$ is an antichain in P. In fact, given any antichain A in P, the set $\downarrow A$ is a down-set whose maximal elements are the elements of A. Since each down-set I is uniquely determined by its maximal elements $\mathcal{M}_I$, the down map $A \mapsto (\downarrow A)$ is a bijection between antichains A in P and down-sets I of P.

Monotone Maps

Order-preserving maps are defined as follows.

Definition Let P and Q be posets and let $f : P \to Q$.

1) f is **order-preserving** or **monotone** if

$$x \leq y \quad \Rightarrow \quad f(x) \leq f(y)$$

   and **strictly monotone** if

$$x < y \quad \Rightarrow \quad f(x) < f(y)$$

2) f is an **order embedding** if

$$x \leq y \quad \Leftrightarrow \quad f(x) \leq f(y)$$

   Note that such a map must be injective.
3) An order embedding f is an **order isomorphism** if it is also surjective. If f is an order isomorphism, we say that P and Q are **order isomorphic**. □

Note that if $f : P \to Q$ is a monotone bijection, then f^{-1} need not be monotone, that is, f need not be an order isomorphism. (Map two incomparable elements to two comparable elements.)

A Brief Introduction to Graphs and Trees

For those who are not familiar with graphs and trees, here is a very brief introduction—just what we need for this book (well, perhaps a tiny bit more).

Definition 11

1) A **graph** $G = (V, E)$ is an ordered pair, where V is any nonempty finite set and E is a set of two-element subsets of V.
2) The elements of V are called the **vertices** or **nodes** of the graph and the elements of E are called the **edges** of the graph. We denote the set of edges of G by $\mathscr{E}(G)$ and the set of vertices by $\mathcal{V}(G)$. The number $|G| = |\mathcal{V}(G)|$ of vertices of G is called the **size** of the graph.
3) Some definitions of a graph allow for **loops**, that is, subsets of V of size 1. Also, some definitions allow for multiple edges between vertices. A graph with no loops or multiple edges is called a **simple graph**. Our graphs will be simple unless explicitly stated otherwise. □

It is customary to draw graphs in the plane, where each vertex is represented by a point and each edge by a line segment between the two vertices in the edge. For example, the graph in Figure 18 has seven vertices, labeled $v_1, \ldots, v_7$, and eight edges, one of which is $\{v_1, v_2\}$. There is also a loop at vertex v_7.

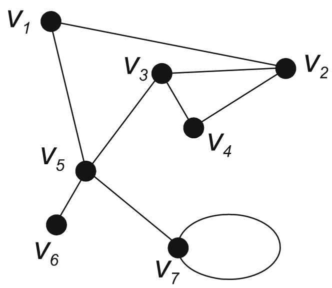


Figure 18 A graph with seven vertices

One of the most important classes of graphs are the complete graphs. The **complete graph** of order n, denoted by K_n, is the graph that has n vertices and one edge connecting each pair of distinct vertices. Some examples of complete graphs are shown in Figure 19.

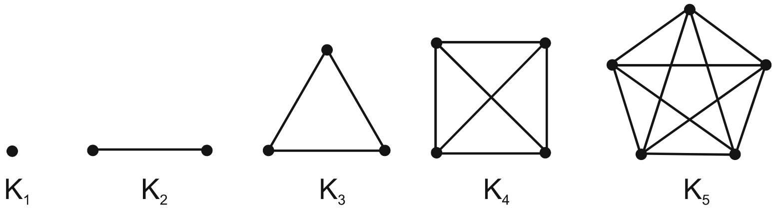
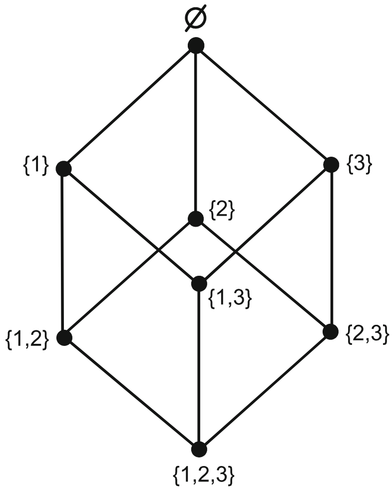


Figure 19 Complete graphs

Graphs are extremely useful in applications ranging from mathematics to urban planning, but we will give only two examples.

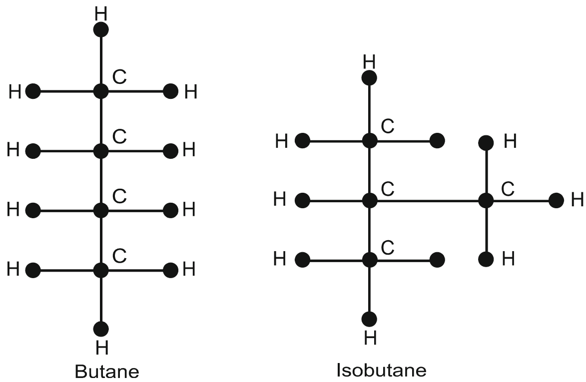
Example 2 Graphs can be used to describe finite partially ordered sets, such as the power set $\wp(S)$ of a set S. As an example, the graph in Figure 20 describes the power set of the set $S = \{1, 2, 3\}$. Note that the edges of the graph describe the covering relation on $\wp(S)$. □

Figure 20 The lattice structure of the power set $\wp(\{1, 2, 3\})$



Example 3 In 1857, the mathematician Arthur Cayley (1821–1895) used graphs to help describe and enumerate the number of isomers of the hydrocarbon molecules C_nH_{2n+2}. (*Isomers* are compounds that have the same chemical formula but different structural formulas.) As an example, the molecule C_4H_{10} has two isomers, called *butane* and *isobutane*, as shown in Figure 21. □

Figure 21 Isomers of C_4H_{10}



Adjacency, Incidence, and Degree

Definition 12 Let G be a graph, with vertex set $\mathcal{V}(G)$ and edge set $\mathcal{E}(G)$. Two vertices $v, w \in \mathcal{V}(G)$ are **adjacent** if $\{v, w\} \in \mathcal{E}(G)$. Two edges of G are **adjacent** if they share a common vertex. A vertex $v \in \mathcal{V}(G)$ and an edge $e \in \mathcal{E}(G)$ are **incident** if $v \in e$. □

Definition 13 The **degree** of a vertex v, denoted by deg(v), is simply the number of times an edge of G meets v. This is not quite the same as the number of edges that are incident with v, since we must count any loops at v twice when determining the degree of v. A vertex of degree zero has no edges incident with it and so it is called an **isolated vertex**. □

You might enjoy proving that if G is a graph and if $\mathcal{V}(G) = \{v_1, v_2, \ldots, v_n\}$ then

$$\deg(v_1) + \deg(v_2) + \cdots + \deg(v_n) = 2|\mathscr{E}(G)|$$

Subgraphs

A graph H is called a **subgraph** of a graph G if H is obtained from G by removing some (or no) edges and some (or no) vertices. However, if a vertex is removed, then all edges incident with that vertex must also be removed. Of course, a graph is a subgraph of itself. A subgraph H of G that is not equal to G is called a **proper subgraph** of G.

Walks, Trails, and Paths

Let u and v be vertices of a graph G. A **walk** from u to v is a sequence of edges of G, each edge incident with the previous edge

$$\{v_1, v_2\}, \{v_2, v_3\}, \ldots, \{v_{n-1}, v_n\}$$

where $v_1 = u$ and $v_n = v$. The edges in a walk need not be distinct. A walk is **closed** if the first and last vertices are the same; otherwise it is **open**.

A **trail** is a walk in which repeated edges are not allowed. However, a vertex may be crossed more than once. A **path** is a trail in which no vertex is crossed more than once. A closed trail is called a **circuit** and a closed path is called a **cycle**. The **length** of a path is the number of *edges* in the path.

A graph is **acyclic** if it has no cycles. An acyclic graph is often called a **forest**.

Connectedness

Two vertices u and v in a graph G are **connected** if there is a path in G from u to v. A graph G in which *every* pair of vertices is connected is called a **connected graph**. A graph that is not connected is said to be **disconnected**. The relation of connectedness on the vertex set $\mathcal{V}(G)$ is an equivalence relation and thus partitions the vertices into blocks, called the **connected components** of the graph. Thus, the connected components are the maximal connected subgraphs of G. The following can be proved by induction.

Theorem 2
Any connected graph with n vertices must have at least $n - 1$ edges. □

Trees

Recall that an acyclic graph is called a forest. An acyclic, *connected* graph is called a **tree**. Figure 22 is an example.

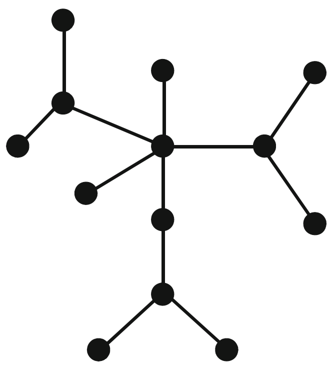


Figure 22 A tree

There are several ways to characterize trees. A **bridge** in a connected graph is an edge e whose removal results in a disconnected graph. Here is a variety of facts about trees.

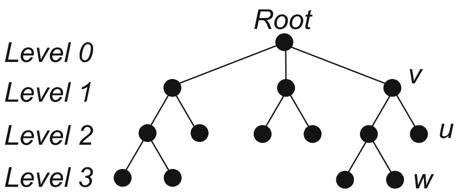
Theorem 3

1) *Any tree T with more than one vertex must have at least two end vertices.*
2) *A tree with n vertices has exactly $n-1$ edges.*
3) *A connected graph G with n vertices is a tree if and only if it has exactly $n-1$ edges.*
4) *An acyclic graph G with n vertices is a tree if and only if it has exactly $n-1$ edges.*
5) *A connected graph G is a tree if and only if every edge of G is a bridge.*
6) *A graph G with no loops is a tree if and only if it has the property that any two distinct vertices in G are connected by exactly one path.*
7) *A graph G is a tree if and only if it is acyclic and has the property that the addition of any new edge to G creates exactly one cycle.* □

Rooted Trees

Trees are often arranged in levels as shown in Figure 23.

Figure 23 A rooted tree



A single vertex is selected and called the **root**. It is placed at the highest level, which is level 0. The vertices adjacent to the root are placed at level 1 and so on. When the tree is drawn in this fashion, it corresponds to a family tree and so it is customary to borrow terminology from that concept. For example, vertex u is a **child** of vertex v, vertex v is a **parent** of vertex u, and vertex w is a **descendant** of vertex v. A tree drawn in this form is called a **rooted tree**. For a *rooted* tree, a vertex with no children is called a **leaf**.

Subtrees

If v is a nonleaf vertex, then the graph consisting of a single child of v and all of the descendants of that child is called a **subtree** of v. For instance, the root r of the tree on the left in Figure 24 has three subtrees, also shown in the figure. The vertex v has two subtrees.

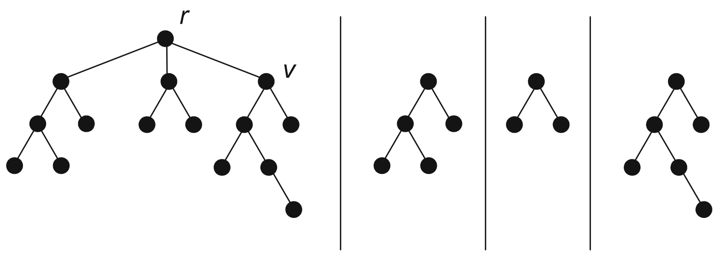
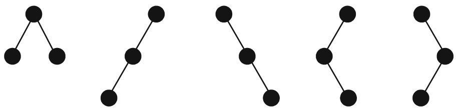


Figure 24 Subtrees

Binary Trees

A **binary tree** is a rooted tree for which each vertex has *at most* two children and in which we distinguish between left and right subtrees. Thus, for example, the five binary trees with 3 vertices are shown in Figure 25. Note that the second and third trees are distinct as binary trees because the second binary tree has a left subtree but no right subtree whereas the third binary tree has a right subtree and no left subtree.

Figure 25 The five binary trees with 3 vertices

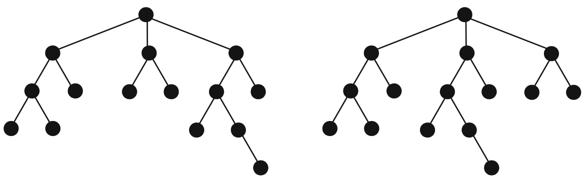


A binary tree in which every nonleaf has exactly two children is called a **full binary tree**. It is also convenient to consider the empty tree as a binary tree.

Ordered Trees

It is often desirable to distinguish order in a rooted tree (as we do in a binary tree, for example). For instance, we may want to think of the two trees in Figure 26 as different because the subtrees of the root are in a different order.

Figure 26 Different or the same?



To do this, we define the concept of an **ordered tree**, which is a rooted tree drawn in the plane in such a way that edges can intersect only at vertices, where we consider the subtrees of each nonleaf as being ordered from left to right. For instance, we would refer to the subtrees in Figure 24, reading from left to right, as the **first subtree**, **second subtree**, and **third subtree** of the root vertex. Using this concept, the two *ordered trees* in Figure 26 are different because, for example, the second subtrees of their root vertices are different. Ordered trees are also referred to as **plane trees**.

Note that there is a distinction between binary trees and ordered trees. For instance, the second and third trees in Figure 25 are distinct as binary trees, which distinguish *left* from *right* but the same as ordered trees, which distinguish only *leftmost* from *rightmost*.

Index

S. Roman, *An Introduction to Catalan Numbers*, Compact Textbooks in Mathematics, DOI 10.1007/978-3-319-22144-1