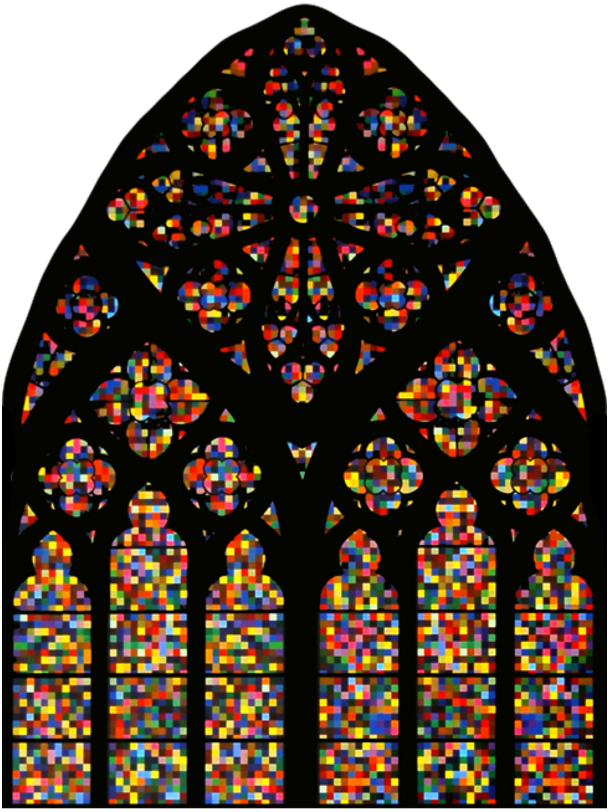
Combinatorics

Enumeration and Structure

MATH 501/2

Alexander Hulpke

Spring 2021



LaTeX on May 7, 2021

Alexander Hulpke
Department of Mathematics
Colorado State University
1874 Campus Delivery
Fort Collins, CO, 80523

Title graphics:

Window in the Southern
Transept of the Cathedral in
Cologne (detail)
Gerhard Richter

These notes are accompanying my course MATH 501/2, Combinatorics, held Fall 2020 and Spring 2021 at Colorado State University.

Contents

Preface

Counting is innate to man.

History of India
Abū Rayhān Muhammad ibn Ahmad
Al-Bīrūnī

This are lecture notes I prepared for a graduate Combinatorics course which I teach in Fall 2020 at Colorado State University.

They started many years ago from an attempt to supplement the book [Cam94] (which has great taste in selecting topics, but sometimes is exceedingly terse) with further explanations and topics, without aiming for the encyclopedic completeness of [Sta12, Sta99].

In compiling these notes, In addition to the books already mentioned, I have benefitted from: [Cam94, CvL91, GR01, LN94, vLW01, Hal86, GKP94, Knu98, Rei84].

You are welcome to use these notes freely for your own courses or students – I'd be indebted to hear if you found them useful.

Fort Collins, Spring 2021
Alexander Hulpke
`hulpke@math.colostate.edu`

Chapter

I Introduction

I.1 What is Combinatorics?

If one looks at university classes in Mathematics that were taught a hundred years ago, we already find the basic pattern of many of the current classes. Single-variable analysis already had most of itfbs current shape. Algebra had started with the formalization of groups, rings and fields, that just a few years later looked very close to what is taught today. Much of the theory of differential equations was known, and numerical methods were just lacking the availability of fast computers. But there was little of combinatorics beyond basic counting formulas used for examples in statistics. Combinatorics suddenly started to grow in the 1950s and 1960s , in part motivated by the existence of computers, and arguably did not get a standard list of topics until the 1980s or 1990s. It differs from many other areas of mathematics in that it was not driven by a small number of deep (often even today not fully solved) problems, but by the observation that problems from seemingly different areas, actually follow similar patterns and can be studied using similar methods. Its scope (in the broadest sense) is the study of the different ways objects can be put in relation to each other, including counting the number of possibilities. This course looks at combinatorics split into two main areas, roughly corresponding to semesters: the first is enumerative combinatorics, the study of counting the different ways configurations can be set up. The second is the study of properties of combinatorial structures that consists of many objects subject to certain prescribed conditions.

I.2 Prerequisites

This being a graduate class we shall assume knowledge of some topics that have been covered in undergraduate classes. In particular we shall use:

Sets, Functions, Relations We assume the reader is comfortable with the concept of sets and standard constructs such as Cartesian products. We denote the set of all subsets of X by $\mathcal{P}(X)$, it is called the *power set* of X.

A relation on a set X is a subset of $X \times X$, functions can be considered as a particular class of relations. We thus might consider a function $f\colon X \to Y$ as a subset of $X \times Y$.

Another important class of relations are equivalence relations. Via equivalence classes they correspond to partitions of the set.

Induction The technique of proof by induction is intimately related to the concept of recursion. It is assumed, that the reader is comfortable with the various variants (different starting values, referring to multiple previous values, postulating a smallest counterexample) of finite induction. We also might sometimes just state that a proof follows by induction, if base case or inductive step are obvious or standard.

Abstract Algebra

Abstract algebra is often useful in providing a formal framework for describing objects. We assume the reader is familiar with the standard concepts from an undergraduate abstract algebra class – groups, permutations (we multiply permutations from left to right), cycle form, polynomial rings, finite fields, and linear algebra.

Graph Terminology

This being a graduate class, the assumption is that the reader has encountered the basic definitions of graph theory – such as: vertex, edge, degree, directed/undirected, path, tree – already in an undergraduate class.

Calculus

It often comes as a surprise to students , that combinatorics – this epitome of discrete mathematics fb uses techniques from calculus. Some of it is the classical use of approximations to estimate growth, but also the toolset for manipulating power series. Still, there is no need to worry about messy approximations, convergence tests, or bathtubs that are filled while simultaneously draining.

I.3 OEIS

A problem that arises often in combinatorics is that we can easily describe small examples, but that it is initially hard to see the underlying patterns. For example, we might be able to count the total number of objects of small size, but will be unable to count how many there are of larger size. In investigating such situations, the *Online Encyclopedia of Integer Sequences* (OEIS, at `oeis.org`) is an invaluable tool that allows to look up number sequences that fit the particular pattern Given by a few values , and for many of these gives a huge number of connections and references . Sequences in this encyclopedia have a fbstorage numberfb starting with the letter ”A”, and we will refer to these at times by an indicator box OEIS A002106 .

Chapter

II

Basic Counting

One of the basic tasks of combinatorics is to determine the cardinality of (finite) classes of objects. Beyond basic applicability of such a number – for example to estimate probabilities – the actual process of counting may be of interest, as it gives further insight into the problem:

- If we cannot count a class of objects, we cannot claim that we know it.

- The process of enumeration might – for example by giving a bijection between classes of different sets – uncover a relation between different classes of objects.

- The process of counting might lend itself to become constructive, that is allow an actual construction of (or iteration through) all objects in the class.

The class of objects might have a clear mathematical description – e.g. all subsets of the numbers $\{1, \ldots, 5\}$. In other situations the description itself needs to be translated into proper mathematical language:

Definition II.1 (Derangements): Given n letters and n addressed envelopes, a *derangement* is an assignment of letters to envelopes that no letter is in the correct envelope. How many derangement of n exist?

(This particular problem will be solved in section III.5)

We will start in this chapter by considering the enumeration of some basic constructs – sets and sequences. More interesting problems, such as the derangements here, arise later if further conditions restrict to sub-classes, or if obvious descriptions could have multiple sequences describe the same object.

II.1 Basic Counting of Sequences and Sets

> I am the sea of permutation.
> I live beyond interpretation.
> I scramble all the names and the combinations.
> I penetrate the walls of explanation.
>
> Lay My Love
> Brian Eno

The number of elements of a set A, denoted by $|A|$ (or sometimes as $\#A$) and called its *cardinality*, is defined[1] as the unique n, such that there is a bijective function from A to $\{1, 2, \ldots, n\}$.

There are three basic principles that underlie counting:

Disjoint Union If $A = A_1 \cup A_2$ with $A_1 \cap A_2 = \varnothing$, then $|A| = |A_1| + |A_2|$.

Cartesian product $|A \times B| = |A| \cdot |B|$.

Equivalence classes If we can represent each element of A in m different ways by elements of B, then $|A| = |B| / m$.

A sequence (or tuple) of length k is simply an element of the k-fold cartesian product. Entries are chosen independently, that is if the first entry has a choices and the second entry b, there are $a \cdot b$ possible choices for a length two sequence. Thus, if we consider sequences of length k, entries chosen from a set A of cardinality n, there are n^k such sequences.

This allows for duplication of entries, but in some cases – arranging objects in sequence – this is not desired. In this case we can still choose n entries in the first position, but in the second position need to avoid the entry already chosen in the first position, giving $n - 1$ options. (The number of options is always the same, the actual set of options of course depends on the choice in the first position.) The number of sequences of length k thus is $(n)_k = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot (n-k+1) = \frac{n!}{(n-k)!}$, called[2] "*$n$ lower factorial k*".

This could be continued up to a sequence of length n (after which all n element choices have been exhausted). Such a sequence is called a *permutation* of A, there are $n! = (n)_n = n \cdot (n-1) \cdot \cdots \cdot 2 \cdot 1$ such permutations.

Next we consider sets of elements. While every duplicate-free sequence describes a set, sequences that have the same elements arranged in different order describe the same set. Every set of k elements from n thus will be described by $k!$ different duplicate-free sequences. To enumerate sets, we therefore need to divide

[1] We only deal with the finite case here, there are generalizations for infinite sets

[2] Warning: The notation $(n)_k$ has different meaning in other areas of mathematics!

by this factor, and get for the number of k-element sets from n the count given by the *binomial coefficient*

$$\binom{n}{k} = \frac{(n)_k}{k!} = \frac{n!}{(n-k)!k!}$$

Note (using a convention of $0! = 1$) we get that $\binom{n}{0} = \binom{n}{n} = 1$. It also can be convenient to define $\binom{n}{k} = 0$ for $k < 0$ or $k > n$.

Binomial coefficients also allow us to count the number of *compositions* of a number n into k parts, that is to write n as a sum of exactly k positive integers with ordering being relevant. For example $4 = 2 + 2 = 1 + 3 = 3 + 1$ are the 3 possible compositions of 4 into 2 parts.

To get a formula of the number of possibilities, write the maximum composition

$$n = 1 + 1 + 1 + \cdots + 1$$

which has $n - 1$ plus-signs. We obtain the possible compositions into k parts by grouping summands together to only have k summands. That is, we designate $k - 1$ plus signs from the given $n - 1$ possible ones as giving us the separation. The number of possibilities thus is $\binom{n-1}{k-1}$.

If we also want to allow summands of 0 when writing n as a sum of k terms, we can simply assume that we temporarily add 1 to each summand. This guarantees that each summand is positive, but adds k to the sum. We thus count the number of ways to express $n + k$ as a sum of k summands which is by the previous formula $\binom{n+k-1}{k-1} = \binom{n+k-1}{n}$.

Example: Check this for $n = 4$ and $k = 2$.

This count is particular relevant as it also counts multisets, that is sets of n elements from k, in which we allow the same element to appear multiple times. Each such set is described as a sum of k terms expressing n, the i-th term indicating how often the i-th element is in the set.

Swapping the role of n and k, We denote by $\left(\binom{n}{k}\right) = \binom{n+k-1}{k}$ the number of k-element multisets chosen from n possibilities.

The results of the previous paragraphs are summarized in the following theorem:

Theorem II.2: The number of ways to select k objects from a set of n is given by the following table:

	Repetition	No Repetition
Order significant (sequences)	n^k	$(n)_k$
Order not significant (sets)	$\binom{n+k-1}{k} = \left(\binom{n}{k}\right)$	$\binom{n}{k}$

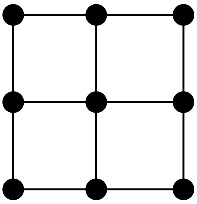


Figure II.1: A 3 × 3 grid

Note II.3: Instead of using the word *significant* some books talk about *ordered* or *unordered* sequences. I find this confusing, as the use of ordered is opposite that of the word *sorted* which has the same meaning in common language. We therefore use the language of *significant*.

II.2 Bijections and Double Counting

> As long as there is no double counting, Section 3(a) adopts the principle of the recent cases allowing recovery of both a complainants actual losses and a misappropriator's unjust benefit…
>
> Draft of Uniform Trade Secrets Act
> American Bar Association

In this section we consider two further important counting principles that can be used to build on the basic constructs.

Instead of counting a set A of objects directly, it might be easiest to establish a bijection to another set B, that is a function $f: A \to B$ which is one-to-one (also called *injective*) and onto (also called *surjective*). Once such a function has been established we know that $|A| = |B|$ and if we know $|B|$ we thus have counted $|A|$.

As an example, consider the following problem: We have an $n \times n$ grid of points with horizontal and vertical connections (depicted in figure II.1 for 3×3) and want to count the number of different paths from the bottom left, to the top right corner, that only go right or up.

Each such path thus has exactly $n-1$ right steps, and $n-1$ up steps. We thus (this already could be considered as one bijection) could count instead 0/1 sequences (0 is right, 1 is up) of length $2n-2$ that contain exactly $n-1$ ones (and zeros). Denote the set of such sequences by A.

To determine $|A|$, we observe that each sequence is determined uniquely by the *positions* of the ones and there are exactly $n-1$ of them. Thus let B the set of all $n-1$ element subsets of $\{1,\ldots,2n-2\}$.

We define $f\colon A \to B$ to assign to a sequence the positions of the ones, $f\colon a_1, \ldots, a_{2n-2} \mapsto \{i \mid a_i = 1\}$.

As every sequence has exactly $n-1$ ones, indeed f goes from A to B. As the positions of the ones define the sequence, f is injective. And as we clearly can construct a sequence which has ones in exactly $n-1$ given positions, f is surjective as well. Thus f is a bijection.

We know that $|B| = \binom{2n-2}{n-1}$, this is also the cardinality of A.
Example: Check this for $n = 2$, $n = 3$.

The second useful technique is to "double counting", counting the same set in two different ways (which is a bijection from a set to itself). Both counts must give the same result, which can often give rise to interesting identities. The following lemma is an example of this paradigm:

Lemma II.4 (Handshaking Lemma): At a convention (where not everyone greets everyone else but no pair greets twice), the number of delegates who shake hands an odd number of times is even.

Proof: we assume without loss of generality that the delegates are given by $\{1, \ldots, n\}$. Consider the set of handshakes

$$S = \{(i, j) \mid i \text{ and } j \text{ shake hands.}\}.$$

We know that if (i, j) is in S, so is (j, i). This means that $|S|$ is even, $|S| = 2y$ where y is the total number of handshakes occurring.

On the other hand, let x_i be the number of pairs with i in the first position. We thus get $\sum x_i = |S| = 2y$. If a sum of numbers is even, there must be an even number of odd summands.

But x_i is also the number of times that i shakes hands, proving the result. □

> About binomial theorem I'm teeming with a lot o' news,
> With many cheerful facts about the square of the hypotenuse.
>
> The Pirates of Penzance
> W.S. Gilbert

The combinatorial interpretation of binomial coefficients and double counting allows us to easily prove some identities for binomial coefficients (which typically are proven by induction in undergraduate classes):

Proposition II.5: Let n, k be nonnegative integers with $k \leq n$. Then:

a) $\binom{n}{k} = \binom{n}{n-k}$.

b) $k\binom{n}{k} = n\binom{n-1}{k-1}$.

c) $\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$ (Pascal's triangle).

d) $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.

e) $\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}$.

f) $(1+t)^n = \sum_{k=0}^{n} \binom{n}{k} t^k$ (Binomial Theorem).

Proof: a) Instead of selecting a subset of k elements we could select the $n-k$ elements not in the set.
b) Suppose we want to count committees of k people (out of n) with a designated chair. We can do so by either choosing first the $\binom{n}{k}$ committees and then for each team the k possible chairs out of the committee members. Or we choose first the n possible chairs and then the remaining $k-1$ committee members out of the $n-1$ remaining persons.
c) Suppose that I am part of a group that contains $n+1$ persons and we want to determine subsets of this group that contain k people. These either include me (and $k-1$ further persons from the n others), or do not include me and thus k from the n other people.
d) We count the total number of subsets of a set with n elements. Each subset can be described by a $0/1$ sequence of length n, indicating whether the i-th element is in the set.
e) Suppose we have n men and n women and we want to select groups of n persons. This is the right hand side. The left hand side enumerates separately the options with exactly k men, which is $\binom{n}{k}\binom{n}{n-k} = \binom{n}{k}^2$ by a).
f) Clearly $(1+t)^n$ is a polynomial of degree n. The coefficient for t^k gives the number of possibilities to choose the t-summand when multiplying out the product

$$(1+t)(1+t)\cdot\dots\cdot(1+t)$$

of n factors so that there are k such summands overall. This is simply the number of k-subsets, $\binom{n}{k}$. □

Example: Prove the theorem using induction. Compare the effort. Which method gives you more insight?

II.3 Stirling's Estimate

Since the factorial function is somewhat unhandy in estimates, it can be useful to have an approximation in terms of elementary functions. The most prominent of such estimates if probably given by Stirling's formula:

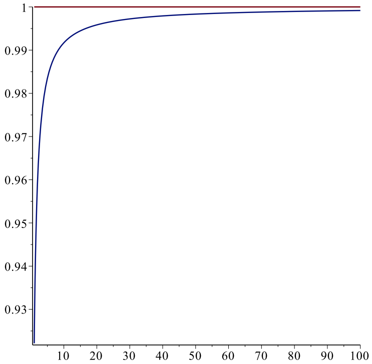


Figure II.2: Plot of $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n/n!$

Theorem II.6:

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(1+O\left(\frac{1}{n}\right)\right). \tag{II.7}$$

Here the factor $1+O(1/n)$ means that the quotient of estimate and real value is bounded by $1 \pm 1/n$. Figure II.2 shows a plot of the ratio of the estimate to $n!$.
Proof: Consider the natural[3] logarithm of the factorial:

$$\log(n!) = \log 1 + \log 2 + \cdots + \log n$$

If we define a step function $L(x) = \log(\lfloor x + 1/2 \rfloor)$, we thus have that $\log(n!)$ is an integral of $L(x)$. We also know that

$$I(x) = \int_0^x \log(t)dt = x\log x - x.$$

We thus need to consider the integral over $L(x) - \log(x)$. To avoid divergence issues, we consider this in two parts. Let

$$a_k = \frac{1}{2}\log k - \int_{k-\frac{1}{2}}^{k} \log x dx = \int_{k-\frac{1}{2}}^{k} \log(k/x)dx,$$

and

$$b_k = \int_k^{k+\frac{1}{2}} \log x dx - \frac{1}{2} logk = \int_k^{l+\frac{1}{2}} \log(x/k)dx.$$

Then

$$S_n = a_1 - b_1 + a_2 - b_2 + \cdots + a_n = \log n! - \frac{1}{2}\log n + I(n) - I(\frac{1}{2}).$$

[3] all logarithms in this book are natural, unless stated differently

A substitution gives

$$a_k = \int_0^{\frac{1}{2}} \log \frac{1}{1-(t/k)} dt, b_k = \int_0^{\frac{1}{2}} \log(1+t/k) dt,$$

from which we see that $a_k > b_k > a_{k+1} > 0$. By the Leibniz' criterion thus S_n converges to a value S and we have that

$$\log n! - (n + \frac{1}{2}) \log n + n \to S - I(\frac{1}{2}).$$

Taking the exponential function we get that

$$n! \to e^C \sqrt{n} \left(\frac{n}{e}\right)^n$$

with $C = \sum_{k=1}^{\infty}(a_k - b_k) - I(\frac{1}{2})$.

Using standard analysis techniques, one can show that $e^C = \sqrt{2\pi}$, see [Fel67]. Alternatively for concrete calculations, we could simply approximate the value to any accuracy desired. □

The appearance of e might seem surprising, but the following example shows that it arises naturally in this context:

Proposition II.8: The number s_n of all sequences (of arbitrary length) without repetition that can be formed from n objects is $\lfloor e \cdot n! \rfloor$.

Proof: If we has a sequence of length k, there are $\binom{n}{k}$ choices of objects each in $k!$ possible arrangements, thus $k!\binom{n}{k} = \frac{n!}{(n-k)!}$ such sequences. Summing over all values of $n-k$ we get

$$s_n = \sum_{k=0}^{n} \frac{n!}{k!} = n! \sum_{k=0}^{n} \frac{1}{k!}.$$

Using the Taylor series for e^x we see that

$$\begin{aligned} e \cdot n! - s_n &= \frac{1}{n+1} + \frac{1}{(n+1)(n+2)} + \cdots \\ &< \frac{1}{n+1} + \frac{1}{(n+1)^2} + \cdots \\ &= \frac{1}{n} < 1. \end{aligned}$$

□

This is an example (if we ignore language meaning) of the popular problem how many words could be formed from the letters of a given word, *if no letters are duplicate.*

Example: If we allow duplicate letters the situation gets harder. For example, consider words (ignoring meaning) made from the letters of the word COLORADO. The letter O occurs thrice, the other five letters only once. If a word contains at most one O, the formula from above gives $\sum_{k=0}^{6}\frac{6!}{k!} = 720+720+360+120+30+5+1 = 1957$ such words.

If the word contains two O, and k other letters there are $\binom{5}{k}$ options to select these letters and $(k+2)!/2$ possibility to arrange the letters (the denominator 2 making up for the fact that both O cannot be distinguished). Thus we get

$$\frac{2!}{2}+\frac{5\cdot 3!}{2}+\frac{10\cdot 4!}{2}+\frac{10\cdot 5!}{2}+\frac{5\cdot 6!}{2}+\frac{7!}{2}=1+15+120+600+1800+2520=5056$$

such words.

If the word contains three O, and k other letters we get a similar formula, but with a cofactor $(k+3)!/6$, that is

$$\frac{3!}{6}+\frac{5\cdot 4!}{6}+\frac{10\cdot 5!}{6}+\frac{10\cdot 6!}{6}+\frac{5\cdot 7!}{6}+\frac{8!}{6}=1+20+200+1200+4200+6720=12341$$

possibilities, summing up to 19354 possibilities in total.

Lucky we do not live in MISSISSIPPI!

II.4 The Twelvefold Way

A generalization of counting sets and sequences is given by considering functions between finite sets. We shall consider functions $f\colon N \to X$ with $|N| = n$ and $|X| = x$. These functions could be arbitrary, injective, or surjective.

What does the concept of order (in)significant mean in this context? If the order is not significant, we actually only care about the *set of values* of a function, but not the values on particular elements. That is, all the elements of N are equivalent, we call them *indistinguishable* or *unlabeled* (since labels will force objects to be different). Otherwise we talk about *distinguishable* or *unlabeled* objects.

Formally, we are counting equivalence classes of functions, in which two functions $f, g\colon N \to X$ are called N-equivalent, if there is a bijection $u\colon N \to N$ such that $f(u(a)) = g(a)$ for all $a \in N$.

Similarly, we define an X-equivalence of function, calling sf and g equivalent, if there is $v\colon X \to X$ such that $v(f(a)) = g(a)$ for all $a \in N$. If we say the elements of X are indistinguishable, we count functions up to X-equivalence.

We can combine both equivalences to get even larger equivalence classes, which are the case of elements of both N and X being indistinguishable.

(The reader might feel this to be insufficiently stringent, or wonder about the case of different classes of equivalent objects. We will treat such situations in Section VII.6 under the framework of group actions.)

This setup (N and X distinguishable or not and functions being injective, surjective or neither) gives in total $3\cdot 2\cdot 2 = 12$ possible classes of functions. This

set of counting problems is called the *Twelvefold Way* in [Sta12, Section 1.9] (and attributed there to Gian-Carlo Rota).
Example: For each of the 12 categories, give an example of a concrete counting problem in common language. Say, we have n balls and x boxes (either of them might be labeled), and we might require that no box has more than one ball, or that every box contains at least one ball.

Say, we have $N = \{1,2\}$ and $X = \{a,b,c\}$. Then we have the following functions $N \to X$:

All functions: There are 9 functions, namely (giving functions as sequences of values): $[a,a]$, $[a,b]$, $[a,c]$, $[b,a]$, $[b,b]$, $[b,c]$, $[c,a]$, $[c,b]$, $[c,c]$.

Injective functions: There are 6 functions, namely $[a,b]$, $[a,c]$, $[b,a]$, $[b,c]$, $[c,a]$, $[c,b]$.

Surjective functions: There are no such functions, but there are 6 functions from $\{1,2,3\}$ to $\{a,b\}$, namely: $[a,a,b]$, $[a,b,a]$, $[a,b,b]$, $[b,a,a]$, $[b,a,b]$, $[b,b,a]$.

Up to permutation of N So we consider only the sets of values, which gives 6 possibilities: $\{a,a\}$, $\{a,b\}$, $\{a,c\}$, $\{b,b\}$, $\{b,c\}$, $\{c,c\}$.

Injective, up to permutation of N The values need to be different, so 3 possibilities: $\{a,b\}$, $\{a,c\}$, $\{b,c\}$.

Surjective, up to permutation of N There are no such functions, but there are 2 such functions from $\{1,2,3\}$ to $\{a,b\}$, namely: $[a,a,b]$, $[a,b,b]$.

Up to permutations of X Since $|N| = 2$, the question is just whether the two values are the same or not: $[a,a]$, $[a,b]$.

Injective, up to permutations of X Here $[a,b]$ is the only such function.

Surjective, up to permutations of X Again, no such function, but from $\{1,2,3\}$ to $\{a,b\}$ there are 3 such functions namely $[a,a,b]$, $[a,b,a]$, $[b,a,a]$.

Up to permutations of N **and** X Again two possibilities, $[a,a]$, $[a,b]$; but if $N = \{1,2,3\}$, there are three possibilities, namely $[a,a,a]$, $[a,a,b]$, $[a,b,c]$.

Injective, up to permutations of N **and** X Again, $[a,b]$ is the only such function.

Surjective, up to permutations of N **and** X Again, no such function, but from $\{1,2,3\}$ to $\{a,b\}$ there is one, namely $[a,a,b]$.

To give expressions for each count, we introduce the following definitions. Determining closed formulae for these is not always easy, and will require further work in subsequent chapters.

A *partition* of a set A is a collection $\{A_i\}$ of subsets (called *parts* or *cells*) $\emptyset \neq A_i \subset A$ such that for all i:

- $\bigcup\limits_i A_i = A$

- $A_i \cap A_j = \varnothing$ for $j \neq i$.

Note that a partition of a set gives an equivalence relation and that any equivalence relation on a set defines a partition into equivalence classes.

Definition II.9: We denote the number of partitions of $\{1, \dots, n\}$ into k (non-empty) parts by $S(n,k)$. It is called the *Stirling number of the second kind*[4] OEIS A0082h7. The total number of partitions of $\{1, \dots, n\}$ is given by the *Bell number* OEIS A000110

$$B_n = \sum_{k=1}^{n} S(n,k).$$

Example: There are $B_3 = 5$ partitions of the set $\{1,2,3\}$.

Again we might want to set $S(n,k) = 0$ unless $1 \leq k \leq n$.

We will give a formula for $S(n,k)$ in Lemma III.5 and study $B(n)$ in section III.6.

In some cases we shall care not which numbers are in which parts of a partition, but only the size of the parts. We denote the number of partitions of n into k parts by $p_k(n)$, the total number of partitions by $p(n)$ OEIS A000041.

Again we study the function $p(n)$ later, however here we shall not achieve a closed formula for the value.

The twelvefold way theorem

We now extend the table of theorem II.2:

Theorem II.10: if N, X are finite sets with $|N| = n$ and $|X| = x$, the number of (equivalence classes of) functions $f\colon N \to X$ is given by the following table. (In the first two columns, d/i indicates whether elements are considered distinguishable or indistinguishable. The boxed numbers refer to the explanations in the proof.):

N	X	f arbitrary	f injective	f surjective
d	d	$\boxed{1}$ x^n	$\boxed{2}$ $(x)_n$	$\boxed{3}$ $x!S(n,x)$
i	d	$\boxed{4}$ $\left(\!\binom{x}{n}\!\right)$	$\boxed{5}$ $\binom{x}{n}$	$\boxed{6}$ $\binom{n-1}{x-1} = \left(\!\binom{x}{n-x}\!\right)$
d	i	$\boxed{7}$ $\sum_{k=0}^{x} S(n,k)$	$\boxed{8}$ 1 if $n \leq x$ 0 if $n > x$	$\boxed{9}$ $S(n,x)$
i	i	$\boxed{10}$ $\sum_{k=0}^{x} p_k(n)$	$\boxed{11}$ 1 if $n \leq x$ 0 if $n > x$	$\boxed{12}$ $p_x(n)$

[4] There also is a Stirling number of the first kind

Proof: If N is distinguishable, we can simply write the elements of N in a row and consider a function on N as a sequence of values. In 1), we thus have sequences of length n with x possible values, in 2) such sequences without repetition, both formulas we already know.

If such a sequence takes exactly x values, each value $f(n)$ can be taken to indicate the cell of a partition into x parts, into which n is placed. As we consider a partition as a set of parts, it does not distinguish the elements of X, that shows that the value in 9) has to be $S(n,x)$. If we distinguish the elements of X we need to account for the $x!$ possible arrangements of cells, yielding the value in 3).

Similarly to 9), if we do not require f to be surjective, the number of different values of f gives us the number of parts. Up to x different parts are possible, thus we need to add the values of the Stirling numbers.

To get 12) from 9) and 10) from 7) we notice that making the elements of N indistinguishable simply means that we only care about the sizes of the parts, not which number is in which part. This means that the Stirling number $S(n,x)$ gets replaced by the partition (shape) count $p_x(n)$.

If we again start at 1) but now consider the elements of N as indistinguishable, we go from sequences to sets. If f is injective we have ordinary sets, in the general case multisets, and have already established the results of 4) and 5).

For 6), we interpret the x distinct values of f to separate the elements of N into x parts. This is a composition of n into x parts, for which the count has been established.

In 8) and 11), finally, injectivity demands that we assign all elements of n to different values which is only possible if $n \leq x$. As we do not distinguish the elements of X it does not matter what the actual values are, thus there is only one such function up to equivalence. □

Chapter

III

Recurrence and Generating Functions

Finding a close formula for a combinatorial counting function can be hard. It often is much easier to establish a recursion, based on a reduction of the problem. Such a reduction often is the principal tool when constructing all objects in the respective class.

An easy example of such a situation is given by the number of partitions of n, given by the Bell number B_n:

Lemma III.1: For $n \geq 1$ we have:

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

Proof: Consider a partition of $\{1, \ldots, n\}$. Being a partition, it must have 1 in one cell. We group the partitions according to how many points are in the cell containing 1. If there are k elements in this cell, there are $\binom{n-1}{k-1}$ options for the other points in this cell. And the rest of the partition is simply a partition of the remaining $n - k$ points. □

III.1 Power Series – A Review of Calculus

A powerful technique for working with recurrence relations is that of generating functions. The definition is easy, for a counting function $f : \mathbb{Z}_{\geq 0} \to \mathbb{Q}$ defined on the nonnegative integers we define the associated generating function as the power series

$$F(t) = \sum_{n=0}^{\infty} f_n t^n \in \mathbb{R}[[t]]$$

Here $\mathbb{R}[[t]]$ is the ring of formal power series in t, that is the set of all formal sums $\sum_{n=0}^{\infty} a_n t^n$i with $a_n \in \mathbb{R}$.

> *When writing down such objects, we ignore the question whether the series converges, i.e. whether F can be interpreted as a function on a subset of the real numbers.*

Addition and multiplication are as one would expect: If $F(t) = \sum_{n\geq 0} f_n t^n$ and $G(t) = \sum_{n\geq 0} g_n t^n$ we define new power series $F + G$ and $F \cdot G$ by:

$$\begin{aligned}(F+G)(t) &:= \sum_{n\geq 0} (f_n + g_n)t^n \\ (F \cdot G)(t) &:= \sum_{n\geq 0} \left(\sum_{a+b=n} f_a g_b \right) t^n.\end{aligned}$$

With this arithmetic $R[[t]]$ becomes a commutative ring. (There is no convergence issue, as the sum over $a + b = n$ is always finite.)

We also define two operators, called *differentiation* and *integration* on $\mathbb{R}[[t]]$ by

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(\sum_{n\geq 0} a_n t^n \right) = \sum_{n\geq 1} (n a_n) t^{n-1}$$

and

$$\int \left(\sum_{n\geq 0} a_n t^n \right) = \sum_{n\geq 0} \frac{a_n}{n+1} t^{n+1},$$

Two power series are equal if and only if all coefficients are equal, an identity involving a power series as a variable is called a *functional equation*.

As a shorthand we define the following names for particular power series:

$$\begin{aligned}\exp(t) &= \sum_{n\geq 0} \frac{t^n}{n!} \\ \log(1+t) &= \sum_{n\geq 1} \frac{(-1)^{n-1} t^n}{n}\end{aligned}$$

and note[1] that the usual functional identities $\exp(a+b) = \exp(a)\exp(b), \log(ab) = \log(a) + \log(b)$, $\log(\exp(t)) = t$ and differential identities $\frac{\mathrm{d}}{\mathrm{d}t}\exp(t) = \exp(t)$, $\frac{\mathrm{d}}{\mathrm{d}t}\log(t) = 1/t$ hold.

[1] All of this can either be proven directly for power series; alternatively, we choose t in the interval of convergence, use results from Calculus, and obtain the result from the uniqueness of power series representations.

A telescoping argument shows that for $r \in \mathbb{R}$ we have that $(1-rt)\left(\sum_n r^n t^n\right) = 1$, that is we can use the *geometric series* identity

$$\sum_n r^n t^n = \frac{1}{1-rt}$$

to embed rational functions (whose denominators factor completely; this can be taken as given if we allow for complex coefficients) into the ring of power series.

For a real number r, we *define*[2]

$$\binom{r}{n} = \frac{r(r-1)\cdots(r-n+1)}{n!} t^n$$

as well as

$$(1+t)^r = \sum_{n\geq 0} \binom{r}{n}.$$

For obvious reasons we call this definition the *binomial formula*. Note that for integral r this agrees with the usual definition of exponents and binomial coefficients, so there is no conflict with the traditional definition of exponents.

We also notice that for arbitrary r, s we have that $(1+t)^r(1+t)^s = (1+t)^{r+s}$ and $\frac{\mathrm{d}}{\mathrm{d}t}(1+t)^r = r(1+t)^{r-1}$.

Up to this point, generating functions seem to be a formality for formalities sake. They however come into their own with the following observations: Operations on the entries of a sequence, relations amongst its entries (such as recursion), or a sequence having been built on top of other sequences often have natural analogues with generating functions. Recursive identities amongst a coefficient sequence then become functional equations or differential equations for the generating functions.

If these equations have solutions, known start values typically give uniqueness of the solution, and (assuming it converges in an open interval), its power series representation must be identical to the generating function. Methods from Analysis, such as Taylor's theorem, then can be used to determine expressions for the terms of the sequence.

Before describing this more formally, lets look at a toy example:

We define a function recursively by setting $f_0 = 1$ and $f_{n+1} = 2f_n$. (This recursion comes from the number of subsets of a set of cardinality n – fix one element and distinguish between subsets containing this element and those that don't.) The associated generating function is $F(t) = \sum_{n\geq 0} f_n t^n$. We now observe that

$$2tF(t) = \sum_{n\geq 0} 2f_n t^{n+1} = \sum_{n\geq 0} f_{n+1} t^{n+1} = F(t) - 1.$$

[2] Unless r is an integer, this is a definition in the ring of formal power series.

We solve this functional equation as

$$F(t) = \frac{1}{1-2t}$$

and (geometric series!) obtain the power series representation

$$F(t) = \sum_{n\geq 0} 2^n t^n.$$

this allows us to conclude that $f_n = 2^n$, solving the recursion (and giving us the combinatorial result we knew already that a set of cardinality n has 2^n subsets.)

Operations on Generating Functions

Lets start with basic linearity. If f_n and g_n are sequences, associated to generating functions $F(t)$ and $G(t)$, respectively, then $\lambda F(t)+G(t)$ is the generating function associated to the sequence $\lambda f_n + g_n$. The case $\lambda = 1$ is sometimes called the *sum rule*, describing the case that f and g count the cardinality of disjoint sets and $f_n + g_n$ the cardinality of their unions.

Shifting Shifting the indices corresponds to multiplication by powers of t.

Differentiation If $F(t) = \sum_{i\geq 0} f_i t^i$, the derivative of $F(t)$ corresponds to the sequence $s_n = (n+1)f_{n+1}$, thus allowing for variable coefficients. This can often be used effectively together with shifts. For example, starting with the sequence $f_n = 1$, it corresponds (geometric series) to the generating function $\frac{1}{1-t}$. Its derivative, $\frac{1}{(1-t)^2}$ thus corresponds to the sequence $1, 2, 3, 4, \ldots$ and $\frac{t}{(1-t)^2}$ then to the sequence $0, 1, 2, 3, \ldots$. Taking a further derivative yields a sequence of squares $1, 4, 9, 16, \ldots$ at generating function $\frac{1+t}{(1-t)^3}$. We shift once more to get the sequence $s_n = n^2$, associated to the generating function $\frac{t(1+t)}{(1-t)^3}$.

Products If we define a new sequence as sum of terms whose indices add up to the desired value:

$$c_n = f_0 g_n + f_1 g_{n-1} + f_2 g_{n-2} + \cdots + f_n g_0$$

(this is sometimes called *convolution*) its generating function is $F(t)\cdot G(t)$. This is called the *product rule.*

One particular case of this is the *summation rule*: Suppose we define $c_n = \sum_{i=0}^n f_i$, we can interpret this as convolution with the constant sequence $g_i = 1$, whose generating function is $\frac{1}{1-t}$. The generating function for the summatory sequence thus is $C(t) = \frac{F(t)}{1-t}$.

Continuing the previous example, the sequence $s_n = \sum_{i=0}^n i^2$ thus has the generating function $S(t) = \frac{t(1+t)}{(1-t)^4}$.

We now make use of these tools by looking at general tools to find closed-form expressions for recursively defined sequences:

III.2 Linear recursion with constant coefficients

Then, at the age of forty, you sit,
theologians without Jehovah,
hairless and sick of altitude,
in weathered suits,
before an empty desk,
burned out, oh Fibonacci,
oh Kummer[1]*, oh Gödel, oh Mandelbrot*
in the purgatory of recursion.

Dann, mit vierzig, sitzt ihr,
o Theologen ohne Jehova,
haarlos und höhenkrank
in verwitterten Anzügen
vor dem leeren Schreibtisch,
ausgebrannt, o Fibonacci,
o Kummer, o Gödel, o Mandelbrot,
im Fegefeuer der Rekursion.

Die Mathematiker
Hans Magnus Enzensberger

Suppose that f_n satisfies a recursion of the form

$$f_n = a_1 f_{n-1} + a_2 f_{n-2} + \cdots + a_k f_{n-k}.$$

That is, there is a fixed number of recursion terms and each term is just a scalar multiple of a prior value (we shall see below that easy cases of index dependence also can be treated this way). We also assume that k initial values $f_0, \cdots, f_{k-1}$ have been established.

The most prominent case of this are clearly the Fibonacci numbers OEIS A000045 with $k = 2$, recursion $f_n = f_{n-1} + f_{n-2}$ and initial values $f_0 = f_1 = 1$. We shall use these as an example.

Step 1: Get the functional equation Using the recursion, expand the coefficient f_n in the generating function $F(t) = \sum_n f_n t^n$ with terms of lower index. Note that for $n < k$ the recursion does not hold, you will need to look at the initial values to see whether the given formula suffices, or if you need to add explicit multiples of powers of t to get equality.

Separate summands into different sums, factor out powers of t to get f_n combined with t^n.

Replace all expressions $\sum f_n t^n$ back with the generating function $F(t)$. The whole expression also must be equal to $F(t)$, this is the functional equation.

Example: in the case of the Fibonacci numbers, the recursion is $f_n = f_{n-1} + f_{n-2}$ for

[1] also means: "grief"

$n > 1$. Thus we get (using the initial values $f_0 = f_1 = 1$ that

$$\begin{aligned}
\sum_{n\geq 0} f_n t^n &= \sum_{n>1} (f_{n-1}t^n + f_{n-2}t^n) + f_1 t + f_0 \\
&= \sum_{n>1} f_{n-1}t^n + \sum_{n>1} f_{n-2}t^n + t + 1 \\
&= t\sum_{n>1} f_{n-1}t^{n-1} + t^2\sum_{n>1} f_{n-2}t^{n-2} + t + 1 \\
&= t\sum_{n>0} f_n t^n + t^2\sum_{n\geq 0} f_n t^n + t + 1 \\
&= t(\sum_{n\geq 0} f_n t^n - f_0) + t^2\sum_{n\geq 0} f_n t^n + t + 1 \\
&= t\cdot F(t) - t\cdot f_0 + t^2F(t) + t + 1 = t\cdot F(t) + t^2F(t) + 1.
\end{aligned}$$

The functional equation is thus $F(t) = t\cdot F(t) + t^2F(t) + 1$.

Step 2: Partial Fractions We can solve the functional equation to express $F(t)$ as a rational function in t. (This is possible, because the functional equation will be a linear polynomial in $F(t)$.) Then, using partial fractions (Calculus 2), we can write this as a sum of terms of the form $\frac{a_i}{(t-\alpha_i)^{e_i}}$.

Example: We solve the functional equation to give us $F(t) = \frac{1}{1-t-t^2}$. For a partial fraction decomposition, let $\alpha = \frac{-1+\sqrt{5}}{2}$, $\beta = \frac{-1-\sqrt{5}}{2}$ the roots of the equation $1 - t - t^2 = 0$. Then

$$F(t) = \frac{1}{1-t-t^2} = \frac{a}{t-\alpha} + \frac{b}{t-\beta}$$

We solve this as $a = -1/\sqrt{5}$, $b = 1/\sqrt{(}5)$.

Step 3: Use known power series to express each summand as a power series The geometric series gives us that

$$\frac{a}{t-\alpha} = \sum_{n\geq 0} \frac{-a}{\alpha^{n+1}} t^n$$

If there are multiple roots, denominators could arise in powers. For this we notice that

$$\frac{1}{(t-\alpha)^2} = \sum_{n\geq 0} \frac{(n+1)}{\alpha^{n+2}} t^n$$

and for integral $c \geq 1$ that

$$\frac{1}{(t-1)^c} = -1^c \sum_{n\geq 0} \binom{c+n-1}{n} t^n$$

Using these formulae, we can write each summand of the partial fraction decomposition as an infinite series.

Example: In the example we get

$$\begin{aligned} F(t) &= \frac{-1}{\sqrt{5}}\frac{1}{t-\alpha} + \frac{1}{\sqrt{5}}\frac{1}{t-\beta} \\ &= \frac{-1}{\sqrt{5}}\sum_{n\geq 0}\frac{-1}{\alpha^{n+1}}t^n + \frac{1}{\sqrt{5}}\sum_{n\geq 0}\frac{-1}{\beta^{n+1}}t^n \end{aligned}$$

Step 4: Combine to one sum, and read off coefficients We now take this (unique!) power series expression and read off the coefficients. The coefficient of t^n will be f_n, which gives us an explicit formula.

Example: Continuing the calculation above, we get

$$F(t) = \sum_{n\geq 0}\left(\frac{1}{\sqrt{5}\cdot\alpha^{n+1}} + \frac{-1}{\sqrt{5}\cdot\beta^{n+1}}\right)t^n$$

and thus a closed formula for the Fibonacci numbers:

$$\begin{aligned} f_n &= \frac{1}{\sqrt{5}\cdot\alpha^{n+1}} + \frac{-1}{\sqrt{5}\cdot\beta^{n+1}} \\ &= \frac{1}{\sqrt{5}\cdot\left(\frac{-1+\sqrt{5}}{2}\right)^{n+1}} + \frac{-1}{\sqrt{5}\cdot\left(\frac{-1-\sqrt{5}}{2}\right)^{n+1}} \\ &= \frac{1}{\sqrt{5}}\left(\left(\frac{2}{\sqrt{5}-1}\right)^{n+1} + \left(\frac{2}{\sqrt{5}+1}\right)^{n+1}\right) \end{aligned}$$

We notice that $\frac{2}{\sqrt{5}-1} > \frac{2}{\sqrt{5}+1} > 0$, thus asymptotically

$$f_{n+1}/f_n = \frac{2}{\sqrt{5}-1} = \phi \approx 1.618$$

the value of the golden ratio.

Another example

We try another example. Take the (somewhat random) recursion given by

$$\begin{aligned} g_0 &= g_1 = 1 \\ g_n &= g_{n-1} + 2\cdot g_{n-2} + (-1)^n, \qquad \text{for } n\geq 2 \end{aligned}$$

We get for the generating function

$$\begin{aligned} G(t) &= \sum_n g_n t^n = \sum_{n\geq 1} g_{n-1}t^n + 2\sum_{n\geq 2} g_{n-2}t^n + \sum_{n\geq 0}(-1)^n t^n + t \\ &= tG(t) + 2t^2G(t) + \frac{1}{1+t} + t. \end{aligned}$$

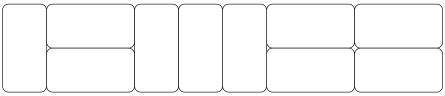


Figure III.1: A 2 × 10 domino tiling

(you should verify that the addition of t was all that was required to resolve the initial value settings.)

We solve this functional equation as

$$G(t) = \frac{1+t+t^2}{(1-2t)(1+t)^2}$$

and get the partial fraction decomposition

$$G(t) = \frac{-7}{18(t-\frac{1}{2})} - \frac{1}{9(t+1)} + \frac{1}{3(t+1)^2}.$$

We can read off the power series representation

$$\begin{aligned} G(t) &= \frac{7}{18}\sum_n 2^{n+1}t^n + \frac{1}{9}\sum_n (-1)^{n+1}t^n + \frac{1}{3}\sum_n -1^n(n+1)t^n \\ &= \sum_n \left(\frac{7}{9}2^n - \frac{1}{9}(-1)^n + \frac{1}{3}(-1)^n(n+1)\right)t^n \\ &= \sum_n \left(\frac{7}{9}2^n + \left(\frac{1}{3}n + \frac{2}{9}\right)(-1)^n\right)t^n, \end{aligned}$$

solving the recursion as $g_n = \frac{7}{9}2^n + \left(\frac{1}{3}n + \frac{2}{9}\right)(-1)^n$.

III.3 Nested Recursions: Domino Tilings

> The Domino Theory had become conventional wisdom and was rarely challenged.
>
> Diplomacy
> HENRY KISSINGER

We apply the method of generating functions to some counting problems.

Suppose we have tiles that have dimensions 1 × 2 (in your favorite units) and we want to tile a corridor. Let f_n the number of possible tilings of a corridor that has dimensions $2 \times n$. We can start on the left with a vertical domino (thus leaving to the right of it a tiling of a corridor of length $n-1$) or with two horizontal dominos (leaving to the right of it a corridor of length $n-2$). this gives us the recursion

$$d_n = d_{n-1} + d_{n-2}, \qquad n > 1$$

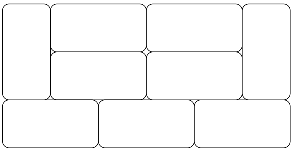


Figure III.2: A tiling pattern that has no vertical cut

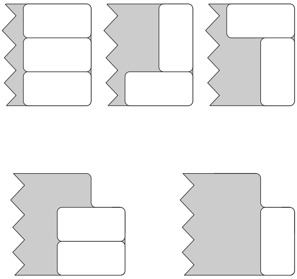


Figure III.3: Variants for a 3 × n tiling

with $d_1 = 1$ and $d_2 = 2$ (and thus $d_0 = 1$ to fit the recursion). This is just again the Fibonacci numbers we have already worked out.

If we assume that the tiles are not symmetric, there are actually 2 ways to place a horizontal tile and 2 ways to place a vertical tile. We thus get a recursion with different coefficients,

$$d_n = 2d_{n-1} + 4d_{n-2}, \qquad n > 1$$

with $d_1 = 2$, $d_2 = 8$ (and thus $d_0 = 1$).

If we expand the corridor to dimensions $3 \times n$, the recursion seems to be problematic – it is possible to have patterns or arbitrary length that do not reduce to a shorter length, see figure III.2.

We thus instead look only at the right end of the tiling and persuade ourselves (going back to the assumption of symmetric tiles) that every tiling has to end with one of the patterns depicted in the top row of figure III.3.

Removing these end stones either produces a tiling of length $n-2$, or a tiling of length n in which the top (or bottom) right corner is missing. We thus introduce a count e_n for tilings of length n with the top right corner missing, the count for the bottom right corner will by symmetry be the same. This gives us the recursion

$$d_n = d_{n-2} + 2 \cdot e_{n-1}$$

The introduction of the count e_n forces us to also get a description for these. Consider a tiling with the top right corner missing. Its right end must be a vertical tile or two horizontal tiles and thus look as in the bottom row of figure III.3.

This gives us the recursion

$$e_n = d_{n-1} + e_{n-2}$$

We use the initial values $d_0 = 1$, $d_1 = 0$, $d_2 = 3$, $e_0 = 0$, $e_1 = 1$, and thus get the following identities for the associate generating functions

$$\begin{aligned}
D(t) &= \sum_n d_n t^n = \sum_{n>1} (d_{n-2} + 2e_{n-1})\, t^n + d_0 + d_1 t \\
&= t^2 \sum_{n>1} d_{n-2} t^{n-2} + 2t \sum_{n>1} e_{n-1} t^{n-1} + 1 \\
&= t^2 \sum_{n\geq 0} d_n t^n + 2t\Big(\sum_{n\geq 0} e_n t^n - e_0 t^0\Big) + 1 \\
&= t^2 D(t) + 2tE(t) + 1
\end{aligned}$$

and

$$\begin{aligned}
E(t) &= \sum_n e_n t^n = \sum_{n>1} (d_{n-1} + e(n-2))\, t^n + e_0 + e_1 t \\
&= t(D(t) - d_0) + t^2 E(t) + t = tD(t) + t^2 E(t)
\end{aligned}$$

We solve this second equation as

$$E(t) = \frac{t}{1-t^2} D(t)$$

and substitute into the first equation, obtaining the functional equation

$$D(t) = t^2 D(t) + \frac{2t^2}{1-t^2} D(t) + 1$$

which we solve for

$$D(t) = \frac{1-t^2}{1-4t^2+t^4}$$

This is a function in t^2, indicating that $d_n = 0$ for odd n (indeed, this must be, as in this case $3 \times n$ is odd and cannot be tiled). We thus can consider instead the function

$$R(t) = \frac{1-t}{1-4t+t^2} = \sum_n r_n t^n$$

with $d_{2n} = r_n$. Partial fraction decomposition, geometric series, and final collection of coefficients gives us the formula

$$d_{2n} = r_n = \frac{(2+\sqrt{3})^n}{3-\sqrt{3}} + \frac{(2-\sqrt{3})^n}{3+\sqrt{3}}$$

and the sequence of r_n given by OEIS A001835

$$1, 3, 11, 41, 153, 571, 2131, 7953, 29681, 110771, 413403, \ldots$$

III.4 Catalan Numbers

> *The induction I used was pretty tedious, but I do not doubt that this result could be obtained much easier. Concerning the progression of the numbers 1, 2, 5, 14, 42, 132, etc. . . .*

> Die Induction aber, so ich gebraucht, war ziemlich mühsam, doch zweifle ich nicht, dass diese Sach nicht sollte weit leichter entwickelt werden können. Ueber die Progression der Zahlen 1, 2, 5, 14, 42, 132, etc. . . .

Letter to Goldbach
September 4, 1751
Leonard Euler

Next, we look at an example of a recursion which is not just linear and we effectively use the product rule:

Definition III.2: The n-th *Catalan number*[3] C_n OEIS A000108 is defined[4] as the number of ways a sum of $n+1$ variables can be evaluated by inserting parentheses.

Example III.3: We have $C_0 = C_1 = 1$, $C_2 = 2$: $(a+b)+c$ and $a+(b+c)$, and $C_3 = 5$:

$$\begin{array}{l} ((a+b)+c)+d \\ (a+(b+c))+d \\ a+((b+c)+d) \\ a+(b+(c+d)) \\ (a+b)+(c+d) \end{array}$$

To get a recursion, we consider the "outermost" addition, assume thus comes after $k+1$ of the variables. This addition combines two proper parenthetical expressions, the one on the left on $k+1$ variables, the one on the right on $(n+1)-(k+1) = n-k$ variables. We thus get the recursion

$$C_n = \sum_{k=0}^{n-1} C_k C_{n-k-1}, \qquad \text{if } n > 0,\ C_0 = 1.$$

in which we sum over the products of lower terms.

This is basically the pattern of the product rule, just shifted by one.

We thus get the functional equation

$$\mathcal{C}(t) = t \cdot \mathcal{C}(t)^2 + 1.$$

(which is easiest seen by writing out the expression for $t\mathcal{C}(t)^2$ and collecting terms). The factor t is due to the way we index and 1 is due to initial values.

[3] Named in honor of Eugène Charles Catalan (1814-1894) who first stated the standard formula. The naming after Catalan only stems from a 1968 book, see http://www.math.ucla.edu/~pak/papers/cathist4.pdf. Catalan himself attributed them to Segner, though Euler's work is even earlier.

[4] Careful, some books use a shifted index, starting at 1 only!

This is a quadratic equation in the variable $C(t)$ and we thus have

$$C(t) = \frac{1 - \sqrt{1-4t}}{2t},$$

where we choose the branch of the root, for which for $t = 0$ we get the desired value of C_0.

The binomial series then gives us that

$$\sqrt{1-4t} = \sum_{k\geq 0} \binom{1/2}{k} (-4t)^k = 1 + \sum_{k\geq 1} \frac{1}{2k} \binom{-1/2}{k-1} (-4t)^k.$$

We conclude that

$$\begin{aligned} \frac{1 - \sqrt{1-4t}}{2t} &= \sum_{k\geq 1} \frac{1}{k} \binom{-1/2}{k-1} (-4t)^{k-1} \\ &= \sum_{n\geq 0} \binom{2n}{n} \frac{t^n}{n+1}, \end{aligned}$$

in other words:

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

Catalan numbers have many other combinatorial interpretations, and we shall encounter some in the exercises. Exercise 6.19 in [Sta99] (and its algebraic continuation 6.25, as well as an online supplement) contain hundreds of combinatorial interpretations.

III.5 Index-dependent coefficients and Exponential generating functions

A recursion does not necessarily have constant coefficient, but might have a coefficient that is a polynomial in n. In this situation we can use (formal) differentiation, which will convert a term $f_n t^n$ into $n f_n t^{n-1}$. The second derivative will give a term $n(n-1) f_n t^{n-2}$; first and second derivative thus allow us to construct a coefficient $n^2 f_n$ and so on for higher order polynomials.

The functional equation for the generating function then becomes a differential equation, and we might hope that a solution for it can be found in the extensive literature on differential equations.

Alternatively, (that is we use the special form of derivatives for a typical summand), such a situation often can be translated immediately to a generating function by using the power series

$$\frac{1}{(1-t)^{k+1}} = \sum_n \binom{n+k}{n} t^n.$$

For an example of variable coefficients, we take the case of counting *derangements* OEIS A000166, that is permutations that leave no point fixed. We denote by d_n the number of derangements on $\{1, \ldots, n\}$.

To build a recursion formula, suppose that π is a derangement on $\{1, \ldots, n\}$. Then $n^\pi = i < n$. We now distinguish two cases, depending on how i is mapped by n:
a) If $i^\pi = n$, then π swaps i and n and is a derangement of the remaining $n-2$ points, thus there are d_{n-2} derangements that swap i and n. As there are $n-1$ choices for i, there are $(n-1)d_{n-2}$ derangements that swap n with a smaller point.
b) Suppose there is $j \neq i$ such that $j^\pi = n$. In this case we can "bend" π into another permutation ψ, by setting

$$k^\psi = \begin{cases} k^\pi & \text{if } k \neq j \\ i & \text{if } k = j \end{cases}.$$

We notice that ψ is a derangement of the points $\{1, \ldots, n-1\}$.

Vice versa, if ψ is a derangement on $\{1, \ldots, n-1\}$, and we choose a point $j < n$, we can define π on $\{1, \ldots, n\}$ by

$$k^\pi = \begin{cases} k^\psi & \text{if } k \neq i, n \\ n & \text{if } k = j \\ j^\psi & \text{if } k = n \end{cases}.$$

We again notice that π is a derangement, and that different choices of ψ and i result in different π's. Furthermore, the two constructions are mutually inverse, that is every derangement that is not in class a) is obtained by this construction.

There are d_{n-1} possible derangements ψ and $n-1$ choices for j, so there are $(n-1)d_{n-1}$ derangements in this second class.

We thus obtain a recursion

$$d_n = (n-1)\left(d_{n-1} + d_{n-2}\right)$$

and hand-calculate[5] the initial values $d_0 = 1$ and $d_1 = 0$.

From this recursion we could now construct a differential equation for the generating function of d_n, but there is a problem: Because of the factor n in the recursion, the values d_n grow roughly like $n!$. The resulting series thus will have convergence radius 0, making it unlikely that a function satisfying this differential equation can be found in the literature.

We therefore introduce the *exponential generating function* which is defined simply by dividing the i-th coefficient by a factor of $i!$, thus keeping coefficient growth bounded.

[5] The reader might have an issue with the choice of $d_0 = 1$, as it is unclear what a derangement on no points is. By going backwards to the recursion (using d_2 and d_1 to calculate d_0), it turns out that this is the right number to make the recursion work in case d_0 is referred to.

In our example, we get

$$D(t) = \sum_n \frac{d_n}{n!} t^n$$

and thus

$$\frac{\mathrm{d}}{\mathrm{d}t} D(t) = \sum_{n\geq 1} \frac{n \cdot d_n}{n!} t^{n-1} = \sum_{n\geq 1} \frac{d_n}{(n-1)!} t^{n-1} = \sum_{n\geq 0} \frac{d_{n+1}}{n!} t^n.$$

We also have

$$\begin{aligned} t \cdot D(t) &= \sum_{n\geq 0} d_n \frac{t^{n+1}}{n!} = \sum_{n\geq 0} (n+1) d_n \frac{t^{n+1}}{(n+1)!} \\ &= \sum_{n\geq 1} n \cdot d_{n-1} \frac{t^n}{n!} = \sum_{n\geq 0} n \cdot d_{n-1} \frac{t^n}{n!} \\ t \cdot D'(t) &= \sum_{n\geq 0} \frac{n \cdot d_n}{n!} t^n \end{aligned}$$

From this, the recursion (in the form $d_{n+1} = n(d_n + d_{n-1})$) gives us:

$$\begin{aligned} t \cdot D(t) + t \cdot D'(t) &= \sum_{n\geq 0} (n \cdot d_{n-1} + n \cdot d_n) \frac{t^n}{n!} \\ &= \sum_{n\geq 0} d_{n+1} \frac{t^n}{n!} = D'(t), \end{aligned}$$

and thus the separable differential equation

$$\frac{D'(t)}{D(t)} = \frac{t}{1-t}.$$

with $D(0) = d_0 = 1$.

Standard techniques from Calculus give us the solution

$$D(t) = \frac{e^{-t}}{1-t}.$$

Looking up this function for Taylor coefficients (respectively determining the formula by induction) shows that

$$D(t) = \sum_{n\geq 0} \left(\sum_{i=0}^{n} \frac{(-1)^i}{i!} \right) t^n$$

and thus (introducing a factor $n!$ to make up for the denominator in the generating function) that

$$d_n = n! \left(\sum_{i=0}^{n} \frac{(-1)^i}{i!} \right).$$

This is simply $n!$ multiplied with a Taylor approximation of e^{-1}. Indeed, if we consider the difference, the theory for alternating series gives us for $n \geq 1$ that:

$$\begin{aligned} \left| d_n - \frac{n!}{e} \right| &= n! \left| \sum_{i=n+1}^{\infty} \frac{(-1)^i}{i!} \right| \\ &< n! \left| \frac{(-1)^{n+1}}{(n+1)!} \right| = \frac{1}{n+1} \leq \frac{1}{2} \end{aligned}$$

We have proven:

Lemma III.4: d_n is the integer closest to $n!/e$.

That is asymptotically, if we put letters into envelopes, the probability is $1/e$ that no letter is in the correct envelope.

III.6 The product rule, revisited

Exponential generating functions have one more trick up their sleeve, and arguably this is their most important contribution. For this, lets return to the product rule. It corresponds to the situation that the objects of a certain weight can be described in terms of combining objects of lower weights (that sum up to the desired weight) in all possible ways.

This splitting-up however only considers the *number* of (sub)objects in each part, not *which particular* ones are in each part. In other words, we consider the constituent objects as indistinguishable.

Suppose now, however, that we are counting objects whose parts have identifying labels. In the example of the Catalan numbers this would be for example, if we cared not only about the parentheses placement, but also about the symbols we add, that is $(a + b) + c$ would be different from $(c + a) + b$.

In such a situation, the recursion formula will have to account for each k, *which* k elements are chosen to be in the "left side", with the rest being in the"right side". That is, the recursion becomes:

$$d_n = \sum_{k=0}^{n} \binom{n}{k} a_k b_{n-k}.$$

We can write this as

$$\frac{d_n}{n!} = \sum_{k=0}^{n} \frac{a_k}{k!} \frac{b_{n-k}}{(n-k)!},$$

which is the formula for multiplication of the *exponential* generating functions!

Lets look at this in a pathetic example, the number of functions from $N = \{1, \ldots, n\}$ to $\{1, \ldots, r\}$ (which we know already well as r^n).

Let a_n count the number of constant functions on an n-element set, that is $a_n = 1$. The associated exponential generating function thus is

$$A(t) = \sum_n \frac{t^n}{n!} = \exp(t)$$

(which, incidentally, shows why these are called "exponential" generating functions).

If we take an arbitrary function f on N, we can partition N into r (possibly empty) sets $N_1, \ldots, N_r$, such that f is constant on N_i and the N_i are maximal with this property.

We get all possible functions f by combining constant functions on the possible N_i's for all possible partitions of N. Note that the ordering of the partitions is significant – they indicate the actual values.

We are thus exactly in the situation described, and get as exponential generating function (start with $r = 2$, then use induction for larger r) the r-fold product of the exponential generating functions for the number of constant functions:

$$D(t) = \underbrace{\exp(t) \cdot \cdots \exp(t)}_{r \text{ factors}} = \exp(rt)$$

The coefficient for t^n in the power series for $\exp(rt)$ of course is simply $\frac{r^n}{n!}$, that is the counting function is r^n, as expected.

Bell Numbers

$$\int_1^{\sqrt[3]{3}} z^2 dz \times \cos \frac{3\pi}{9} = \log(\sqrt[3]{e})$$

The integral z-squared dz,
From one to the cube root of three,
Times the cosine,
Of three pi over nine
Equals log of the cube root of e.

Anon.

Another example of this approach wil allow up to give an exponential generating function for the Bell numbers (though not a closed form expression for its coefficients):

Recall that the Bell numbers B_n are defined as giving the total number of partitions of $\{1, \ldots, n\}$ and are (Lemma III.1) satisfying the recursion:

$$B_n = \sum_{k=1}^{n} \binom{n-1}{k-1} B_{n-k} = \sum_{k=0}^{n-1} \binom{n-1}{k} B_{n-1-k}$$

In light of the product rule, we insert a factor 1 and write this (after reindexing) as

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} 1 \cdot B_{n-k}$$

and thus

$$\sum_t \frac{B_{n+1}}{n!} t^n = \sum_t \left(\sum_{k=0}^{n} \frac{1}{k!} \cdot \frac{B_{n-k}}{(n-k)!} \right) t^n$$

If we denote the exponential generating function of the B_n by $F(t) = \sum_n B_n t^n / n!$, the right hand side thus will give the product of the generating function of the constant sequence $a_n = 1$ (which we just saw is $\exp(t)$) with $F(t)$. This reflects the split that gave us the recursion – into a set of size k containing the number 1 (the total number of such sets being 1 once the numbers are chosen), and a partition of the remaining numbers.

The left hand side is the exponential generating function of B_{n+1} which is just the derivative of $F(t)$, thus we have that

$$\frac{\mathrm{d}}{\mathrm{d}t} F(t) = \sum_{n \geq 1} \frac{B_n t^{n-1}}{(n-1)!} = \sum_{n \geq 0} \frac{B_{n+1} t^n}{n!} = \exp(t) F(t).$$

This is again a separable differential equation, its solution is

$$F(t) = c \cdot \exp(\exp(t))$$

for some constant c. As $F(0) = 1$ we solve for $c = \exp(-1)$ and thus get the exponential generating function

$$\sum_n \frac{B_n t^n}{n!} = \exp(\exp(t) - 1).$$

There is no nice way to express the power series coefficients of this function in closed form, a Taylor approximation is (with denominators being deliberately kept in the form of $n!$ to allow reading off the Bell numbers):

$$1 + t + \frac{2t^2}{2!} + \frac{5t^3}{3!} + \frac{15t^4}{4!} + \frac{52t^5}{5!} + \frac{203t^6}{6!} + \frac{877t^7}{7!} + \frac{4140t^8}{8!} + \frac{21147t^9}{9!} + \frac{115975t^{10}}{10!}.$$

One, somewhat surprising application of Bell numbers is to consider rhyme schemes. Given a sequence of n lines, the lines which rhyme form the cells of a partition of $\{1, \dots, n\}$. For example, the partition $\{\{1, 2, 5\}, \{3, 4\}\}$ is the scheme used by Limericks.

We can read off that $B_5 = 52$. The classic 11th century Japanese novel *Genji monogatari* (The Tale of Genji) has 54 chapters of which first and last are considered "extra". The remaining 52 chapters are introduced each with a poem in one of the 52 possible rhyme schemes and a symbol illustrating the scheme. These symbols, see figure III.4, the *Genji-mon*, have been used extensively in Art. See `https://www.viewingjapaneseprints.net/texts/topics_faq/genjimon.html`

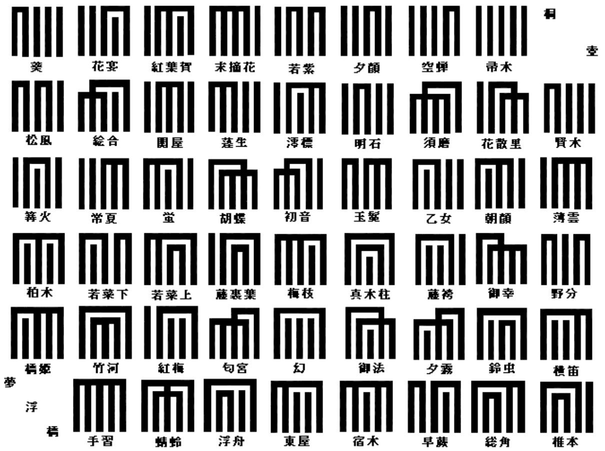


Figure III.4: The Genji-mon

ja.wikipedia.org

Stirling Numbers

We apply the same idea to the Stirling numbers of the second kind, $S(n,k)$ denoting the number of partitions of n into k (non-empty) parts. According to II.10, part 3) there are $k!S(n,k)$ order-significant partitions of n into k parts.

We denote the associated exponential generating function (for order-significant partitions) by

$$\mathcal{S}_k(t) = \sum_n k!S(n,k)\frac{t^n}{n!}.$$

We also know that there is – apart from the empty set – exactly one partition into one cell. That is

$$\mathcal{S}_1(t) = t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots = \exp(t) - 1$$

If we have a partition into k-parts, we can fix the first cell and then partition the rest further. Thus, for $k > 1$ we have that

$$\mathcal{S}_k(t) = \mathcal{S}_1(t) \cdot \mathcal{S}_{k-1}(t),$$

which immediately gives that

$$\mathcal{S}_k(t) = (\exp(t) - 1)^k$$

We deduce that the Stirling numbers of the second kind have the exponential generating function

$$\sum_n S(n,k)\frac{t^n}{n!} = \frac{(\exp(t) - 1)^k}{k!}.$$

Using the fact that $B_n = \sum_{k=1}^{n} S(n,k)$, we thus get the exponential generating function for the Bell numbers as

$$\begin{aligned}\sum_n B_n \frac{t^n}{n!} &= \sum_k \sum_n S(n,k)\frac{t^n}{n!}\\ &= \sum_k \frac{(\exp(t)-1)^k}{k!} = \exp(\exp(t)-1)\end{aligned}$$

in agreement with the above result.

We also can use the exponential generating function for the Stirling numbers to deduce a coefficient formula for them:

Lemma III.5:

$$S(n,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}\binom{k}{i}i^n.$$

Proof: We first note that – as $0^n = 0$ – the i-sum could start at 1 or 0 alternatively without changing the result.

Then multiplying through with $k!$, we know by the binomial formula that

$$\begin{aligned}k!\sum_n S(n,k)\frac{t^n}{n!} &= (\exp(t)-1)^k = \sum_{i=0}^{k}\binom{k}{i}\exp(t)^i(-1)^{k-i}\\ &= \sum_{i=0}^{k}\binom{k}{i}\exp(i\cdot t)(-1)^{k-i} = \sum_{i=0}^{k}\binom{k}{i}\left(\sum_n \frac{(it)^n}{n!}\right)(-1)^{k-i}\\ &= \sum_n\left(\sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i}i^n\right)\frac{t^n}{n!}.\end{aligned}$$

and we read off the coefficients. □

Involutions

Finally, lets use this in a new situation:

Definition III.6: A permutation π on $\{1,\ldots,n\}$ is called[6] an *involution* OEIS A000085 if $\pi^2 = 1$, that is $(i^\pi)^\pi = i$ for all i.

We want to determine the number s_n of involutions on n points.

The reduction we shall use is to consider the number of cycles of the involution, also counting 1-cycles. Let $s_r(n)$ the number of involutions that have exactly r cycles. Clearly $s_0(0) = 1$, $s_1(0) = 0$, $s_1(1) = 1$, $s_1(2) = 1$ which gives the exponential generating function $S_1(t) = t + \frac{t^2}{2}$.

[6] Group theorists often exclude the identity, but it is convenient to allow it here.

When considering an arbitrary involution, we can split off a cycle, seemingly leading to a formula

$$\left(t + \frac{t^2}{2}\right)^r$$

for $S_r(t)$. However doing so distinguishes the cycles in the order they were split off, that is we need to divide by $r!$ to avoid overcounting. Thus

$$S_r(t) = \frac{1}{r!}\left(t + \frac{t^2}{2}\right)^r$$

and thus for the exponential generating function of the number of involutions a sum over all possible r:

$$S(t) = \sum_{r=0}^{\infty} S_r(t) = \sum_r \frac{1}{r!}\left(t + \frac{t^2}{2}\right)^r = \exp(t + \frac{t^2}{2}) = \exp(t)\exp\left(\frac{t^2}{2}\right).$$

We easily write down power series for the two factors

$$\begin{aligned} \exp(t) &= \sum_n \frac{t^n}{n!} \\ \exp\left(\frac{t^2}{2}\right) &= \sum_n \frac{t^{2n}}{2^n \cdot n!} \end{aligned}$$

and multiply out, yielding

$$\begin{aligned} S(t) &= \left(\sum_n \frac{t^{2n}}{2^n \cdot n!}\right) \cdot \left(\sum_n \frac{t^n}{n!}\right) \\ &= \sum_m \sum_{k=0}^{\lfloor m/2 \rfloor} \frac{t^{2k} t^{m-2k}}{2^k k!(m-2k)!} \end{aligned}$$

and thus (again introducing a factor of $n!$ for making up for the exponential generating function)

$$s_n = \sum_{k=0}^{\lfloor m/2 \rfloor} \frac{n!}{2^k k!(n-2k)!}.$$

Chapter

IV

Inclusion, Incidence, and Inversion

It is calculus exam week and, as every time, a number of students were reported to require alternate accommodations:

- 14 students are sick.
- 12 students are scheduled to play for the university Quiddich team at the county tournament.
- 12 students are planning to go on the restaurant excursion for the food appreciation class.
- 5 students on the Quiddich team are sick (having been hit by balls).
- 4 students are scheduled for the excursion and the tournament.
- 3 students of the food appreciation class are sick (with food poisoning), and
- 2 of these students also planned to go to the tournament, i.e. have all three excuses.

The course coordinator wonders how many alternate exams need to be provided.

Using a Venn diagram and some trial-and-error, it is not hard to come up with the diagram in figure IV.1, showing that there are 28 alternate exams to offer.

IV.1 The Principle of Inclusion and Exclusion

> By the method of exclusion, I had arrived at this result, for no other hypothesis would meet the facts.
>
> A Study in Scarlet
> Arthur Conan Doyle

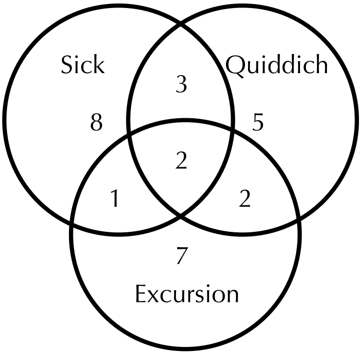


Figure IV.1: An example of Inclusion/Exclusion

The *Principle of Inclusion and Exclusion* (PIE) formalizes this process for an arbitrary number of sets:

Let X be a set and $\{A_1, \ldots, A_n\}$ a family of subsets. For any subset $I \subset \{1, \ldots, n\}$ we define

$$A_I = \bigcap_{i \in I} A_i,$$

using that $A_\varnothing = X$. Then

Lemma IV.1: The number of elements that lie in none of the subsets A_i is given by

$$\sum_{I \subset \{1,\ldots,n\}} (-1)^{|I|} |A_I| .$$

Proof: Take $x \in X$ and consider the contribution of this element to the given sum. If $x \notin A_i$ for any i, it only is counted for $I = \varnothing$, that is contributes 1.

Otherwise let $J = \{1 \leq a \leq n \mid x \in A_a\}$ and let $j = |J|$. We have that $x \in A_I$ if and only if $I \subset J$. Thus x contributes

$$\sum_{I \subset J} (-1)^{|I|} = \sum_{i=0}^{j} \binom{j}{i} (-1)^i = (1-1)^j = 0$$

□

As a first application we determine the number of derangements on n points in an different way:

Let $X = S_n$ be the set of all permutations of degree n, and let A_i be the set of all permutations π with $i^\pi = i$. Then $S_n \setminus \bigcup_i A_i$ is exactly the set of derangements, there are $\binom{n}{i}$ possibilities to intersect i of the A_i's, and the formula gives us:

$$d(n) = \sum_{i=0}^{n} (-1)^i \binom{n}{i} (n-i)! = n! \sum_{i=0}^{n} \frac{(-1)^i}{i!}.$$

For a second example, we calculate the number of surjective mappings from an n-set to a k-set (which we know already from II.10 to be $k!S(n,k)$):

Let X be the set of all mappings from $\{1,\ldots,n\}$ to $\{1,\ldots,k\}$, then $|X| = k^n$. Let A_i be the set of those mappings f, such that i is not in the image of f, so $|A_i| = (k-1)^n$. More generally, if $I \subset \{1,\ldots k\}$ we have that $|A_I| = (k-|I|)^n$. The surjective mappings are exactly those in X outside any of the A_i, thus the formula gives us the count

$$\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^n,$$

using again that there are $\binom{k}{i}$ possible sets I of cardinality i.

The factor $(-1)^i$ in a formula often is a good indication that inclusion/exclusion is to be used.

Lemma IV.2:

$$\sum_{i=0}^{n}(-1)^i\binom{n}{i}\binom{m+n-i}{k-i} = \begin{cases} \binom{m}{k} & \text{if } m \geq k, \\ 0 & \text{if } m < k. \end{cases}$$

Proof: To use PIE, the sets A_i need to involve choosing from an n, set, and after choosing i of these we must choose from a set of size $m+n-i$.

Consider a bucket filled with n blue balls, labeled with $1,\ldots,n$, and m red balls. How many selections of k balls only involve red balls? Clearly the answer is the right hand side of the formula.

Let X be the set of all k-subsets of balls and A_i those subsets that contain blue ball number i, then PIE gives the left side of the formula. □

We finish this section with an application from number theory. The *Euler function* $\varphi(n)$ counts the number of integers $1 \leq k \leq n$ with $\gcd(k,n) = 1$.

Suppose that $n = \prod_{i=1}^{r} p_i^{e_i}$, $X = \{1,\ldots,n\}$ and A_i the integers in X that are multiples of p_i. Then (inclusion/exclusion)

$$\varphi(n) = n - \sum_{i=1}^{r}\frac{n}{p_i} + \sum_{1\leq i,j\leq r}\frac{n}{p_ip_j} - \cdots = n\prod\left(1-\frac{1}{p_i}\right).$$

with the second identity obtained by multiplying out the product,

We also note – exercise **??** – that $\sum_{d|n}\varphi(d) = n$. That is, the sum over one function over a nice index set is another (easier) function. We will put this into larger context in later sections.

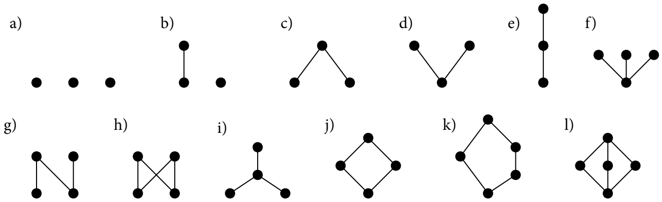


Figure IV.2: Hasse Diagrams of Small Posets and Lattices

IV.2 Partially Ordered Sets and Lattices

> The doors are open; and the surfeited grooms
> Do mock their charge with snores:
> I have drugg'd their possets,
> That death and nature do contend about them
>
> Macbeth, Act II, Scene II
> William Shakespeare

A *poset* or *partially ordered set* is a set A with a relation $R \subset A \times A$ on the elements of A which we will typically write as $a \le b$ instead of $(a, b) \in R$, such that for all $a, b, c \in A$:

(reflexive) $a \le a$.

(antisymmetric) $a \le b$ and $b \le a$ imply that $a = b$.

(transitive) $a \le b$ and $b \le c$ imply that $a \le c$.

The elements of a poset thus are the elements of A, not those of the underlying relation and is cardinality is that of A.

For example, A could be the set of subsets of a particular set, and $\le$ with be the "subset or equal" relation.

A convenient way do describe a poset for a finite set A is by its *Hasse-diagram*. Say that *a covers b* if $a \ge b$, $a \ne b$ and there is no $a \ne c =\ne b$ with $a \ge c \ge b$. The Hasse diagram of the poset is a graph in the plane which connects two vertices a and b only if a covers b, and in this case the edge from b to a goes upwards.

Because of transitivity, we have that $a \le b$ if and only if one can go up along edges from a to reach b.

Figure IV.2 gives a number of examples of posets, given by their Hasse diagrams, including all posets on 3 elements.

An isomorphism of posets is a bijection that preserves the $\le$ relation.

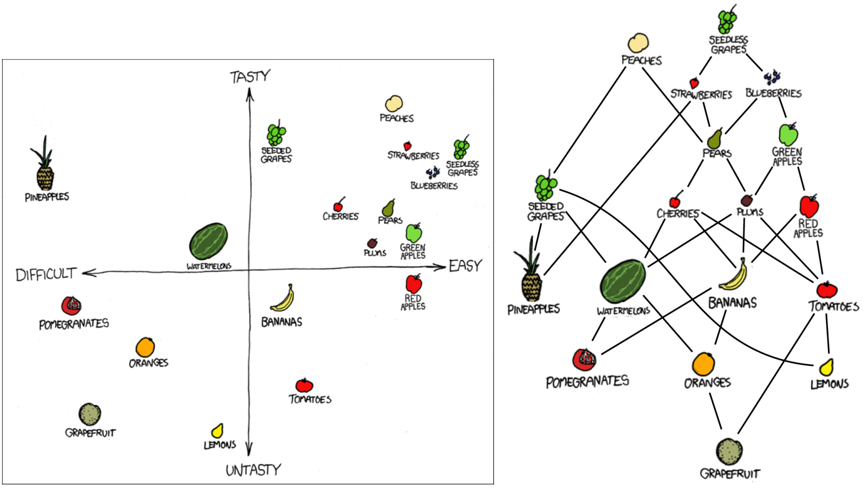


Figure IV.3: Fruits, arranged by subjective taste and ease-of-use

R. Munroe, F*** Grapefruit, https://xkcd.com/388/, and [AU17], p.108

An element in a poset is called *maximal* if there is no larger (wrt. ≤) element, *minimal* is defined in the same way. Posets might have multiple maximal and minimal elements.

For another cute example [1] consider in Figure IV.3, left, the arrangement of fruits according to convenience and taste, as given by[2] https://xkcd.com/388/.

We can read of a partial order from this by declaring a fruit as "better" than another, if it both more tasty and easier to consume. The resulting Hasse diagram is on the right side of Figure IV.3. It is easily seen that not every pair of fruits is comparable this way, and that there is no universal "best" or "worst" fruit.

Linear extension

> My scheme of Order gave me the most trouble
>
> Autobiography
> BENJAMIN FRANKLIN

A partial order is called a *total order*, if for every pair $a, b \in A$ of elements we have that $a \leq b$ or $b \leq a$.

While this is not part of our definition, we can always embed a partial order into a total order.

[1] Taken from lecture notes [AU17]

[2] pardon the title of the cartoon

Proposition IV.3: Let $R \subset X \times X$ be a partial order on X. Then there exists a total order (called a *linear extension*) $T \subset X \times X$ such that $R \subset T$.

To avoid set acrobatics we shall prove this only in the case of a finite set X. Note that in Computer science the process of finding such an embedding is called a *topological sorting*.

Proof: We proceed by induction over the number of pairs a, b that are incomparable. In the base case we already have a total order.

Otherwise, let a, b such an incomparable pair. We set (arbitrarily) that $a < b$. Now let

$$L = \{x \in X \mid x \leq_R a\}, U = \{x \in X \mid b \leq_R x\}$$

We claim that $S = R \cup \{(l, u) \mid l \in L, u \in u\}$ is a partial order. As $(a, b) \in S$ it has fewer incomparable pairs, this shows by induction that there exists a total order $T \supset S \supset R$, proving the theorem.

Since R is reflexive, S is. For antisymmetry, suppose that for $x \neq y$ we have that $(x, y), (y, x) \in S$. Since R is a partial order, not both can be in R. Suppose that

$$(x, y), (y, x) \in S \setminus R = \{(l, u) \mid l \in L, u \in u\}.$$

This implies that $x \leq_R a$ and $b \leq_R x$, thus by transitivity $b \leq_R a$, contradicting the incomparability.

If $(x, y) \in R$, $(y, x) \in S \setminus R$, we have that $b \leq_R x \leq_R y \leq_R a$, again contradicting incomparability.

For transitivity, suppose that $(x, y) \in S \setminus R$ and $(y, z) \in S$. Then $(y, z) \in R$, as otherwise $b \leq_R y \leq_R a$. But then $b \leq_R y \leq_R z$, implying that $(x, z) \in S$. The other case is analog. □

This theorem implies that we can always label the elements of a countable poset with positive integers, such that the poset ordering implies the integer ordering. Such an embedding is in general not unique, see Theorem IV.16.

The case of a totally ordered subset gets a special name:

Definition IV.4: A *chain* in a poset P is a subset of P such that any two elements of it are comparable. (That is, restricted to the chain the order is total.)

Lattices

Definition IV.5: Let A be a poset and $a, b \in A$.

- A *greatest lower bound* of a and b is an element $c \leq a, b$ which is maximal in the set of elements with this property.
- A *least upper bound* of a and b is an element $c \geq a, b$ which is minimal in the set of elements with this property.

A is a *lattice* if any pair $a, b \in A$ have a unique greatest lower bound, called the *meet* and denoted by $a \wedge b$; as well as unique least upper bound, called the *join* and denoted by $a \vee b$.

Amongst the Hasse diagrams in figure IV.2, e,j,k,l) are lattices, while the others are not. Lattices always have unique maximal and minimal elements, sometimes denoted by 0 (minimal) and 1 (maximal).

Other examples of lattices are:

1. Given a set X, let $A = \mathcal{P}(X) = \{Y \subseteq X\}$ the power set of X with $\leq$ defined by inclusion. Meet is the intersection, join the union of subsets.

2. Given an integer n, let A be the set of divisors of n with $\leq$ given by "divides". Meet and join are gcd, respectively lcm.

3. For an algebraic structure S, let A be the set of all substructures (e.g. group and subgroups) of S and $\leq$ given by inclusion. Meet is the intersection, join the substructure spanned by the two constituents.

4. For particular algebraic structures there might be classes of substructures that are closed under meet and join, e.g. normal subgroups. These then form a (sub)lattice.

Using meet and join as binary operations, we can axiomatize the structure of a lattice:

Proposition IV.6: Let X be a set with two binary operations $\wedge$ and $\vee$ and two distinguished elements $0, 1 \in X$. Then $(X, \wedge, \vee, 0, 1)$ is a lattice if and only if the following axioms are satisfies for all $x, y, z \in X$:

Associativity: $x \wedge (y \wedge z) = (x \wedge y) \wedge z$ and $x \vee (y \vee z) = (x \vee y) \vee z$;

Commutativity: $x \wedge y = y \wedge x$ and $x \vee y = y \vee x$;

Idempotence: $x \wedge x = x$ and $x \vee x = x$;

Inclusion: $(x \vee y) \wedge x = x = (x \wedge y) \vee x$;

Maximality: $x \wedge 0 = 0$ and $x \vee 1 = 1$.

Proof: The verification that these axioms hold for a lattice is left as exercise to the reader.

Vice versa, assume that these axioms hold. We need to produce a poset structure and thus define that $x \leq y$ iff $x \wedge y = x$. Using commutativity and inclusion this implies the dual property that $x \vee y = (x \wedge y) \vee y = y$.

To show that $\leq$ is a partial order, idempotence shows reflexivity. If $x \leq y$ and $y \leq x$ then $x = x \wedge y = y \wedge x = y$ and thus antisymmetry. Finally suppose that $x \leq y$ and $y \leq z$, that is $x = x \wedge y$ and $y = y \wedge z$. Then

$$x \wedge z = (x \wedge y) \wedge z = x \wedge (y \wedge z) = x \wedge y = x$$

and thus $x \leq z$. Associativity gives us that $x \wedge y \leq x, y$ if also $z \leq x, y$ then

$$z \wedge (x \wedge y) = (z \wedge x) \wedge y = z \wedge y = z$$

and thus $z \leq x \wedge y$, thus $x \wedge y$ is the unique greatest lower bound. The least upper bound is proven in the same way and the last axiom shows that 0 is the unique minimal and 1 the unique maximal element. □

DEFINITION IV.7: An element x of a lattice L is *join-irreducible* (JI) if $x \neq 0$ and if $x = y \vee z$ implies that $x = y$ or $x = z$.

For example, figure IV.4 shows a lattice in which the black vertices are JI, the others not.

Example: If we take the lattice of subsets of a set, the join-irreducibles are the 1-element sets. If we take divisors of n, the join-irreducibles are prime powers.

When representing elements of a finite lattices, it is possible to do so by storing the JI elements once and representing every element based on the JI elements that are below. This is used for example in one of the algorithms for calculating the subgroups of a group.

Product of posets

The cartesian product provides a way to construct new posets (or lattices) from old ones: Suppose that X, Y are posets with orderings $\leq_X$, $\leq_Y$, we define a partial order on $X \times Y$ by setting

$$(x_1, y_1) \leq (x_2, y_2) \quad \text{if and only if} \quad x_1 \leq_X x_2 \quad \text{and} \quad y_1 \leq_Y \leq y_2.$$

PROPOSITION IV.8: This is a partial ordering, so $X \times Y$ is a poset. If furthermore both X and Y are lattices, then so is $X \times Y$.

The proof of this is exercise **??**.

This allows us to describe two familiar lattices as constructed from smaller pieces (with a proof also delegated to the exercises):

PROPOSITION IV.9: a) Let $|A| = n$ and $\mathcal{P}(A)$ the power-set lattice (that is the subsets of A, sorted by inclusion). Then $\mathcal{P}(A)$ is (isomorphic to) the direct product of n copies of the two element lattice $\{0, 1\}$.
b) For an integer $n = \prod_{i=1}^{r} p_i^{e_i} > 1$ written as a product of powers of distinct primes, let $\mathcal{D}(n)$ be the lattice of divisors of n. Then $\mathcal{D}(n) \cong \mathcal{D}(p_1^{e_1}) \times \cdots \times \mathcal{D}(p_r^{e_r})$.

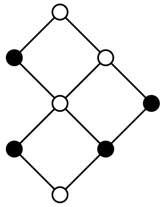


Figure IV.4: Order-ideal lattice for the “N” poset.

IV.3 Distributive Lattices

A lattice L is *distributive*, if for any $x, y, z \in L$ one (and thus also the other) of the two following laws hold:

$$\begin{aligned} x \vee (y \wedge z) &= (x \vee y) \wedge (x \vee z) \\ x \wedge (y \vee z) &= (x \wedge y) \vee (x \wedge z) \end{aligned}$$

Example: These laws clearly hold for the lattice of subsets of a set or the lattice of divisors of an integer n.

Lattices of substructures of algebraic structures are typically not distributive, the easiest example (diagram l) in figure IV.2) is the lattice of subgroups of $C_2 \times C_2$ which also is the lattice of subspaces of F_2^2.

If $P = (X, \leq)$ is a poset, a subset $Y \leq X$ is an *order ideal*, if for any $y \in Y$ and $z \in X$ we have that $z \leq y$ implies $z \in Y$.

Lemma IV.10: The set of order ideals is closed under union and intersection.

Proof: Let A, B be order ideals and $y \in A \cup B$ and $z \leq y$. Then $y \in A$ or $y \in B$. In the first case we have that $z \in A$, in the second case that $z \in B$, and thus always $z \in A \cup B$. The same argument also works for intersections. □

This implies:

Lemma IV.11: The set of order ideals of P, denoted by $J(P)$ is a lattice under intersection and union.

As a sublattice of the lattice of subsets, $J(P)$ is clearly distributive.

For example, if P is the poset on 4 elements with a Hasse diagram given by the letter N (figure IV.2, g) then figure IV.4 describes the lattice $J(P)$.

In fact, any finite distributive lattice can be obtained this way

THEOREM IV.12 (Fundamental Theorem for Finite Distributive Lattices, BIRKHOFF): Let L be a finite distributive lattice. Then there is a unique (up to isomorphism) finite poset P, such that $L \cong J(P)$.

To prove this theorem we use the following definition:

DEFINITION IV.13: For any element x of the poset P, let $\downarrow x = \{y \mid y \leq x\}$ be the principal order ideal generated by x.

LEMMA IV.14: An order ideal of a finite poset P is join irreducible in $J(P)$ if and only it is principal.

Proof: First consider a principal order ideal $\downarrow x$ and suppose that $\downarrow x = b \vee c$ with b and c being order ideals. Then $x \in b$ or $x \in c$, which by the order ideal property implies that $\downarrow x \subset b$ or $\leq x \subset c$.

Vice versa, suppose that a is a join irreducible order ideal and assume that a is not principal. Then for any $x \in a$, $\downarrow x$ is a proper subset of a. But clearly $a = \bigcup_{x \in a} \downarrow x$. □

COROLLARY IV.15: Given a finite poset P, the set of join-irreducibles of $J(P)$, considered as a subposet of $J(P)$, is isomorphic to P.

Proof: Consider the map that maps x to $\downarrow x$. It maps P bijective to the set of join-irreducibles, and clearly preserves inclusion. □

We now can prove theorem IV.12
Proof: Given a distributive lattice L, let X be the set of join-irreducible elements of L and P be the subposet of L formed by them. By corollary IV.15, this is the only option for P up to isomorphism, which will show uniqueness.

Let $\phi: L \to J(P)$ be defined by $\phi(a) = \{x \in X \mid x \leq a\}$, that is it assigns to every element of L the JI elements below it. (Note that indeed $\phi(a)$ is an order ideal.). We want to show that ϕ is an isomorphism of lattices.

Step1: Clearly we have that $a = \bigvee_{x \in \phi(a)} x$ for any $a \in L$ (using the join over the empty set equal to 0). Thus ϕ is injective.

Step 2: To show that ϕ is surjective, let $Y \in J(P)$ be an order ideal of P, and let $a = \bigvee_{y \in Y} y$. We aim to show that $\phi(a) = Y$: Clearly every $y \in Y$ also has $y \leq a$, so $Y \subset \phi(a)$. Next take a join irreducible $x \in \phi(a)$, that is $x \leq a$. Then $x \leq \bigvee_{y \in Y} y$ and thus

$$x = x \wedge \left(\bigvee_{y \in Y} y \right) = \bigvee_{y \in Y} (x \wedge y)$$

by the distributive law. Because x is JI, we must have that $x = x \wedge y$ for some $y \in Y$, implying that $x \leq y$. But as Y is an order ideal this implies that $x \in Y$. Thus $\phi(a) \subset Y$ and thus equality, showing that ϕ is surjective.

Step 3: We finally need to show that ϕ maps the lattice operations: Let $x \in X$. Then $x \le a \wedge b$ if and only if $x \le a$ and $x \le b$. Thus $\phi(a \wedge b) = \phi(a) \cap \phi(b)$.

For the join, take $x \in \phi(a) \cup \phi(b)$. Then $x \in \phi(a)$, implying $x \le a$, or (same argument) $x \le b$; therefore $x \le a \vee b$. Vice versa, suppose that $x \in \phi(a \vee b)$, so $x \le a \vee b$ and thus

$$x = x \wedge (a \vee b) = (x \wedge a) \vee (x \wedge b).$$

Because x is JI that implies $x = x \wedge a$, respectively $x = x \wedge b$.

In the first case this gives $x \le a$ and thus $x \in \phi(a)$; the second case similarly gives $x \in \phi(b)$. □

We close with a further use of the order ideal lattice:

Theorem IV.16: Let P be a poset of size m. The number of different linear orderings of P is equal to the number of different chains of (maximal) length m in $J(P)$.

Proof: Exercise. □

IV.4 Chains, Antichains, and Extremal Set Theory

> Man is born free;
> and everywhere he is in chains
>
> The Social Contract
> Jean-Jacques Rousseau

We start with a definition dual to that of a chain:

Definition IV.17: An *antichain* in a poset is a subset, such that any two (different) elements are incomparable.

We shall consider partitions of (the elements of) a poset into a collection of chains (or of antichains).

Clearly a chain C and antichain A can intersect in at most one element. This gives the following duality:

Lemma IV.18: Let P be a poset.
a) If P has a chain of size r, then it cannot be partitioned in fewer than r antichains.
b) If P has an antichain of size r, then it cannot be partitioned in fewer than r chains.

A stronger version of this goes usually under the name of Dilworth's theorem[3].

[3] proven earlier by Gallai and Milgram

Theorem IV.19 (Dilworth, 1950): The minimum number m of chains in a partition of a finite poset P is equal to the maximum number M of elements in an antichain.

Proof: The previous lemma shows that $m \geq M$, so we only need to show that we can partition P into M chains. We use induction on $|P|$, in the base case $|P| = 1$ nothing needs to be shown.

Consider a chain C in P of maximal size. If every antichain in $P \setminus C$ contains at most $M - 1$ elements, we apply induction and partition $P \setminus C$ into $M - 1$ chains and are done.

Thus assume now that $\{a_1, \ldots, a_M\}$ was an antichain in $P \setminus C$. Let

$$\begin{aligned} S^- &= \{x \in P \mid x \leq a_i \text{ for some } i\} \\ S^+ &= \{x \in P \mid x \geq a_i \text{ for some } i\} \end{aligned}$$

Then $S^- \cup S^+ = P$, as there otherwise would be an element we could add to the antichain and increase its size.

As C is of maximal size, the largest element of C cannot be in S^-, and thus we can apply induction to S^-. As there is an antichain of cardinality M in S^-, we partition S^- into M disjoint chains.

Similarly we partition S^+ into M disjoint chains. But each a_i is maximal element of exactly one chain in S^- and minimal element of exactly one chain of S^+. We can combine these chains at the a_i's and thus partition P into M chains. □

Corollary IV.20: If P is a poset with $nm + 1$ elements, it has a chain of size $n + 1$ or an antichain of size $m + 1$.

Proof: Suppose not, then every antichain has at most m elements and by Dilworth's theorem we can partition P into m chains of size $\leq n$ each, so $|P| \leq mn$. □

Corollary IV.21 (Erdős-Szekeres, 1935): Every sequence of $nm + 1$ distinct integers contains an increasing subsequence of length at least $n + 1$, or a decreasing subsequence of at length least $m + 1$.

Proof: Suppose the sequence is $a_1, \ldots, a_N$ with $N = nm + 1$. We construct a poset on N elements $x_1, \ldots, x_N$ by defining $x_i \leq x_j$ if and only if $i \leq j$ and $a_i \leq a_j$. (Verify that it is a partial order!) □

The theorems of this section in fact belong into a bigger context that has its own chapter, chapter V, devoted to.

A similar argument applies in the following two famous theorems:

THEOREM IV.22 (SPERNER, 1928): Let $N = \{1, \dots, n\}$ and $A_1, \dots A_m \subset N$, such that $A_i \not\subset A_j$ if $i \neq j$. Then $m \leq \binom{n}{\lfloor n/2 \rfloor}$.

Proof: Consider the poset of subsets of N and let $\mathcal{A} = \{A_1, \dots, A_m\}$. Then $\mathcal{A}$ is an antichain.

A maximal chain $\mathcal{C}$ in this poset will consist of sets that iteratively add one new point, so there are $n!$ maximal chains, and $k!(n-k)!$ maximal chains that involve a particular k-subset of N.

We now count the pairs $(A, \mathcal{C})$ such that $A \in \mathcal{A}$ and $\mathcal{C}$ is a maximal chain with $A \in \mathcal{C}$. As a chain can contain at most one element of an antichain this is at most $n!$.

On the other hand, denoting by a_k the number of sets A_i with $|A_i| = k$, we know there are

$$n! \geq \sum_{k=0}^{n} k!(n-k)! a_k = n! \left(\sum_{k=0}^{n} \frac{a_k}{\binom{n}{k}} \right)$$

such pairs. As $\binom{n}{k}$ is maximal for $k = \lfloor n/2 \rfloor$ we get

$$\binom{n}{\lfloor n/2 \rfloor} \geq \binom{n}{\lfloor n/2 \rfloor} \sum_{k=0}^{n} \frac{a_k}{\binom{n}{k}} \geq \sum_{k=0}^{n} a_k = m.$$

□

We note that equality is achieved if $\mathcal{A}$ is the set of all $\lfloor n/2 \rfloor$-subsets of N.

THEOREM IV.23 (ERDŐS-KO-RADO, 1961): Let $\mathcal{A} = \{A_1, \dots, A_m\}$ a collection of m distinct k-subsets of $N = \{1, \dots, n\}$, where $k \leq n/2$, such that any two subsets have nonempty intersection. Then $m \leq \binom{n-1}{k-1}$.

Proof: Consider "cyclic k-sequences" $\mathcal{F} = \{F_1, \dots, F_n\}$ with $F_i = \{i, i+1, \dots, i+k-1\}$, taken "modulo n" (that is each number should be $((x-1) \bmod n) + 1$).

Note that $\mathcal{A} \cap \mathcal{F} \leq k$, since if some F_i equals A_j, then any only other $F_l \in \mathcal{A}$ must intersect F_i, so we only need to consider (again considering indices modulo n) F_l for $i - k + 1 \leq l \leq i + k - 1$. But F_l will not intersect F_{l+k}, allowing at most for a set of k subsequent F_i's to be in $\mathcal{A}$.

As this holds for an arbitrary $\mathcal{A}$, the result remains true after applying any arbitrary permutation π to the numbers in $\mathcal{F}$. Thus

$$z := \sum_{\pi \in S_n} |\mathcal{A} \cap \mathcal{F}^{\pi}| \leq k \cdot n!$$

We now calculate the sum z by fixing $A_j \in \mathcal{A}$, $F_i \in \mathcal{F}$ and observe that there are $k!(n-k)!$ permutations π such that $F_i^{\pi} = A_j$. Thus $z = m \cdot n \cdot k!(n-k)!$, proving the theorem. □

IV.5 Incidence Algebras and Möbius Functions

A common tool in mathematics is to consider instead of a set S the set of functions defined on S. To use this paradigm for (finite) posets, define an *interval* on a poset P as a set of elements z such that $x \leq z \leq y$ for a given pair $x \leq y$, and denote by $\mathrm{Int}(P)$ the set of all intervals.

For a field K, we shall consider the set of functions on the intervals:

$$I(P) = I(P, K) = \{f \colon \mathrm{Int(P)} \to K\}$$

and call it the *incidence algebra* of P. This set of functions is obviously a K-vector space under pointwise operations. We shall denote intervals by their end points x, y and thus write $f(x, y) \in I(P)$.

We also define a multiplication on $I(P)$ by defining, for $f, g \in I(P)$ a function fg by

$$(fg)(x, y) = \sum_{x \leq z \leq y} f(x, z)g(z, y)$$

In exercise **??** we will show that with this definition $I(P)$ becomes an associative K-algebra[4] with a one, given by

$$\delta(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}.$$

We could consider $I(P)$ as the set of formal K-linear combinations of intervals $[x, y]$ and a product defined by

$$[x, y][a, b] = \begin{cases} [x, b] & a = y \\ 0 & a \neq y \end{cases},$$

and extended bilinearily.

If P is finite, we can, by theorem IV.3, arrange the elements of P as $x_1, \ldots, x_n$ where $x_i \leq x_j$ implies that $i \leq j$. Then $I(P)$ is, by exercise **??** isomorphic to the algebra of upper triangular matrices $M = (m_{i,j})$ where $m_{i,j} = 0$ if $x_i \nleq x_j$.

LEMMA IV.24: Let $f \in I(P)$. Then f has a (two-sided) inverse if and only if $f(x, x) \neq 0$ for all $x \in P$.

Proof: The property $fg = \delta$ is equivalent to:

$$f(x, x)g(x, x) = 1 \qquad \text{for all } x \in P,$$

(implying the necessity of $f(x, x) \neq 0$) and

$$g(x, y) = -f(x, x)^{-1} \sum_{x < z \leq y} f(x, z)g(z, y).$$

[4] An *algebra* is a structure that is both a vector space and a ring, such that vector space and ring operations interact as one would expect. The prototype is the set of matrices over a field.

If $f(x,x) \neq 0$ the second formula will define the values f^{-1} uniquely, depending only on the interval $[x,y]$. Reverting the roles of f and g shows the existence of a left inverse and standard algebra shows that both have to be equal. □

The *zeta function* of P is the characteristic function of the underlying relation, that is $\zeta(x,y) = 1$ if and only if $x \leq y$ (and 0 otherwise).

This implies that

$$\zeta^2(x,y) = \sum_{x\leq z\leq y} \zeta(x,z)\zeta(z,y) = \sum_{x\leq z\leq y} 1 = |\{z \mid z \leq z \leq y\}|.$$

By Lemma IV.24, ζ is invertible. The inverse $\mu = \zeta^{-1}$ is called the *Möbius function* of the lattice P. The identities

$$\mu(x,x) = 1 \tag{IV.25}$$

$$\mu(x,y) = -\sum_{x\leq z<y} \mu(x,z) \tag{IV.26}$$

follow from $\mu\zeta = \delta$ and allow for a recursive computation of values of μ and imply that μ is integer-valued.

For illustration, we shall compute the Möbius function for a number of common posets.

Lemma IV.27: Let P be the total order on the numbers $\{1,\ldots,n\}$. Then for any $x,y \in P$ we have:

$$\mu(x,y) = \begin{cases} 1 & \text{if } x = y \\ -1 & \text{if } x+1 = y \\ 0 & \text{otherwise} \end{cases}$$

Proof: The case of $x = y$ is trivial. If $x + 1 = y$, the sum in IV.26 has only one summand, and the result follows. Thus assume that $x \leq y$ but $y \notin \{x, x+1\}$. Then

$$\mu(x,y) = -\mu(x,x) - \mu(x,x+1) - \sum_{x+2\leq z<y} \mu(x,z) = -\sum_{x+2\leq z<y} \mu(x,z)$$

and the result follows by induction on $y - x$. □

Lemma IV.28: If P, Q are posets, the Möbius function on $P \times Q$ satisfies

$$\mu((x_1,y_1),(x_2,y_2)) = \mu_P(x_1,x_2)\mu_Q(y_1,y_2)$$

Proof: It is sufficient to verify that the right hand side of the equation satisfies IV.26. □

Together with Theorem IV.9 and Lemma IV.27 we get

Corollary IV.29: a) For $X, Y \in \mathcal{P}(A)$, we have that $\mu(X, Y) = -1^{|Y|-|X|}$ if $X \subseteq Y$, and 0 otherwise.
b) If x, y are divisors of n, then in $\mathcal{D}(n)$ we have that $\mu(x, y) = (-1)^d$ if y/x is the product of d different primes, and 0 otherwise.

Part b) explains the name: $\mu(1, n)$ is the value of the classical number theoretic Möbius function.

Part a) connects us back to section IV.1: The Möbius function gives the coefficients for inclusion/exclusion over an arbitrary poset. We will investigate and clarify this further in the rest of this section.

Möbius inversion

The property of being inverse of the incidence function can be used to invert summation formulas with the aid of the Möbius function:

Theorem IV.30 (Möbius inversion formula): Let P be a finite poset, and $f, g\colon P \to K$, where K is a field. Then

$$g(t) = \sum_{s \le t} f(s) \qquad \text{for all } t \in P$$

is equivalent to

$$f(t) = \sum_{s \le t} g(s)\mu(s, t) \qquad \text{for all } t \in P.$$

Proof: Let K^P be the K-vector space of functions $P \to K$. The incidence algebra $I(P)$ acts linearly on this vector space by

$$(f\xi)(t) = \sum_{s \le t} f(s)\xi(s, t).$$

The two equations thus become

$$g = f\zeta, \qquad \text{respectively} \qquad f = g\mu,$$

and their equivalence follows from the fact that μ is the inverse of ζ. □

The classical Möbius inversion formula from Number Theory follows as a special case of this.

If we consider the linear poset $\{1, \ldots, n\}$ with total ordering, Lemma IV.27 gives us that (assuming $f(0) = g(0)$)

$$g(n) = \sum_{i=0}^{n} f(i)$$

is (unsurprisingly) equivalent to

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

the finite difference analog of the fundamental theorem of Calculus!

Going back to the start of the chapter, consider n subsets A_i of a set X. We take the the poset $\mathcal{P}(\{1, \ldots, n\})$ and define, for $I \subset \{1, \ldots, n\}$

$$\begin{aligned} f(I) &= -1^{|I|} \left| X \setminus \bigcup_{i \in I} A_i \right| \\ g(I) &= \left| \bigcap_{i \in I} A_i \right|. \end{aligned}$$

Then $g(I) = \sum_{J \subset I} f(J)$ by Inclusion/Exclusion over the complements $X \setminus A_i$, while

$$f(I) = \sum_{J \subset I} g(J)\mu(J, I) = \sum_{J} -1^{|I|-|J|} \left| \bigcap_{j \in J} A_j \right| = -1^{|I|} |X \setminus \bigcup A_i|$$

is the ordinary inclusion/exclusion formula.

Chapter

V

Connections

Mathematicians are like Frenchmen.
When you talk to them, they translate
it into their own language,
and then it is something quite different.

Die Mathematiker sind eine Art Franzosen;
redet man zu ihnen, so übersetzen sie es
in ihre Sprache, und dann ist es
alsobald ganz etwas anders.

Maximen und Reflexionen:
Über Natur und Naturwissenschaft
Johann Wolfgang von Goethe

With so many joints and connections, leaks were plentiful. As the magazine *The Builder* remarked, in 1856: "The fate of a theater is to be burned. It seems simply a question of time."

Connections
James Burke

In this chapter we will look at connections – both in an applied sense of modeling situations of connected objects, in the abstract sense of connecting mathematical theorems that initially seem to be unrelated, and in connecting concepts that might seem to be hopelessly abstract to practical applications. One of the joys of mathematics is to discover such connections and see the unity of the discipline.

This practical relevance of the results places much of this chapter also in close contact to the realm of (discrete) optimization.

We shall investigate a gaggle of theorems (which each in their area are fundamental), but which turn out to be equivalent in the sense that they are all equivalent in the sense that we deduce each theorem from any other. Furthermore this kind of derivation is often easier than to derive the theorem from scratch. The description owes much to [Rei84]

Many theorems in this chapter are formulated in the language of graphs: We typically denote a graph by $\Gamma = (V, E)$ with vertices V and edges E being 2-element sets of vertices. A *digraph* is a graph with directed edges (that is we consider edges as elements of $V \times V$ instead of 2-element subsets of V). In a digraph, we would allow distinct edges (x, y) and (y, x). Weighted edges means we have a weight function $w: E \to \mathbb{R}$.

Our start will be Dilworth's Theorem IV.19 that we have proven already:

> The minimum number m of chains in a partition of a finite poset P is equal to the maximum number M of elements in an antichain.

V.1 Halls' Marriage Theorem

> *He was, methinks, an understanding fellow who said, 'twas a happy marriage betwixt a blind wife and a deaf husband.*

> Celuy là s'y entendoit, ce me semble, qui dict qu'un bon mariage se dressoit d'une femme aveugle avec un mary sourd.
>
> Essais, Livre III
> Michel de Montaigne

Consider a family of subsets $A_1, \ldots, A_n \subseteq X$. A *system of distinct representatives* (SDR) for these sets is an n-tuple of *distinct* elements $x_1, \ldots, x_n$ such that $x_i \in A_i$.
Example: SDRs do not have to exist, for example consider $A_1 = A_2 = \{1\}$.

We define, for a subset $J \subseteq \{1, \ldots, n\}$ of indices, a set

$$A(J) = \bigcup_{j \in J} A_j.$$

We immediately see a reason for the above example failing: For an SDR to exist, by necessity $|A(J)| \geq |J|$ for any such subset J, since there are otherwise not sufficiently many elements available to pick distinct representatives. The following Theorem[1] shows this condition is not only necessary, but also sufficient.

Theorem V.1 (Halls' Marriage Theorem): The family $\{A_1, \ldots, A_n\}$ of finite sets has a system of distinct representatives if and only if

$$|A(J)| \geq |J| \quad \text{for all} \quad J \subseteq \{1, \ldots, n\}. \tag{V.2}$$

[1] Proven by the British Mathematician Philip Hall and extended by the unrelated American Mathematician Marshall Hall for the infinite case. Thus the apostrophe placement in the section title.

The name "Marriage Theorem" comes from the following interpretation: Suppose we have a set of m men and n women. We let A_i be the set of men that women i would consider for a potential marriage[2]. Then every women can marry a suitable man if and only if every group of k women together considers at least k men as suitable.

Proof:[DILWORTH⇒HALL] As the necessity of the condition is trivial, we show sufficiency:

Given sets A_i satisfying condition (V.2), Let $Y = \{y_1, \ldots, y_n\}$ be n symbols representing the sets. We create a poset P in the following way: The elements of P is the disjoint union $X \cup Y$. The only relations are that $x \leq y_i$ iff $x \in A_i$.

Clearly X is an antichain in P. Suppose S is another arbitrary antichain and let $J = \{1 \leq j \leq n \mid y_j \in S\}$. The antichain criterion imposes that $A(J) \cap S = \varnothing$, so

$$|S| \leq |J| + (|X| - |A(J)|) \leq |X|$$

because of (V.2). That means X is a maximal antichain, and by DILWORTH's theorem, P can be partitioned into $|X|$ chains. As a chain cannot contain more than one point from X or more than one point from Y, the pigeonhole principle implies that each chain contains exactly one point from X and at most one point from Y. Suppose that $x_i \in X$ is the element that is together with $y_i \in Y$ in a chain. Then $x_i \leq y_i$, and thus $x_i \in A_i$, so $\{x_1, \ldots, x_n\}$ is a SDR. □

As an example of the use of this theorem we consider the following theorem due to G. BIRKHOFF. A matrix $M \in \mathbb{Z}_{\geq 0}^{n\times n}$ with nonnegative entries is called *doubly statistic* if every row and every column of M has the same row/column sum.

COROLLARY V.3: A doubly statistic matrix $M = (m_{ij})$ with row/column sum l can be written as the sum of l permutation matrices.

Proof: We define sets A_i, corresponding to the row indices by $A_i = \{j \mid m_{ij} > 0\}$. For any k-tuple K of indices, the sum of the corresponding rows of M is kl. As every column of M has column sum l, this means that these k rows must have nonzero entries in at least k columns, that is $|\bigcup_{i\in K} A_i| \geq k$. Thus V.2 is satisfied and there is a SDR. We set $p_{i,j} = 1$ if j is the chosen representative for A_i (and $p_{i,j} = 0$ otherwise). Then $P = (p_{i,j})$ is a permutation matrix and $M - P$ is doubly statistic with sum $l - 1$. The statement follows by induction on l. □

[2] The theorem goes back to times of more restrictive social mores. The concerned reader might want to consider k applicants for n jobs and A_i being the set of candidates that satisfy the conditions for job i.

V.2 Kőnig's Theorems – Matchings

> Eventually everything connects - people, ideas, objects…the quality of the connections is the key to quality per se.
>
> CHARLES EAMES

Let A be an $m \times n$ matrix with 0/1 entries. A *line* of A is a row or column. A set of lines *covers* A, if every nonzero entry of A lies on at least one of the lines. Nonzero entries are called *independent* if no two lie on a common line. The *term rank* of A is the maximum number of independent entries of A.

The following theorem is reminiscent of the properties of a basis in linear algebra:

THEOREM V.4 (KŐNIG-EGERVÁRY): The minimum number of lines covering A equals the term rank of A.

Example: Let

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

Then A can be covered with 3 lines, and has an independent set of cardinality 3.

Proof:[HALL⇒KŐNIG-EGERVÁRY] Given $A \in \{0,1\}^{m\times n}$, let p be the term rank of A and q the minimum number of lines covering A. We have to show that $p = q$.

We first show that $p \leq q$: A cover of l lines can cover at most l independent 1's (no line covers more than one) but we can cover *all* ones with q lines, so $p \leq q$.

To see that $p \geq q$, without loss of generality permute the rows and columns of A such that a minimal cover involves the first r rows and the last s columns of the matrix. We aim to find an independent set that has r entries in the first r rows and in columns $1, \dots, n-s$, and s entries in the last s columns in rows $r+1, \dots, m$:

For a row index $1 \leq i \leq r$ let $N_i = \{1 \leq j \leq n-s \mid A_{i,j} = 1\}$. Then the union of any k of these N_i's contains at least k column indices – otherwise we could replace these k rows with $< k$ columns in a minimal cover. By Hall's theorem we thus have an SDR $\{x_1, \dots, x_r\}$. By definition $x_i \in N_i$ implies that $A_{i,x_i} = 1$. Let $S = \{(i, x_i) \mid i = 1, \dots, r\}$. Since the x_i are distinct, this is an independent set.

A dual argument for the last s columns gives an independent set T of positions in the last s columns and rows $r+1, \dots, n$. Since no position in S shares row or column index with any position in T, $S \cup T$ also is an independent set, of cardinality $r + s = q$. □

We reformulate this theorem in the language of graphs. A graph is called *bipartite*, if its vertex V set can be written as a disjoint union $V = V_1 \cup V_2$, so that no

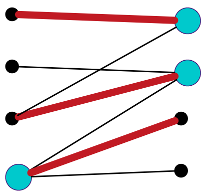


Figure V.1: A bipartite graph with maximum matching and minimum vertex cover.

vertex in V_i has a neighbor in V_i (but only neighbors in V_{3-i}). A *vertex cover* in a graph is a set of vertices that every edge is incident to at least one vertex in the cover. A *matching* in a graph is a set of edges such that no two edges are incident to the same vertex. A matching is called maximal, if no matching with a larger number of edges exists.

We assume that the graphs we consider here have no isolated vertices.

Theorem V.5 (Kőnig): In a bipartite graph without isolated vertices, the number of edges in a maximum matching equals the number of vertices in a minimum vertex cover.

Example: Figure V.1 shows a bipartite graph with a maximum matching (bold).
Proof:[Kőnig-Egerváry⇒Kőnig] Let $W, U \subset V$ be the two parts of the bipartition of vertices. Suppose $|W| = m$, $|U| = n$. We describe the adjacency in an $m \times n$ matrix A with 0/1 entries, $A_{i,j} = 1$ iff w_i is adjacent to u_j. (Note that the examples for this and the previous theorem are illustrating such a situation.)

A matching in the graph corresponds to an independent set — edges sharing no common vertices. A vertex cover — vertices such that every edge is incident — correspond to a line cover of the matrix. The result follows immediately from the Kőnig-Egerváry theorem. □

Matchings in bipartite graphs can be interpreted as creating assignments between tasks and operators, between customers and producers, between men and women, etc., and thus clearly have practical implications (Compare the Marriage Theorem interpretation of Hall's theorem!)

Due to this practical relevance, there is obvious interest in an algorithm for finding maximum matchings in a bipartite graph, the *Assignment problem*. The first algorithm published uses Kőnig's theorem as a criterion whether an existing matching can be improved. Due to the origin of the theorem authors, it has been named the *Hungarian Method*. (It turns out that this method had been discovered before independently by Jacobi [Jac90].)

Let $M \subset E$ be a (not necessarily maximal) matching in a graph $\Gamma = (V, E)$. We call an edge *matched* if it is in M, and *free* otherwise. Similarly, vertices incident to

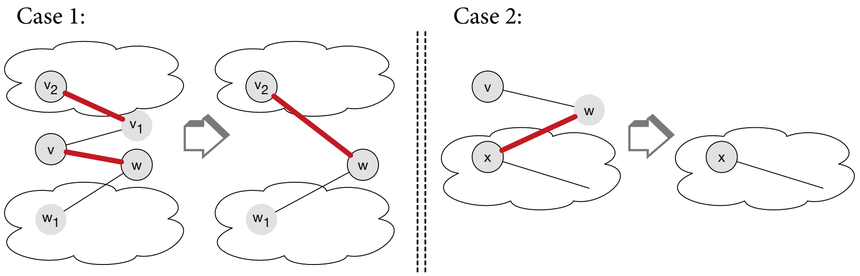


Figure V.2: The two reduction in Corollary V.6.

matched edges are called matched, vertices not incident to any matched edge are free.

An *augmenting path* for M is a path whose end points are free and whose edges are alternatingly matched and free. This means all vertices but the end points are matched.

If the cardinality of M is that of a vertex cover C, no augmenting paths can exist: Every edge of M will be incident with exactly one vertex in C, leaving one of the two free edges at the end of the path without cover. Otherwise there must be an augmenting path:

Corollary V.6: If $|M| < |C|$ for a vertex cover C, there either is a vertex cover of smaller size, or there is an augmenting path for M.

Proof: We prove the statement by induction over the number of vertices $|V|$, the base case being trivial.

Case 1: Assume that two vertices $v, w \in C$ are connected by an edge in M. Then both v and w must be incident to other edges, as we otherwise could drop one of them from C and obtain a strictly smaller matching.

Furthermore one such edge $\{v, v_1\}$ must be to a vertex $v_1 \notin C$ (as we otherwise could drop v from the vertex cover). Clearly $\{v, v_1\}$ is free. We similarly find a free edge $\{w, w_1\}$ to a vertex $w_1 \notin C$. If both v_1 and w_1 are free, then v_1, v, w, w_1 is an augmenting path. Otherwise at least one vertex, WLOG v_1, is matched thus incident to an edge $\{v_1, v_2\} \in M$. As we assumed that $v_1 \notin C$ this implies that $v_2 \in C$, as the graph is bipartite we know that $v_2 \neq w$. Now consider the smaller graph Γ' obtained from Γ by deleting v and v_1 (and the adjacent edges) and adding an edge $\{v_2, w\}$ if it did not yet exist. Then $M' = (M - \{\{v, w\}, \{v_1, v_2\}\}) \cup \{\{v_2, w\}$ is a matching in Γ' and $C' = C - \{v\}$ a vertex cover in Γ' (see Figure V.2, left). By induction, we can either obtain a strictly smaller vertex cover D for Γ' (in which case $D \cup \{v\}$ or $D \cup \{v_1\}$ is a strictly smaller cover for Γ), or an augmenting path for Γ'. If this path uses the new matching edge $\{v_2, w\}$, we replace this edge by the edge sequence $\{v_2, v_1\}, \{v_1, v\}, \{v, w\}$ and obtain an augmenting path in Γ. This completes case 1.

In Case 2, we now may assume that no two vertices of the cover are connected by an edge in the matching, as $|M| \leq |C|$ there must be an unmatched vertex $v \in C$.

If v is adjacent to any other unmatched vertex this is an augmented path of length 1.

Otherwise v is adjacent to a matched vertex $w \notin C$ (if all neighbors are in C, we could remove v from C). By a matched edge, w is adjacent to another vertex x, and to cover that edge we must have $x \in C$. We furthermore may assume that w was chosen so that x is also incident to an edge not in the matching, as we otherwise could replace all x's with the respective w's in the cover and remove v from C, reducing the cover size.

Now consider the graph Γ', obtained by removing the vertices v and w and their incident edges and set $M' = M \setminus \{\{x, w\}\}$ and $C' = C \setminus \{v\}$ (Figure V.2, right).

Then C' is a vertex cover of Γ' (all edges that v covered are gone). Also M' is a matching with $|M'| < |C'|$. By induction, either Γ' has a smaller cover D (in which case $D \cup \{v\}$ is a smaller cover for Γ), or there is an augmenting path. If this path imvolves x, we extend it by xwv and thus obtain an augmenting path for Γ. □

The existence of an augmenting path allows us to replace M with a larger matching by replacing the matched edges of an augmenting path in M with the free edges in the path.

To find a maximal matching thus simply start with some matching and searches for augmenting paths.

It is not hard to generalize this approach to (complete) bipartite graphs with weighted edges to obtain a (perfect: all vertices are matched) matching of maximal weight (profit):

Given an arbitrary matching with weighted edges, replace an edge $\{a, b\}$ with (assume: integral) weight w by a set of edges $\{a, b_1\}, \{b_1, a_2\}, \{a_2, b_2\}, \ldots, \{a_w, b\}$ with newly introduced vertices a_i, b_i that are not connected in any other way. Selecting the edge $\{a, b\}$ in the original graph thus now allows selection of w edges. Thus the maximum weight of a matching in the original graph corresponds to the cardinality of a maximum matching in the new graph.

Stable Matchings

> ...to Alvin E. Roth and Lloyd S. Shapley "for the theory of stable allocations and the practice of market design".
>
> Citation for the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2012

Also of practical interest is the concept of a stable matching (or stable marriage). Imagine a matching in a bipartite graph represents an assignment of students to college places. Every student has a personal preference ranking of colleges, every college (place – assume for simplicity that every college just has one student) has a personal ranking of students. A matching is *instable* (otherwise: *stable*), if there is a student s and a college c such that s ranks c higher than their current college, and c ranks s higher than their current student.

A priori it is unclear that stable matchings exist, the following algorithm not only proves that but also gives a way of producing one:

ALGORITHM V.7 (Deferred Acceptance, GALE, SHAPLEY): Every student has a temporarily assigned college, colleges may have a temporarily assigned student. Once a college has been assigned a student, it will replace it only by one it ranks higher. A student may be moved down from higher ranked to lower ranked colleges.

For bookkeeping, students carry a *rejected* label that will change status as the algorithm goes on, and also and carry for each college an indicator whether they have been rejected by this college.

1. [Initialize] Label every student as rejected

2. [Complete] If no student is rejected, terminate.

3. [Student Choice] Every student marked as rejected applies to the college she ranks highest amongst those who have not yet rejected her.

4. [College Choice] Every college tentatively picks from the students who chose it the one it ranks highest (and removes that students rejection status). It rejects all other students who have selected it,even if they had been picked tentatively before.

5. Go back to step 2.

Proof: The only way the process terminates is if no student is rejected, that is they were tentatively accepted by a college and this college has no student applying it would rank higher. All colleges which the student ranks higher have rejected her because they were able to accept a student they rank higher, so the matching is stable.

In every round, students select from colleges and are either tentatively accepted, or are left with colleges of lower preference. As there is a finite set of students and preferences, this process must terminate. □

This algorithm is used for example in the US to match graduating medical doctors to hospital training positions, www.nrmp.org. It also was instrumental for the award of the 2012 Nobel prize in Economics to SHAPLEY and ROTH.

V.3 Menger's theorem

> Live in fragments no longer. Only connect, and the beast and the monk, robbed of the isolation that is life to either, will die.
>
> Howards End
> E. M. Forster

Kőnig's aim in proving the Theorem V.5 was as a tool towards a proof for a more general theorem by Menger that deals with connections in a graph (and gets us in statement and proof back close to Dilworth's theorem), but whose proof, when first published, turned out to not work in the case of bipartite graphs.

Let u, w be nonadjacent vertices in a graph Γ. A *uw-vertex path* in Γ is a path from u to w. A collection of uw-vertex paths is called *independent*, if the paths are pairwise disjoint except for u and w. A set S of vertices, excluding u and w, is *uw-vertex separating*, if every uw path must contain a vertex of S. Clearly the minimum cardinality of a uw vertex separating set must be at least as large as that of a number of uw independent paths. It turns out that they are equal:

Theorem V.8 (Menger): Let u, w be nonadjacent vertices in a graph. Then the maximum number of independent uw vertex paths equals the minimum cardinality of a uw vertex separating set.

As Kőnig's proof is comparatively long, we will instead give a direct proof due to G. Dirac. Note the similarity to the proof of Dilworth's theorem!
Proof: Let m be the minimum cardinality of a separating set, and M the maximum number of independent paths. We have seen that $m \geq M$.

Assume that the theorem is false and $\Gamma = (V, E)$ be a graph with a minimum number of edges for which the theorem fails, that is we have two nonadjacent vertices u and w and fewer than m independent vertex paths.

If e is an edge not incident to u or w, we can remove e and obtain a smaller graph Δ for which the theorem holds. Since Γ, and thus Δ has at most $m - 1$ uw-independent paths this means that in Δ there is an uw-separating set T of size $m - 1$. If we add one of the vertices incident to e to this set T, we obtain a set S of cardinality m which is uw separating – any path connecting in Γ but not in Δ clearly would have to use the edge e. We thus may assume that S contains a vertex not adjacent to u or w.

Next, we note that there is no path usw with $s \in S$ of length 2 – if there was we could remove s and have a graph with fewer edges and $m - 1$ separating vertices and less than $m - 1$ independent paths, contradicting the minimality of Γ.

For the final contradiction, denote by P_u the set of vertex paths between u and exactly one vertex in S and P_w ditto for w. The paths in P_u and P_w have only vertices of S in common (otherwise we could circumvent S, contradicting that it is separating).

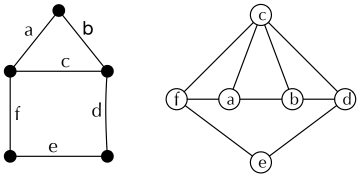


Figure V.3: A graph and its line graph

We claim that all vertices of S are adjacent to u, or that all vertices in S are adjacent to w. This will contradict the choice of S and prove the theorem:

Consider the graph consisting of P_u, w and edges sw for $s \in S$. If w is not adjacent to all vertices in s, this graph has fewer edges than G, so by assumption has m independent paths from u to w. If we leave out the sw-step from these paths we obtain a set R_u of m independent paths from u to the m elements of S. Similarly, we get a set of independent paths R_w from w to the elements of S. Combining these paths at the elements of s yields m independent paths from u to w, contradiction. □

A dual version of the theorem holds for disjoint paths:

Theorem V.9 (Menger's theorem, edge version): The maximum number of edge-disjoint paths from u to w equals the minimum number of edges, whose removal would put u and v into disjoint components.

The proof of the theorem is by the ordinary version, applied to the *line graph* $L(\Gamma)$ of Γ. The vertices of $L(\Gamma)$ are the edges of Γ, two vertices are adjacent if the edges in Γ are incident to the same vertex. See figure V.3 for an example.

Menger's theorem also generalizes in the obvious way to directed graphs, though we shall not give a statement or proof here.

V.4 Max-flow/Min-cut

> And he thought of himself floating on his back down a river, or striking out from one island to another, and he felt that that was really the life for a Tigger.
>
> The house at Pooh corner
> A. A. Milne

A *network* is a digraph (V, E) with weighted edges with two distinguished vertices, the *source* s and the *target* t. We assume that the weights (which we call *capacities*) $c: E \to \mathbb{R}_{\geq 0}$ are nonnegative. (Below we shall assume that weights are in fact

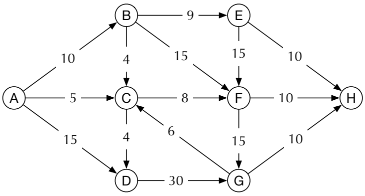


Figure V.4: A network

rational – clearly we can approximate real weights to arbitrary precision. We shall investigate the issue of irrational capacities in the exercises.) An example is given in Figure V.4. For a (directed) edge $e = (a, b)$ we denote by $\iota(e) = a$ the initial, and by $\tau(e) = b$ the terminal vertex of this edge.

The reader might want to consider this as a network of roads with weights being the maximal capacity. (It is possible to have different capacities in different directions on the same road.) We want to transport vehicles from s to t (and implicitly ask how many we can transport through the network).

Note that we are looking for a steady state in a steady-state transport situation; that is we do not want to transport once, but forever, and ask for the transport capacity per time unit.

A *flow* is a function f defined on the edges (indicating the number of units we transport along that edge) such that

- $0 \leq f(e) \leq c(e) \forall e \in E$ (do not exceed capacity).
- $\sum_{\iota(e)=v} f(e) = \sum_{\tau(e)=v} f(e) \forall v \in V \setminus \{s, t\}$ (No intermediate vertex can buffer capacity or create new flow. This is Kirchhoff's law in Physics.)

The *value of a flow* is the net transport out of the source:

$$\operatorname{val}(f) = \sum_{\iota(e)=s} f(e) - \sum_{\tau(e)=s} f(e).$$

This definition might seem asymmetric in focusing on the source, however Lemma V.10 will show that this is a somewhat arbitrary choice.

The question we ask thus is for the maximal value of a flow. The main tool for this is to consider the flow passing "by a point" in the network:

A *cut* $C \subset V$ in a network is a set of vertices such that $s \in C$, $t \notin C$. We do not require the cut to be connected, though in practice it usually will be.

We note that the value of any flow is equal to the net flow out of a cut. (This implies that the value of a flow also is equal to the net transport into the target):

Lemma V.10: Let C be a cut. We denote the edges crossing the cut by

$$\begin{aligned} C^o &= \{e \in E \mid \iota(e) \in C, \tau(e) \notin C\}, \\ C^i &= \{e \in E \mid \tau(e) \in C, \iota(e) \notin C\}. \end{aligned}$$

Then for any flow f we have that

$$\mathrm{val}(f) = \sum_{e \in C^o} f(e) - \sum_{e \in C^i} f(e).$$

Proof: We prove the statement by induction on the number of vertices in C. For $|C| = 1$ we have that $C = \{s\}$ and the statement is the definition of the value of a flow.

Now suppose that $|C| > 1$, let $s \neq v \in C$ and $D = C \setminus \{v\}$. By induction, the statement holds for D.

We now consider the following disjoint sets of edges:

$$\begin{aligned} P^o &= \{e \in E \mid \iota(e) \in D, \tau(e) \notin C\}, \\ P^i &= \{e \in E \mid \tau(e) \in D, \iota(e) \notin C\}, \\ Q^o &= \{e \in E \mid \iota(e) \in D, \tau(e) = v\}, \\ Q^i &= \{e \in E \mid \tau(e) \in D, \iota(e) = v\}, \\ R^o &= \{e \in E \mid \iota(e) = v, \tau(e) \notin C\}, \\ R^i &= \{e \in E \mid \tau(e) = v, \iota(e) \notin C\}. \end{aligned}$$

Any edge that starts at v must be in Q^i or in R^o, any edge ending at v must be in Q^o or in R^i. Kirchhoffs law thus gives us that

$$\sum_{e \in Q^i} f(e) + \sum_{e \in R^o} f(e) = \sum_{e \in Q^o} f(e) + \sum_{e \in R^i} f(e).$$

It is easily seen that $C^o = P^o \cup R^0$, $C^i = P^i \cup R^i$ and that $D^o = P^o \cup Q^o$ and $D^i = P^i \cup Q^i$. Therefore, using the previous expression to replace a difference of Q-sums by a difference of R-sums:

$$\begin{aligned} \mathrm{val}(f) &= \sum_{e \in D^o} f(e) - \sum_{e \in D^i} f(e) \\ &= \sum_{e \in P^o} f(e) - \sum_{e \in P^i} f(e) + \underbrace{\sum_{e \in Q^o} f(e) - \sum_{e \in Q^i} f(e)}_{=\sum_{e \in R^o} f(e) - \sum_{e \in R^i} f(e)} \\ &= \sum_{e \in C^o} f(e) - \sum_{e \in C^i} f(e). \end{aligned}$$

The claim follows by induction. □

We define the *capacity* of a cut as the sum of the capacity of the edges leaving the cut:

$$\operatorname{cap}(C) = \sum_{e \in C^o} c(e).$$

Lemma V.10 thus implies that the value of *any* flow must be bounded by the capacity of *any* cut and thus (unsurprisingly) the maximal flow value is bounded by the minimum cut capacity – a chain is only as strong as its weakest link.

Similar to the dualities we considered before, equality is attained:

Theorem V.11 (Max-Flow Min-Cut, integer version): Suppose the capacity function c in integer valued. Then the maximum value of a flow in a network is equal to the minimum capacity of a cut. Furthermore, there is a maximum flow f that is integer valued.

Note V.12: If the capacity function is rational valued, we can simply scale with the lcm of the denominators and obtain an integer valued capacity function. In the case of irrational capacities, the theorem holds by a boring approximation argument.

Proof:[Menger⇒Max-flow] Given a directed network, replace every arc with integral weight c by c disjoint[3] directed edges.

Clearly, if one edge of a c-fold multi-edge set is in a minimum separating set, the other $c-1$ edges have to be.

Thus the cardinality of a minimum edge-separating set is a minimum cut, while k independent paths give a flow of k. By the edge version of Menger's theorem, the minimum cardinality of an edge-separating set equals the maximum number of independent edge paths, proving the theorem. □

The algorithm of Ford and Fulkerson finds a maximal flow for a given network. It start with a valid flow (say $f(e) = 0$ for all edges) and then improves this flow if possible, using the following main step:

Algorithm V.13 (increase flow): Given a flow f, return a flow with higher value, or show that no such flow exists.

1. Let $A := \{s\}$. *(A is a cut of vertices that can be supplied at higher rate from the source).* $I := \{\}$. *(I is a set of edges that are not yet at capacity.)* $R := \{\}$. *(R is a set of edges whose flow should be reduced.)*

2. If $t \in A$, go to step 6.

3. If there is an edge $e \in A^o$ with $f(e) < c(e)$, set $A := A \cup \{\tau(e)\}$, $I := I \cup \{e\}$. Go to step 2.

4. If there is an edge $e \in A^i$ with $f(e) > 0$ *(This is flow into A which we could reduce)*, set $A := A \cup \{\iota(e)\}$, $R := R \cup \{e\}$. Go to step 2.

[3] Or edges → ∘ → with intermediate vertices to ensure disjointness

5. If neither of these two properties hold, terminate with a message that the flow is maximal.

6. By tracing back how t got added to A, we construct an (undirected) path (the *augmenting path*) $P = (e_1, e_2, \ldots)$ from s to t. Let

$$d = \min\left(\min_{e\in P\cap I}(c(e) - f(e)), \min_{e\in P\cap R} f(e)\right).$$

Increase the flow on $P \cap I$ by d, reduce the flow on $P \cap R$ by d. *(This satisfies Kirchhoffs law.)* Return the larger flow.

Proof: We shall show termination in the case of integral capacities: Every time we adjust a flow, we may assume that the flow on one edge increases by an integral value. This can only happen a finite number of times.

Once the algorithm ends in step 5, we have a cut A such that $\sum_{e\in A^i} f(e) = 0$, $\sum_{e\in A^o} f(e) = \sum_{e\in A^o} c(e)$, thus $\mathrm{val}(f) = \mathrm{cap}(A)$ and the flow must be maximal by Theorem V.11. □

Example: We illustrate the algorithm in the example from above. Figure V.5, a) shows some flow (in blue). The total flow is 24.

We start building the list of under-supplied vertices and mark edges on which to increase, respectively reduce the flow:

Vertices	Increase	Reduce
s		
B	sB	
A		AB
D	AD	
t	DT	

We find an augmenting path sBADt (red in Figure b) and note that we can increase the flow by 2 (the limiting factor being edge sB) to a total of 26.

Starting again, we obtain (Figure c) the augmenting path sCFBAEt, on which we can increase the flow by 2 again, resulting is a flow of 28.

Finally we build the under-supplied cities as C,F,B and then cannot increase the set. Thus (Figure d), shaded) $\{s, C, F, B\}$ is a cut whose capacity of 28 is reached, proving that we have a maximal flow.

Note V.14: As given, the algorithm is not necessarily in polynomial time, as there might be a huge number of small incremental steps. Take the network in Figure V.6. Starting with a flow of 0, we choose the augmenting path sABt on which we can increase the flow by 1. Next, we take the augmenting path sBAt, reducing the flow on AB to 0 again and have a total flow of 2. We can iterate this, increasing the flow by 1 each step until we get the maximum flow of 200.

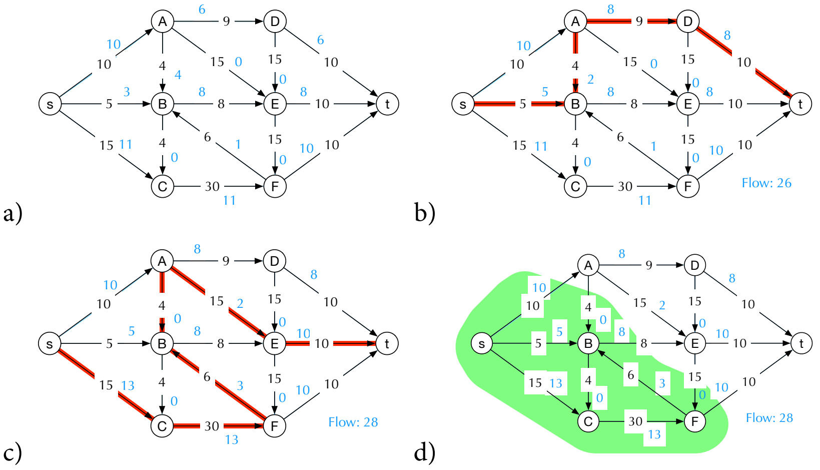


Figure V.5: Increasing a given flow to maximum

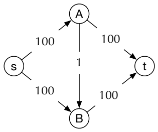


Figure V.6: Example for bad augmenting paths

Polynomial time in this algorithm however can be achieved by a more careful choice of the augmenting path (e.g. a shortest path, Edmonds-Karp algorithm), that analysis however is beyond the scope of these notes.

Note V.15: To close the circle, observe that the Max-flow/Min-cut theorem implies Dilworth's theorem: Given a poset, we model it as a directed network with flow from bottom to top and arc capacity one. Thus all theorems mentioned in this section (and in fact a few more we have not mentioned) are in a sense mutually equivalent and form the fundamental theorem of discrete optimization.

Braess' Paradox

> On Earth Day this year, New York City's Transportation Commissioner decided to close 42d Street, which as every New Yorker knows is always congested. "Many predicted it would be doomsday," said the Commissioner [...] But to everyone's surprise [...] Traffic flow actually improved when 42d Street was closed.
>
> New York Times, 12/25/1990, p.38
> Gina Kolata

> From the time of Say and Ricardo the classical economists have taught that supply creates its own demand
>
> The General Theory of Employment, Interest, and Money
> John Maynard Keynes

Often the question of maximal flow turns up not just for an existing network, but already at the stage of network design, aiming to maximize flow. Braess' paradox shows that this can be very nonintuitive.

While we shall give a theoretical example, concrete instances of this effect have been observed in the real world in several cities (New York; Stuttgart, Germany; Winnipeg, Canada;...) when roads had been temporarily blocked because of building work, or after new roads were built. The interested reader might observe obvious implications to society and politics.

Suppose we have four cities, A, B, C, D with connecting roads as shown in figure V.7, left. The reader may observe that this configuration is similar to the one in Note V.14.

Cities A and B as well as C and D are connected by a minor road which easily clogs up. The driving time thus depends strongly on the number of cars, it is

$$t_{AB} = t_{CD} = 10x.$$

with x the number of drivers in thousands. An obstacle blocks direct connections from B to C.

High capacity roads have been build to connect A and C, as well as B and D. While overall time is longer, the impact for extra cars is less, time with x thousand drivers on the road is

$$t_{AC} = t_{BD} = 50 + x.$$

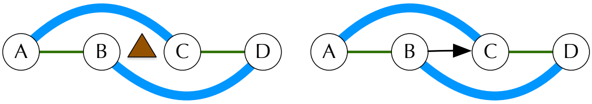


Figure V.7: Braess' paradox: Before and after a new road is built.

Assume 6000 drivers want to travel from A to D. A symmetry argument shows that half (3000) travel via B and half via C. Their individual travel time is $10 \cdot 3 + 50 + 3 = 83$. It is not hard to show that this is a stable equilibrium, i.e. no traveler can gain by changing their route. Furthermore, the system will settle in this equilibrium by itself, as long we assume drivers want to minimize their travel time and when starting their journey have perfect knowledge of how many cars are on each road.

After time, a new connection is built from B to C through the obstacle (right image). (For simplicity of the argument we shall assume that it is one-way, but that is not crucial for the paradox.) It has high capacity and is short, so travel time on this route is

$$t_{BC} = 10 + x.$$

Assuming perfect information, drivers will move towards a distribution in which travel time along all possible routes is equal. In the example this is 2000 drivers using ABD, 2000 drivers ACD and 2000 drivers using ABCD. (This means that 4000 drivers are traveling AB and CD, respectively, 2000 travel BC.) Their travel times are (again, one can show that this is a stable equilibrium)

$$t_{ABD} = 10 \cdot 4 + 10 + 2 + 10 \cdot 4 = 92 = 10 \cdot 4 + 50 + 2t_{ABD} = t_{ACD},$$

which is more than before.

Chapter

VI

Partitions, Tableaux and Permutations

Some of the hardest, but also most intriguing, counting problems are those of partitions.

VI.1 Partitions and their Diagrams

An *order-irrelevant partition* (from now on simply: partition) of n into k parts is an expression

$$n = x_1 + x_2 + \cdots + x_k, \qquad x_1 \geq x_2 \geq \cdots\cdots x_k \geq 1$$

with all x_i being integers. We will typically denote partitions by small Greek letters and write $\lambda \vdash n$ to mean that λ is a partition of n.

As in II.4, we denote by $p_k(n)$ the number of partitions of n into k parts and by $p(n)$ the total number of partitions of n (into any number of parts). OEIS A000041 For example

$$4 = 4 = 3 + 1 = 2 + 2 = 2 + 1 + 1 = 1 + 1 + 1 + 1,$$

so $p(4) = 5$ and

$$7 = 5 + 1 + 1 = 4 + 2 + 1 = 3 + 3 + 1 = 3 + 2 + 2,$$

giving $p_3(7) = 4$. Instead of writing a sum, we might give a count of the part sizes, that is instead of $7 = 3 + 2 + 2$ we would write $3^1 2^2$.

A convenient way to depict a partition is graphically by what is called a *Young diagram* (boxes) or *Ferrers' diagram* (dots): Each part is depicted by a row of dots or of empty squares, in descending[1] order. Figure VI.1 gives two versions of the diagram for the partition $3^1 2^2 1^1$:.

[1] this is the convention in English speaking countries. The French build instead from the bottom up.

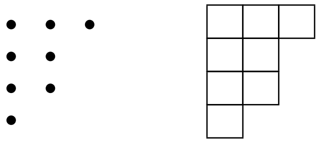


Figure VI.1: The Ferrers and Young diagram for $3^1 2^2 1^1$

The *conjugate* λ^* of a partition λ is the partition with the "transposed" diagram, for example if $\lambda = 3^1 2^2 1^1$, then $\lambda^* = 4^1 3^1 1^1$. Clearly $\lambda^{**} = \lambda$.

VI.2 Some partition counts

The number $p_k(n)$ is clearly equal to the number of ways to write $n-k = y_1 + \cdots + y_k$ with $y_1 \geq \cdots \geq y_k \geq 0$. If exactly s of these integers y_i are nonzero, there will be $p_s(n-k)$ of this type. This gives the recursion

$$p_k(n) = \sum_{s=1}^{k} p_s(n-k)$$

which, together with the initial values $p_k(n) = 0$ for $n < k$ and $p_k(k) = 1$ gives us a way to determine $p_k(n)$ recursively.

We start by giving a generating function for the partition function $p(n)$:

Proposition VI.1:

$$\sum_n p(n) t^n = \prod_{i \geq 1} (1 - t^i)^{-1}$$

Proof: Expanding the right hand side, and using that $(1-t^i)^{-1} = (1 + t^i + t^{2i} + \cdots)$, we get

$$\prod_i (1-t^i)^{-1} = (1 + t + t^2 + \cdots)(1 + t^2 + t^4 + \cdots)\cdots$$

In this product, t^n is obtained as a product of t^{x_1} from the first factor, t^{2x_2} from the second, and so on, such that $1^{x_1} 2^{x_2} \cdots \vdash n$. Vice versa, every partition of n gives a different way to form t^n, that is $p(n)$ is the coefficient of t^n. □

This infinite product does not have a nice closed form, but we will use it below in Corollary VI.6 to obtain a recursion formula for $p(n)$.

If we consider $p_k(n)$ for a fixed k, we only need to consider a partial product in the expansion. This can be used to calculate the value directly:

Lemma VI.2: $p_3(n)$ is the integer closest to $\frac{n^2}{12}$.

Proof: Let $a(n) = p_3(n+3)$ the number of solutions of $n = x_1 + x_2 + x_3$ with $x_1 \geq x_2 \geq x_3 \geq 0$. We set $y_3 = x_3$, $y_2 = x_2 - x_3$ and $y_1 = x_1 - x_2$ and thus get that

$a(n)$ is the number of solutions of $n = y_1 + 2y_2 + 3y_3$. As in the previous proof, we get a generating function

$$\sum_n a(n)t^n = (1-t)^{-1}(1-t^2)^{-1}(1-t^3)^{-1}.$$

A partial fraction decomposition gives us

$$\begin{aligned}\sum_n a(n)t^n &= \frac{1}{6}(1-t)^{-3} + \frac{1}{4}(1-t)^{-2} + \frac{17}{72}(1-t)^{-1}\\ &\quad + \frac{1}{8}(1+t)^{-1} + \frac{1}{9}(1-\zeta t)^{-1} + \frac{1}{9}(1-\zeta^2 t)^{-1}\end{aligned}$$

with ζ a primitive 3-rd root of unity. Using the series expression

$$(1-t)^{-a-1} = \sum_j \binom{a+j}{j} t^j,$$

collecting and telescoping, we get

$$a(n) = \frac{1}{12}(n+3)^2 - \frac{7}{72} + \frac{(-1)^n}{8} + \frac{1}{9}(\zeta^n + \zeta^{2n}).$$

We write this in the form

$$\left|a(n) - \frac{1}{12}(n+3)^2\right| \le \frac{7}{72} + \frac{1}{8} + \frac{2}{9} < \frac{1}{2}$$

giving the desired result. □

More generally, we have the following approximation:

Proposition VI.3: For fixed k, we have that

$$p_k(n) \sim \frac{n^{k-1}}{k!(k-1)!} \qquad (n \to \infty).$$

Proof: Suppose that $n = x_1 + \cdots + x_k$ with $x_1 \ge \cdots \ge x_k \ge 1$. If we permute the variables x_i, we get (not necessarily different) compositions of n into k parts and each composition can be obtained by a permutation of a partition. By Section II.1 there are $\binom{n-1}{k-1}$ such compositions, giving us

$$k!p_k(n) \ge \binom{n-1}{k-1}.$$

We now set $y_i = x_i + (k-i)$ for $1 \le i \le k$. Then the y_i are all different, as

$$x_i + k - i = y_i = y_j = x_j + k - j$$

implies that $x_i + j = x_j + i$. Assuming WLOG that $i < j$ we have that $x_i \leq x_j$, contradiction. Clearly $y_1 \geq y_2 \geq \cdots y_k$.

The summation formula over i gives (after an index swap) that $y_1 + \cdots + y_k = n + \frac{k(k-1)}{2}$ is a partition of $n + \frac{k(k-1)}{2}$. As the parts are different, every permutation gives a different composition, but we do not necessarily obtain all compositions this way. Thus

$$k!p_k(n) \leq \binom{n + \frac{k(k-1)}{2} - 1}{k-1}.$$

Combining the two inequalities, and dividing by $k!$ gives (with $a_k = \frac{k(k-1)}{2}$ only dependent on k)

$$\frac{(n-1)(n-2)\cdots(n-k+1)}{k!(k-1)!} \leq p_k(n) \leq \frac{(n+a_k-1)\cdots(n+a_k-k+1)}{k!(k-1)!}.$$

As k (and thus a_k) are both fixed, both numerators approach n^{k-1} if $n \to \infty$. □

VI.3 Pentagonal Numbers

> Our chief weapon is surprise...
> surprise and fear...
> fear and surprise...
> Our **two** weapons are fear and surprise...
> and ruthless efficiency...
>
> The Spanish Inquisition
> Monty Python

For reasons we shall see soon, we divert in a surprising way.

A number is *pentagonal* if it is of the form $k(3k-1)/2$ or generalized pentagonal OEIS A001318 if it is of the form $k(3k+1)/2$ for some positive integer k. Alternatively, one can simply use $k(3k-1)/2$ for arbitrary integers k. The first few pentagonal numbers for positive index are $1(3\cdot 1-1)/2 = 1$, $2(3\cdot 2-1)/2 = 5$, $3(3\cdot 3-1)/2 = 12$, $4(3\cdot 4-1)/2 = 22$ as depicted in Figure VI.2, respectively $0(3\cdot 0-1)/2 = 0$, $-1(3\cdot(-1)-1)/2 = 2$, $-2(3\cdot(-2)-1)/2 = 7$, $-3(3\cdot(-3)-1)/2 = 15$ for other indices.

The pentagonal numbers give the following extraordinary[2] theorem.

Theorem VI.4 (Euler's Pentagonal Numbers Theorem): a) If n is not a pentagonal number, the number of partitions of n into an even and an odd number of distinct parts are equal.

[2] as in the title: *une loi tout extraordinaire*, [Eul15]

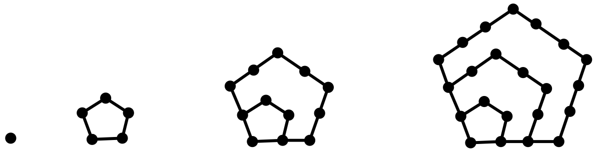


Figure VI.2: The first few pentagonal numbers: 1,5,12,22

b) If $n = k(3k-1)/2$ for $k \in \mathbb{Z}$ even, then the number of partitions of n into an even number of distinct parts is one more than the number of partitions into an odd number of distinct parts.
c) If $n = k(3k-1)/2$ for $k \in \mathbb{Z}$ odd, then the number of partitions of n into an even number of distinct parts is one less than the number of partitions into an odd number of distinct parts.

Example: There are 2 partitions of 6 into an even number of distinct parts and 2 into an odd number of distinct parts. There are 3 partitions of 7 into an even number of distinct parts, and 2 into an odd number of distinct parts. There are 7 partitions of 12 into even parts, and 8 into odd parts. Proof: The proof will consist of constructing a bijection between partitions with an even number of distinct parts and partitions with an odd number of distinct parts. In the case that n is pentagonal this will have leave out exactly one partition.

For a partition $\lambda \vdash n$ into different parts, we consider two subsets of the cells, as depicted in Figure VI.3:

The *base* is the bottom row of cells (the shortest row), the *slope* is the cells starting at the end of the top row and descending straight down-left as far as possible (this might be only of length 1).

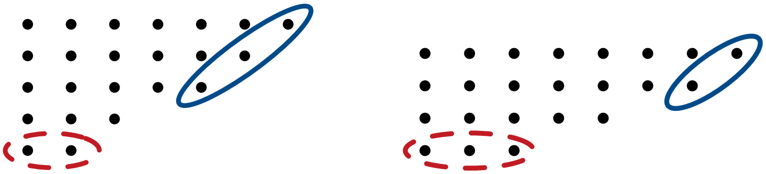


Figure VI.3: Base (dashed) and Slope of two partitions

Base and slope might intersect in one cell, we call those cells of the base that are not in the slope the *pure base*; similarly the *pure slope* are those slope cells that are not also in the base.

We now divide the partitions of n with distinct parts into three classes:

Class 1 are those partitions for which the pure base contains more cells than the slope.

Class 2 are those partitions for which the pure slope is at least as long as the base.

Class 3 are all remaining partitions[3].

For a partition λ in Class 1, we create a new partition μ by removing the slope (possibly including the base point) and placing these cells in a new row at the bottom of the diagram. As the pure base is larger than the slope this results in a partition into different parts. Furthermore, the slope of μ is at least as long as the slope of λ (which is the base of μ) if base and slope of μ are distinct; if they intersect the slope of μ is strictly larger. This means μ is in Class 2.

Vice versa, if μ is in Class 2, we remove the base (possibly including the lowest point of the slope) and attach it from top right as new slope. This results in a partition λ into different parts which lies in Class 1.

For example, the left partition in Figure VI.3 is in Class 2 and will be transformed into the partition on the right side in Class 1, and vice versa.

It is easily seen that these two operations are mutually inverse, that is Class 1 and Class 2 contain the same number of partitions. Furthermore the operations change the number of parts by exactly one. That means that in Class 1 ∪ Class 2, there are exactly as many partitions with an even number of parts, as with an odd number of parts.

Finally consider Class 3: A partition in this class must have base and slope intersecting, and the base contains the same number of cells, or exactly one more, than the slope, see Figure VI.4.

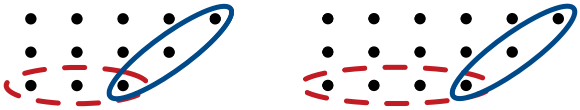


Figure VI.4: Possibilities for partitions in Class 3

This means that, if there are k parts, it consists of a $k \times k$ rectangle and a $1, 2, \ldots, k$ triangle that might or might not overlap. In the first case there are $k^2 + k(k-1)/2 = k(3k-1)/2$ parts, respectively $k^2 + k(k+1)/2 = k(3k+1)/2 = (-k)(3(-k)-1)/2$ parts. For a given n only one of these two options is possible, and it implies that n is a pentagonal number. This proves the theorem. □

[3] Classifications become easy if we allow an "all the rest" category

To simplify notation we write $\omega(k) = k(3k-1)/2$ for $k \in \mathbb{Z}$ to state the following consequence for the inverse of the generating function of partitions.

Corollary VI.5:

$$\prod_{n\geq 1}(1-t^n) = \sum_{k=-\infty}^{\infty}(-1)^k t^{\omega(k)} = 1 + \sum_{k>0}(-1)^k\left(t^{\omega(k)} + t^{\omega(-k)}\right).$$

Proof: The second equality simply relies on the now familiar two ways to describe generalized pentagonal numbers.

Let $\mathrm{even}(n)$ be the number of partitions of n into an even number of distinct parts, $\mathrm{odd}(n)$ similarly for an odd number of parts. By the Pentagonal Numbers Theorem, the right hand side of the equation is the generating function for $\mathrm{even}(n) - \mathrm{odd}(n)$. We aim to show that this is true also for the left hand side:

The coefficient for t^n will be made up by contributions from terms of the form t^{n_i}, all distinct, such that $n = n_1 + \cdots + n_l$. This partition contributes $(-1)^l$. That is, every partition into an even number of distinct parts contributes 1, every partition into an odd number of distinct parts contributes −1, which was to be shown. □

This now permits us to formulate a recurrence formula for $p(n)$:

Corollary VI.6:

$$\begin{aligned} p(n) &= \sum_{k>0}(-1)^{k-1}\left(p(n-\omega(k)) + p(n-\omega(-k))\right) \\ &= p(n-1) + p(n-2) - p(n-5) - p(n-7) + p(n-12) + \cdots \end{aligned}$$

Proof: We know that $\sum_n p(n)t^n = \prod(1-t^n)^{-1}$, and therefore

$$\left(\sum_n p(n)t^n\right)\cdot\left(1 + \sum_{k>0}(-1)^k\left(t^{\omega(k)} + t^{\omega(-k)}\right)\right) = 1.$$

We now consider the coefficient of t^n is this product (which is zero for $n > 0$). This gives

$$0 = p(n) + \sum_{k>0}(-1)^k\left(p(n-\omega(k)) + p(n-\omega(-k))\right).$$

By using that $p(n) = 0$ for $n < 0$ and substituting values for the pentagonal numbers we get the explicit recursion. □

This formula can be used for example to calculate values for $p(n)$ effectively, or to estimate the growth of the function $p(n)$.

While it might not seems so we have by now wandered deep into Number theory. In this area we get for example Ramanujan's famous congruence identities

$$\begin{aligned} p(5n+4) &\equiv 0 \pmod{5} \\ p(7n+5) &\equiv 0 \pmod{7} \\ p(11n+6) &\equiv 0 \pmod{11} \end{aligned}$$

that recently have been generalized by Ono and collaborators.

Euler's theorem and its consequence are a special case of the *Jacobi triple product identity*

$$\prod_{n=1}^{\infty}(1-q^{2n})(1+q^{2n-1}t)(1+q^{2n-1}t^{-1}) = \sum_{r=-\infty}^{\infty} q^{r^2} t^r$$

which has applications in the theory of theta-functions – the territory of Number Theory – and Physics.

VI.4 Tableaux

> *It's like the story of the drawing on the blackboard, remember?*

> C'est comme pour l'histoire du dessin sur le tableau noir, tu te rappelles?
>
> Histoires inédites du Petit Nicolas,
> Tome 1
> René Goscinny

Drawing a partition as Young diagram with boxes allows us to place numbers in them.

1	3	4	8	12
2	6	9	11	
5	7			
10				

Definition VI.7: Let $\lambda \vdash n$. A (standard) Young tableau[4] is a Young diagram of shape λ into whose boxes the numbers $1,\ldots,n$ have been entered, so that every row and every column is (strictly) increasing.

Beyond their intrinsic combinatorial interest, tableaux are a crucial tool for studying permutations, and for the representation theory of symmetric groups.

An obvious question with a surprising answer is to count the number f_λ of Young tableaux of shape λ. In the case that $\lambda = n^2$ as partition of $2n$, we observe, exercise **??**, that $f_\lambda = C_{n+1}$, the counts f_λ thus can be considered as generalizing Catalan numbers.

Our aim is to derive the famous hook formula (Theorem VI.13) for f_λ. While this formula is inherently combinatorial, the proof we shall give is mainly algebraic – indeed, until surprisingly recently [GNW79, Zei84] no purely combinatorial proof was known.

We start with a recursively defined function f on m-tuples of natural numbers (for arbitrary m). (As it will turn out that this function is f_λ, we use the same letter f.)

[4] French: "(black)board". Plural: *tableaux*

1. $f(n_1,\ldots,n_m) = 0$, unless $n_1 \geq n_2 \geq \cdots \geq 0$.
2. $f(n_1,\ldots,n_m,0) = f(n_1,\ldots,n_m)$.
3. $f(n_1,\ldots,n_m) = f(n_1-1,n_2,\ldots,n_m) + f(n_1,n_2-1,\ldots,n_m) + \cdots$ $+f(n_1,n_2,\ldots,n_m-1)$, if $n_1 \geq n_2 \geq \cdots n_m \geq 0$.
4. $f(n) = 1$ if $n \geq 0$.

Properties i) and iv) are base cases, property ii) allows for an arbitrary number of arguments, Property iii) finally gives a recursive formula that defined the values of f in terms of smaller arguments; thus f is uniquely defined.

Lemma VI.8: If $\lambda = (n_1,\ldots,n_m)$, then $f_\lambda = f(n_1,\ldots,n_m)$ is the number of Young tableaux of shape λ.

Proof: Conditions i), ii) and iv) are obviously satisfied by f_λ. For condition iii), take a tableau of shape $(n_1,\ldots,n_m)$ and consider the location of n. It needs to be at the end of a row (say row j) and the bottom of a column, and if we remove that entry, we obtain a valid tableau with $n-1$ entries and of shape $(n_1,\ldots,n_{i-1},n_i-1,n_{i+1},\ldots,n_m)$. (Note that some of the terms could be zero if $n_i - 1 < n_{i+1}$.) □

Next we shall need an identity in the polynomial ring in variables $x_1,\ldots,x_m,y$.

Definition VI.9: The *root of the discriminant* of $x_1,\ldots,x_n$ is defined as

$$\Delta(x_1,\ldots,x_n) := \prod_{1\leq i<j\leq m} (x_i - x_j)$$

The eagle-eyed reader will recognize the right hand side as a Vandermonde determinant. The next, auxiliary lemma is simply an identity about a particular multivariate polynomial.

Lemma VI.10: Define a polynomial $g \in \mathbb{Q}[x_1,\ldots,x_m,y]$ as

$$\begin{aligned} g(x_1,\ldots,x_m;y) &:= x_1\Delta(x_1+y,x_2,\ldots,x_m) \\ &+ x_2\Delta(x_1,x_2+y,\ldots,x_m) + \cdots + x_m\Delta(x_1,x_2,\ldots,x_m+y). \end{aligned}$$

Then

$$g(x_1,\ldots,x_m;y) = \left(x_1 + \cdots + x_m + \binom{m}{2}y\right)\Delta(x_1,\ldots,x_m).$$

Proof: Clearly g is a homogeneous (every monomial in g has the same total degree) polynomial of degree $1 + \deg\Delta(x_1,\ldots,x_m)$. If we interchange x_i and x_j, then g changes its sign, thus, if we set $x_i = x_j$ the polynomial evaluates to 0. Thus (consider g as univariate in x_i with all other variables part of a rational function field) $x_i - x_j$ divides g. Hence $\Delta(x_1,\ldots,x_m)$ divides g.

We now concentrate instead on the variable y. The degree of g as a polynomial in y is at most one, and for $y = 0$ the statement is obvious. We therefore only need to show that the coefficient of y in g is $\binom{m}{2}\Delta$.

If we expand g, the terms in y are $\frac{x_i y}{x_i - x_j}\Delta(x_1, \dots, x_m)$ and $-\frac{x_j y}{x_i - x_j}\Delta(x_1, \dots, x_m)$ for pairs i, j with $1 \leq i < j \leq m$. These two terms add up to $y\Delta$, thus the $\binom{m}{2}$ pairs $(i < j)$ yield in sum $\binom{m}{2}\Delta(x_1, \dots, x_m)$. □

Using the defining recurrences, we can now verify an (otherwise unmotivated) formula for f_λ:

Lemma VI.11: If $\lambda = (n_1, \dots, n_m)$, then

$$f_\lambda = \frac{\Delta(n_1 + m - 1, n_2 + m - 2, \dots, n_m)n!}{(n_1 + m - 1)!(n_2 + m - 2)! \cdots n_m!}.$$

In fact this formula satisfies properties i-iv) as long as $n_1 + m - 1 \geq n_2 + m - 2 \geq \cdots n_m$.

Proof: We shall show that the right hand side satisfies conditions i)-iv). To make the recursion in iii) work we need to include the case that $n_{i+1} = n_i + 1$, but in that case $n_{i+1} + m - i + 1 = n_i + 1 + m - i + 1 = n_i + m - i$, that is two arguments of Δ become equal and the right hand side evaluates to 0.

With this, properties i) and iv) become trivially true.

Concerning property ii), if we replace m by $m + 1$ and have $n_{m+1} = 0$, the denominator changes by a factor $(n_1 + m + 1 + 1) \cdots (n_m + 1)$ which is the same factor by which the numerator changes.

Concerning property iii), this is the statement of VI.10 for $x_i = n_i + m - i$ and $y = -1$. □

The Hook Formula

> Thus, although the term "hooker" did not originate during the Civil War, it certainly became popular […] in tribute to the proclivities of General Joseph Hooker […]
>
> Studies in Etymology and Etiology
> David Gould

A much nicer way to describe the values of f_λ is given by the following concept:

Definition VI.12: Given a cell in a Young diagram, its *hook* consists of the cell itself as well as the cells to the right in the same row, and those in the same column below. The *hook length* is the number of cells in the hook.

Theorem VI.13 (Hook formula, Frame, de Robinson, Thrall (1954)): The number of tableaux of shape $\lambda \vdash n$ is given by $n!$, divided by the product of all hook lengths.

Example: Consider $\lambda = 6^1 4^2 2^1$. The numbers in the diagram for λ indicate hook lengths:

9	8	6	5	2	1
6	5	3	2	●	
5	4	2	1	●	
2	1	●			

Thus

$$f_\lambda = \frac{16!}{1^3 \cdot 2^4 \cdot 3 \cdot 4 \cdot 5^3 \cdot 6^2 \cdot 8 \cdot 9} = \frac{20922789888000}{62208000} = 336336.$$

Proof: Suppose that $\lambda = (n_1 m \ldots, n_m) \vdash n$. Then the hook length in the top left corner is $n_1 + m - 1$. The hook length of the second cell is $n_1 + m - 2$, and so on, seemingly down to hook length 1 in the last cell. As in the example above, certain lengths are skipped. This happens if a lower row is shorter than the top row. This is indicated by the $\bullet$-dots in the above diagram.

We are thus skipping $(n_1 + m - 1) - (n_m)$, in the column after the last row ends, $(n_1 + m - 1) - (n_{m-1} + 1)$ after the second last row ends and so on.

Going to the second row, we get hook lengths $n_2 + m - 2$, $n_2 + m - 3$, and so on, but excluding $(n_2 + m - 2) - (n_m)$ for the last row, $(n_2 + m - 2) - (n_{m-1} + 1)$ for the second last row, etc.

Combining this, we get as product of the hook lengths a fraction with numerator $(n_i + m - i)!$ for the i-th row, and the denominator combining the canceled terms to $\Delta(n_1 + m - 1, n_2 + m - 2, \ldots, n_m)$, proving the theorem with use of Lemma VI.11. □

We close this section with the remark that the numbers f_λ are also the degrees of the irreducible (ordinary) representations of the symmetric group. Proving this requires a substantial amount of abstract algebra and cannot be done here..

A consequence is the nice identity that $\sum_{\lambda \vdash n} f_\lambda^2 = n!$, which we will also prove below in VI.39

VI.5 Symmetric Functions

> *Reality favors symmetries and slight anachronisms*

> A la realidad le gustan las simetrías y los leves anacronismos
>
> El Sur
> Jorge Luis Borges

For a commutative ring R, consider the ring $A = R[x_1, \ldots, x_N]$ of polynomials in N variables[5]

[5] This is the posh way of defining this set of polynomials. The reader will be perfectly well suited if she simply sets $R = \mathbb{Q}$ and considers A as the set of rational polynomials in n variables.

For writing polynomials, it will be convenient to introduce the following notation. A composition (see II.1) $\alpha = (n_1, n_2, \cdots, n_k)$ of a number M is a sequence of numbers such that $M = n_1 + \cdots + n_k$. We also write $\alpha_i = n_i$ to denote the i-th part of a composition α. We can consider partitions of M to be a special case of compositions. We shall use a vector notation for monomials, writing

$$\underline{x}^\alpha := x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k},$$

this way an arbitrary polynomial may be written as $\sum_\alpha c_\alpha \underline{x}^\alpha$.

Definition VI.14: A polynomial $f \in A$ is a *symmetric function*, if it stays the same after any permutation of the arguments: $f(x_1, \ldots x_n) = f(x_{1^\pi}, \ldots, x_{n^\pi})$ for $\pi \in S_n$.

Some literature instead talks about *symmetric polynomials*, meaning the same property.

The set of symmetric functions clearly forms a subring of A, we denote this subring by Λ.

Note VI.15: One could, in the spirit of chapter VII consider this as an action of S_n on the polynomial ring and ask what happens if we restrict to a subgroup. This is the topic of *Invariant Theory* – a beautiful subject but not the one we study here.

As a permutation of the indices leaves the degree of each term invariant we can consider a symmetric function as a sum of *homogeneous* (that is each term has the same total degree) parts. It therefore is sufficient to consider homogeneous symmetric functions of degree n, say.

As a vector space, we can consider $\Lambda = \Lambda^0 \oplus \Lambda^1 \oplus \cdots$ as a direct sum of subspaces Λ^i consisting of homogeneous symmetric functions of degree i.

The following definition describes some particular classes of symmetric functions:

Definition VI.16: Let $n > 0$ and $\lambda = n_1 + n_2 + \cdots + n_k \vdash n$.

a) The *Monomial Symmetric Function* m_λ is the sum $\sum_\alpha \underline{x}^\alpha$ where α runs over all *distinct* permutations of (the entries of) λ.

b) The *Elementary Symmetric Function* e_m is the sum over all products of m (distinct) indeterminates, that is $e_m = m_{1^m}$. We define $e_\lambda = e_{n_1} \cdot e_{n_2} \cdots e_{n_k}$.

c) The *Complete Homogeneous Symmetric Function* h_m is the sum over all products of m (possibly repeated) indeterminates: $h_n = \sum_{\lambda \vdash n} m_\lambda$. We define $h_\lambda = h_{n_1} \cdot h_{n_2} \cdots h_{n_k}$.

d) The *Power Sum Function* p_m is the sum $x_1^m + x_2^m + \cdots + x_N^m$ over the m-th power of all N indeterminates. We define $p_\lambda = p_{n_1} \cdot p_{n_2} \cdots p_{n_k}$.

It is easily seen that all of these functions are indeed symmetric.

Example: Suppose that $N = 3$ and $\lambda = 2 + 1$. Then

$$\begin{array}{rcl} m_\lambda &=& x_1^2x_2 + x_2^2x_1 + x_1^2x_3 + x_3^2x_1 + x_2^2x_3 + x_3^2x_2, \\ e_\lambda &=& (x_1x_2 + x_1x_3 + x_2x_3)(x_1 + x_2 + x_3), \\ p_\lambda &=& (x_1^2 + x_2^2 + x_3^2)(x_1 + x_2 + x_3), \\ h_\lambda &=& e_\lambda + p_\lambda \end{array}$$

Note VI.17: It can be convenient to consider N arbitrary large or even infinite – the identities amongst symmetric functions still hold.

Each of these classes of symmetric functions allows to generate all:

Theorem VI.18: Suppose that $N \geq n$ and f a homogeneous symmetric function of degree n in $x_1, \ldots, x_N$. Then

a) $f = \sum_{\lambda \vdash n} c_\lambda m_\lambda$ for suitable coefficients c_λ.

b) $f = \sum_{\lambda \vdash n} c_\lambda e_\lambda$ for suitable coefficients c_λ.

c) $f = \sum_{\lambda \vdash n} c_\lambda h_\lambda$ for suitable coefficients c_λ.

d) $f = \sum_{\lambda \vdash n} c_\lambda p_\lambda$ for suitable coefficients c_λ.

In each case the coefficients c_λ are unique, furthermore in cases a,b,c), if f has integral coefficients, then all c_λ are integral.

Corollary VI.19: Any symmetric function $f(x_1, \ldots, x_N)$ can be written as a polynomial $g(z_1, \ldots, z_N)$, where z stands for one of the symbols e,h, p.

Note VI.20: For algebraists we remark that one can show that the ring of symmetric polynomials in N variables has transcendence degree N and that the $\{e_i\}$, the $\{h_i\}$ and the $\{p_i\}$ each form a transcendence basis of this ring.

Proof:[of Theorem VI.18, part a)] If $\boldsymbol{f = \sum_\alpha c_\alpha \underline{x}^\alpha}$ is homogeneous of degree n, then $f = \sum_{\lambda \vdash n} c_\lambda m_\lambda$. □

Corollary VI.21: The set $\{m_\lambda\}_{\lambda \vdash n}$ is a vector space basis of Λ^n, thus $\dim(\Lambda^n) = p(n)$.

We shall see below that each of the other sets, $\{e_\lambda\}$, $\{h_\lambda\}$, $\{p_\lambda\}$, also forms a basis.

VI.6 Base Changes

The proof we shall give for the remaining parts of Theorem VI.18 will be based on linear algebra: We shall consider the coefficients of expressing one set of symmetric functions in terms of another set, and obtain a combinatorial interpretation of these coefficients. From this we shall deduce that the matrix of coefficients is invertible (and possibly has an integral inverse). This implies that the other sets also form a basis and thus the remaining parts of the theorem.

As we will argue with matrices, it involves an arrangement of basis vectors, we thus need to define an ordering on partitions:

Definition VI.22: Let $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ and Let $\mu = (\mu_1, \mu_2, \ldots, \mu_l)$ be partitions of n. We assume WLOG that $k = l$ by allowing cells of size 0.

We say that $\mu \leq \lambda$ in the *natural order*, (also called *dominance order* or *majorization order*) if $\mu \neq \lambda$ and for any $i \geq 1$ we have that

$$\mu_1 + \mu_2 + \cdots + \mu_i \leq \lambda_1 + \lambda_2 + \cdots + \lambda_i.$$

Following Theorem IV.3, this partial order has a total order as a linear extension, one such total order is given in exercise **??**.

Definition VI.23: Let $A = (a_{ij})$ a matrix. We consider row sums $r_i = \sum_j a_{ij}$ and column sums $c_j = \sum_i a_{ij}$ and call $\mathrm{row}(A) = (r_1, r_2, \ldots)$ the *row sum vector*, respectively $\mathrm{col}(A) = (c_1, c_2, \ldots)$ the *column sum vector*.

Since the m_λ form a basis there are integral coefficients $M_{\lambda\mu}$, such that

$$e_\lambda = \sum_{\mu \vdash n} M_{\lambda\mu} m_\mu.$$

In fact $M_{\lambda\mu}$ is simply the coefficient of $\underline{\boldsymbol{x}}^{\boldsymbol{\mu}}$ in e_λ.

Lemma VI.24: $M_{\lambda\mu}$ equals the number of $(0,1)$-matrices (that is, matrices whose entries are either 0 or 1) $A = (a_{ij})$ satisfying $\mathrm{row}(A) = \lambda$ and $\mathrm{col}(A) = \mu$.

Proof: Let $\lambda = (\lambda_1, \lambda_2, \ldots)$. Any term of e_λ is a product of a term of e_{λ_1}, e_{λ_2}, etc. and any term of e_{λ_i} is the product of λ_i different variables. We shall describe any term using a matrix. Let

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots \\ x_1 & x_2 & x_3 & \cdots \\ \vdots & & \vdots & \end{pmatrix}$$

Then each monomial $\underline{\boldsymbol{x}}^{\boldsymbol{\alpha}}$ of e_λ is the product of λ_1 entries from the first row, λ_2 entries from the second row and so on. We describe the selection of the factors by a $(0,1)$-matrix A with 1 indicating the selected factors.

The products contributing to a particular monomial $\underline{\boldsymbol{x}}^{\boldsymbol{\alpha}}$ correspond exactly to the matrices with row sums λ and column sums α. The statement follows. □

We denote by $M = (M_{\lambda\mu})$ the matrix of these coefficients.

As the transpose of a matrix exchanges rows and columns we get

COROLLARY VI.25: The matrix M is symmetric, that is $M_{\lambda\mu} = M_{\mu\lambda}$.

We now establish the fact that the matrix M is, for a suitable arrangement of the partitions, triangular with diagonal entries 1. This shows[6] that $\det M = 1$, that is M is invertible over $\mathbb{Z}$. This in turn implies Theorem VI.18, b).

PROPOSITION VI.26: Let $\lambda, \mu \vdash n$. Then $M_{\lambda\mu} = 0$, unless $\mu \leq \lambda^*$. Also $M_{\lambda\lambda^*} = 1$. Thus, if the λ are arranged in (a linear extension of) the natural ordering, and the μ according to ordering of its duals, then M is upper triangular with diagonal 1.

Proof: Suppose that $M_{\lambda\mu} \neq 0$, so by Lemma VI.24 there is a $(0,1)$-matrix A with $\mathrm{row}(A) = \lambda$ and $\mathrm{col}(A) = \mu$. Let A^* be the matrix with $\mathrm{row}(A^*) = \lambda$ and the 1's left justified, that is $A^*_{ij} = 1$ if and only if $1 \leq j \leq \lambda_i$. By definition of the dual, we have that $\mathrm{col}(A^*) = \lambda^*$. On the other hand, for any j the number of 1's in the first j columns of A is not less that the number of 1's in the first j columns of A, so in the natural order of partitions we have that

$$\lambda^* = \mathrm{col}(A^*) \geq \mathrm{col}(A) = \mu.$$

To see that $M_{\lambda\lambda^*} = 1$ we observe that A^* is the *only* matrix with $\mathrm{row}(A^*) = \lambda$ and $\mathrm{col}(A^*) = \lambda^*$. □

NOTE VI.27: This approach of proving that the e_λ form a basis — expressing the e_λ in terms of the m_μ, giving a combinatorial interpretation of the coefficients in this expression, and deducing from there that the coefficient matrix must be invertible — could also be applied to the other two bases, h_λ and p_λ. We shall instead give other proofs, in part because they illustrate further aspects of symmetric functions.

For the reader who feels that we have moved far away from enumerating combinatorial structures, we close this section with the following observation (whose proof is left as exercise **??**):

Suppose we have n balls in total, of which λ_i balls are labeled with the number i. We also have boxes 1, 2,.... Then $M_{\lambda\mu}$ is the number of ways to place the balls into the boxes such that box j contains exactly μ_j balls, but no box contains more than one ball with the same label.

Complete Homogeneous Symmetric Functions

For the complete homogeneous symmetric functions h_λ we get the following analog to Lemma VI.24:

[6] A consequence of the Cayley-Hamilton theorem is that the inverse of a matrix A is a polynomial in A with denominators being divisors of $\det(A)$.

Lemma VI.28: Define $N_{\lambda\mu}$ as the coefficient of $\underline{\boldsymbol{x}}^{\mu}$ in h_λ, that is

$$h_\lambda = \sum_{\mu\vdash n} N_{\lambda\mu} m_\mu.$$

Then $N_{\lambda\mu}$ equals the number of $\mathbb{N}_0$-matrices $A = (a_{ij})$ satisfying $\operatorname{row}(A) = \lambda$ and $\operatorname{col}(A) = \mu$.

Proof: Exercise **??**

We now establish a duality between the e-polynomials and the h-polynomials.

Definition VI.29: We define a map $\omega\colon \Lambda^n \to \Lambda^n$ by setting

$$\omega\colon \begin{array}{l} e_i \to h_i \\ e_\lambda \to h_\lambda \end{array}$$

and extending linearly.

As the $\{e_\lambda\}_{\lambda\vdash n}$ for a basis of Λ^n this defines ω uniquely and shows that ω is a linear map. Since the e_i are also algebraically independent (a fact we shall take as given), ω is also an algebra endomorphism, that is it preserves products.

Proposition VI.30: The endomorphism ω is an involution, that is ω^2 is the identity. In particular $\omega(h_n) = e_n$ and $\omega(h_\lambda) = e_\lambda$.

We note that Theorem VI.18, c) is an immediate consequence of this proposition.
Proof: We go in the ring $\Lambda[[t]]$ of formal power series over Λ and set

$$\begin{aligned} E(t) &:= \sum_{n\geq 0} e_n t^n, \\ H(t) &:= \sum_{n\geq 0} h_n t^n. \end{aligned}$$

We have that

$$\begin{aligned} E(t) &= \prod_{r=1}^{N} (1 + x_r t), \\ H(t) &= \prod_{r} (1 + x_r t + x_r^2 t^2 + \cdots) = \prod_{r=1}^{N} (1 - x_r t)^{-1}, \end{aligned}$$

as can be seen by expanding the products on the right hand side. This implies that $H(t)E(-t) = 1$. Comparing coefficients of t^n on both sides yields

$$0 = \sum_{i=0}^{n} (-1)^i e_i h_{n-i}, \qquad n \geq 1.$$

We now apply ω, use $\omega(e_i) = h_i$, and re-index and thus get

$$0 = \sum_{i=0}^{n} (-1)^i h_i \omega(h_{n-i}) \qquad \text{and} \qquad 0 = \sum_{i=0}^{n} (-1)^i h_{n-i} \omega(h_i)$$

We consider this as a linear system of equations with the $\omega(h_i)$ as variables. The coefficient matrix of this system is lower triangular with diagonal entries $h_0 = 1$, thus has full rank. The solution therefore must be unique, but we know already that the e_i form a solution, proving the theorem. □

We finally come to the power sum functions p_λ. Define

$$P(t) := \sum_{n\geq 1} p_n t^{n-1}$$

Proposition VI.31: $\frac{\mathrm{d}}{\mathrm{d}t} H(t) = P(t)H(t)$ and $\frac{\mathrm{d}}{\mathrm{d}t} E(t) = P(-t)E(t)$.

<u>Proof:</u>

$$\begin{aligned} P(t) &= \sum_{r\geq 1} p_r t^{r-1} = \sum_{r\geq 1} \sum_{i=1}^{N} x_i^r t^{r-1} \\ &= \sum_{i=1}^{N} \frac{x_i}{1 - x_i t} = \frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=1}^{N} \log(1 - x_i t)^{-1} \\ &= \frac{\mathrm{d}}{\mathrm{d}t} \log\left(\prod_{i=1}^{N} (1 - x_i t)^{-1}. \right) \end{aligned}$$

The result for H follows by logarithmic differentiation. The argument for E is analog. □

By considering the coefficients of the power series, we can write this result in the form $nh_n = \sum_{r=1}^{n} p_r h_{n-r}$. This allows us to express the h_i in terms of the p_i, albeit at the cost of introducing denominators. For example, $h_2 = \frac{1}{2}(p_1^2 + p_2)$.

Therefore the p_λ generate as vector space and we get Theorem VI.18 d); albeit not necessarily with integral coefficients.

We close this section with the mention of another basis of symmetric polynomials which is of significant practical relevance, but more complicated to define. Given a partition $\lambda = (\lambda_1, \lambda_2, \ldots)$, define

$$snum_\lambda(x_1, \ldots, x_n) = \det \begin{pmatrix} x_1^{\lambda_1+n-1} & x_2^{\lambda_1+n-1} & \cdots & x_n^{\lambda_1+n-1} \\ x_1^{\lambda_2+n-2} & x_2^{\lambda_2+n-2} & \cdots & x_n^{\lambda_2+n-2} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{\lambda_n} & x_2^{\lambda_n} & \cdots & x_n^{\lambda_n} \end{pmatrix}.$$

By the properties of the determinant, *snum* is invariant under all even permutations of the variables, but is not symmetric. It thus must be divisible by the Vandermonde determinant $\prod_{i<j}(x_i - x_j)$, and the quotient

$$s_\lambda(x_1,\ldots,x_n) = \frac{snum_\lambda(x_1,\ldots,x_n)}{\prod_{i<j}(x_i - x_j)}$$

is symmetric. We call the set of s_λ the *Schur polynomials*. They form another basis of Λ^n, but we will not prove this here.

Amongst the many interesting properties of Schur functions, one can show, for example, that in the expression $s_\lambda = \sum_\mu K_{\lambda\mu} m_\mu$, the coefficient $K_{\lambda\mu}$ (called a *Kostka Number*) is given by the number of semistandard tableaux (that is, we allow duplicate entry and enforce strict increase in columns, but allow equality or increase in rows) of shape λ. Also the base change matrix from $\{p_\lambda\}$ to $\{s_\lambda\}$ is the character table of the symmetric group S_n.

VI.7 The Robinson-Schensted-Knuth Correspondence

> From ghoulies and ghosties
> And long-leggedy beasties
> And things that go bump in the night,
> Good Lord, deliver us!
>
> Scottish Prayer
> TRADITIONAL

We now investigate a miraculous connection between permutations and pairs of tableaux. For this we relax the definition of a standard tableau to not requires entries from $\{1,\ldots,n\}$, but allow larger entries (thus skipping some smaller numbers). It also will be convenient to consider permutations as sequences of images.

The fundamental tool s an *insert* routine that takes a tableau and a number and inserts the number in the tableau, adding one cell and moving some entries, so that the tableau property is preserved.

NOTE VI.32: The web page `https://integral-domain.org/lwilliams/Applets/index.php` by L. WILLIAMS has a number of calculators for the algorithms in this chapter.

The routine proceeds by inserting the entry a in the first row. Let x be the smallest entry in this row for which $x \geq a$. If no such element exists, add a (in a new cell) at the end of the row. Otherwise a "bumps out" x and takes its place in this row. Then x is inserted into by the same process into the second row, possibly bumping out another number which then is inserted into the third row and so on.

Example: Suppose we want to insert 8 into the tableau

1	3	5	9	12	16
2	6	10	15		
4	13	14			
11					
17					

The smallest entry of the first row larger than 8 is 9, which gets bumped out, resulting in the row 1, 3, 5, 8, 12, 16 and entry 9 being inserted into the second row.

There 10 is the smallest entry larger than 9 and gets bumped out, resulting in row 2, 6, 9, 15.

Inserting 10 into row three bumps out 13. Finally 13 is added at the end of row four, obtaining the tableau

1	3	5	8	12	16
2	6	9	15		
4	10	14			
11	13				
17					

To be able to prove statements, we now describe this process in a formal way for a tableau $T = (T_{ij})$. For convenience we assume that the tableau is bordered by 0's to the top and left, and by ∞ to the bottom and right, so that T_{ij} is defined for all $i, j \geq 0$.

We also define a relation $\prec$ on entries of T by

$$a \prec b \text{ if and only if } a < b \text{ or } a = b = 0 \text{ or } a = b = \infty,$$

with this the property of being a tableau is given by the following characterization

$$T_{ij} = 0 \text{ if and only if } i = 0 \text{ or } j = 0; \qquad \text{(VI.33)}$$

$$T_{ij} \prec T_{i(j+1)} \text{ and } T_{ij} \prec T_{(i+1)j}, \text{ for all } i, j \geq 0. \qquad \text{(VI.34)}$$

We write $x \notin T$ if $x = \infty$ or $x \neq T_{ij}$ for all $i, j \geq 0$.

Algorithm VI.35 (Insert): Let $T = (T_{i,j})$ a tableau and $x \notin T$ a positive integer. This algorithm transforms T into another tableau that contains x in addition to the elements of T and adds one cell in a position (s, t) determined by the algorithm.

1. [Input x.] Set $i := 1$, $x_1 := x$, and j smallest such that $T_{1j} = \infty$.
2. [Find x_{i+1}.] *(At this point $T_{(i-1)j} < x_i < T_{ij}$ and $x_i \notin T$.)* If $x_i < T_{i(j-1)}$ then decrease j and repeat the step. Otherwise set $x_{i+1} := T_{ij}$ and $r_i := j$.
3. [Replace by x_i.] *(Now $T_{i(j-1)} < x_i < x_{i+1} = T_{ij} \prec T_{i(j+1)}$, $T_{(i-1)j} < x_i < x_{i+1} = T_{ij} \prec T_{(i+1)j}$, and $r_i = j$.)* Set $T_{ij} := x_i$.

4. [At end?] *(Now $T_{i(j-1)} < T_{ij} = x_i < x_{i+1} \prec T_{i(j+1)}$, $T_{(i-1)j} < T_{ij} = x_i < x_{i+1} \prec T_{(i+1)j}$, $r_i = j$, and $x_{i+1} \nmid nT$.)* If $x_{i+1} \neq \infty$ then increase i by 1 and return to step 2.

5. return $s := i$, $t := j$ and terminate. *(At this point $T_{st} \neq \infty$ and $T_{(s+1)t} = T_{s(t+1)} = \infty$.)*

Proof: The parenthetical remarks in the algorithm ensure that T remains a tableau at each step. □

The algorithm implicitly defines a "bumping sequence"

$$x = x_1 < x_2 < \cdots < x_s < x_{s+1} = \infty,$$

as well as column indices

$$r_1 \geq r_2 \geq \cdots \geq r_s = t.$$

We now consider a slight generalization of permutations, namely *two-line arrays* of integers A=

$$\begin{pmatrix} q_1 & q_2 & \cdots & q_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix} \qquad \text{(VI.36)}$$

such that $q_1 < q_2 \cdots < q_n$ and that all p_i are different. Permutations thus are the special case of $q_i = i$ and $\{p_1, p_2, \ldots, p_n\} = \{1, \ldots, n\}$.

The following algorithm takes such an array and produces a set of two tableaux of the same shape

ALGORITHM VI.37 (Robinson-Schensted-Knuth): Given a two-line array of the form (VI.36), the following algorithm produces two tableaux P, Q of the same shape with entries from the p_i, respectively the q_i:

1. Initialize P and Q to the empty tableau. Let $i := 1$.

2. Insert p_i into P, obtaining a position (s, t).

3. Place q_i into a new cell at position (s, t) in Q.

4. Increment i. If $i \leq n$ go to step 2. Otherwise terminate and return P and Q.

Proof: The proof of Algorithm VI.35 shows that P is a tableau. Furthermore, the position s, t returned is always a valid position to add a cell to the Young diagram. As the entries of q_i are increasing this shows that Q always is a tableau as well. □

We abbreviate this process as RSK and note that if the array is permutation both P and Q are standard tableaux with entries $1, \ldots, n$.

Example: We illustrate the RSK process with the example of the array $\begin{pmatrix} 1 & 3 & 5 & 7 & 9 \\ 5 & 2 & 3 & 6 & 4 \end{pmatrix}$.

Insert 5 and 1 $P = \boxed{5}$, $Q = \boxed{1}$

Insert 2 and 3 $P = \begin{array}{|c|}\hline 2\\\hline 5\\\hline\end{array}\ , Q = \begin{array}{|c|}\hline 1\\\hline 3\\\hline\end{array}$

Insert 3 and 5 $P = \begin{array}{|c|c|}\hline 2&3\\\hline 5&\multicolumn{1}{c}{}\\\cline{1-1}\end{array}\ , Q = \begin{array}{|c|c|}\hline 1&5\\\hline 3&\multicolumn{1}{c}{}\\\cline{1-1}\end{array}$

Insert 6 and 7 $P = \begin{array}{|c|c|c|}\hline 2&3&6\\\hline 5&\multicolumn{2}{c}{}\\\cline{1-1}\end{array}\ , Q = \begin{array}{|c|c|c|}\hline 1&5&7\\\hline 3&\multicolumn{2}{c}{}\\\cline{1-1}\end{array}$

Insert 4 and 9 $P = \begin{array}{|c|c|c|}\hline 2&3&4\\\hline 5&6&\multicolumn{1}{c}{}\\\cline{1-2}\end{array}\ , Q = \begin{array}{|c|c|c|}\hline 1&5&7\\\hline 3&9&\multicolumn{1}{c}{}\\\cline{1-2}\end{array}$

Amazingly this process is a proper bijection

Theorem VI.38 (RSK Correspondence): There is a bijection between the set of permutations of $\{1, \ldots n\}$ and the set of ordered pairs (P, Q) of tableaux of the same shape, formed from $\{1, \ldots, n\}$. Under this bijection, if g corresponds to the part (P, Q), then g^{-1} corresponds to (Q, P).

Example: The permutation $(1, 3, 2, 5, 4, 6) \in S_7$ has an array with second row 3, 5, 2, 6, 4, 1 and gives the tableaux

$$P = \begin{array}{|c|c|c|c|}\hline 1&4&6&7\\\hline 2&5&\multicolumn{2}{c}{}\\\cline{1-2} 3&\multicolumn{3}{c}{}\\\cline{1-1}\end{array}\ , Q = \begin{array}{|c|c|c|c|}\hline 1&2&4&7\\\hline 3&5&\multicolumn{2}{c}{}\\\cline{1-2} 6&\multicolumn{3}{c}{}\\\cline{1-1}\end{array}\ .$$

Before proving this theorem we notice a number of consequences concerning the number f_λ of tableaux:

Corollary VI.39:

$$\sum_{\lambda \vdash n} f_\lambda^2 = n!$$

Example: For $n = 4$ there are 5 partitions and respective tableau counts of $f_{1^4} = 1$, $f_{1^2 2^1} = 3$, $f_{2^2} = 2$, $f_{1^1 3^1} = 3$, $f_{4^1} = 1$. Also $1^2 + 3^2 + 2^2 + 3^2 + 1^2 = 24 = 4!$.

Corollary VI.40: The number of involutions (permutations that are self-inverse, see III.6 for another formula) is given by $\sum_{\lambda \vdash n} f_\lambda$.

Proof: A permutation g is an involution if and only if $g = g^{-1}$. That means that involutions correspond to pairs with $P = Q$. □

Example: S_4 has one identity, three elements of shape $(1, 2)(3, 4)$ and $\binom{4}{2} = 6$ elements of shape $(1, 2)$, thus $10 = 1 + 3 + 2 + 3 + 1$ involutions.

Proof of the RSK correspondence

> The unusual nature of these coincidences
> might lead us to suspect that some sort
> of witchcraft is operating behind the scenes.
>
> The Art of Computer Programming,
> Section 5.1.4: Tableaux.
> DONALD E. KNUTH

The first part of the proof is to observe that Algorithm VI.35 can be reversed:

We say that a cell of a tableau is "*at the edge*", if there is no cell to the right or below. The insertion algorithm adds a cell which is on the edge of the new tableau.

Now suppose we have a tableau into which an element x has been inserted, resulting in a new cell at the border of the resulting tableau P, say at coordinates (s,t). If we are in row $s=1$, the entry of the cell is simply the element x.

Otherwise the entry x_s must have been bumped out of the previous row. It was bumped out by an element x_{s-1} and x_s was the smallest number in the row that was larger than x_{s-1}. That is, we can identify x_{s-1} in P as the largest number in row $s-1$ that is smaller than x_s. We thus can simply put x_s back into its old position and now have to deal with x_{s-1}. If $s-1=1$ it was the element x inserted into the tableau, otherwise we repeat the process for $s-2$ and so on.

We can consider this process as an "uninsert" algorithm that takes an (arbitrary) tableau P, a position (s,t) and returns a tableau P' missing a cell in position $s,t)$ as well as an element x. (It is easily see that at all steps the tableau properties are maintained.)

The description of the uninsert process also ensures that, after uninserting, we could obtain P back by inserting x into P' – insert and uninsert are mutual inverses.

Proof:(of Theorem VI.38, part 1) To establish a bijection, we describe an inverse process to Algorithm VI.37: Given two tableaux P, Q, we produce a two-line array A.

1. Let A be the empty two-line array.

2. Identify the position (s,t) of the largest entry q of Q. *((s,t) must be at the edge of* Q*.)* Remove that entry q from Q.

3. Apply the uninsert process to P for position (s,t), resulting in the removal of an element x.*(and reducing* P *to a tableau as the same shape as the reduced* Q*.)*

4. Add the column $\binom{q}{x}$ *at the front* of A.

5. Unless P and Q are empty, go back to step 2.

The result of this process is a two-line array A with the first row consisting of the numbers $1, 2 \ldots, n$ in canonical order, and the second row containing each of these numbers exactly once, that is a permutation.

It is easily seen that this process and the RSK algorithm are mutually inverse, thus establishing the bijection. □

To prove the second part, inversion corresponding to swap of the tableaux, we need further analysis of the RSK process:

If we consider the first row of the P-tableau, the construction process inserts some numbers, and bumps some out. Once a number has been bumped it will never affect row 1 again.

Furthermore, the lower rows of P and Q simply are tableaux corresponding to the "bumped" two-line arrays, that is two-line arrays that consist of the entries bumped out of the first row and the associated q-entries. (This bumped array is the reason why we do not require the first row to be the canonical sequence 1, 2, ...)

In the example of the permutation $\left(\begin{smallmatrix} 1 & 3 & 5 & 7 & 9 \\ 5 & 2 & 3 & 6 & 4 \end{smallmatrix}\right)$ used above, we inserted 2, 3, 4 in the top row and bumped 5 and 6, resulting in the two-line array $A_{\text{bump}} = \left(\begin{smallmatrix} 3 & 9 \\ 5 & 6 \end{smallmatrix}\right)$ describing the second row and below.

This shows that the RSK algorithm can be split into two parts – constructing the first row only, and constructing the bumped two-line array. We consider both parts:

Construction of the first row To study the construction of the first row – remember that every entry of P is inserted first into this to row – we define:

DEFINITION VI.41: Consider a two-line array as given in (VI.36). A column (q_i, p_i) of this array is *in class* t, if after inserting $p_1, \ldots, p_i$ into an empty tableau, p_i is in column t of the tableau.

Example: Looking at the example above, $(3, 2)$ and $(1, 5)$ are in class 1; $(5, 3)$ is in class 2; and $(7, 6)$ and $(9, 4)$ are in class 3. We can easily characterize the class of a column:

LEMMA VI.42: The column (q_i, p_i) belongs to class t if and only if t is the length of the longest increasing subsequence $p_{i_1} < p_{i_2} < \ldots < p_{i_t} = p_i$ ending at p_i.

Proof: For p_i to be in class t, that is inserted into column t, there must be $t - 1$ elements in the first row that are smaller. These elements must be p_j's for $j < i$ showing that these p_j's together with p_i give an increasing subsequence of length t. We thus need to show the converse and will do so by induction on the length t of the longest increasing subsequence ending at p_i:

For $t = 1$, p_i is the smallest element and clearly will be inserted in column 1. Now assume that $t > 1$ and that the theorem holds for class $t - 1$. Let p_j the predecessor of p_i in an increasing subsequence of length j. Then by induction p_j belongs to class $t-1$, that is p_j was inserted into the $t-1$-st column. That means that at the point of inserting p_i, the entry in this position is not larger than p_j, implying that p_i must be inserted in column t or higher. But if it was higher there would be an increasing subsequence of length $> t$, contradiction. □

Corollary VI.43: The length of row 1 of the tableau is the length of a longest increasing subsequence of $\{p_i\}$.

Classes and bumped tableaux We observe that the tableaux resulting from the RSK process are completely determined by the columns of the two-line array and their classes, regardless of the order in which these columns are given:

First observe, that if $(q_{i_1}, p_{i_1}), \dots, (q_{i_k}, pi_k)$ form the set of columns in one particular class, the way the insertion process and bumping works means that we can assume without loss of generality that the indices have been arranged so that $q_{i_1} < q_{i_2} < \cdots < q_{i_k}$ and simultaneously $p_{i_1} > p_{i_2} > \cdots > p_{i_k}$.

If we consider the first row of tableau P, the entry in column t must be the p-part of a column (q, p) in class t. Because of bumping, it will be the (smallest) value p_{i_k}. Similarly the corresponding entry in Q must be the q-part of such a column, but it is the q-part used when the entry is created for the first time, that is the minimal entry q_{i_1}. This describes the first row of P and Q fully.

For the further rows consider the bumped tableaux. To describe these it is sufficient to see what becomes of columns within class t (as entries from different classes never bump one another). The entry p_{i_1} gets bumped by p_{i_2} which comes with q-part q_{i_2}, resulting in a column (q_{i_2}, p_{i_1}). Similarly p_{i_2} gets paired with q_{i_3} and so on, giving the (partial) two-line array arising from class t as

$$A_{\text{bump}} = \begin{pmatrix} q_{i_2} & q_{i_3} & \cdots q_{i_k} \\ p_{i_1} & p_{i_2} & \cdots p_{i_{k-1}} \end{pmatrix}. \tag{VI.44}$$

We then proceed from the bumped array anew to obtain new classes and from these the second rows of P and Q and so on.

With this observation, we are now able to prove the second part of the theorem, concerning the swap of P and Q. As the two tableaux are constructed by very different methods, this is a most surprising symmetry.

Proof:(of Theorem VI.38, part 1) On the level of two-line arrays, inversion means to swap the two rows and sort the columns according to the entries in the first row. Denote this array by

$$A^{-} = \begin{pmatrix} p_1' & p_2' & \cdots p_n' \\ q_1' & q_2' & \cdots q_n' \end{pmatrix}. \tag{VI.45}$$

If we ignore the arrangement of columns, Lemma VI.42 states that a column (q_i, p_i) is in class t if and only if t is the maximal size of an index set $\{i_1, \ldots, i_t\}$ such that

$$q_{i_1} < q_{i_2} < \cdots < q_{i_t} \quad \text{and}$$
$$p_{i_1} < p_{i_2} < \cdots < p_{i_t}.$$

This means that (q, p) is of class t in (VI.36), if and only if (p, q) is of class t in (VI.45).

That implies that the RSK algorithm applied to A^- produces tableaux whose first rows are exactly swapped from the tableaux produced for the original two-line array A.

Furthermore, following the formula (VI.44) for the bumped array – within one class, remove the smallest p and the smallest q entries and "shift the entries together" – we have that A^-_{bump} is obtained from A^-_{bump} by swapping the two rows and sorting according to the first row.

Applying the same argument again thus gives us that the second rows etc. also are swapped, as was to be proven. □

There is much more to the RSK construction. For example *reverting* the permutation (as a sequence of numbers) results in a transposition of P (but a somewhat more complicated transformation of Q namely the transpose of a transformed tableau.)

VI.8 Sliding

Another surprising connection of tableaux arises if we consider semistandard tableaux – allowing duplicate entries and allowing row entries to be equal, while insisting on strict increase within columns. We also generalize from Young diagrams to *skew diagrams*, these are diagrams $\lambda \setminus \mu$ with μ a smaller diagram that fits into the diagram for λ, and $\lambda \setminus \mu$ consisting of the boxes of λ that do not lie in μ. A *skew tableau* then is a skew diagram filled according to the rules of semistandard tableaux. For example, for $\lambda = (5, 4, 4, 3, 2)$ and $\mu = (4, 3, 1)$, Figure VI.5 shows a) the skew diagram $\lambda \setminus \mu$, and b) a skew tableau of that shape.

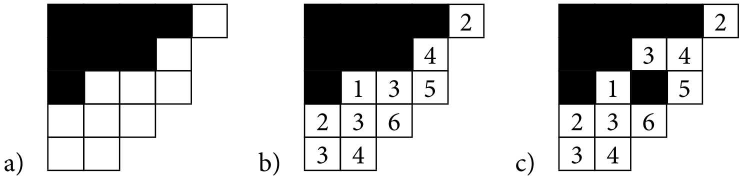


Figure VI.5: A skew diagram, a skew tableau of that shape, and a slide

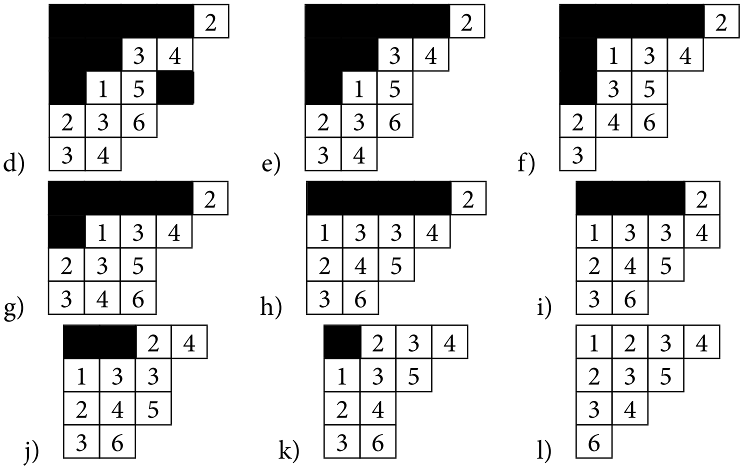


Figure VI.6: Further sliding moves

A box of $\lambda \setminus \mu$ that was deleted is called an *inside corner*, if the boxes to the right and to the bottom of it are both not in μ (and maybe not even in λ). In the example, the boxes in positions 1/4, 2/3, and in position 3/1, are inside corners.

We now describe a process that will transform a skew tableau into a semistandard tableau through a process, due to M. Schützenberger, that goes by the name of *Sliding* or *Jeu de Taquin*[7]. Consider the labeled boxes of the skew tableau as tiles that can move. Starts by selecting an inside corner.

The basic step now takes a selected, empty, box and slides in (thus the name) the box on the right, or below, depending on which holds the smaller number. (If both numbers are the same, the box below is chosen.) Afterwards the position, from which the box slid, is selected.

In the example, let's select the inside corner at position 2/3. The box below, with 3 has the smaller number, and thus slides up, resulting in the tableau in Figure VI.5 c).

We now repeat the process for the newly selected box, sliding in 5 from the right,resulting in Figure VI.6, d). If, as in this example, the selected box has no further boxes to the right or below, we simply delete it from the tableau e) and select a new inside corner, if any exists. The process repeats, until the tableau is not skew any longer.

In the example we select inside corner 2/2 and slide up (in turn) 1,3,4 (f). Next, select 3/1 and slide up (g) 2 and 3. We then select 2/1 and slide in 1, 3 (from below!), 4, 6 (h). From now on there always is only one inner corner left. We slide in 2 (i). Then we slide in 2, 4 (j); then 2, 3, 5 (k); and finally 1, 2, 3, 6 (l).

[7] The French name for the 15-puzzle.

Note VI.46: Convince yourself that after each sliding step, rows remain (when ignoring empty boxes) weakly increasing, and columns strictly increasing. This process thus results in a semistandard tableau.

In the description of the process, we had a choice of inside corners. Somewhat surprisingly, this choice does not affect the result:

Theorem VI.47: The resulting semistandard tableau does not depend on the choice of inside corners.

We thus call the resulting tableau the *rectification* of the original skew tableau (denoted by $\mathrm{rect}(S)$).

Using this process, we now define a multiplication on tableaux: Given two tableaux S and T, we define the product $S \cdot T$ as the rectification of the skew tableau $\begin{array}{cc} \blacksquare & T \\ S & \end{array}$ obtained by placing S and T, connected at a corner.

For example, we have that

$$\begin{array}{ccc} 1 & 2 & 4 \\ 3 & 5 & \end{array} \cdot \begin{array}{cc} 6 & 8 \\ 7 & \end{array} = \mathrm{rect}\left(\begin{array}{ccccc} \blacksquare & \blacksquare & \blacksquare & 6 & 8 \\ \blacksquare & \blacksquare & \blacksquare & 7 & \\ 1 & 2 & 4 & & \\ 3 & 5 & & & \end{array} \right) = \begin{array}{ccccc} 1 & 2 & 4 & 6 & 8 \\ 3 & 5 & 7 & & \end{array}$$

Surprisingly, this product also can be obtained from the insertion process:

Theorem VI.48: The product $S \cdot T$ can also be obtained by inserting the elements of T, bottom row to top row, subsequently into S.

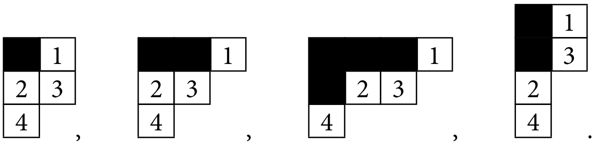
Finally, maybe the biggest surprise, and with no obvious reason, we have the main structural statement:

Theorem VI.49: This multiplication of tableaux is associative.

VI.9 The plactic monoid

The tool to show these three statements will be algebraic (after all, associativity is an algebraic property). We associate to every tableau T a *row word* (or simply "word") $w(T)$, which is the sequence of numbers in the tableau, read left to right, bottom row to top row. For example, the tableau $T = \begin{array}{ccc} 1 & 2 & 4 \\ 3 & 5 & \end{array}$ has word $w(T) = 35124$ (considered as a sequence of symbols.

Conversely, if T is a semistandard tableau, the first entry of row i must be smaller than the last entry of row $i + 1$. Given the word $w(T)$, we thus can recognize the places of row changes, as the places where the entries in the word become strictly smaller. Not every word corresponds to a semistandard tableau, for example 4231 would force a violation of the tableau shape. However each word corresponds to (multiple) skew tableaux. A few examples for 4231 are



Formally, we have an *alphabet* $A = \{1, 2, 3, \dots, k\}$ (or even infinite) and consider the set A^* of words (= sequences) in A, including the empty word. With concatenation as product it has the structure of a *monoid*, i.e. the multiplication is associative and there is an identity element. It is in fact a free monoid.

The words of (proper) tableaux form a subset of A^*. (Such a subset is called a *formal language*, and the elements of A are called *letters*[8].) We want to understand this subset and in particular understand how the sliding process produces words in this subset. We will do so by describing transformation rules for words that represent the bumping and sliding processes.

We first look at the situation of Theorem VI.48 and consider the insertion of element q into tableau S, resulting in tableau T. If q is larger than every element in the first row of S, then $w(T) = w(S) \cdot q$. Otherwise the first row of S has the form $a \cdot p \cdot b$ with p being bumped out by q. This means that every factor if a is smaller than q, and every factor of pb at least as large as q.

After insertion, the first row of T will have the form $a \cdot q \cdot b$, and with p being inserted into the second row we can consider this as a word transformation

$$apbq \to paqb$$

We want to describe this transformation in terms of simpler rules. First among these is that we pull q past letters in b which are larger, as long as the letter before is also larger than q. We write these rules as

$$yzx \mapsto yxz, \text{for letters } x < y \leq z. \tag{VI.50}$$

This set of rules will (iteratively) transform $apbq$ into $apqb$. Next we move p past the letters in a, which all are smaller than p. This is not a reversal of the rules (VI.50), since the letter following p is smaller than p itself. (In the first instance $apqb$ we might have $q = p$, but then $apqb = aqpb$ trivially.) This gives rules:

$$xzy \mapsto zxy, \text{for letters } x \leq y < z. \tag{VI.51}$$

The set of these rules now transform $apbq$ into $paqb$, which is a word tail that represents the insertion of p into the second row of the tableau.

One can now consider the equivalence relation on words (transitive closure) defined by the relations of type (VI.50) and (VI.51). These equivalence classes form a monoid, the factor of the free monoid A^* by these relations. It is called the *plactic monoid*.

[8] though in our case they are numbers

- Slide moves do not change the equivalence class of skew tableau words.
- There is exactly one semistandard tableau word in each equivalence class.

The multiplication we defined thus is the multiplication in the plactic monoid, whose elements are in bijection to semistandard tableaux.

Chapter

VII

Symmetry

Tyger Tyger, burning bright,
In the forests of the night;
What immortal hand or eye,
Could frame thy fearful symmetry?

The Tyger
William Blake

The use of symmetries is one of the tools of science that crosses into many disciplines. So far we have only touched upon this, in a rather perfunctory way, by considering objects as identical or non-identical, or by counting labeled versus non-labeled objects. In this chapter we shall go into more detail. In combinatorics, there are at least three uses of symmetry:

- We might want to consider objects only up to symmetry, thereby reducing the total number of objects. For example when considering bracelets made from 10 green and one golden pearl, we want to consider this is one bracelet, rather than 11 possible positions of the golden pearl. We will look at counting up to symmetry in Section VII.5.

- Symmetries can be used to classify what is sometimes a huge number of objects into equivalence classes. For example, relabeling points in a permutation preserves the cycle structure.

- Finally, in a related way, we might want to use symmetries as a proxy for objects being interesting. When combining parts, there might be a huge number of possible objects. For example there are

$$64549012279579984185616463849074274944\boxed{\text{OEIS A000088}}$$

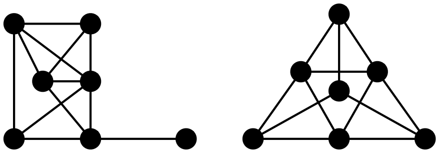


Figure VII.1: Two graphs with different symmetries

different graphs on 20 vertices. But, be it graphs or other objects, it turns out that many of them are just stuck together from smaller objects and do not amount to more than the sum of their parts. On the other hand symmetries that do not keep the parts fixed often indicate something interesting. Figure VII.1 shows two graphs with 7 vertices and 12 edges. Clearly the graph on the right (which has 5 nontrivial symmetries) is more interesting than the left one (with only trivial symmetries).

Of the formal setting in mathematics for symmetries is a group. Instead of going all abstract, we will consider first, what we want symmetries to be, and then get the formal axioms as an afterthought. So, what are symmetries? If we have an object (which can be a set of smaller objects), a symmetry is a map from the object to itself, may be mixing up the constituent parts, that preserves some structural properties.

VII.1 Automorphisms and Group Actions

> Nothing will make me believe that a man who arranged all the rest of that room with that exaggerated symmetry left that one feature of it lopsided.
>
> The Worst Crime in the World
> G. K. Chesterton

We call such a map an automorphism, from the Greek autos for self and morphe for shape. There always is at least one such map, namely the identity which doesn't do anything. In an abuse of notation we shall write 1 for the identity map. By preserving the larger collection, the maps have to be bijective, and thus afford an inverse. Finally we can compose these maps, that is apply one after the other. As always, composition of maps is associative. we thus get the formal axioms of a group, though this axiomatic theory shall be of little concern to us. Following conventions introduced in [Wie64], we denote groups by upper case letters, group elements by lower case letters, and objects the group acts on by Greek letters. If the objects we act on have numbered parts or are themselves numbered, it can be convenient to represent maps as permutations. Indeed, if we act on a set of objects with numbered parts, permutations of these numbers yield automorphisms.

Formally, we can distinguish maps and the permutations induced by them, in this case an action of a group G on n numbered objects yields a group homomorphism $G \to S_n$, assigning to every map the permutation induced by it.

We shall write the image of a point ω under a map g as ω^g. That is our groups act on the right. That way the image of ω under the product of g with h is simply gh. Similarly permutations multiply as $(1,2,3)\cdot(2,3) = (1,3)$. (While this is clearly the natural way for languages that are written from left to right, the reader who had abstract algebra might notice that in some classes the convention used is an action from the left, which then requires the introduction of a special product operation for composition of maps.)

In some situations we might want to consider what maps do to parts, collections, or structures derived from the original objects. For this, it is most convenient to not consider new maps, but a new action of the original maps. The rules for this are simply that the identity must act trivially, and that products act from left to right, that is

$$\omega^1 = \omega \quad \forall \omega \in \Omega, \qquad \text{and} \qquad \omega^{(gh)} = \left(\omega^g\right)^h \quad \forall \omega \in \Omega \forall g, h \in G$$

Such an action, given by a group G, a domain Ω, and a map $\Omega \times G \to \Omega$, $(\omega, g) \mapsto \omega^g$, is called a group action. Group actions are the basic objects we shall consider.

Examples

Before delving into further details of the theory, lets consider a few examples.

The set of all permutations of n points form a group, called the *symmetric group* S_n. It acts on the points $\{1, \dots n\}$, but we also can have it act on other structures that involve these numbers, for example all subsets of size k, sequences of length k, or partitions with k cells.

If we consider a bracelet with 6 pearls as depicted on the left in Figure VII.2, then automorphisms are given for example by a rotation along the bracelet (which we shall denote by the permutation $(1,2,3,4,5,6)$), or by flips along the dashed reflection axes. For example, the flip along the axis going between 1 and 2 is given by the permutation $(1,2)(3,6)(4,5)$. Together with the trivial rotation this gives $1+5+6 = 12$ symmetries. We also observe that there are 6 positions where 1 can go, and then 2 positions where 2 goes with which an automorphism is fully determined. The whole symmetry group thus has order 12. (For those who had abstract algebra, it is the dihedral group of order 12.)

Next consider symmetries of a cube that are physically possible without disassembling, that is only rotations and no reflections. (However in other cases one might make another choice, indeed when talking about automorphisms one needs to decide what exactly is to be preserved.) Besides the identity there are 3 rotations each for the three pairs of opposite faces, 2 rotations for each of the 4 diagonal axes,

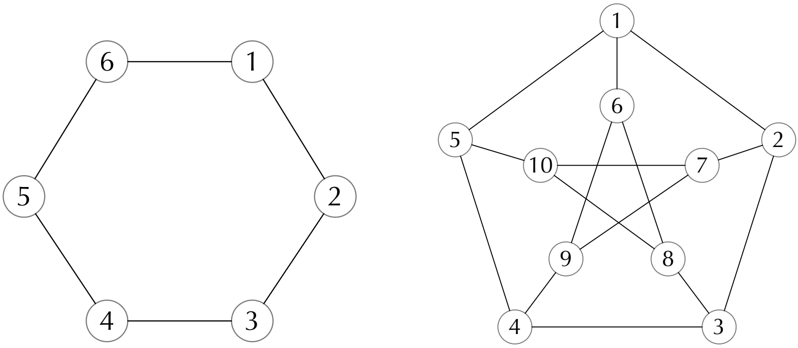


Figure VII.2: A circle graph and the Petersen graph

as well as 1 rotation on axes that go through the middle points of 6 pairs of opposite faces, for a total of $1 + 9 + 8 + 6 = 24$ symmetries. (Again, the images of neighboring faces gives us at most $6 \cdot 4 = 24$ possibilities, thus we found all.)

The set of automorphisms of an object X is called its *automorphism group* and denoted by $\mathrm{Aut}(X)$. (As noted that might depend on what structure is to be preserved. The automorphism group of the 6 faces would be S_6 if we did not want to preserve the cube structure.)

It is not always that easy to determine the number of automorphisms (also called the order of the automorphism group). Take the graph on the right side of Figure VII.2 (which is called the Petersen-graph). An automorphism of a graph is a permutation of the vertices that preserves adjacency. By checking this property, it is easy to verify that the permutation $(1,2,3,4,5)(6,7,8,9,10)$ is an automorphism, as is (with a bit more effort) $(3,7)(4,10),(8,9)$. But is is not trivial to show that there are 120 automorphisms in total, and that any automorphism can be formed as a suitable product of the two permutations given. We will look at this in Exercise **??**.

For an example with a larger set of objects, consider the order-significant partitions of n into k parts and S_k acting by permuting the parts.

Finally, we consider the set of all possible graphs on three vertices, labeled with $1,2,3$, as depicted in Figure VII.3. There are $\binom{3}{2} = 3$ possible edges and thus $2^3 = 8$ different graphs. The group S_3 acts on this set of graphs by permuting the vertices. (When doing so, the *number of edges* (and more — On four vertices we cannot map a triangle to a Y) is preserved, so not every graph may be mapped to any other.

In general, for two given graphs on n (labeled) vertices, a permutation in S_n that maps one graph to the other is called a *graph isomorphism*, if such a map exists the graphs are called *isomorphic*.

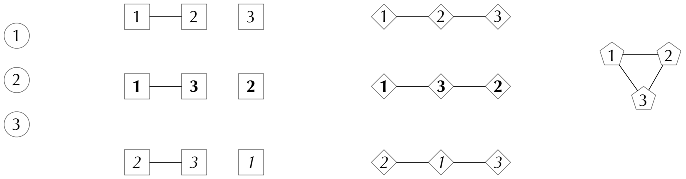


Figure VII.3: The labeled graphs on 3 vertices.

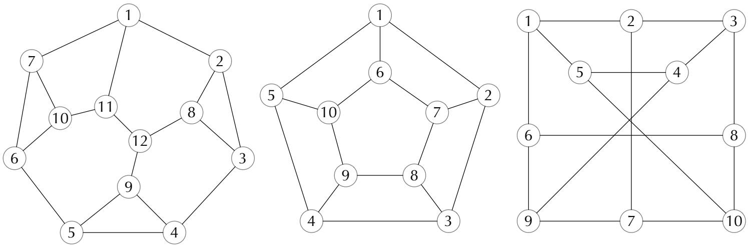


Figure VII.4: Three nonisomorphic graphs that are regular of degree 3.

Note VII.1: The question of determining graph isomorphisms is a prominent [Bab16] and hard problem. While local properties — number of edges or degrees of vertices — can eliminate the existence of automorphisms, one can show that for any set of local properties, there are graphs in different orbits which agree on all of the invariants.

Figure VII.4 shows three graphs, each of them regular of degree 3. The left graph has trivial automorphism group, the middle graph an automorphism group of order 20. The right graph is in fact, with the given labeling, isomorphic to the Petersen graph (with automorphism group of order 120), though this isomorphism is not obvious without the labeling – there are 10! possible maps, only 120 of them, that is 1 in 30240 induce an isomorphism.

We close this section with the definition of an equivalence of actions:

Definition VII.2: Let G be a group acting on a set Ω and H a group acting on a set Δ. We call these actions *equivalent (actions)*, if there is a group homomorphism $\alpha\colon G \to H$ and a bijection $\psi\colon \Omega \to \Delta$, such that

$$\psi(\omega^g) = \psi(\omega)^{\alpha(g)}$$

That means if a group G acts on a set Ω of size n, inducing a homomorphism $\varphi\colon G \to S_n$, the action of G on Ω and the action of $\varphi(G)$ on $\{1,\ldots,n\}$ are equivalent.

Orbits and Stabilizers

> Ever he would wander, selfcompelled, to the extreme limit of his cometary orbit.
>
> Ulysses
> JAMES JOYCE

The two main concepts for working with group actions are orbits and stabilizers.

DEFINITION VII.3: Let G act on Ω and $\omega \in \Omega$.
a) The *Orbit* of ω is the set of all possible images:

$$\omega^G = \{\omega^G \mid g \in G\}.$$

b) The *stabilizer* of ω is

$$\mathrm{Stab}_G(\omega) = \{g \in G \mid \omega^G = \omega\}.$$

Since group elements are invertible, the orbits of a group form equivalence classes. They represent "equivalence up to symmetry". If we want to describe objects up to symmetry, we really ask to describe the orbits of an appropriate symmetry group.
Example: Determine orbits in the examples of the previous section?

DEFINITION VII.4: The action of G on Ω is *transitive* if there is only one orbit, that is $\Omega = \omega^G$ for (arbitrary) $\omega \in \Omega$. An element $\omega \in \Omega$, that is the only one in its orbit — i.e. $\omega^G = \{\omega\}$ — is called a fixed point of G.

The stabilizer of a point is closed under products and inverses and thus always forms a group itself. Indeed,the automorphism group of an object can be considered as a stabilizer of the object in a larger "full" symmetry group: In the case of the bracelet we can consider the set of all permutations of the pearls and want only those that preserve the neighbor relations amongst pearls. The automorphism group of a cube (centered at the origin) is the stabilizer of the cube in the orthogonal group of 3-dimensional space.

The key theorem about orbits and stabilizers is the bijection between orbit elements and stabilizer cosets. Two elements g, h map ω in the same way if and only if their quotient lies in the stabilizer:

$$\omega^g = \omega^h \Leftrightarrow \omega = \omega^{(g \cdot g^{-1})} = (\omega^g)^{g^{-1}} = (\omega^h)^{g^{-1}} = \omega^{h/g} \Leftrightarrow h/g \in \mathrm{Stab}_G(\omega).$$

This means that if g is one element such that $\omega^g = \delta$ then the set of all elements that map ω in the same way is (what is called a *coset*):

$$\mathrm{Stab}_G(g) \cdot g = \{s \cdot g \mid s \in \mathrm{Stab}_G(\omega)\}.$$

For any orbit element $\delta \in \omega^G$ there therefore are exactly $|\mathrm{Stab}_G(\omega)|$ elements of G that map ω to δ. We therefore have

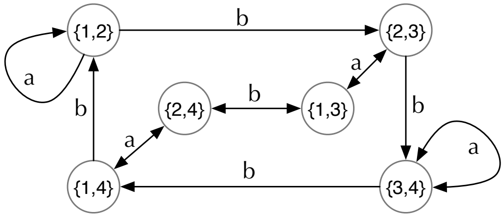


Figure VII.5: Schreier graph for the action of $\langle a = (1,2), b = (1,2,3,4)\rangle$ on pairs.

Theorem VII.5 (Orbit-Stabilizer): There is a bijection between elements of ω^G and cosets of $\mathrm{Stab}_G(\omega)$. In particular:

$$|G| = \left|\omega^G\right| \cdot |\mathrm{Stab}_G(\omega)|.$$

In particular, this means that the length of an orbit must divide the order of the acting group.

Example: A group of order 25, acting on 12 points thus must have fixed points.

We note some consequences from this bijection. They show that the possible transitive actions of a group G are prescribed by its structure:

Lemma VII.6: Let G act on Ω and $\omega, \delta \in \Omega$ with $g \in G$ such that $\omega^g = \delta$. Then

a) $\mathrm{Stab}_G(\omega)^g = \mathrm{Stab}_G(\delta)$.

b) The set of elements of G mapping ω to δ is

$$\mathrm{Stab}_G(\omega) \cdot g = g \cdot \mathrm{Stab}_G(\delta) = \mathrm{Stab}_G(\omega) \cdot g \cdot \mathrm{Stab}_G(\delta).$$

c) The action of G on ω^G is equivalent to the action of G on the cosets of $S = \mathrm{Stab}_G(\omega)$ by $(Sx, g) \mapsto S(xg)$: If $\omega^x = \delta$, $\omega^y = \gamma$ and $\delta^g = \gamma$ then $S(xg) = Sy$.

We now define a directed graph that describes the action of certain elements $A \subset G$ on Ω. (Typically A will be a generating set for G, that is every element of G is a product of elements of A and their inverses.)

Definition VII.7: Let $\mathrm{Sch}_A(\Omega) = (\Omega, E)$ the digraph with vertex set Ω and edges $E \subset \Omega \times \Omega$ labeled by elements of A: We have that $(\omega, \delta) \in E$ if and only if there exists an $a \in A$ such that $\omega^a = \delta$. In this case we label the edge (ω, δ) with a. We call $\mathrm{Sch}_A(\Omega)$ the *Schreier graph*[1] for the action of A on Ω.

[1] after Otto Schreier, 1901-1929

For example, Figure VII.5 gives the Schreier graph for the action of S_4, generated by the permutations $a = (1,2)$ and $b = (1,2,3,4)$ on sets of order 2 (which form just one orbit).

A Schreier graph is connected (and, for a finite group then also strongly connected, that is connected with respect to directed edges) if and only if all of Ω is one single orbit.

By tracing paths between vertices and multiplying up the elements of A on the edges (inverses, if we go an arrow in reverse), we obtain a factorization of an element, mapping one point to another, as a word in A. In the example above, we can read off that b^2 will map $\{2,3\}$ to $\{1,4\}$, as would aba, b^{-2}, bab, and so on.

Puzzles, such as Rubik's cube are essentially just about finding such factorizations, in examples in which the group is far too large to draw the Schreier graph.

We can determine the orbit of a point as the connected component of the Schreier graph, starting from a vertex and continuing to add images under group generators, until no new images are found.images under group generators, until no new images are found. (One needs to store only vertices encountered this way, not the edges.) A spanning tree of a connected Schreier graph is called a Schreier tree, it is easy to see that often there will be many possible Schreier trees, of different depth.

We note that loops in the Schreier graph images yield stabilizer elements, that is if $\omega^g = \delta$, $\delta^a = \gamma$, and $\omega^h = \gamma$, then ga and h are in the same coset of $S = \mathrm{Stab}_G(\omega)$ and $ga/h \in S$. Schreier's theorem (which we shall not prove here) shows that if we do so systematically for all loops that get created from elements of A, we obtain a generating set for $\mathrm{Stab}_G(\omega)$.

This in fact provides the tool — called the Schreier-Sims algorithm — by which computers calculate the order of a permutation group: They calculate the orbit of a point and generators for its stabilizer, then recurse to the stabilizer (and another point). By the Orbit-Stabilizer theorem the group order then is the product of the orbit lengths.

Example: An icosahedron has 20 faces that are triangle shaped. Rotations can map any face to any other face, yielding an orbit of length 20. The stabilizer of a face can clearly have only 3 rotations, yielding a rotational automorphism group of order 60. Similarly we obtain a rotation/reflection automorphism group of order 120.

VII.2 Cayley Graphs and Graph Automorphisms

> Spread oot, lads, and pretend ye're enjoying the cailey.
>
> The Wee Free Men
> Terry Pratchett

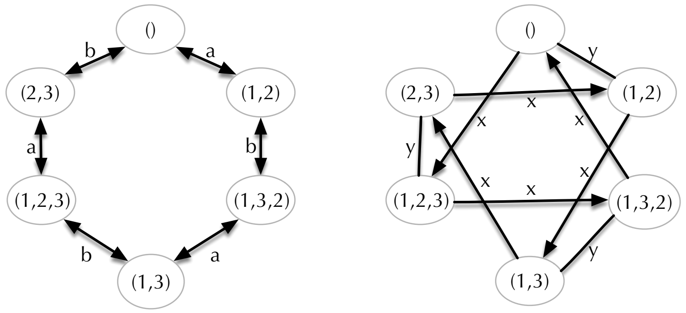


Figure VII.6: Two Cayley graphs for S_3.

Definition VII.8: An action of G on Ω is *semiregular* if the stabilizer of any $\omega \in \Omega$ is trivial. If the action is also transitive, it is called *regular*.

By Lemma VII.6, a regular action of G is equivalent to the action on the cosets of the trivial subgroup, that is the action of G on its elements by right multiplication: $(x, g) \mapsto xg$. If we have a generating set A of G, the Schreier graph for this action gets a special name, it is called the *Cayley graph* of G (with respect to A).

Figure VII.6 shows the Cayley graphs for S_3 for the generating set $a = (1, 2)$, $b = (2, 3)$, as well as for the generating set $x = (1, 2, 3)$, $y = (1, 2)$.

We notice (Exercise **??**) that G induces automorphisms of the Cayley graph $\mathrm{Cay}_A(G)$. To be consistent with the way we defined actions and with the labeling of edges, the action of $g \in G$ is defined by $(v, g) \mapsto g^{-1}v$ for a vertex $v \in G$. We denote this automorphism by μ_g.

However Cayley graphs can, as graphs that ignore labels and directions, have other automorphisms. For example, the graph in Figure VII.6, left, has a dihedral group of order 12 of symmetries, no all of which make sense for S_3 (a rotation for example would change element orders).

Such extra automorphisms however will not preserve edge labels:

Lemma VII.9: Let $\Gamma = \mathrm{Cay}_A(G)$ and $\alpha \in \mathrm{Aut}(\Gamma)$. Then α preserves the labels of all edges, if and only if there exists $g \in G$ such that $\alpha = \mu_g$, that is $\alpha(v) = g^{-1}v$ for all $v \in G$.

Proof: If $x \cdot a = y$ then $g^{-1} \cdot (x \cdot a) = g^{-1} \cdot y$. Thus all μ_g preserve the edge labels. Let 1 be the identity of G and $g := \alpha(1)$. Then $\alpha \cdot \mu_g$ is an automorphism that maps 1 to 1. Furthermore (as μ_g preserves edge labels), $\alpha \cdot \mu_G$ preserves edge labels if and only if α does. We thus may assume, without loss of generality, that $\alpha(1) = 1$. (We shall show that in this case α must be the identity map that is μ_1. As $\alpha \cdot \mu_g = \mu_1$ we get that $\alpha = \mu_{g^{-1}}$.)

By definition, for every $a \in A$, there exists a unique vertex (namely the vertex a) that is connected by an edge from 1 with label a. That means that this edge must be fixed by α and thus $\alpha(a) = a$. The same argument for edges *to* 1 shows that a^{-1} must be fixed by α for any $a \in A$. This shows that every vertex at distance one from 1 must be fixed by α.

The same argument now can be used by induction to show that all elements at distance $2, 3 \ldots$ from 1 must be fixed by α. As A is a generating set of G, this is all elements, thus α must be the identity, completing the proof. □

This result implies that all finite groups can be considered as the automorphism group of a suitable graph.

Theorem VII.10 (Frucht): Let G be a finite group. Then there exists a finite graph Γ such that $\mathrm{Aut}(\Gamma) \cong G$.

Proof: We have seen already that every group is the automorphism group of a directed graph with labeled edges, for example its Cayley graph. What we will simply do is to simulate direction and labeling by extra vertices and edges. First for direction: For each directed edge $a \to b$ between vertices a, b, we introduce three new vertices x, y, z and replace the directed edge by undirected edges $a - x - y - b$, as well as $y - z$. That is, a and b are still connected, and z indicates the arrow direction. The z-vertices of this new graph are the only ones of degree 1, so the automorphisms must map a z-vertex to a z-vertex, and thus an y-vertex to a y-vertex. We also add a "universal" vertex $\underline{X}$ which we connect to all x-vertices. Then (ensured if necessary by adding further vertices we connect to $\underline{X}$) the vertex $\underline{X}$ is the only vertex of its degree and thus must remain fixed under any automorphism, implying that the x-vertices must be permuted amongst each other as well. Thus the "original" vertices of the graph cannot be mixed up with any of the newly introduced vertices representing edges. Also distance arguments show that each group of $x/y/z$ vertices must stay together as a group. The automorphisms of this new undirected graph thus are the same as the automorphisms of the original directed graph.

To indicate edge label number i, we insert $i - 1$ extra vertices into the $y - z$ edges, replacing $y - z$ by $y - z_1 - z_2 - c \cdots - z_{i-1} - z$. Automorphisms then need to preserve these distances and thus are restricted to preserving the prior labeling. □

(Frucht in fact proved a stronger theorem, namely that every finite group is the automorphism group of a 3-regular graph, i.e. a graph in which every vertex has degree 3.)

Note VII.11: A final caveat: The same, *unlabeled*, directed graph can be a Cayley graph for two different, non-isomorphic, groups. Figure VII.7 depicts the Cayley graphs for the cyclic group $\langle a \mid a^9 \rangle$ of order 9 and the generating set $\{a, b = a^4, c = b^7\}$, as well as the Cayley graph for the noncyclic group $\langle x, y \mid x^3, y^3, xy = yx \rangle$ of the same order and the generating set $\{a = x, b = xy, c = x/y\}$. The graphs are isomorphic, when labels are ignored.

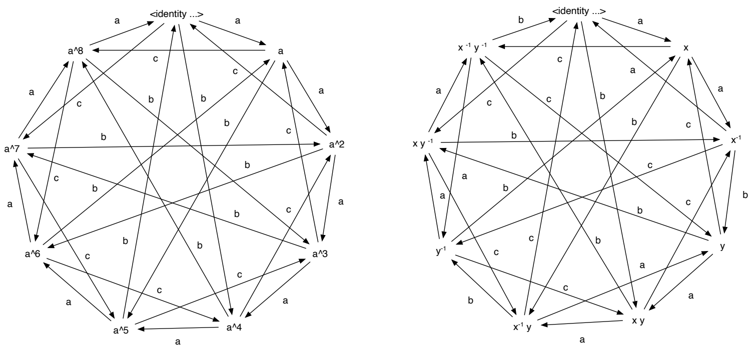


Figure VII.7: Isomorphic Cayley graphs for nonisomorphic groups

VII.3 Permutation Group Decompositions

> **Permutation**:
> The theory of how hairdos evolved
>
> The New Uxbridge Dictionary
> Brook-Taylor et.al.

Finding the automorphism group of an object can be hard. In many practically relevant cases, however the objects have substructures that need to be preserved by automorphisms. In this case we can relate automorphisms of the structure with automorphisms of the substructures in a meaningful way.

In this section we shall look at two kinds of decompositions that lend themselves to nice group theoretic structures.

By labeling relevant parts of the object acted on by G we can assume without loss of generality that the automorphism group group on a set Ω.

We start with the situation that Ω can be partitioned into two subsets that must be both preserved as sets, that is $\Omega = \Delta \cup \Lambda$ with $\Delta \cap \Lambda = \emptyset$. This could be distinct objects — say vertices and edges of a graph — or objects that structurally cannot be mapped to each other – for example vertices of a graph of different degree. (The partition could be into more than two cells, in which case one could take a union of orbits and iterate, taking the case of two as base case.)

In this case G must permute the elements of Δ amongst themselves, as well as the elements of Λ. We thus can write every element of G as a product of a permutation on Δ and a permutation on Λ. Since the sets are disjoint, these two permutations commute.

Group theoretically, this means that we get a homomorphism $\varphi: G \to S_\Delta$, as well as a homomorphism $\psi: G \to S_\Lambda$. The intersection $\mathrm{Kern}\,\varphi \cap \mathrm{Kern}\,\psi$ is clearly trivial,

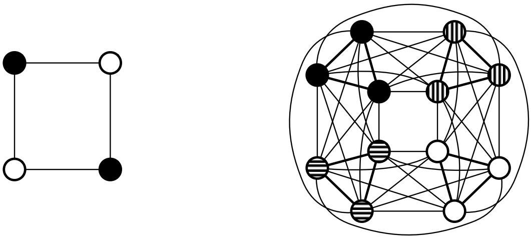


Figure VII.8: Two graphs with an imprimitive automorphism group.

as elements of G are determined uniquely by their action on $\Omega = \Delta \cup \Lambda$. We thus can combine these maps to an embedding (an injective group homomorphisms)

$$\iota\colon G \to S_\Delta \times S_\Lambda, \qquad g \mapsto (\varphi(g), \psi(g))$$

If we consider the direct product as acting on the disjoint union of Δ and Λ, this map is simply the identity map on permutations.

The images of φ and ψ are typically not the whole symmetric groups. Thus, setting $A = \varphi(G)$ and $B = \psi(G)$, we can consider G as a subgroup of $A \times B$, and call it a *subdirect product* of A and B. All intransitive permutation groups are such subdirect products.

Note that a subdirect product does not need to be a direct product itself, but can be a proper subgroup, see Exercise **??**. There is a description on possible groups arising this way, based on isomorphic factor groups. Doing this however requires a bit more of group theory than we are prepared to use here.

We also note that it is not hard to classify subdirect products abstractly, and that this indeed has been done to enumerate intransitive permutation groups.

Block systems and Wreath Products

The second kind of decomposition we shall investigate is the case of a group G that is transitive on Ω, but permutes a partition into subsets (which then need to be of equal size).

For example, in the two graphs in Figure VII.8, one such partition would be given by vertices that are diagonally opposed. Neither diagonal is fixed, but a diagonal must be mapped to a diagonal again.

In the graph on the right, constructed from four triangles that are connected in all possible ways with the two neighbors, such a partition would be given by the triangles.

We formalize this in the following definition:

Definition VII.12: Let G act transitively on Ω. A *block system* (or *system of imprimitivity*) for G on Ω is a partition $\mathcal{B}$ of Ω that is invariant under the action of G.

Example: For example, taking $G = \mathrm{GL}_n(F)$ for a finite field F, acting on the nonzero vectors of F^n, the sect of vectors that span the same 1-dimensional space, that is those that are equivalent under multiplication by F^*, form a block system.

We note a few basic properties of block systems, the proof of which is left as an exercise.

Lemma VII.13: Let G act transitively on Ω with $|\Omega| = n$.

a) There are two *trivial* block systems, $\mathcal{B}_0 = \{\{\omega\}\}_{\omega\in\Omega}$, as well as $\mathcal{B}_\infty = \{\Omega\}$.

b) All blocks in a block system must have the same size.

c) If $\mathcal{B}$ is a block system, consisting of a blocks of size b, then $n = ab$.

d) If $\mathcal{B}$ is a block system, then G acts transitively on the blocks in $\mathcal{B}$.

e) A block system is determined uniquely by one of its blocks.

We note — See section **??** — that part a) of the lemma is the best possible – there are groups which only afford the trivial block systems.

A connection between blocks and group structure is given by the following proposition which should be seen as an extension of Theorem VII.5:

Proposition VII.14: Let G act transitively on Ω and let $\omega \in \Omega$ with $S = \mathrm{Stab}_G(\omega)$. There is a bijection between block systems of G on Ω and subgroups $S \le U \le G$.

Proof: We will establish the bijection by representing each block system by the block B containing ω. For a block B with $\omega \in B$, let $\mathrm{Stab}_G(B)$ be the set-wise stabilizer of B. We have that $S \le \mathrm{Stab}_G(B)$, as S maps ω to ω and thus must fix the block B. Since G is transitive on Ω there are elements in G that map ω to an arbitrary $\delta \in B$, as B is a block this means that these elements must lie in $\mathrm{Stab}_G(B)$. This shows that $\mathrm{Stab}_G(B)$ acts transitively on B and that $B = \omega^{\mathrm{Stab}_G(B)}$. The map from blocks to subgroups therefore is injective.

Vice versa, for a subgroup $U \ge S$, let $B = \omega^U$. Clearly B is, as a set, stabilized by U. We note that in fact $U = \mathrm{Stab}_G(B)$, as any element $x \in \mathrm{Stab}_G(B)$ must map ω to $\omega^x inB$ and by definition there is $u \in U$ such that $\omega = (\omega^x)^u = \omega^{xu}$, so $xu \in S \le U$ and therefore $x \in U$. Thus the map from subgroups containing S to subsets containing ω is injective.

We claim that B is the block in a block system. Since G is transitive, the images of B clearly cover all of Ω. We just need to show that they form a partition. For that, assume that $B^g \cap B^h \not= \emptyset$, that is there exists $\delta \in B^g \cap B^h$. This means that $\delta^{g^{-1}}, \delta^{h^{-1}} \in B = \omega^U$ and thus there exists $u_g, u_h \in U$ such that $\omega^{u_g g} = \delta = \omega^{u_h h}$. But then $u_g g(u_h h)^{-1} = u_g g h^- 1 u_h^{-1} \in \mathrm{Stab}_G(\omega) \le U$, the $gh^{-1} \in U$ and (as $U =$

$\mathrm{Stab}_G(B)$) we have that $B^g = B^h$. This shows that the images of B form a partition of Ω.

The properties shown together establish the bijection. □

Corollary VII.15: The blocks systems for G on Ω form a lattice under the "subset" (that is blocks are either subsets of each other or intersect trivially) relation. Its maximal element is $\mathcal{B}_\infty$, its minimal element is $\mathcal{B}_0$.

We call a transitive permutation group *imprimitive*, if it affords a nontrivial block system on Ω. We want to obtain an embedding theorem for imprimitive groups, similar to what we did for direct products. That is, we want to describe a "universal" group into which imprimitive groups embed.

The construction for this is called the wreath product"

Definition VII.16: Let A be a group, b an integer, and $B \leq S_b$ a permutation group. The *wreath product* $A \wr B$ is the semidirect product of $N = A^b = \underbrace{A \times \cdots \times A}_{b}$ with B where B acting on N by permuting the components.

If $A \leq S_a$ is also a permutation group, we can represent $A \wr B$ as an imprimitive group on $a \cdot b$ points: Consider the numbers $1, \ldots, ab$, arranged as in the following diagram.

1	2	$\cdots$	a
$a+1$	$a+2$	$\cdots$	$2a$
		$\vdots$	
$a(b-1)+1$	$a(b-1)+2$	$\cdots$	ab

Then the direct product $N = A^b$ can be represented as permutations of these numbers, the i-th copy of A acting on the i-th row. Now represent the permutations of B by acting *simultaneously* on the columns, permuting the b rows. The resulting group W is the wreath product $A \wr B$. Points in the same row of the diagram will be mapped by W to points in the same row, thus W acts imprimitively on $1, \ldots, ab$ with blocks according to the rows of the diagram. W is therefore called the *imprimitive action* of the wreath product.

Lemma VII.17: Let G act imprimitively on $1, \ldots n = ab$ with b blocks of size a. Then G can be embedded into a wreath product $S_a \wr S_b$ in

Proof: By renumbering the points we may assume without loss of generality that the blocks of G are exactly $\{1, \ldots, a\}$, $\{a+1, \ldots, 2a\}$ and so on, as given by the rows of the above diagram. Let $g \in G$. Then g will permute the blocks according to a permutation $b \in S_b$. By considering S_b as embedded into $S_a \wr S_b$, we have that g/b fixes all rows of the diagram as sets and thus is in $N = (S_a)^b$. □

(Again, it is possible — for example under the name of *induced representations* — to give a better description, reducing to a wreath product of the block stabilizers action on its block with the groups action on all the blocks, but doing so requires a bit more work.)

Product Action

There is a different permutation action of wreath products, called the *product action*: For a group G acting on a set Ω and a group H acting on a set Δ of order n, we define an action of $G \wr H$ on the cartesian product Ω^n: The normal subgroup G^n acts with the i-th copy of G acting in the i-th dimension, H then acts by permuting dimensions.

Somewhat surprisingly, given that the degree is larger than that of the imprimitive action and thus has a smaller point stabilizer, this action is often primitive. We give the proof in an easy case (in fact the theorem remains true if we only require G to be primitive but not regular).

LEMMA VII.18: Suppose that G acts 2-transitively on Ω, with $|\Omega| > 2$, and that H acts transitively on $\{1, \ldots, n\}$. Then $G \wr H$ acts primitively on Ω^n

Proof: With G and H acting transitively, $G \wr H$ acts transitively on Ω^n. Assume this action is imprimitive and B is a block that contains the point $p = (\omega, \ldots, \omega) \in \Omega^n$, as well as another point $q = (\delta_1, \ldots, \delta_n)$ with at least one δ_i, say δ_1, different from ω. By exercise **??**, B must be invariant under

$$S = \mathrm{Stab}_{G \wr H}(\omega, \ldots, \omega) = \mathrm{Stab}_G(\omega) \wr H.$$

Thus B must also contain all points $(\gamma, \delta_2, \ldots, \delta_n)$. As $|\Omega| > 2$ it thus contains one such point with $\gamma \neq \delta_1$. But the stabilizer of q will (similar argument using doubly transitivity) map $(\gamma, \delta_2, \ldots, \delta_n)$ to $(\omega, \delta_2, \ldots, \delta_n)$.

Using the action of B to permute components we thus get that $\Omega^n \subset B$, that is the block system is trivial. □

VII.4 Primitivity and Higher Transitivity

> *In action be primitive; in foresight, a strategist.*
>
> Agir en primitif et prévoir en stratège.
>
> Feuillets d'Hypnos #72
> RENÉ CHAR

A transitive permutation group $G \leq S_\Omega$ is called *primitive*, if the only block systems it affords are the trivial ones $\mathcal{B}_0$ and $\mathcal{B}_\infty$. Since block sized need to divide the degree, every transitive group of prime degree is primitive.

Example in arbitrary degree are given by symmetric groups:

Example: For $n > 1$, the symmetric group S_n is primitive: Assume that B is a block in a nontrivial block system, then $2 \le |B| \le n - 1$. Thus there are $\omega, \delta \in B$ and $\gamma \in \Omega \setminus B$. But there is $g \in S_n$ such that $\omega^g = \omega$ and $\delta^g = \gamma$, thus B is mapped by g to a set that has proper, nontrivial, intersection with B contradicting the block property.

This argument in fact only requires that the action on pairs of points is transitive — map $\{\omega, \delta\}$ to $\{\omega, \gamma\}$, motivating the following definitions:

Definition VII.19: Let G be a permutation group that is transitive on Ω.
a) If G acts transitively on all k-tuples of distinct elements of Ω, then G is called k-transitive.
b) If G acts transitively on all k-sets of elements of Ω, then G is called k-homogeneous.

These properties are related:

Theorem VII.20: a) If G is k-transitive for $1 < k$ then G is also $k - 1$-transitive
b) If G is k-homogeneous for $1 < k \le |\Omega|/2$, then G is also $k - 1$-homogeneous
c) If G is k-transitive, then G is k-homogeneous
d) If G is k-transitive or k-homogeneous for $k > 1$, then G is primitive.

Proof: a and c) are clear. b) follows from Proposition **??** below. For d), it is sufficient to show that a 2-homogeneous group must be primitive: Assume that B is a nontrivial block and choose $\alpha, \beta \in B$ and $\gamma \notin B$. An element g mapping $\{\alpha, \beta\}$ to $\{\alpha, \gamma\}$ will map B to an image B^g with nontrivial intersection, contradiction.
□

Note VII.21: Apart from the symmetric and alternating groups there are very few groups that are more than quadruply (or higher) homogeneous. See also Section X.7.

Observe that 2-homogeneity is a sufficient, but not a necessary condition for primitivity. Consider for example the group

$$\begin{aligned} & \langle (1,2,3)(4,5,6)(7,8,9), (1,5,9)(2,6,7)(3,4,8), (2,7,3,4)(5,8,9,6) \rangle \\ \cong \quad & \mathbb{F}_3^2 \rtimes \left\langle \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \right\rangle \end{aligned}$$

with $\mathbb{F}_3$ the field with three elements and the matrix considered modulo 3. It has two orbits of length 18 on sets of cardinality 2.

We now give the proof for part b) of Theorem VII.20, following [Cam76]: For a set Ω and $s \ge 0$, we denote by $\binom{\Omega}{s}$ the set of all s-element subsets of Ω.

Proposition VII.22: Let $G \le S_\Omega$ and $s, t \in \mathbb{Z}$ with $0 \le s \le t$ and $s + t \le |\Omega|$. Then G has at least as many orbits on $\binom{\Omega}{t}$, as it has orbits on $\binom{\Omega}{s}$.

Proof: Let F be the vector space of rational-valued functions from $\binom{\Omega}{t}$. We can consider this set as a $\binom{|\Omega|}{t}$-dimensional vector space over $\mathbb{Q}$ with a basis indexed by $\binom{\Omega}{t}$. We similarly define E to be the vector space of functions on $\binom{\Omega}{s}$. We now define a linear map $\theta\colon E \to F$ on basis elements as

$$\theta(f)(\Delta) = \sum_{\substack{\Lambda\subset\Delta \\ \Lambda\in\binom{\Omega}{s}}} f(\Lambda),$$

using the subset relation of s-element subsets in t-element subsets. If we take bases corresponding to the s-element, respectively t-element, subsets of Ω, then the matrix for θ is the incidence matrix for the sets. That is the matrix entry at a row corresponding to $A \subset \Omega$ and a column corresponding to $B \subset \Omega$ is one if and only if $A \subset B$ (and zero otherwise). We claim that Kern $\theta = \{0\}$, that is that θ is injective. For that, let $f \in$ Kern θ.

First consider the special case that $|\Omega| = s + t$. Then we can identify any t-set with its complement, an s-set with inclusion becoming disjointness. We define, for $\Delta \subset \Omega$

$$F(\Delta) = \sum_{\substack{S\in\binom{\Omega}{s} \\ S\cap\Delta=\varnothing}} f(S).$$

and then have for $\Lambda \in \binom{\Omega}{s}$, that $F(\Lambda) = 0$, because this is the value of $\theta(f)$ on the complement of Λ.

Now assume that $|\Delta| = j$ with $0 \le j \le s$. Then any $S \in \binom{\Omega}{s}$ with $S \cap \Delta = \varnothing$ will complement $\binom{t-j}{s-j}$ different s-element sets $\Lambda \supset \Delta$ (there are $t - j$ elements in $\Omega \setminus (\Delta \cup S)$, of which we can add $s - j$ elements to Δ to form such a Λ). If we sum the values of $F(\Lambda)$ over all these sets Λ, we thus add the value $f(S)$ for each set S disjoint from Δ exactly $\binom{t-j}{s-j}$ times. Thus

$$\binom{t-j}{s-j}F(\Delta) = \binom{t-j}{s-j} \sum_{\substack{S\in\binom{\Omega}{s} \\ S\cap\Delta=\varnothing}} f(S) = \sum_{\substack{\Lambda\in\binom{\Omega}{s} \\ \Delta\subset\Lambda}} \sum_{\substack{S\in\binom{\Omega}{s} \\ S\cap\Lambda=\varnothing}} f(S) = \sum_{\substack{\Lambda\in\binom{\Omega}{s} \\ \Delta\subset\Lambda}} F(\Lambda) = 0,$$

and thus $F(\Delta) = 0$.

We now change the scope of subsets in the summation, one element at a time. For disjoint subsets $\Lambda \in \binom{\Omega}{i}$, $\Delta \in \binom{\Omega}{j}$, with $i + j \le s$ (and $\Lambda \cap \Delta = \varnothing$) let

$$F(\Lambda, \Delta) = \sum_{\substack{S\in\binom{\Omega}{s} \\ \Lambda\subset S \\ S\cap\Delta=\varnothing}} f(S)$$

be the sum over $f(S)$ of sets containing Λ but disjoint to Δ. Then for $\Lambda \in \binom{\Omega}{s}$ we have that $F(\varnothing, \Lambda) = F(\Lambda) = 0$.

If we have $\omega \notin \Lambda$, then

$$F(\Lambda \cup \{\omega\}, \Delta) = F(\Lambda, \Delta) - F(\Lambda, \Delta \cup \{\omega\}),$$

since the sum on the left hand side disallows sets not containing ω. By induction over $|\Lambda|$ we thus get that $F(\Lambda, \Delta) = 0$ and in particular that $f(S) = F(S, \emptyset) = 0$. This proves the statement for $|\Omega| = s + t$.

In the general case, given $S \in \binom{\Omega}{s}$ pick $\Omega' \in \Omega s + t$ with $S \subset Y$ and define θ' analog to θ for subsets of Ω' and f' as the restriction of f to (the span of) s-element subsets of Ω'. Then $\theta'(f')$ is the restriction of $\theta(f)$ to (the span of) $\binom{\Omega'}{t}$, and thus zero. That is $f' \in \operatorname{Kern} \theta'$, and by the argument for the case $|\Omega| = s + t$ this implies that $f' = 0$. But $f(S) = f'(S)$, and so f is zero on S. Since S was arbitrary, this shows that $\operatorname{Kern} f = \{0\}$.

The action of G on $\binom{\Omega}{s}$ induces an action on E (with $f^g(S) := f(S^{g^{-1}})$ for $g \in G$, the inverse required for the product property), and similarly on F. For a vector space X we set

$$\operatorname{Fix}_G(X) = \{f \in X \mid f^g = f \forall g \in G\}$$

and observe that $f \in \operatorname{Fix}_G(E)$, if and only if f is constant on orbits of G on $\binom{\Omega}{s}$. The number of orbits of G on $\binom{\Omega}{s}$ thus equals the dimension of $\operatorname{Fix}_G(E)$, and similarly for F.

We now note that the G-actions are compatible with θ, we have that $\theta(f^g) = \theta(f)^g$, since for $\Delta \in \binom{\Omega}{t}$:

$$\begin{aligned} \theta(f)^g(\Delta) &= \theta(f)(\Delta^{g^{-1}}) = \sum_{\Lambda \subset \Delta^{g^{-1}}} f(\Lambda) = \sum_{(\Lambda^g)^{g^{-1}} \subset \Delta^{g^{-1}}} f((\Lambda^g)^{g^{-1}}) \\ &= \sum_{\Gamma^{g^{-1}} \subset \Delta^{g^{-1}}} f(\Gamma^{g^{-1}}) = \sum_{\Gamma \subset \Delta} f^g(\Gamma) = \theta(f^g)(\Delta). \end{aligned}$$

That means that $\operatorname{Fix}_G(E)$ under θ is mapped into $\operatorname{Fix}_G(F)$. But since θ is injective, this implies $\dim(\operatorname{Fix}_G(F)) \geq \dim(\operatorname{Fix}_G(E))$, which yields the claimed statement. □

By proposition VII.14, G is primitive if and only if the point stabilizer is maximal, respectively the action of G on the cosets of $S \leq G$ is primitive if and only if S is maximal in G. This indicates the importance of the concept of primitivity to pure algebra.

Exercise **??** shows that normal subgroups in primitive groups must act transitively. This gives, under the name O'Nan-Scott theorem, an entry point towards classifying primitive groups based on information about simple groups. This has been done explicitly so far for degree up to 4095 [CQRD11].

VII.5 Enumeration up to Group Action

So far we have seen several ways of enumerating objects by taking them apart and reducing the enumeration to that of the constituents. This however might involve choices and thus arbitrarily restrict symmetries. A typical example would be to enumerate all graphs on n vertices: For example, figure VII.3 showed all possible labeled graphs on 3 vertices clustered into orbits of S_3. Typically we do not want to enumerate the *labeled* graphs (in which the vertices have fixed numbers, and where there clearly are $2^{\binom{n}{2}}$ of them), but *unlabeled* ones, allowing for an arbitrary labeling of the vertices. That is, we want to enumerate orbits.

If the set Ω of objects is small, this is easy to do. We thus will assume now that Ω is large. We will use instead the structure of the acting group G to count the number of orbits of G on Ω without explicitly listing the orbits (or obtaining orbit representatives). While this might initially look as if we just replaces one problem with another, these groups are often highly regular and useful structural information is known.

The basic tool is the following lemma which sometimes is attributed to Burnside, Frobenius and Cauchy:

Lemma VII.23: Let G act on Ω and denote by $\mathrm{fix}(g) = |\{\omega \in \Omega \mid \omega^g = \omega\}|$ the number of fixed points of a group element $g \in G$. Then the number of orbits of G on Ω is given by

$$\frac{1}{|G|}\sum_{g\in G}\mathrm{fix}(g)$$

Proof: We double count the number of pairs $(\omega, g) \in \Omega \times G$ for which $\omega^g = \omega$. Summing over group elements this is clearly $\sum_{g\in G}\mathrm{fix}(g)$.

On the other hand we can sum over elements of Ω, according to the n orbits $X_1, \dots, X_n$, and get

$$\begin{aligned}\sum_{\omega\in\Omega}|\mathrm{Stab}_G(\omega)| &= \sum_i\sum_{\omega\in X_i}|\mathrm{Stab}_G(\omega)| = \sum_i\sum_{\omega\in X_i}\frac{|G|}{|\omega^G|} = \sum_i\sum_{\omega\in X_i}\frac{|G|}{|X_i|}\\ &= |G|\sum_i\frac{1}{|X_i|}\sum_{\omega\in X_i}1 = |G|\sum_i\frac{|X_i|}{|X_i|} = |G|\cdot n.\end{aligned}$$

Setting the two expressions equal yields the desired result. □

In applying this lemma we note that "action equivalent" elements — for example[2] elements conjugate in G, that is equivalent under the relation $g \sim g^h = h^{-1}gh$ for $h \in G$ — have equal fix point numbers, allowing a summation over a significantly smaller set.

Example: We want to count the ways in which we can color the six faces of a cube with green or gold. With two color choices, there are $2^6 = 64$ possibilities if we consider the faces of the cube of being labeled.

To count orbits under the symmetry group we need to enumerate its element types (e.g. conjugacy classes) and determine for each the number of their fixed points.

[2] In some situations we can also consider conjugacy under a larger group normalizing G

To enumerate the elements we shall simply exhibit some elements that clearly must exist and then (as we know already the group order) show that we have found all 24 elements of this group.

To count fixed points, we notice that a coloring fixed under a group element g must have the same color on faces in the same orbit of $\langle g\rangle$. That is, if we write the action of g on the faces of the cube in cycle form, there are 2^c fixed colorings, where c is the number of cycles (including cycles of length 1 which we do not write down).

The possible non-identity rotations will have axes that go through the middle of a face, an edge or a corner. This gives the following tabulation:

Kind	Order	#	Cycle Structure	$\mathrm{fix}(g) = 2^c$	$\# \cdot \mathrm{fix}(g)$
Identity	1	1	1^6	$2^6 = 64$	64
Face,180°	2	3	$1^2 2^2$	$2^4 = 16$	48
Face,90°	4	6	$1^2 4^1$	$2^3 = 8$	48
Edge	2	6	2^3	$2^3 = 8$	48
Corner	3	8	3^2	$2^2 = 4$	32
Sum		24			240

There thus are $240/24 = 10$ different colorings. If we permitted three colors, we only would need to replace 2^c by 3^c.

What makes this method powerful is that the acting group in many cases has a structure that allows us to determine the conjugacy classes easily, even if the action is not the natural one.

Example: We want to count the number of (unlabeled) graphs on four vertices. We consider edges as pairs of vertices, the group action is thus the one by S_4 on pairs. (There are $\binom{4}{2} = 6$ such pairs. In fact these two examples utilize the two different transitive actions of S_4 on 6 points.) Again for a graph to be fixed under a group element all edges in one element orbit must either be set or not set. An element with o orbits on pairs thus has 2^o fixed graphs.

This gives the following table:

Kind	Order	#	o = #orbits on pairs	$\# \cdot 2^o$
$()$	1	1	6	64
$(1,2)$	2	6	4	96
$(1,2)(3,4)$	2	3	4	48
$(1,2,3)$	3	8	2	32
$(1,2,3,4)$	4	6	2	24
Sum		24		264

There thus are $264/24 = 11$ different unlabeled graphs on 4 vertices.

VII.6 Pólya Enumeration Theory

In many cases, such as the graphs we just considered, the action we want to consider is one on composed objects, such as sets or sequences. That is, we have an action of G on a set X and based on this want to count orbits on a more complicated set Ω.

In the case of ordinary enumeration (Section III.6), the product of generating functions was a tool do allow composition of objects. We will introduce a similar tool for enumeration up to symmetry in this section. After its discoverers this often goes under the name of Pólya or Pólya-Redfield enumeration theory.

We start with a partition of $n = |X|$, given by the cycle structure of a permutation g. Assuming a partition $1^{c_1}2^{c_2}\cdots n^{c_n}$, that is c_1 fixed points, c_2 2-cycles etc..

Definition VII.24: Let $G \le S_n$. For $g \in G$ the *cycle index* of g is

$$z(g; s_1, \ldots, s_n) = s_1^{c_1} s_2^{c_2} \cdots s_n^{c_n}$$

where the s_i are indeterminates.

The cycle index of a finite group G is defined as

$$Z(G; s_1, \ldots, s_n) = \frac{1}{|G|} \sum_{g \in G} z(g; s_1, \ldots, s_n) = \frac{1}{|G|} \sum_{g \in G} \prod_i s_i^{c_i(g)}$$

where the $c_i(g)$ indicate the cycle structure of $g \in G$.

If G is not a permutation group itself, but acts on X we consider the permutations induced on X.

If G acts on X, it induces an action on k-sets, or k-sequences of elements of X. The following proposition indicates how the cycle index connects to these actions. Note the connection to the multiplication principles for generating functions — a tuple of distinct elements is a set with labeled positions for the entries.

Proposition VII.25: a) For $1 \le k \le n$ let f_k the number of orbits of G on k-sets of elements of X. Then

$$\sum_{k=0}^{n} f_k t^k = Z(G, 1+t, 1+t^2, \ldots, 1+t^n).$$

b) For $1 \le k \le n$ let F_k the number of orbits of G on k-tuples of (distinct) elements of X. Then

$$\sum_{k=0}^{n} F_k \frac{t^k}{k!} = Z(G, 1+t, 1, \ldots, 1).$$

<u>Proof:</u> a) By Lemma VII.23 we have that

$$\sum_{k=0}^{n} f_k t^k = \frac{1}{|G|} \sum_{g \in G} \sum_{k=0}^{n} \mathrm{fix}_k(g) t^k$$

with fix_k denoting the number of fixed-points on k-sets. To determine this number for $g \in G$, we note that a fixed k-set must be the union of cycles of g. On the other hand, for any choice of numbers $b_i \le c_i(g)$ with $\sum ib_i = k$, we can find k-sets that are fixed under g by selecting b_i cycles of length i. There are $\binom{c_i(g)}{b_i}$ choices for such cycles. Thus

$$\mathrm{fix}_k(g) = \sum_{(b_1,\ldots,b_n)\in B_k(g)} \prod_{i=1}^{n} \binom{c_i(g)}{b_i}$$

where

$$B_k(g) = \left\{(b_1,\ldots,b_n) \mid 0 \le b_i \le c_i(g), \sum_i b_i i = k\right\}.$$

Therefore

$$\begin{aligned}\sum_{k=0}^{n} \mathrm{fix}_k(g)t^k &= \sum_k \left(\sum_{(b_1,\ldots,b_n)\in B_k(g)} \prod_{i=1}^{n} \binom{c_i(g)}{b_i} \right) t^k \\ &= \prod_{i=1}^{n} \sum_{b_i=0}^{c_i(g)} \binom{c_i(g)}{b_i} t^{ib_i} = \prod_{i=1}^{n} \sum_{b=0}^{c_i(g)} \binom{c_i(g)}{b} t^{ib} \\ &= \prod_{i=1}^{n} (1+t^i)^{c_i(g)} = z(g; 1+t, 1+t^2, \ldots 1+t^n).\end{aligned}$$

by the binomial theorem. Averaging over G gives the desired result.
b) Now $\mathrm{Fix}_k(g)$ shall denote the number of fixed points on k-tuples. For a tuple to be fixed, all entries must be fixed, that is the entries may be chosen only from the fixed points of g. Using the notation of Theorem II.2, we thus get that:

$$\mathrm{Fix}_k(g) = c_1(g)(c_1(g)-1)\cdots(c_1(g)-k+1) = (c_1(g))_k.$$

As

$$\sum_{k=0}^{n} \frac{(c_1(g))_k}{k!} t^k = \sum_{k=0}^{n} \binom{c_1(g)}{k} t^k = (1+t)^{c_1(g)} = z(g; 1+t, 1, \ldots, 1),$$

we get again from Lemma VII.23 and averaging that:

$$\begin{aligned}\sum_{k=0}^{n} F_k \frac{t^k}{k!} &= \frac{1}{|G|} \sum_{g\in G} \sum_{k=0}^{n} \frac{(c_1(g))_k}{k!} t^k = \frac{1}{|G|} \sum_{g\in G} z(g; 1+t, 1, \ldots, 1) \\ &= Z(G, 1+t, 1, \ldots, 1).\end{aligned}$$

□

The Cycle Index Theorem

Stepping up from sets and tuples, we now generalize, as in Section II.4 to functions. This is a generalization, since a subset of X can be considered as a function from X to $\{0,1\}$.

For this, consider a set C of objects we shall call "colors" (though they could simply be different integers). For a set X, we consider the set C^X of functions from $X \to C$. If G acts on X, it also acts on the functions $f \in C^X$ by

$$(f^g)(x) := f(x^{g^{-1}})$$

The inverse is needed, as we are permuting the arguments. It ensures that

$$((f^g)^h)(x) = (f^g)(x^{h^{-1}}) = f\left((x^{h^{-1}})^{g^{-1}}\right) = f(x^{h^{-1}g^{-1}}) = f(x^{(gh)^{-1}}) = f^{gh}(x).$$

We also introduce a "weight" function $w\colon C \to \mathbb{Z}_{\geq 0}$ and define the weight of a function to be the sum of the weights of its values:

$$w(f) := \sum_{x \in X} w(f(x))$$

The action of G, permuting X, clearly preserves weights.

Many enumerations up to symmetry then can be phrased as counting the orbits of G on a set of such functions (of particular weight(s)).
Example:

1. For subsets of X, we let $C = \{0, 1\}$ and assign weight 0 to 0 and weight 1 to 1. A subset of size k then corresponds to a function $f \in C^X$ of weight k and the action of G on the set of such functions is equivalent to the action of G on k-subsets of X.

2. Let X be the faces of a cube and G the group of rotational symmetries. A coloring of the cube with colors from a finite set C then corresponds to a function in C^X, and the action of G on C^X is the rotational equivalence of colorings. We could simply give all colors the same weight, or use weights to count separately the colorings in which certain colors arise a certain number of times.

3. A graph on n (labeled) vertices can be considered as a function from the set X of possible edges (that is $|X| = \binom{n}{2}$) to the indicators $C = \{0, 1\}$ that indicate whether a possible edge exists. The action of $G = S_n$ permutes vertex labels so that the induced action on the functions C^X gives isomorphic graphs. One could use weights on C to consider graphs with a particular number of edges.

Theorem VII.26 (Cycle Index Theorem): Let a_i be the number of colors of weight i and let $a(t) = \sum_{i \geq 0} a_i t^i$. Let b_i the number of orbits of G on the set $\Omega_i \subset C^X$ of functions of weight i and let $b(t) = \sum_{i \geq 0} b_i t^i$. Then

$$b(t) = Z(G; a(t), a(t^2), \ldots, a(t^n)).$$

Corollary VII.27: If we assign weight 0 to all of C then $a(t) = |C|$. Thus the total number of orbits of G on $\Omega = C^X$ is given by

$$Z(G; |C|, |C|, \ldots, |C|).$$

Corollary VII.28: If $C = \{0,1\}$ with weight $0,1$ respectively, then $a(t) = 1+t$. The number of orbits of G on k-subsets of X is (in agreement with Proposition VII.25 a)) given by

$$Z(G, 1+t, 1+t^2, \ldots, 1+t^n).$$

Example: We consider once more the example of coloring the faces of a cube. Going back to the example above, we can read off the cycle index from the table as:

$$Z(G; s_1, s_2, s_3, s_4) = \frac{1}{24}(s_1^6 + 3s_1^2s_2^2 + 6s_1^2s_4 + 6s_2^3 + 8s_3^2)$$

We give green weight 1 and gold weight 0, so we have $a(t) = 1 + t$. The theorem gives us that the series $b(t)$, counting the colorings of particular weight is

$$b(t) = Z(G, 1+t, 1+t^2, \ldots, 1+t^n) = t^6 + t^5 + 2t^4 + 2t^3 + 2t^2 + t + 1.$$

We thus can read off that there is one coloring each with $0, 1, 5$ or 6 green faces, and two colorings each with $2, 3$ or 4 green faces.

Proof: (of Theorem VII.26) We start by considering the function-counting series for the trivial group, that is counting functions $X \to C$ according to weight. We claim that this series is $a(t)^n$ with $n = |X|$:

The coefficient of t^e in this series is the number of tuples $(c_1, \ldots, c_n) \in C^n$ of weight sum $e = \sum_i w(c_i)$. For a particular distribution of weights, $w(c_i) = w_i$ this gives a_{w_i} color choices for c_i, that is $\prod_i a_{w_i}$ possibilities of tuples. Summing over all possible weight distributions gives

$$\sum_{(w_i), \sum w_i = e} \prod_i a_{w_i}$$

which is also the coefficient of t^e in $a(t)^n = \left(\sum a_i t^i\right)^n$, showing that the series must be equal.

Next we consider functions that are fixed by a particular permutation g. This means that the functions must be constant on the cycles of g. For a single cycle of length i this gives the weight of the chosen color i-fold, thus the weight counting series for the single cycle of length i is $a(t^i)$. If there are k cycles it thus is $a(t^i)^k$.

We now consider all cycles, assuming that there are k_i cycles of length i. By induction we get the enumeration function

$$\prod_i a(t^i)^{k_i} = z(g; a(t), a(t^2), \ldots, a(t^n)).$$

The number of orbits of G on functions of weight w is, by Lemma VII.23, the average number of fixed functions of weight w. By using the enumeration function we can do so simultaneously for all weights, getting

$$\frac{1}{|G|} \sum_{g \in G} z(g; a(t), a(t^2), \ldots, a(t^n)) = Z(G; a(t), a(t^2), \ldots, a(t^n))$$

by definition. □

We consider another example of enumerating graphs:
Example: Consider the graphs on 5 vertices. The group acting on the vertices is S_5. The conjugacy classes are parameterized by partitions of 5, and it is an easy counting argument to get the class sizes. However we will need the cycle index for the action on the 10 pairs of points. So we first need to get the cycle structure for this action, which is done in Exercise **??**. This gives us the following table of conjugacy classes:

Element	Number	Cycle structure on pairs
()	1	1^{10}
(1,2)	10	$1^4 2^3$
(1,2)(3,4)	15	$1^2 2^4$
(1,2,3)	20	$1 \cdot 3^3$
(1,2,3)(4,5)	20	$1 \cdot 3 \cdot 6$
(1,2,3,4)	30	$2 \cdot 4^2$
(1,2,3,4,5)	24	5^2

This gives the cycle index:

$$\frac{1}{120}\left(s_1^{10} + 10 \cdot s_1^4 s_2^3 + 15 \cdot s_1^2 s_2^4 + 20 \cdot s_1 s_3^3 + 20 \cdot s_1 s_3 s_6 + 30 \cdot s_2 s_4^2 + 24 \cdot s_5^2\right)$$

We consider edges as weight 1 and non-edges as weight 0, giving again the series $a(t) = 1 + t$ The theorem gives us:

$$\begin{aligned} b(t) &= Z(G; 1+t, 1+t^2, 1+t^3, 1+t^4, 1+t^5, 1+t^6) \\ &= t^{10} + t^9 + 2t^8 + 4t^7 + 6t^6 + 6t^5 + 6t^4 + 4t^3 + 2t^2 + t + 1, \end{aligned}$$

telling us that there are one graphs each with $0, 1, 9$ and 10 edges, two graphs with 2 and 8 edges, four each with 3 and 7 edges, as well as six each with $4, 5$ or 6 edges. Knowing these separate numbers makes it far easier to enumerate them explicitly.

If we were only interested in the total number of graphs, we could have given the same weight to both edges and obtained

$$Z(G; 2; 2; 2; 2; 2; 2) = 34.$$

If we wanted to enumerate colored edges, say with 3 possible edge colors, we would give each color a different weight and thus get

$$Z(G; 4; 4; 4; 4; 4; 4) = 10688$$

possible graphs.

We notice that the cycle index behaves nicely with respect to group products:

Lemma VII.29: Let G, H be permutation groups. Then
a) $Z(G \times H) = Z(G)Z(H)$ (in the intransitive action of $G \times H$),
b) $Z(G \wr H) = Z(H; Z(G; s_1, s_2, \ldots), Z(G; s_2, s_4, \ldots), Z(G; s_3, s_6, \ldots), \ldots)$ (in the imprimitive action of the wreath product).

The proof of part a) is exercise **??**, part b) an unappealing direct calculation.

Chapter

VIII

Finite Geometry

One must be able to say at all times–instead of points, straight lines, and planes–tables, chairs, and beer mugs

Man muß jederzeit an Stelle von “Punkte, Geraden, Ebenen” “Tische, Stühle, Bierseidel” sagen können.

Lebensgeschichte, in:
David Hilbert, Gesammelte mathematische Abhandlungen
O. Blumenthal

Mathematics in the Ancient World arose from counting and from geometry. Having dealt with counting problems in the first part, it thus seems appropriate to begin this second part with geometric ideas.

We consider a *geometry* as a collections of two sets of objects – points and lines, together with an *incidence* relation of a point being on a line, or a line containing a point. (We shall encounter this dichotomy amongst two classes of objects again in later chapters.) A *finite geometry* then is simply a geometry for which the sets of points and lines both are finite. (In such a setting we may consider a line simply as the set of its points.)

We shall start by describing the construction of some (finite) structures consisting of points and lines. For this we shall need a few basic facts about finite fields:

Intermezzo: Finite Fields

A *field* is a set that is closed under commutative addition and multiplication and that has a multiplicative one (a commutative ring with one) such that every nonzero element has a multiplicative inverse. (There are some axioms which ensure the usual arithmetic rules apply, we will not need to delve into these here.)

Typical examples of fields are the Rational, the Real, or the Complex Numbers, but we will be more interested in finite fields.

The standard example of a finite field is given by modulo arithmetic: For a prime p, we take the numbers $0, \ldots, p-1$ under addition and multiplication modulo p. Inverses exist as for any nonzero $1 \leq a \leq p-1$ we have that $\gcd(a, p) = 1$, thus (extended Euclidean algorithm) there exist x, y such that $xa + yp = 1$, that is $x \cdot a \equiv 1 \pmod{p}$ and thus x is the multiplicative inverse of a. We denote this field with p elements by $\mathbb{F}_p$.

A source of further finite fields comes from polynomials. For an irreducible polynomial $f(x) \in \mathbb{F}_p[x]$ of degree n, we consider the polynomials $\in \mathbb{F}_p[x]$ of degree $< n$ with arithmetic modulo $f(x)$. These polynomials form an n-dimensional vector space over $\mathbb{F}_p$ which thus contains p^n elements. We denote this set with the modulo arithmetic by $\mathbb{F}_p(\alpha)$ where α, the element represented by the remainder x, is a root of f.

Since f is irreducible every nonzero polynomial has a multiplicative inverse, so $\mathbb{F}_p(\alpha)$ is a field.

Example: Working modulo 2, the polynomial $x^2 + x + 1$ is irreducible over the field with 2 elements. The associated field $\mathbb{F}_2(\alpha)$ thus has elements we can denote by $\{0, 1, \alpha, \alpha+1\}$ and arithmetic given by

$+$	0	1	α	$\alpha+1$
0	0	1	α	$\alpha+1$
1	1	0	$\alpha+1$	α
α	α	$\alpha+1$	0	1
$\alpha+1$	$\alpha+1$	α	1	0

$\cdot$	0	1	α	$\alpha+1$
0	0	0	0	0
1	0	1	α	$\alpha+1$
α	0	α	$\alpha+1$	1
$\alpha+1$	0	$\alpha+1$	1	α

Example: In the field $\mathbb{F}_3$ there is no element a such that $a^2 = -1$, thus the polynomial $x^2 + 1$ is irreducible. This yields a field $\mathbb{F}_3(\alpha)$ with 9 elements.

The following theorem (which summarizes results that are proven in a standard graduate abstract algebra class) summarizes properties of these fields.

Theorem VIII.1: a) Any finite field has order p^n for a prime p and an integer $n > 1$. We call p the *characteristic* of the field.

Now let p be a prime and $n > 1$ an integer

b) There exists an irreducible polynomial of degree n over $\mathbb{F}_p$ and thus a field with p^n elements.

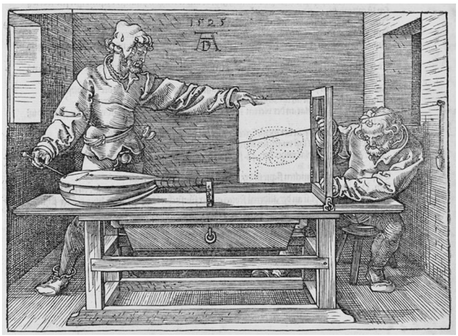
c) Regardless of the choice of polynomial, any two fields with p^n elements are isomorphic (since the polynomial $x^{p^n} - x$ splits over either field into p^n different linear factors). We thus can talk of *the* field with p^n elements; we denote this field by $\mathbb{F}_{p^n}$.

d) The field $\mathbb{F}_{p^a}$ is isomorphic to a subfield of $\mathbb{F}_{p^b}$ if and only if $a \mid b$.

e) The map $x \mapsto x^p$ is an automorphism (i.e. a map that preserves the arithmetic operations) of the field $\mathbb{F}_{p^a}$, called the *Frobenius automorphism*. If has order a (thus is the identity if $a = 1$) and generates the group of field automorphisms (also called the Galois group) of $\mathbb{F}$. f) The multiplicative group of $\mathbb{F}$ is cyclic (of order $p^n - 1$).

This theorem in particularly implies that the choice of a different irreducible polynomial (while affecting the coordinates of how we describe an extension) has no impact on the abstract structure of the field extension we obtain.

VIII.1 Projective Geometry

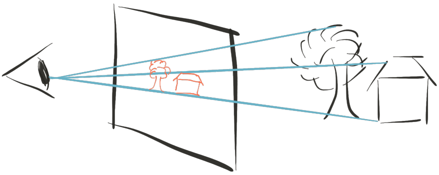


Underweysung der Messung
Albrecht Dürer

In standard 3-dimensional space, lines may intersect, be parallel, or be skew. This is a certain asymmetry which we can avoid by adding further points and lines. We can describe this nicely in a systematic way by employing a tool of Renaissance Art:

When looking along two parallel lines – say the two metal bars of a long straight railway track, they actually do not seem parallel to us, but to approach asymptotically in the distance. A picture of such a situation thus only looks realistic if it shows a similar narrowing in the distance.

We can describe this phenomenon by imagining the picture being located between the viewer's eye (or a virtual focus point) and the situation in reality:



A point a in reality then is depicted by a point b in the picture, by taking the line connecting a with the eye, and taking for b the intersection of this line with the picture plane. That is, all points on a line from the eye would project to the same point in the picture (and usually only the closest point is drawn).

Mathematically, we can interpret this situation as the vector space $\mathbb{R}^3$ with the eye at an origin. The lines are one-dimensional subspaces, and the picture is a hyperplane offset from the origin (that is, a coset of a 2-dimensional subspace). Every point in this hyperplane intersects a 1-dimensional subspace.

Two 1-dimensional subspaces span a 2-dimensional space, the 1-dimensional subspaces within all project onto a line in the hyperplane.

We thus define $\mathrm{PG}(2,\mathbb{R})$ as the set of all 1-dimensional subspaces of $\mathbb{R}^3$, each of which we call a projective "point". The "lines" of $\mathrm{PG}(2,\mathbb{R})$ are given the sets of points, arising from 1-dimensional subspaces that lie in a two-dimensional space.

With this definition we get more than the Renaissance artist bargained for:

1. There are 1-dimensional subspaces (those in the 2-dimensional subspace that yielded the hyperplane) which do not intersect in a projection point. We call these the "points at infinity". All of these lie in the "line at infinity", the projection of the 2-dimensional space they lie in.

2. If two lines are given, they arise from two 2-dimensional spaces. In a three-dimensional space they must intersect nontrivially in a 1-dimensional subspace that will project to a point. That is *any two (different) lines intersect in exactly one point*. (That point will be "at infinity" for lines that are parallel, it would be the "vanishing point" of these lines.)

Analogous constructions are of course possible in higher dimensions (where we might get multiple lines at infinity and non-intersecting lines) and over arbitrary fields F, yielding the projective space $\mathrm{PG}(n,F)$. The projection image of a $k+1$-dimensional subspace of F^{n+1} in $\mathrm{PG}(n,F)$ is called a k-flat.

Taking finite fields we get finite structures of points and lines that are examples of finite geometries. We shall write $\mathrm{PG}(n,q)$ for $\mathrm{PG}(n,\mathbb{F}_q)$.
Example: The smallest example possible is $\mathrm{PG}(2,2)$ with 7 points and 7 lines. It is called the *Fano plane*. Its incidence of points and lines is depicted in Figure VIII.3.

VIII.2 Gaussian Coefficients

To count points in these spaces, we introduce a relative of the binomial coefficients:

Definition VIII.2: The *Gaussian coefficient* $\begin{bmatrix} n \\ k \end{bmatrix}_q$ is defined to be the number of k-dimensional subspaces of $\mathbb{F}_q^n$.

and get immediately:

Corollary VIII.3: $\mathrm{PG}(n,q)$ has $\begin{bmatrix} n+1 \\ 1 \end{bmatrix}$ points. It has $\begin{bmatrix} n+1 \\ k+1 \end{bmatrix}_q$ k-flats, each containing $\begin{bmatrix} k+1 \\ 1 \end{bmatrix}_q$ points.

Proposition VIII.4:

$$\begin{bmatrix} n \\ k \end{bmatrix}_q = \frac{(q^n-1)(q^{n-1}-1)\cdots(q^{n-k+1}-1)}{(q^k-1)(q^{k-1}-1)\cdots(q-1)}$$

Proof: A k-dimensional subspace will be spanned by a set of k linear independent vectors. Counting these sets we get q^n-1 choices for the first vector, $q^n-1=q(q^{n-1}-1)$ choices for the second vector (we need to leave out multiples of the first vector) and, leaving out vectors in the span of the first $i-1$ vectors, $q^n-q^{i-1}=q^{i-1}(q^{n-i+1}-1)$ choices for the i-th basis vector.

The same counting argument gives us that an k-dimensional subspace has

$$|\mathrm{GL}_k(q)|=(q^k-1)(q^k-q)\cdots(q^k-q^{k-1})=(q^k-1)q(q^{k-1}-1)q^2(q^{k-2}-1)\cdots q^{k-1}(q-1)$$

different bases. To count subspaces we thus need to take the quotient of these two expressions. □

VIII.3 Automorphisms of PG$(n,\mathbb{F})$

We want to describe the group of automorphisms of PG$(n,\mathbb{F})$, that is of those permutations of points that preserve the structure of lines, respectively permutations of points and lines that preserve incidence. (Such maps also are called *collineations*.) The resulting groups are prominent, not only in geometry, but also in many other areas of mathematics.

We start by describing two classes of automorphisms of PG$(n,\mathbb{F})$:

The group $\mathrm{GL}_{n+1}(\mathbb{F})$ acts linearly on the vectors of $\mathbb{F}^{n+1}$ and thus on subspaces of a given dimension. It thus induces automorphisms of PG$(n,\mathbb{F})$.

Example: The Fano plane, PG$(2,2)$ arises from $\mathbb{F}_2^3$ and thus inherits an action of $\mathrm{GL}_3(2)$ of order 168. On the other hand, automorphisms that fix one of the 7 lines (as a set of points) are completely determined by what they do on the 4 points not on that line, thus giving at most 168 automorphisms.

If $\mathbb{F}$ is not a prime field (i.e. it contains another field as strict subset – for finite $\mathbb{F}$ this means the order of $\mathbb{F}$ is not a prime), there typically will be field automorphisms of $\mathbb{F}$.

According to Theorem VIII.1, for $\mathbb{F}=\mathbb{F}_{p^a}$ this groups is cyclic of order a and generated by the Frobenius automorphism $x\mapsto x^p$.

On the vector space $\mathbb{F}_q^{n+1}$ these field automorphisms induce maps that map subspaces to subspaces, though they are not linear (they preserve addition, but scalar multiplication would be mapped to a multiple by the mapped scalar; such maps are called *semilinear*).

If we take these two sets of operations of $\mathbb{F}^{n+1}$ together we get a group $\Gamma L_{n+1}(\mathbb{F})$. As shown in exercise **??**, its structure is that of a semidirect product

$$\Gamma L_{n+1}(\mathbb{F})=\mathrm{GL}_{n+1}(\mathbb{F})\rtimes\langle x\mapsto x^p\rangle$$

and its order is (for $\mathbb{F}=\mathbb{F}_q$ with $q=p^a$)

$$(q^{n+1}-1)(q^{n+1}-q)(q^{n+1}-q^2)\cdots(q^{n+1}-q^n)\cdot a.$$

The action of this group on subspaces induces automorphisms (incidence-preserving bijective maps) of $\mathrm{PG}(n, \mathbb{F})$. Let $\varphi\colon \Gamma L_{n+1}(\mathbb{F}) \to \mathrm{Aut}(\mathrm{PG}(n, \mathbb{F}))$.

Lemma VIII.5:

$$\mathrm{Kern}\,\varphi = \{f \cdot I \in \mathrm{GL}_{n+1}(\mathbb{F}) \mid f \in \mathbb{F}\} = Z(GL_{n+1}(\mathbb{F}))$$

Proof: It is clear that scalar multiplication fixes all subspaces and thus induces trivial automorphisms of PG.

Vice versa, suppose that $\alpha \cdot M \in \mathrm{Kern}\,\varphi$ with $M \in \mathrm{GL}_{n+1}(\mathbb{F})$ and $\alpha \in \mathrm{Aut}(\mathbb{F})$.

Let $\{e_1, \ldots, e_{n+1}\}$ the standard basis of $\mathbb{F}^{n+1}$. As the entries of the e_i are zero or one, α will act trivially on these vectors. By virtue of lying in $\mathrm{Kern}\,\varphi$, $\alpha \cdot M$, and thus M preserves the spaces $\langle e_i \rangle$, thus $e_i \cdot M = \lambda_i \cdot M$.

A similar argument shows that $\langle e_i + e_j \rangle$ must be fixed by M and thus there is $\lambda \in \mathbb{F}$ such that

$$\lambda(e_i + e_j) = (e_i + e_j)M = e_iM + e_jM = \lambda_i e_i + \lambda_j e_j$$

The linear independence of the vectors implies that $\lambda = \lambda_i = \lambda_j$, thus M is scalar.

By considering spaces $\langle \gamma \cdot e_1 \rangle$ for $\gamma \in \mathbb{F}$ we also get that $\alpha = id$.

By considering products with elementary matrices, is seen easily that this kernel indeed the center of the group. □

We thus can consider the factor group

$$P\Gamma L_{n+1}(\mathbb{F}) = \Gamma L_{n+1}(\mathbb{F})/Z(GL_{n+1}(\mathbb{F}))$$

as a subgroup of $\mathrm{Aut}(\mathrm{PG}(n, \mathbb{F}))$. In fact we have equality:

Theorem VIII.6 (Fundamental theorem of projective Geometry):

$$\mathrm{Aut}(\mathrm{PG}(n, \mathbb{F})) \cong P\Gamma L_{n+1}(\mathbb{F}).$$

We do not give a proof here; while it is not overly hard, it is lengthy in nature.

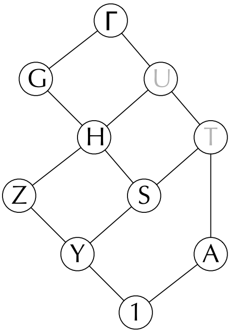
Instead we briefly describe the structure of $\Gamma L_n(\mathbb{F})$, see Figure VIII.1

We set $G = \mathrm{GL}_n(\mathbb{F})$ and $A = \mathrm{Aut}(\mathbb{F})$. By definition $\langle G, A \rangle = \Gamma L_n(\mathbb{F})$ and $G \cap A = \langle 1 \rangle$. We let $S = \mathrm{SL}_n(\mathbb{F})$ the kernel of the determinant map and $Z = Z(\mathrm{GL}_n(\mathbb{F}))$ the group of scalar matrices and set $H = \langle Z, S \rangle$ and $Y = Z \cap S$. Then $Y = Z \cap S = \mathbb{Z}(\mathrm{SL}_n(\mathbb{F}))$, since the elementary matrices lie in SL the proof that $Z(\mathrm{GL})$ is scalar essentially repeats, showing that $Z(\mathrm{SL})$ also must be scalar.

If $\mathbb{F}$ is finite, we can determine some indices and orders:

As matrices in G have nonzero determinant we get that $[G : S] = q - 1 = |Z|$.

The matrices in $Y = Z(\mathrm{SL})$ are scalar $\lambda{\cdot}I$ and must satisfy that $1 = \det(\lambda{\cdot}I) = \lambda^n$, that is the multiplicative order of λ in the multiplicative group $\mathbb{F}^*$ divides n. As $\mathbb{F}^*$



Γ	$\Gamma L_n(\mathbb{F}) = \langle \mathrm{GL}_n(\mathbb{F}), Aut(\mathbb{F})\rangle$
G	$\mathrm{GL}_n(\mathbb{F})$
A	$\mathrm{Aut}(\mathbb{F})$
S	$\mathrm{SL}_n(F)$, det = 1
Z	$Z(\mathrm{GL}_n(\mathbb{F}))$
Y	$Z(\mathrm{SL}_n(\mathbb{F}))$
T	$\langle \mathrm{SL}_n(\mathbb{F}), \mathrm{Aut}(\mathbb{F})\rangle$
H	$\langle \mathrm{SL}_n(\mathbb{F}), Z(\mathrm{GL}_n(\mathbb{F}))\rangle$
U	$\langle H, \mathrm{Aut}(\mathbb{F})\rangle$

Figure VIII.1: Structure of $\Gamma L_n(\mathbb{F})$

is cyclic of order $q-1$, it contains exactly $\gcd(q-1,n)$ (the largest subgroup whose order divides n) elements of order dividing n, thus $|Y| = \gcd q-1, n$.

Since $S, Z \lhd G$, we have that $[Z:Y] = [H:S]$, this implies that

$$[G:H] = \frac{q-1}{[H:S]} = \frac{q-1}{[Z:Y]} = \frac{q-1}{(q-1)/|Y|} = |Y| = \gcd(q-1,n).$$

Finally, as a consequence of the semidirect product structure and Theorem VIII.1 we get that $|A| = [\Gamma : G] = \log_p(|\mathbb{F}|)$.

(The subgroups U, T are only listed for completion.)

VIII.4 Projective Spaces

> Parallel lines
> Who meet
> Side by side!
>
> Side by Side by Side (from *Company*)
> Stephen Sondheim

Euclidean geometry starts with a set of axioms. We thus now change our point of view and introduce an axiomatic description of geometry that fits the projective geometries we have seen.

Formally, we define a projective geometry as a set of points and a set of lines (as sets of points), satisfying the following conditions:

PG1 For any two distinct points, there is one, and only one, line containing both points

PG2 If A, B, C are three points not on a line, and if $D \neq A$ is a point on the line through A, B, and of $E \neq A$ is a point on the line through A, C, then there is a point F on a line with D and E and also on a line with B and C.

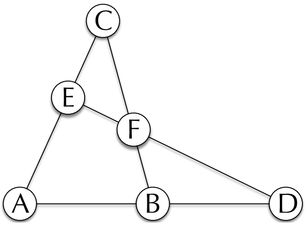


Figure VIII.2: The Pasch axiom

PG3 Every line contains at least three points

Axiom PG1 makes it possible to talk about a "line through two points". It in particular implies that two lines can intersect in but one point (as otherwise there would be two different lines going through both intersection points). Axiom PG3 eliminates some degenerate cases. Axiom PG2, the *Pasch axiom* requires some explanation.

Consider figure VIII.2. The lines $\overline{AB}$ and $\overline{AC}$ intersect in a point, namely A. In Euclidean geometry we would interpret this as the two lines lying in a plane. The configuration (and the point A) thus simply ensures that the lines $\overline{CB}$ and $\overline{DE}$ lie in a common plane. We can thus interpret this axiom instead as: *Two lines in the same plane have a point of intersection.*

We thus do not have a parallel axiom, but always have points (maybe "at infinity") where lines meet.

As we have drawn a picture, a word on what these are is in order: Pictures of configurations only indicate incidence. There are no distances, no angles, no closest points. There are no "intermediate" points. A line does not need to be drawn straight, but neither can we deduce incidence from the intersection of lines. That is, pictures will only ever illustrate proofs, never be proofs in themselves.

It now turns out that these axioms not only match the projective spaces $\mathrm{PG}(n,\mathbb{F})$, but are a complete characterization unless the dimension is 2:

Theorem VIII.7 (Veblen-Young): Consider a projective geometry, satisfying the axioms PG1,2,3 above. If there are two lines that do not intersect (one says that it has dimension > 2), then this configuration is isomorphic to $\mathrm{PG}(n,\mathbb{F})$ for some $n \geq 3$ and a division ring[1] $\mathbb{F}$.

The proof of this is long and consists of geometric constructions, with the axioms, to build the arithmetic in $\mathbb{F}$.

What the theorem does, is to show that the different axiomatic setup is of limited interest for dimensions > 2, but the theorem leaves open the possibility of other structures in dimension 2. We study these next.

[1] That is we do not require commutative multiplication, but otherwise the axioms for a field

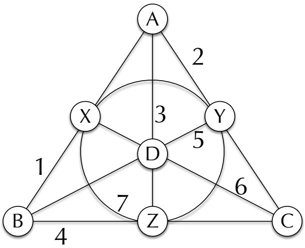


Figure VIII.3: The Fano Plane

Projective Planes

> It's a bird. It's a plane. It's Superman!
>
> The Adventures of Superman (1952-1958)

Motivated by the only exception in the Veblen-Young theorem, we now aim to describe *projective plane*, that is 2-dimensional projective spaces. In this case the axioms simplify. PG2 becomes PP2 below, which is nicely dual to the first axiom.

PP1 For any two distinct points, there is one, and only one, line containing both points

PP2 For any two distinct lines, there is one, and only one, point on both lines.

PP3 There exist four points, no three of which are on a line.

We observe that $\mathrm{PG}(2,q)$ indeed satisfies these axioms. (However it turn out that there are other projective planes.)

Example: The axioms alone let us construct the minimal configuration $\mathrm{PG}(2,2)$:

Let A, B, C, D be four points, no three on a line, that are given by PP3.

Let $L_1 = \overline{AB}$, $L_2 = \overline{AC}$, $L_3 = \overline{AD}$, $L_4 = \overline{BC}$, $L_5 = \overline{BD}$, $L_6 = \overline{CD}$. Because of the choice of the points, these lines must be distinct. By PP2, L_1 and L_6 must intersect in a point that cannot be any of the existing four, we call this point X. Similarly we get $L_2 \cap L_5 = \{Y\}$ and $L_3 \cap L_4 = \{Z\}$.

Finally, there must be a line $L_7 = \overline{XY}$, we may assume that $Z \in L_7$. We thus get the configuration in figure VIII.3, the *Fano plane* $\mathrm{PG}(2,2)$.

We notice, exercise **??**, that in the axioms we could replace PP3 by its dual:

PP3' There exist four lines, no three of which go through the same point.

This means that for a projective plane we can switch the labels of points and planes and again obtain a projective plane, (the *dual*).

Finite projective planes satisfy nice numerical properties:

Theorem VIII.8: Let $n \geq 2$ an integer. In a projective plane π any one of the following properties imply the other five:

1. One line contains exactly $n + 1$ points.
2. One point is on exactly $n + 1$ lines.
3. Every line contains exactly $n + 1$ points.
4. Every point is on exactly $n + 1$ lines.
5. There are exactly $n^2 + n + 1$ points in π.
6. There are exactly $n^2 + n + 1$ lines in π.

Proof:

We assume property 1. Let $L = \{Q_1, \dots Q_{n+1}\}$ be a line with $n + 1$ points and $P \notin L$ a further point. The lines $\overline{PQ_1}, \dots, \overline{PQ_{n+1}}$ then need to be different, for if $\overline{PQ_a} = \overline{PQ_b}$ then $P \in \overline{Q_aQ_b} = L$. By PP2 any line through P must intersect L and thus the $n + 1$ lines $\overline{PQ_i}$ are exactly the lines through P. That is, any point P not on L will lie on (exactly) $n + 1$ lines. By PP3 such a point P must exists, showing that property 1 implies property 2.

Assuming property 2, let P be a point on exactly $n + 1$ lines $K_1, \dots, K_{n+1}$. Then P is the only point common to any pair of these lines. If L any line with $P \notin L$, then the K_i will intersect it in $n + 1$ points Q_i, which must be distinct since P is the only point on any pair of the K_i. If there was another point $Q' \in L$, then $\overline{PQ'}$ would be another line through P, contradiction. That is, any line not containing P must have $n + 1$ points. By PP3 such lines exist, thus property 2 implies property 1.

Now assume 1 and 2 and let P and L as above. We have seen that any line not through P must have $n + 1$ points, and any point not on L lies on $n + 1$ lines. Let K be a line through P and $\{Q\} = K \cap L$. By PP3 there must be two further points in addition to P and Q, neither on K, of which at most one lies on L. Thus there is a point $R \notin L, K$. With the previous arguments we have that R lies on $n + 1$ lines and thus K has $n + 1$ points. Any point $S \notin K$ thus lies on $n + 1$ lines. To show that Q also lies on $n + 1$ lines, we chose another point $Q' \in L$ and $K' = \overline{PQ'}$ and repeat the argument. This shows 3 and 4.

Vice versa, obviously 3 implies 1 and 4 implies 2.

If 3 and 4 hold, pick a point P and sum up the number of (each $n + 1$) points on all $n + 1$ lines through P, yielding $(n + 1)^2 = n^2 + 2n + 1$ Any point except P is on exactly one of these lines, but P is on all lines. This means we are overcounting the number of points by $n + 1 - 1 = n$. Thus in total there are $n^2 + n + 1$ points, yielding 5. A dual argument gives 6.

We notice that $f(x) = x^2 + x + 1$ is one-to-one for $x > 0$, as then $f'(x) = 2x + 1 > 0$. This means, with the implications we know already, that 5 or 6 must imply 1 and 2 (otherwise a different count of points on one line or lines through one point would imply a different total number of points or lines). □

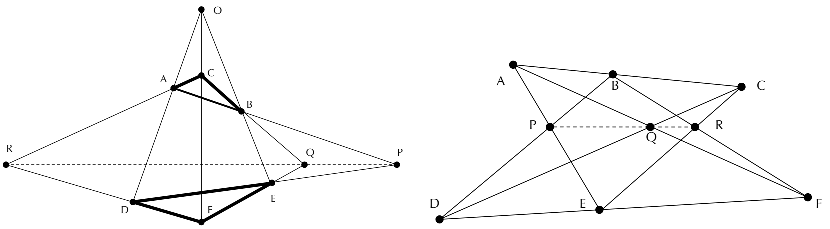


Figure VIII.4: Desargues' theorem and Pappus' theorem

For the planes PG$(2, q)$ we have of course that every line has $\begin{bmatrix}2\\1\end{bmatrix}_q = q+1$ points, motivating the following definition:

Definition VIII.9: A finite projective plane is *of order n* is a line contains $n+1$ points.

It turns out that the axioms for a projective plane alone do not characterize PG$(2, q)$, but that this can be done using two theorems from classical geometry:

Desargues' Theorem Let A, B, C and D, E, F are triangles, such that the lines $\overline{AD}$, $\overline{BE}$, and $\overline{CF}$ intersect in a common point O. Let $P = \overline{AB} \cap \overline{DE}$, $Q = \overline{BC} \cap \overline{EF}$, and $R = \overline{AC} \cap \overline{DF}$. Then P,Q and R are collinear. (Figure VIII.4, left.)

Pappus Theorem Let A,B,C be collinear and D,E,F be collinear and that the triangles ACE and BDF overlap. Then the intersection points $P = \overline{AE} \cap \overline{BD}$, $Q = \overline{AF} \cap \overline{CD}$, and $R = \overline{BF} \cap \overline{CE}$ are collinear. (Figure VIII.4, right.)

Both theorems can be proven using coordinates, these proofs carry over to PG$(2, q)$. It turns out that this in fact characterizes PG$(2, q)$.

Theorem VIII.10: Let Π be a projective plane. The following statements are equivalent:

a) $\Pi \cong$ PG$(2, F)$ (suitable map of points, lines) for a field F. In particular $\Pi \cong$ PG$(2, n)$ if it is finite of order n.

b) Desargues' Theorem holds in Π.

c) Pappus' Theorem holds in Π.

In this situation Π is called *desarguesian*.

We note that there are projective planes that are *not* desarguesian. Some of them are constructed similarly to PG$(2, q)$ but by replacing the field with a different structure. An easy infinite example will be given in section VIII.5.

VIII.5 A Non-Desarguesian Geometry

Since the constructions of finite non-desarguesian planes are somewhat involved, we only give an infinite example here. The description is not a priori a projective plane, but we will see in section VIII.8 how it could be embedded into a projective plane. What we are doing is to “bend” lines of the ordinary geometry locally to ensure that one point does not lie on a particular line:

The *Moulton Plane* is defined by the points $(x, y) \in \mathbb{Q} \times \mathbb{Q}$, together with three kind of lines:

- Vertical lines $x = a$
- Lines with nonnegative slope $y = mx + b$ for parameters (b, m) ,$m \geq 0$.
- Lines with negative slope. We treat these lines, my *doubling* the slope for positive x values, that is

$$y = \begin{cases} mx + b & \text{if } x \leq 0 \\ 2mx + b & \text{otherwise.} \end{cases}$$

Figure VIII.5, left shows some lines in the Moulton plane. Note that for all lines that are not vertical and not of slope 0, the x coordinate of any point on the line is uniquely determined by the y-coordinate.

We claim that any two points line on a unique line. (This will make the structure that of an affine plane (see VIII.8) and we shall see that such planes can be embedded into projective planes.)

Let (a, b) and (c, d) be two different points. If $a = c$, they lie on a (unique) vertical line $x = a$ but no other (since we must have that $b \neq d$ and for non-vertical lines different y-values imply different x-values).

Thus now assume WLOG that $a < c$. If $b < d$ the slope of a line must be positive, and thus there is only one.

If $b < d$ the slope is negative. If $a < c \leq 0$ or $0 < a < c$ we still deal with ordinary lines. If $a \leq 0 < c$ we can still set up linear equations to determine m and b of a line and get a unique solution, thus a unique line.

To show that Desargues’ theorem may fail, we simply take a construction in which (by placing most of it on coordinates with negative x) only one intersection point is affected by the bending lines. See Figure VIII.5, right: The lines $\overline{QR}$ and $\overline{DE}$ have nonnegative slope, thus are not affected. The line $\overline{AB}$ has negative slope and thus clearly has to bend, moving the intersection point P off from the common line with Q and R.

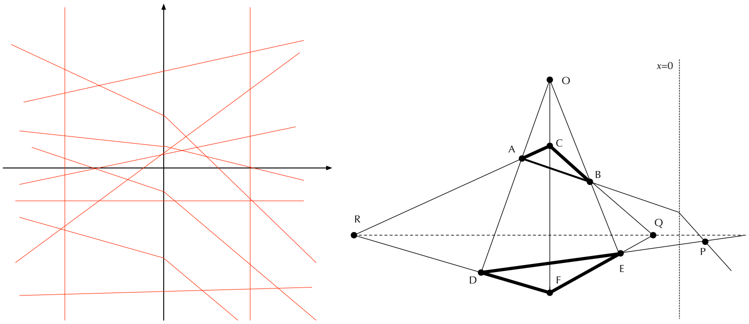


Figure VIII.5: The Moulton plane and failure of Desargues' theorem

VIII.6 Homogeneous coordinates

> Often the algebraic theorem is easy to prove, whereas the search for a "pure geometric" proof requires considerable ingenuity.
>
> Projective Geometry
> H.S.M.Coxeter

A convenient way of working in PG$(2, F)$ is to introduce coordinates, and to use algebraic operations on the coordinates to prove theorems. This is called analytic geometry.

A point, corresponding to the subspace $\langle(x, y, z)\rangle$, will be represented by the *homogeneous coordinate* vector $[x, y, z]$ (not all zero), with the understanding that $[x, y, z] = [cx, cy, cz]$ for $c \neq 0$.

Similarly a line of PG$(2, F)$ stems from a hyperplane $lx + my + nz = 0$. We represent this line by the coordinate vector $[l, m, n]$ (that is we need to specify whether a coordinate vector represents a point or a line) and note that the point $[x, y, z]$ and the line $[l, m, n]$ are incident if and only if the "inner product" $lx + my + nz = 0$.

Lemma VIII.11: a) Let $[x_1, y_1, z_1]$ and $[x_2, y_2, z_2]$ be points. The line through these points has the homogeneous coordinates

$$[l, m, n] = [y_1z_2 - z_1y_2,\ z_1x_2 - x_1z_2,\ x_1y_2 - y_1x_2].$$

b) Three points $[x_1, y_1, z_1]$, $[x_2, y_2, z_2]$, $[x_3, y_3, z_3]$ are collinear if and only if

$$\det\begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{pmatrix} = 0$$

c) The point of intersection of the two lines $[l_1, m_1, n_1]$ and $[l_2, m_2, b_2]$ has coordinates

$$[x, y, z] = [m_1 n_2 - n_1 m_2,\ n_1 l_2 - l_1 n_2,\ l_1 m_2 - m_1 l_2].$$

Proof: Exercise **??** □

Corollary VIII.12: If three points $[x_1, y_1, z_1]$, $[x_2, y_2, z_2]$, $[x_3, y_3, z_3]$ are collinear, we can scale coordinates of one point, so that

$$[x_1, y_1, z_1] + [x_2, y_2, z_2] + [x_3, y_3, z_3] = [0, 0, 0]$$

Using homogeneous coordinates, we can prove geometric theorems by algebraic means:

Theorem VIII.13: Desargues' theorem holds in $\mathrm{PG}(2, F)$.

Proof: We shall use letters $A = [x_A, y_A, z_A]$ to represent the homogeneous coordinates of points. We can represent the lines through O by the equations (possibly rescaling coordinates of D, E, F)

$$\begin{aligned} A + D + O &= 0 \\ B + E + O &= 0 \\ C + F + O &= 0 \end{aligned}$$

which gives that $A + D = B + E - C + F$.

This means that we can choose $P = A - B = E = D$ as coordinates of the intersection of $\overline{AB} \cap \overline{DE}$ (the point with these coordinates is collinear with A, B and with D, E).

Similarly we set $Q = B - C = F - E$ and $R = C - A = D - F$. But then

$$P + Q + R = (A - B) + (B - C) + (C - A) = 0,$$

showing that the three points are collinear. □

VIII.7 The Bruck-Ryser Theorem

An obvious question is for which orders projective planes exist. We know already that for any prime power q we have that $\mathrm{PG}(2, q)$ is a plane of order q.

The main theorem about non-existence is the following

Theorem VIII.14 (Bruck-Ryser): If a projective plane of order n exists, where $n \equiv 1 \text{ or } 2 \pmod 4$, then n is the sum of the squares of two integers.

This theorem (proven below) for example shows that there cannot be a projective plane of order 6. It does not exclude the next interesting degree if $10 = 1^2+3^2 \equiv 2 \pmod 4$, though – the result of a huge computer search [LTS89] – no such plane exists. In fact, as of this writing all projective planes known to exist are of prime-power order.

Proof:(Theorem VIII.14) Suppose there is a projective plane of order n with $n \equiv 1 \text{ or } 2 \pmod 4$. Then the plane has $N = n^2 + n + 1 \equiv 3 \pmod 4$ points and lines. Let $A = (a_{ij}) \in \{0,1\}^{N\times N}$ be the *incidence matrix* of the plane, that is $a_{i,j} = 1$ if and only if the i-th point is on the j-th line.

Then the i,j entry $AA^T = (c_{i,j})$ is $c_{i,j} = \sum_k a_{ik}a_{jk}$ with the product $a_{ik}a_{jk}$ being 1 if and only if the k-th line contains points i and j. Thus $c_{i,j}$ is the number of lines containing the points i and j, which is $n+1$ if $i = j$ and 1 otherwise. Thus $AA^T = nI + J$ where J is the all-one matrix.

We introduce N variables, $x_1, \ldots, x_N$ and let $x = (x_1, \ldots, x_N)$. Then $z = (z_1, \ldots, z_n) := xA$ satisfies

$$z_1^2 + \cdots + z_N^2 = zz^T = xAA^Tx^T = n \cdot xx^T + xJx^T = n(x_1^2 + \cdots + x_N^2) + \omega^2$$

with $\omega = x_1 + \cdots + z_N$. We add nx_{N+1}^2 (for a new variable x_{N+1}) to both sides of the equation, and observe that the number of variables, $N+1$, is a multiple of 4.

Lagrange's four squares theorem from number theory states that we can write $n = b_1^2 + b_2^2 + b_3^2 + b_4^2$ as a sum of four squares. We thus collect the terms nx_i^2 in groups of four and use that

$$\begin{array}{rcl}
n(x_1^2 + \cdots x_4^2) & = & (b_1^2 + \cdots + b_4^2)(x_1^2 + \cdots x_4^2) \\
 & = & y_1^2 + y_2^2 + y_3^2 + y_4^2 \qquad \text{with} \\
y_1 & = & b_1x_1 - b_2x_2 - b_3x_3 - b_4x_4 \\
y_2 & = & b_1x_2 + b_2x_1 + b_3x_4 - b_4x_3 \\
y_3 & = & b_1x_2 + b_3x_1 + b_4x_2 - b_2x_4 \\
y_3 & = & b_1x_4 + b_4x_1 + b_2x_3 - b_3x_2
\end{array}$$

with the y_i's being linear combinations of the x_i's. (This identity, which can be verified by straightforward calculation is the multiplicativity of the norm in the quaternions.) Collecting all terms together gives us

$$z_1^2 + \cdots + z_N^2 + nx_{N+1}^2 = y_1^2 + \cdots + y_{N+1}^2 + \omega^2. \qquad \text{(VIII.15)}$$

We now utilize some freedom in the choice of the variables x_i to simplify this expression:

Recall that we have $z_j = \sum_i a_{i,j}x_i$ (from the incidence matrix) and $y_k = \sum_i b_{i,k}x_i$ (from expanding the four-square sums). Choose indices j, k such that $a_{1,j}$ and $b_{1,k}$ are both nonzero. (Such indices must exist for reasons of matrix rank.) Choose a sign such that $a_{1,j} \pm b_{1,k}$ is nonzero and substitute

$$x_1 = -\frac{(a_{2,j} \pm b_{2,k})x_2 + \cdots + (a_{N,j} \pm b_{N,j})x_N + a_{N+1,j}x_{N+1}}{a_{1,j} \pm b_{1,k}}$$

Then $y_k \pm z_j = 0$ and thus $y_k^2 = z_j^2$, and both terms can be canceled from equation (VIII.15), resulting in an equation in fewer variables that are all linear expressions in $x_2, \ldots, x_{N+1}$.

We can repeat this argument, leaving only one y variable and x_{N+1}, yielding

$$nx_{N+1}^2 = y_k^2 + \omega^2.$$

where y_k and ω are rational multiples of x_{N+1}. Setting x_{n+1} to a suitable integer thus gives an integer equation

$$nx^2 = a^2 + b^2,$$

with a, b, c not all zero. Again a theorem from number theory shows that this is only possible if n is a sum of squares. □

VIII.8 Affine Planes

To build a connection with another areas of combinatorics we first step back to the constructions that motivated PG as an extension of the ordinary points and lines of classical geometry.

Definition VIII.16: A *affine plane* of order q (for $q \geq 2$) is a set X of q^2 points and a set $\mathcal{B}$ of q-element subsets of X, called *lines*, such that any two points lie on a unique line.

The definition implies that two lines may share at most one point. We call lines *parallel*, if they are equal or disjoint.

Proposition VIII.17: Let $(X, \mathcal{B})$ be an affine plane of order q. Then

1. Every point lies on exactly $q + 1$ lines.
2. There are $q(q + 1)$ lines in total.
3. (Parallel postulate) If $p \in X$ and $L \in \mathcal{B}$ there exists a unique $p \in L'$ parallel to L.
4. Parallelity is an equivalence relation on lines.
5. Each class of parallel lines contains q lines, partitioning X.

Proof: For a point p, the $q^2 - 1$ other points distinct from p each lie on a unique line through p that contains $q - 1$ points different from p. So there are $(q^2 - 1)/(q - 1) = q + 1$ lines through p.

If we double-count pairs (p, L) with $p \in L$ we thus get

$$q^2(q+1) = l \cdot q$$

with l the number of lines, yielding $l = q(q+1)$.

If $p \in L$ then L is the unique line parallel to L containing p must be L (as they would share a point). If $p \notin L$, then p lies on $q+1$ lines of which q are needed to connect p to points on L, leaving one unique line L' parallel to L.

By definition, parallelity is reflexive and symmetric. To show transitivity, assume that L, M both are parallel to N. Without loss of generality we may assume that these three lines are different. But if L is not parallel to M, they must share a point p, yielding two lines through p parallel to N, contradiction.

As a partition into cells of equal size, each parallel class must contain $q^2/q = q$ lines. □

By the parallel postulate, the $q+1$ lines through a point must contain representatives for all parallel classes, implying that there are $q+1$ classes.

We now describe two constructions that construct a projective plane from an affine plane by adding a line "at infinity", respectively construct an affine plane from a projective plane by removing a line.

Let $(X, \mathcal{B})$ be an affine plane. Let Y be the set of parallel classes. We add a new point for every parallel class, letting $A = X \cup Y$ be the new point set with q^2+q+1 points. For lines, we add to every line $L \in \mathcal{B}$ the point representing its parallel class. We also add Y as a new line "at infinity". Thus the structure we get is $(A, \mathcal{C})$ with

$$\mathcal{C} = \{L \cup \{y\} \mid L \in \mathcal{B}, L \in y \in Y\} \cup \{Y\}.$$

We note that any new line has $q+1$ points, as we added one point to each new line as well as a new line at infinity with $q+1$ points. To verify the axioms we note that the only new points and new lines are at infinity. That is for two points in X there is a unique line in $\mathcal{B}$ through these points, and this line extends in $\mathcal{C}$. Neither point is on the line at infinity.

For a point $x \in X$ and a point y at infinity, take the line $L \in \mathcal{B}$ in the class y that goes through x, then $L \cup \{y\} \in \mathcal{C}$. x is not on the new line.

For two points at infinity the common line is Y. Thus PP1 holds.

For PP2, we note that two lines on $\mathcal{C}$ that arise from L will intersect in their common point in X if they are not parallel, if they are parallel they intersect in their parallel class y. The line at infinity intersects the other lines each once in the appropriate parallel class.

PP3 is satisfied as we have $q \geq 2$. Thus $(A, \mathcal{C})$ is a projective plane.

Vice versa, we take a projective plane and designate an (arbitrary) line L as being at infinity. Removing this line and its $q+1$ points has q^2 points remaining and any line has q points (one got removed with L). By Axiom PP1 any two points lie on a unique line. Thus the resulting structure is an affine plane.

These two constructions give the obvious consequence

Corollary VIII.18: A projective plane of order q exists, if and only if an affine plane of order q exists.

VIII.9 Orthogonal Latin Squares

> *[...] the sole piece not yet filled in has the almost perfect shape of an X. But the ironical thing, which could have been foreseen long ago, is that the piece the dead man holds between his fingers is shaped like a W.*
>
> Life, a User's Manual

> [...] la seule pièce non encore posée dessine la silhouette presque parfaite d'un X. Mais la pièce que le mort tient entre ses doigts a la forme, depuis longtemps prévisible dans son ironie même, d'un W.
>
> La vie, mode d'emploi
> G. Perec

We now introduce another combinatorial structure:

Definition VIII.19: a) A *Latin Square* of order n is a matrix in $\{1,\ldots,n\}^{n\times n}$, such that every entry occurs once (and thus exactly once) in every row and every column.

b) Two Latin squares $A = (a_{ij})$ and $B = (b_{ij})$ are *orthogonal*, if for any pair of numbers $k,l \in \{1,\ldots n\}$ there are unique indices i,j such that $a_{ij} = k$ and $b_{ij} = l$.

c) A set $\{A_1,\ldots A_l\}$ of Latin squares is called a set of *mutually orthogonal Latin squares* (MOLS) if A_i is orthogonal to A_j for $i \neq j$.

Orthogonal Latin squares used to be written in the same matrix with an ordinary (Latin) letter denoting the one square and a Greek letter designating the other, thus they are sometimes called *Graeco-Latin squares*.

Orthogonal Latin squares have for example application in experimental design – averaging out correlations among pairs of parameters without forcing to try out the total of all n^3 (or more) combinations.

Imagine we'd be testing n possible fertilizers on n kinds of plants and we have n plots (of potentially varying quality for growth). Two orthogonal Latin squares describe a sequence of growing every plant on every plot (square one: columns are the plots, rows are subsequent time units, entries are plant numbers) and using every fertilizer on every plot (square two: entries are fertilizer numbers) with n^2 plantings in total.

An obvious question is on how many squares could be in a set of MOLS. An upper bound is given by $n-1$:

Lemma VIII.20: a) If A, B are orthogonal Latin squares of order n, and we apply any permutation $\pi \in S_n$ to the *entries* of B, the resulting Latin squares are orthogonal as well.

b) The maximal cardinality of a set of MOLS is $n-1$.

Proof: a) is clear as there is a bijection between pairs (k,l) and (k,l^{π}).

For the proof of b) we assume by a) WLOG that all squares have entry 1 in position $(1,1)$. Each of the squares then will have to have an entry of 1 in row two in columns $2,\ldots,n$. By orthogonality no two squares may have this entry 1 in the same column, thus limiting their total number to $n-1$. □

It now turns out (as it will be in other situations later in the course) that a maximal set of MOLS corresponds to an affine plane (or a projective plane).

Proposition VIII.21: An affine plane of order n defines a set of $n-1$ MOLS of order n.

Proof: Given an affine plane, choose two classes $\{H_1,\ldots,H_n\}$ and $\{V_1,\ldots V_n\}$ of parallel lines. As every point p lies on a unique line H_i and V_j, we can identify it with a pair i,j.

For each further class $\{L_1,\ldots,L_n\}$ of parallel lines we define a matrix $A_L = (a_{ij})$ by setting $a_{i,j}=k$, iff the point p corresponding to coordinates i,j is on line L_k.

Then A_L is a Latin square – would we have that $a_{i,j}=k=a_{i,m}$ the point p corresponding to (i,j), and the point q corresponding to (i,m) would be on line L_k (as having the same value) and on line H_i (as being in the same row), contradicting the fact that there is a unique line through p and q. An analog argument holds for same column index.

Similarly, if for another class of lines $\{M_1,\ldots,M_n\}$ and $A_M=(b_{ij})$ we had that A_M was not orthogonal to A_L, there would be a pair of values (k,t) and two positions (i,j) and (r,s), corresponding to points p and q, such that $a_{i,j}=k=a_{r,s}$ and $b_{i,j}=t=b_{r,s}$. But then $p,q\in L_k$, as well as $p,q\in M_t$, again contradicting uniqueness of lines. □

For the converse result we take a set $\{A_1,\ldots A_r\}$ of Latin squares of order n. We consider the set $X=\{1,\ldots,n\}\times\{1,\ldots,n\}$ of coordinates as points and define three classes of lines:

Horizontal: $\{(i,x)\mid x=1,\ldots,n\}$ for fixed i,

Vertical: $\{(x,j)\mid x=1,\ldots,n\}$ for fixed j,

Latin: For each Latin square $A_s=(a_{i,j})$ and value k the lines $\{(i,j)\mid a_{i,j}=k\}$.

This yields n^2 points and $n(r+2)$ lines, each containing n points, and every point lies on $r+2$ lines.

Lemma VIII.22: In this construction any two points lie on at most one line

We note that a structure with these properties is a geometric structure called a *net*.

Proof: Suppose that two points, (i, j) and (s, t), are on two common lines. By definition of the lines they must be of different types.

If one is horizontal and the other vertical we get that $i = s$ and $j = t$, contradiction.

If one line is horizontal and the other Latin, we have that $s = i$ and thus that the Latin square had two equal entries in one row, contradiction.

The argument for vertical and Latin is analog. □

If with $r = n - 1$ the set of MOLS is of maximal size we get $n(n+1)$ lines, each containing n points. As no pair of points lies on two lines, these connect

$$n(n+1)\binom{n}{2} = n(n+1)\frac{n(n-1)}{2} = \frac{n^4 - n^2}{2} = \binom{n^2}{2}.$$

pairs of points. As there are in total only $\binom{n^2}{2}$ pairs, this shows that in this case the net is an affine plane.

Corollary VIII.23: A set of $n-1$ MOLS of order n exists if and only if there exists an affine plane of order n if and only if there exists a projective plane of order n.

Existence of Orthogonal Latin Squares

> *Six different regiments have six officers, each one belonging to different ranks. Can these 36 officers be arranged in a square formation so that each row and column contains one officer of each rank and one of each regiment?*

> Cette question rouloit sur une assemblée de 36 Officiers de six Rêgimens différens, qu'il s'agissoit de ranger dans un quarré, de manière que sur chaque ligne tant horizontale que verticale ils se trouva six Officiers tant de différens caractères que de Régimens différens.
>
> Recherches sur une nouvelle espè ce de quarrés magiques, 1782
> L. Euler

Stepping back from the question of the size of maximal sets of MOLS, one can ask the question for which orders n at least a pair of MOLS exists. The boundaries we've already proven show that this is not possible for $n = 2$ (as $n - 1 = 1$).

With this question we go back to Leonard Euler and the origins of combinatorics. Euler (who introduced the concept of a Graeco-Latin square) constructed examples for any order $n \not\equiv 2 \pmod 4$ and claimed that this would be impossible for any order $n \equiv 2 \pmod 4$, but was not able to prove this impossibility. A proof of the impossibility of order 6 was given in 1900 by G. Tarry, based on case distinctions.

With this pedigree, it came as an enormous surprise[2] that Parker, Bose[3] and Shrikande in 1959 found an example of a pair of orthogonal Latin squares of order 10. They are depicted in the following matrix with the first square given by the first digit, and the second by the second digit of each entry.

$$\begin{pmatrix}
00 & 47 & 18 & 76 & 29 & 93 & 85 & 34 & 61 & 52\\
86 & 11 & 57 & 28 & 70 & 39 & 94 & 45 & 02 & 63\\
95 & 80 & 22 & 67 & 38 & 71 & 49 & 56 & 13 & 04\\
59 & 96 & 81 & 33 & 07 & 48 & 72 & 60 & 24 & 15\\
73 & 69 & 90 & 82 & 44 & 17 & 58 & 01 & 35 & 26\\
68 & 74 & 09 & 91 & 83 & 55 & 27 & 12 & 46 & 30\\
37 & 08 & 75 & 19 & 92 & 84 & 66 & 23 & 50 & 41\\
14 & 25 & 36 & 40 & 51 & 62 & 03 & 77 & 88 & 99\\
21 & 32 & 43 & 54 & 65 & 06 & 10 & 89 & 97 & 78\\
42 & 53 & 64 & 05 & 16 & 20 & 31 & 98 & 79 & 87
\end{pmatrix}$$

In fact it is now known that for any $n \neq 2, 6$ there is a pair of orthogonal Latin squares of order n. We shall give a construction that covers the same orders as Euler did:

We first consider a “direct product” construction for Latin squares:

Definition VIII.24: Let A, B be Latin squares of order m, n respectively. We define the direct product by identifying the pairs in $\{1, \ldots m\} \times \{1, \ldots n\}$ with numbers from 1 to mn by mapping (i, j) to $i + m \cdot (j - 1)$. Then

$$A \times B = \left(c_{x,y}\right) \text{ with } c_{(i,j),(r,s)} = \left(a_{i,r}, b_{j,s}\right).$$

Lemma VIII.25: If A and B are Latin squares, so is $A \times B$.

Proof: It is sufficient to show that every value occurs in every row and every column. Let (x, y) be a value and (i, j) the index of a row. Then $a_{i,r} = x$ and $b_{j,s} = y$ for suitable values (r, s) and thus $A \times B$ has the entry (x, y) in column (r, s) of row (i, j). □

Lemma VIII.26: If A, B are orthogonal Latin squares of order m and C, D are orthogonal of order n, then $A \times C$ and $B \times D$ are orthogonal.

Proof: We use the bijection between numbers and pairs. Suppose that $L = A \times C$ and $M = B \times C$ share the same combination of entries in two positions. That is, we have positions $(i, j), (k, l)$ and $(p, q), (r, s)$ where both L and M have equal entries. That is

$$\begin{aligned}
\left(a_{i,k}, c_{j,l}\right) = L_{(i,j),(k,l)} &= L_{(p,q),(r,s)} = \left(a_{p,r}, c_{q,s}\right)\\
\left(b_{i,k}, d_{j,l}\right) = M_{(i,j),(k,l)} &= M_{(p,q),(r,s)} = \left(b_{p,r}, d_{q,s}\right)
\end{aligned}$$

[2] yielding a front page story in the New York Times, only pushed below the fold by the opening of the St. Lawrence Seaway

[3] at that time at U.NC Chapel Hill but later at CSU!

This implies that $a_{i,k} = a_{p,r}$ and $b_{i,k} = b_{p,r}$, which because of the orthogonality of A and B implies that $(i, j) = (p, q)$. We similarly prove that $(j, l) = (q, s)$, showing that the positions must be equal. □

Corollary VIII.27 (MacNeish's Theorem): Let $n = \prod_i p_i^{e_i}$ a product of prime powers with the p_i being distinct primes. Then there are at least $\min_i \left(p_i^{e_i} - 1\right)$ MOLS of order n.

Proof: The existence of $\mathrm{PG}(2, p_i^{e_i})$ gives us $p_i^{e_i} - 1$ MOLS of order $p_i^{e_i}$. Form direct products of such squares over the different primes. □

Chapter

IX

Designs

> I want your horror,
> I want your design
>
> Bad Romance
> Lady Gaga

In a geometry we had that lines were subsets of points that were defined uniquely defined by the choice of two points. We now generalize this concept:

Definition IX.1: Let X be a set of points with $|X| = v$ and $\mathcal{B}$ a set of k-element subsets of X and $t < k < v$. Then $(X, \mathcal{B})$ is called a $t - (v, k, \lambda)$ *design* (or t-design with parameters (v, k, λ)) for $\lambda > 0$, if any t points are contained in exactly λ blocks.

In the special case of $\lambda = 1$, a $t - (v, k, 1)$ design is also often called a *Steiner system* $S(t, k, v)$.

Example:

- The points and lines of a projective plane of order q are an $2 - (q^2 + q + 1, q + 1, 1)$-design.
- Let X be a set of size v and $\mathcal{B}$ the set of all k-element subsets of X for $k < v$. Then for $t < k$, there are $\binom{v-t}{k-t}$ possibilities to extend a t-set to a k-set, so $(X, \mathcal{B})$ is an $t - (v, k, \binom{v-t}{k-t})$ design.
- A partition of v points into blocks of size k is an $1 - (v, k, 1)$-design.
- Let X be the vertices (labeled $1, \ldots, 7$) of a regular heptagon and consider the triangle $1, 3, 4$ and its images under rotation. Its vertices have distances 1, 2 and 3 respectively and any edge of the complete graph on these vertices

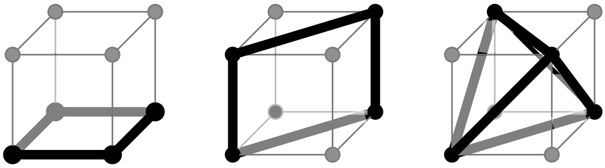


Figure IX.1: A $3-(8,4,1)$ design

lies on exactly one rotation image of this triangle. Thus we get an $2-(7,3,1)$ design. If we also allow reflection images of the triangle we similarly get an $2-(7,3,2)$ design.

- Let X be the 8 corners of a cube. Then three points are either

  1. On a common face, Figure IX.1, a).
  2. Two corners form an edge, but the third is not on a common face. Then the three corners are on a diagonal slice through the common edge, Figure IX.1, b).
  3. No two vertices share a common edge. They then lie on the corners of a tetrahedron, Figure IX.1, c).

  In all three situations three points define uniquely a fourth point. Thus if we make $\mathcal{B}$ the set of all of these 4-corner configurations, we get that $(X,\mathcal{B}$ is a $3-(8,4,1)$ design.

The regularity of a design implies that the parameters determine the number of blocks:

Proposition IX.2: The number of blocks of an $t-(v,k,\lambda)$ design is

$$b=\lambda\binom{v}{t}/\binom{k}{t}.$$

Proof: Double-count the pairs (T,B) with $T\subset B$ and $|T|=t$ and B a block. There are $\binom{v}{t}$ sets T, each in λ blocks, respectively b blocks each with $\binom{k}{t}$ subsets. □

If we have a design, we can remove points and get a smaller design:

Proposition IX.3: Let $(X,\mathcal{B})$ be a $t-(v,k,\lambda)$ design and $S\subset X$ with $|S|=s<t$. We let $Y=X\setminus S$ and

$$\mathcal{C}=\{B\setminus S\mid S\subset B\in\mathcal{B}\}.$$

Then $(Y,\mathcal{C})$ is a $(t-s)-(v-s,k-s,\lambda)$ design, called the *derived design* of $(X,\mathcal{B})$ with respect to S.

Proof: Each block in $\mathcal{C}$ contains $k-s$ points of the $v-s$ points of Y. Let $Z \subset Y$ with $|Z| = t-s$. Then $Y \cup S$ is a t-subset of X and thus lies in λ blocks (that automatically contain S), thus Z lies in λ blocks in $\mathcal{C}$. □

Example: The example above gave a $3-(8,4,1)$ design. Removing one point gives an $2-(7,3,1)$ design.

We note that the number of blocks of this $(t-s)-(v-s,k-s,\lambda)$ design is

$$\lambda_s = \lambda \binom{v-s}{t-s} / \binom{k-s}{t-s}$$

and thus only dependent on s. That means that any s-point subset of X lies in λ_s blocks, that is

Corollary IX.4: Any $t-(v,k,\lambda)$ design is also a $s-(v,k,\lambda_s)$ design for any $s \le t$.

Example: For the $3-(8,4,1)$ design from above and $s=2$ we get

$$\lambda_s = 1\binom{6}{1} / \binom{2}{1} = 3$$

and thus an $2-(8,4,3)$ design.

For the case $s=1$ we are getting similarly an $1-(8,4,7)$ design.

Corollary IX.5: Every point of X lies in

$$r = \lambda_1 = \lambda \binom{v-1}{t-1} / \binom{k-1}{t-1}$$

blocks.

Another operation to construct designs from designs is the *complement design*, that is we replace $\mathcal{B}$ with $\overline{\mathcal{B}} = \{X \setminus B \mid B \in \mathcal{B}\}$.

Proposition IX.6: The complement of a $t-(v,k,\lambda)$ design is a $t-(v,v-k,\bar{\lambda})$ design with

$$\bar{\lambda} = \sum_{s=0}^{t} (-1)^s \binom{t}{s} \lambda_s$$

Proof: The number of blocks in $\overline{\mathcal{B}}$ containing the points $x_1,\ldots,x_t \in X$ is equal to the number of blocks in $\mathcal{B}$ containing none of the x_i. For $I \subset \{1,\ldots,n\}$ let $\mathcal{B}_I$ be the set of blocks containing x_i for all $i \in I$. Then $|\mathcal{B}_I| = \lambda_{|I|}$. By PIE the number of blocks containing none of the x_i thus is

$$\sum_{I \subset \{1,\ldots n\}, |I| \le t} (-1)^{|I|} |\mathcal{B}_I| = \sum_{s=0}^{t} (-1)^s \binom{t}{s} \lambda_s.$$

□

Example: The complement of a $2-(7,3,1)$ design is a $2-(7,7-3=4,2=7-2\cdot 3+1)$ design.

IX.1 2-Designs

Definition IX.7: We fix an indexing of the v points X and the b blocks $\mathcal{B}$ of a design. The *incidence matrix* of the design is $M \in \{0,1\}^{v\times b}$ defined by $m_{ij} = 1$ iff $x_i \in B_j$.

Transposition of the incidence matrix defines a dual structure with the containment relation inverted. This begs the question, whether this structure also could be a design.

To answer this in part, we now consider the case of a 2-design, that is $t = 2$ and every point lies in $r = \lambda\frac{v-1}{k-1}$ blocks

Lemma IX.8: The incidence matrix M of a 2-design satisfies that

$$MM^T = (r-\lambda)I + \lambda J$$

where I is the $v \times v$ identity matrix and J the all-1 matrix.

Proof: The i,k-entry of the product is

$$(MM^T)_{i,k} = \sum_j m_{ij}m_{k,j},$$

that is the number of blocks that contain x_i and x_k. This count is r if $i = j$ and λ if $i \neq j$. □

Lemma IX.9: $\det(MM^T) = rk(r-\lambda)^{v-1}$

Proof: A general linear algebra result, Exercise **??**, shows that $\det(xI + yJ) = (x + ny)x^{n-1}$. Applying this for $n = v$ and $x = r - \lambda$ and $y = \lambda$ gives the determinant as

$$(r-\lambda+v\lambda)(r-\lambda)^{v-1} = (r+(v-1)\lambda)(r-\lambda)^{v-1}.$$

Using that $r + (v-1)\lambda = rk$ then gives the result. □

Theorem IX.10 (Fischer's inequality): A $2-(v,k,\lambda)$ design with $k < v$ has at least as many blocks as points, i.e. $b \geq v$.

Proof: As r is the total number of blocks containing a particular point, and λ the number of blocks containing this point and another, we have that $r = \lambda$ implies that all points must be in one block, contradicting $k < v$. Therefore $r - \lambda \neq 0$. Thus MM^T has full rank v. This is impossible if $b < v$. □

In the case of equality in Fischer's inequality is of particular interest:

Theorem IX.11: Let $(X,\mathcal{B})$ be a $2-(v,k,\lambda)$ design with $k < v$. The following properties are equivalent:

a) $b = v$,

b) $r = k$,

c) Any two blocks meet in λ points

d) Any two blocks meet in a constant number of points

Note that condition d) states, that if we exchange points and blocks, maintaining incidence (i.e. transpose the incidence matrix), the resulting structure is again a design.

Proof: For a 2-design we have by Proposition IX.2 that

$$kb = k\lambda\binom{v}{2}/\binom{k}{2} = \lambda\frac{v(v-1)}{k-1} = vr,$$

thus a) and b) are equivalent.

To see that b) implies c), assume that $r = k$ and thus $b = v$. This means that M is square.

With J being the all-1 matrix, the properties of a design give us that $JM = kJ$ (row sums are the block size), as well as $MJ = rJ$ (column sums are the number of blocks in which one point lies). As $k = r$ this implies that $MJ = JM$. Therefore M will commute with $(r-\lambda)I + \lambda J$ and thus with $((r-\lambda)I + \lambda J)M^{-1} = M^T$.

We thus know that $M^TM = MM^T = (r-\lambda)I + \lambda J$ which implies that any two blocks have λ points in common.

c) trivially implies d)

If d) holds, it means that the dual structure (reversing the roles of points and blocks, as well as the incidence relation) $(\mathcal{B}, X)$ *is* a 2-design. It has b points and v blocks, and thus, by Fischer's inequality, we have that $v \geq b$. But Fischer on the original design implies $b \geq v$, thus $v = b$ and a) is implied. □

Definition IX.12: A 2-design satisfying the conditions of Theorem IX.11 is called a *square design* or a *symmetric design*.

A square design with $\lambda = 1$ is a projective plane.

Note IX.13: As $r = k$ the parameters (v, k, λ) of a square 2-design satisfy $k(k-1) = (v-1)\lambda$.

Note IX.14: The complement of a square 2-design is a square 2-design.

Some restriction (necessary, not sufficient) for the existence of square designs is given by the following theorem:

Theorem IX.15 (Bruck-Ryser-Chowla): Suppose there exists a square $2-(v, k, \lambda)$ design.

a) If v is even, then $k - \lambda$ must be a square.

b) If v is odd, the Diophantine equation

$$z^2 = (k-\lambda)x^2 + (-1)^{(v-1)/2}\lambda y^2$$

has a solution in integers x, y, z not all zero.

Proof: We only prove a), the proof of b) is similar to theorem VIII.14 and can be found, for example in [vLW01]:
By Lemma IX.9, and because $r = k$ we have that

$$\det(M)^2 = \det(MM^T) = rk(r-\lambda)^{v-1} = k^2(k-\lambda)^{v-1}$$

As $\det(M)$ is an integer, the right hand side must be a square. For even v this can only be if $k - \lambda$ is square. □

Note IX.16: The parameter set $2-(111, 11, 1)$ satisfies the conditions of the theorem, but no such design exists.

Note IX.17: A projective plane of order n is a $2-(n^2+n+1, n+1, 1)$ design. Thus

$$r = 1 \cdot \frac{n^2+n}{n} = n+1 = k$$

and the design is square. Note that $v = n^2+n+1$ is odd. If $n \cong 1, 2 \pmod 4$, then $n^2+n+1-1 = n^2+n \cong 1 \pmod 4$, that is $\frac{n^2+n}{2}$ is odd. Thus the Diophantine equation becomes

$$z^2 = (n+1-1)x^2 - y^2$$

which was the equation in the proof of Theorem VIII.14 that gave rise to the fact that n must be a sum of two squares.

A theorem (first proven by Petrenjuk for $t = 4$) by Ray-Chaudhuri and Wilson generalizes Fischer's inequality:

Theorem IX.18: If $t = 2s$ and $k \leq v - s$, a $t-(v, k, \lambda)$ design satisfies that $b \geq \binom{v}{s}$.

Chapter

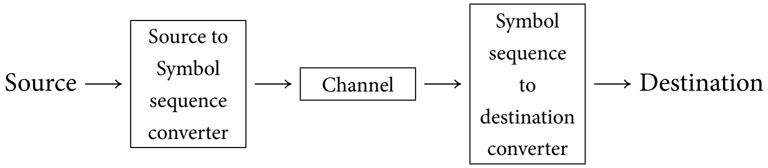
X

Error-Correcting Codes

Codes in the combinatorial sense deal with the *correctness* of transmitted data, not with preventing unauthorized reading:

When transmitting data over a channel (the generic term for any mode of data transfer – a letter, a cable, a wireless transfer, reading from a physical storage medium), it will typically by using symbols from some *alphabet*, be it letters, 0/1, or numbers. choice of alphabet and the translation between message and symbol sequence will depend on the application. For example, music on a CD is encoded by sampling the signal at a fixed rate (44.1kHZ) by 16 bit numbers (this is called Pulse Code Modulation, PCM) and storing these numbers on the CD. A player then reads these numbers and reconstructs an electric signal whose amplitude corresponds to the numbers.

We can describe this by the following diagram:


Source
Source to Symbol sequence converter
Channel
Symbol sequence to destination converter
Destination

Note X.1: We will **not** be concerned with any issues or signal degradations in translating between signal and symbol sequence. Nor are we concerned with data modifications (such as psychoacoustic encoding as utilized for example by MP3 files, or with JPG compression) that produce a signal that is *equally good* or *unperceivably* different from the original. We simply send a sequence of symbols over a channel and need to receive the *same* symbol sequence back on the other side, that is we are interested in the middle part of the diagram

$$\text{Symbol Sequence} \longrightarrow \boxed{\text{Channel}} \longrightarrow \text{Symbol Sequence.}$$

What can (in fact is almost guaranteed) to happen is that errors in the transmission process (transcription errors, noise, dirt, static) affect the data so that the received data differs from the transmitted one. This cannot be plausibly be resolved by better engineering: When the CD format was defined the available manufacturing technology made it impossible to expect that it would be possible to read back the data without a significant number of errors.

Note X.2: The typical form of errors are *bursts* that erase or modify a sequence of subsequent symbols: A smudge of dirt or water on paper, a burst of electromagnetic waves from starting a motor, a manufacturing imperfection in the plastic of a CD.

By *interleaving* data, that is storing a sequence ABCDEFGHI in the form ADGBEHCFI, we can spread out the effect of burst errors to become rather that of individual random symbol errors that occur equidistributed at a given rate for each symbol.

We also will not deal with *channel coding*, the way how symbols are translated to the transmission channel. (An example of this would be Morse code.)

The standard way to combat errors is *redundancy*: The word "houxe" has really only one possible correction in the English language, if a message becomes too garbled we would even ask for it to be repeated in full.

That is we do not transmit the actual information (say the information that the recipient should carry an umbrella) but we *encode* it in a sequence of symbols ("It will be raining") which includes redundant data stored. On the receiver side in a *decode* process the information is restored with small garbles fixed (*error-correction*) or at least that it can be discovered that an error happened (*error-detection*).

$$\text{Message} \longrightarrow \boxed{\text{Encode}} \longrightarrow \boxed{\text{Channel}} \longrightarrow \boxed{\text{Decode}} \longrightarrow \text{Message.}$$

The redundancy of everyday processes however comes at a high cost: Languages typically use three (or more) symbols in average for each symbol transmitted (as one can see by selectively erasing some symbols from a text), repeating information for redundancy further reduces the capacity of the channel.

Another example are check digits or parity conditions (e.g. every CSUID is $\equiv 1 \pmod 7$) that can detect errors.

Our goal is to do reduce the amount of redundancy needed for recovering from a particular error rate.

X.1 Codes

To help with analysis and to avoid practical issues with intermediate storage, we shall assume that the message to be transferred will be cut into units of equal size (possibly by padding the last piece), and each of these pieces will be separately encoded and decoded.

Definition X.3: Given an alphabet Q with $|Q| = q$ and $n > 0$, the *Hamming Space* $H(n,q) = Q^n$ is the set of all n-tuples (called *words*) with entries chosen from Q.

The *Hamming distance* of two words is defined as $d(v,w) = |\{i \mid v_i \neq w_i\}|$

Lemma X.4: Let $v, w \in H(n,q)$.
a) $d(v,w) \geq 0$ and $d(v,w) = 0$ iff $v = w$.
b) $d(v,w) = d(w,v)$
c) $d(u,v) + d(v,w) \geq (u,w)$

Definition X.5: A *code* $\mathcal{C} \subset H(n,q)$ is a subset of the Hamming space, its elements are called *codewords*. We call n the *length* of the code and $\mathcal{C}$ an q-ary code.

The *minimum distance* of $\mathcal{C}$ is $d = \min\{d(v,w) \mid v, w \in \mathcal{C}, v \neq w\}$.

$\mathcal{C}$ is called e-error-correcting if for any $w \in H(n,q)$ there is at most one codeword $v \in \mathcal{C}$ such that $d(v,w) \leq e$.

As an example consider the most pathetic of all codes: The *repetition code* of length n over an alphabet Q consists of the words $(x, x, \cdots, x)$ of length n for $x \in Q$. Its minimum distance is obviously n.

The definition of e-error correcting related to the concept of *nearest neighbor decoding*: When transmitting a codeword c, channel errors might modify it to a word w which is received. If the code is e-error correcting and $\leq e$ errors arose, c will be determined uniquely by w.

Proposition X.6: A code with minimum distance d is e-error correcting if $d \geq 2e + 1$.

Proof: If the code $\mathcal{C}$ is not e-error correcting there will be $w \in H(n,q)$ and $b, c \in \mathcal{C}$ such that $d(w,c), d(w,b) \leq e$ but then

$$d(b,c) \leq d(b,w) + d(w,c) \leq 2e$$

contradicting the definition of minimum distance $d = 2e + 1$. □

If we assume that errors occur independently, with equal probability, and with equal probability for each letter (this is called a *symmetric channel*), the nearest neighbor decoding of a word w returns the codeword that is most likely to result in the received word w.

Dual to the minimum distance is the *rate* of the code $\mathcal{C}$, defined as $\log_q(|\mathcal{C}|)/n$. (That is the symbol overhead in the transmission, compared with the information theoretic minimum required for the ability to transmit $|\mathcal{C}|$ possible messages per unit.

Theorem X.7 (Shannon's channel theorem): Given a symmetric channel with error probability p for a single symbol, there exists a maximal channel capacity M (which can be calculated explicitly):
a) For any $R < M$ and $\epsilon > 0$ there exists a code $\mathcal{C}$ of rate $\geq R$ such that the error probability of nearest neighbor decoding is less than ϵ.

b) There exists $\epsilon_0 > 0$ such that for any code $\mathcal{C}$ with rate $R > M$ the error probability of nearest neighbor decoding is $\geq \epsilon_0$.

We shall not prove this theorem here (which belongs in a class on statistical communication theory), but note that the result of this theorem is theoretical in nature:

- When reducing the error probability, the required length of the codes under a) tends → ∞.
- The proof is an existence proof (the code is constructed by random choices), there is no explicit way to construct an optimal code.
- Without further structure, nearest-neighbor decoding is of a cost exponential in the length and thus is not practical.

It thus is meaningful to ask for the largest code of a given length n and given minimum distance d over an alphabet of size q. Furthermore, the utility of such a code is higher if one can describe an effective method for nearest-neighbor decoding. Mathematicians usually care more about the first, engineers for the second question.

The aspects of decoding however are often (and have been even more in the past) dictating which codes are used in practice. A device that uses codes might not have the resources to perform a complicated decoding process in real time, and thus will have to use a less sophisticated code than theoretically available.

X.2 Minimum Distance Bounds

We start by determining some bounds on the possible quality of codes. The first is an (often pathetic) lower bound that could be obtained by a greedy choice of (otherwise unspecified) code words:

Proposition X.8 (Varshamov-Gilbert bound): Given n, q, d there is a q-ary code of length n and minimum distance $\geq d$ with at least

$$q^n / \left(\sum_{i=0}^{d-1} \binom{n}{i} (q-1)^i \right)$$

codewords.

Proof: For a fixed code word c and a distance i, there are $\binom{n}{i}(q-1)^i$ words w at distance $d(c, w) = i$, as we can choose the i components in which the word differs, and for each component have $q - 1$ choices of different entries.

This means that the choice of every code word eliminates at most

$$\sum_{i=0}^{d-1} \binom{n}{i} (q-1)^i$$

words in $H(n,q)$ as candidates for code words, as they would have distance $< d$. A greedy algorithm, that in each step simply selects a random word that is not at distance $< d$ to any of the words chosen so far, thus will be able to make at least as many choices as claimed. □

We call the set of words at bounded distance $\le r$ the *ball* or (in coding theory) *sphere* of radius r

If $d \ge 2e+1$ these spheres of radius e may not overlap and thus give a similar upper bound:

Theorem X.9 (Hamming bound, sphere-packing bound): Suppose that $d \ge 2e+1$. A q-ary code of length n and minimum distance d has at most

$$q^n / \left(\sum_{i=0}^{e} \binom{n}{i} (q-1)^i \right)$$

codewords.

If this bound is attained, that is the code has as many code words as possible, it is called a *perfect code.*

It is rare that a particular parameter set affords a perfect code.

Proposition X.10 (Singleton bound): A q-ary code C of length n and minimum distance d has at most q^{n-d+1} codewords.

Proof: Two words of C cannot agree in the first $n-d+1$ positions (as they then only would have $d-1 < d$ positions to differ). Thus codewords are determined by their entries in the first $n-d+1$ positions, yielding q^{n-d+1} possibilities. □

A code that achieves the Singleton bound is called *maximum distance separable* (MDS).

We note – Exercise **??** – that depending on the size of the alphabet the Hamming bound or the Singleton bound could be better.

There are further, better, bounds for the size of codes, but we shall not need these here.

X.3 Linear Codes

So far we have not given any examples of codes apart from random selections of codewords. The first class of systematic constructions comes with tools from Linear Algebra. This makes it easier to describe codes and will also provide a better decoding algorithm.

We now consider the case that q is a prime power and $Q = \mathbb{F}_q$. In this case we can identify the Hamming space $H(n,q)$ with the vector space $\mathbb{F}_q^n$.

Definition X.11: A *linear code* is a subspace $C \le \mathbb{F}_q^n$. The *weight* of $c \in C$ is the number $\mathrm{wt}(c)$ of nonzero entries of c. The minimum weight w of C is the minimum weight of nonzero $c \in C$.

If $\mathcal{C}$ is of dimension k, we call it an $[n,k,w]$-code.

Linearity means that distances are based on weights:

LEMMA X.12: Let $\mathcal{C}$ be linear and $v, c \in \mathcal{C}$. Then $d(v,c) = \text{wt}(v-c)$.

If $\mathcal{C}$ is a linear code, say of dimension k, it is the $\mathbb{F}_q$ span of a basis, consisting of k vectors. We can write these vectors as rows of a $k \times n$ matrix G, called the *generator matrix* of $\mathcal{C}$. (That is, $\mathcal{C}$ is the row space of G.) Note that the choice of G is not unique. Since we also can permute columns (replacing $\mathcal{C}$ with an equivalent code) one often assumes that $G = (I \mid P)$ for some matrix P.

Next we consider a basis of the null space (the vectors $G \cdot x = 0$) of G and write the basis vectors as rows in a matrix H, that is $GH^T = 0$. Then H is called a *check matrix* for $\mathcal{C}$. Basic linear algebra gives us that $G \in \mathbb{F}_q^{k\times n}$ and $H \in \mathbb{F}_q^{(n-k)\times n}$, and that the rows of G are a basis of the null space of H.

Note that for $w \in \mathbb{F}_q^n$ we have that $w \in \mathcal{C}$ if and only if $wH^T = 0$, thus the term check matrix.

Generator matrix and check matrix now offer a convenient way of encoding and decoding. As the code is k-dimensional, the messages to be encoded will be simply vectors $m \in \mathbb{F}_q^k$. The encoding process then simply consists of the product $m \cdot G = c \in \mathcal{C}$. This is the code word that will be transmitted.

We receive a word $w = c+u$ with u denoting the (as yet unknown) transmission error. For decoding w we calculate $s = wH^T = uH^T$ (as $cH^T = 0$ by definition). We call s the *syndrome* of w and note that it only depends in the error u. Furthermore, the map from error to syndrome is injective:

Suppose that u_1, u_2 are two error vectors with $u_1H^T = u_2H^T$. Then $(u_1-u_2)H^T = 0$, that is $u_1 - u_2$ is in the null space of H, that is in $\mathcal{C}$.

On the other hand, since they are errors, we have that $\text{wt}(u_1), \text{wt}(u_1) \le e$ and thus $\text{wt}(u_1 - u_2) \le 2e$. The only vector in $\mathcal{C}$ of this weight is the zero vector.

This shows that in principle the error (and thus the correct decoding) can be determined from the syndrome, this is called *syndrome decoding*.

In some situations a clever choice of the check matrix H – e.g. putting it in RREF – allows for construction of an algorithm that computes u from s; in general however (as H^T is not invertible), we might have to rely on a look-up table, listing errors for each syndrome.

Example: In a pathetic example, let $Q = \mathbb{F}_2$ and

$$G = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

We easily find the four vectors in the row space of G and verify that this is a code of minimum weight 3, i.e. an $[5,2,3]$ code. The message $m = (1,1)$ thus gets encoded by $c = (1,1,1,1,0)$.

We also calculate

$$H = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Suppose we receive $w = c + u = (1,1,1,0,0)$. We calculate the syndrome $wH = (1,1,1)$. The only vector u of weight 1 that can give this syndrome is $u = (0,0,0,1,0)$, allowing us to recover the original codeword and message.

If two errors happened and we received $w = c+u = (0,1,1,0,0)$. The syndrome then is $[0,1,1]$, which could come from the errors $(1,0,0,1,0)$ or $(0,1,1,0,0)$.

Proposition X.13: A linear code has minimum weight $\geq d$ if any $d-1$ columns of its check matrix H are linearly independent.

Proof: A product cH^T is a linear combination of the columns of H with coefficients given by c. That is a linear relation amongst m columns of H corresponds to a word in C of weight m and vice versa. Since the minimum weight is $\geq d$ it is impossible for $< d$ columns to be linearly dependent. □

In the special case of 1-error correcting codes this simply means that any two columns of H must be linearly independent. For a given number d of rows, a maximal set of such columns is obtained by representatives of the 1-dimensional subspaces of $\mathbb{F}_q^d$, corresponding to the $\left[\begin{smallmatrix} d \\ 1 \end{smallmatrix}\right]_q$ points of $\mathrm{PG}(d-1,q)$.

Definition X.14: Let H be a $d \times \left[\begin{smallmatrix} d \\ 1 \end{smallmatrix}\right]_q$ matrix whose columns are representatives of $\mathrm{PG}(d-1,q)$, and C the code of length $n = \left[\begin{smallmatrix} d \\ 1 \end{smallmatrix}\right]_q$ whose check matrix is H. It is called the *Hamming code* over $\mathbb{F}_q$ of length n.

The Hamming code of length n thus is a linear $(n, n-d, 3)$ code.

Note X.15: Changing representatives does not change C, changing the order of the columns will change the column order of C, an equivalent code.

Note X.16: Syndrome decoding of Hamming codes is particularly easy. If the syndrome s is nonzero, it is, up to a scalars α, equal to a unique (the i-th) column of H. The (weight one) error vector then is $u = \alpha \cdot e_i$

Example: Taking $d = 3$ and $q = 2$ we get

$$H = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}, \text{ and } G = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

We encode, for example, $m = (1,0,1,0)$ by $mG = (1,0,1,0,1,0,1)$. Suppose a transmission error makes us receive $w = (1,0,0,0,1,0,1)$ and syndrome $(0,1,1)$. This equals the third column of H, so the error was (indeed) in the third position.

Theorem X.17: Hamming codes are perfect 1-error correcting.

Proof: The above discussion shows that a Hamming code is 1-error correcting and its length is $n = \left[\begin{smallmatrix} d \\ 1 \end{smallmatrix}\right]_q = (q^d - 1)/(q-1)$, so $q^d = n(q-1)+1$. The number of code words is $q^{n-d} = q^n/q^d = q^n/(1+n(q-1))$, which is the Hamming bound for $e = 1$. □

TODO: Exercise, Lie

X.4 Code Operations

there are a number of operations that take an existing code and construct new ones. We shall describe these for linear codes, though similar ideas also work for other codes.

Punctured Code

We assume that $C \subset \mathbb{F}_q^n$ is a linear $[n,k,d]$ code. For $1 \le i \le n$, we define the *punctured code* at i to be the set of all code words of C with the i-th entry removed. We denote it by C^*. As image under the linear map omitting the i-th component it is a linear code of length $n-1$.

We obtain a generator matrix for C^* from a generator matrix for C by removing the i-column (and any resulting duplicate or zero rows).

Theorem X.18: a) If $d > 1$ then C^* is of dimension k. If C has a minimum weight codeword with a nonzero entry in the i-th position, then C^* is of minimum distance $d-1$, otherwise of minimum distance d.

b) If $d = 1$ then C^* is of dimension k and minimum distance $d = 1$ if e_i is not a codeword. Otherwise C^* is of dimension $k-1$ and minimum distance ≥ 1.

Proof: Exercise **??** □

Example: Let C be the Hamming code for $d = 3$ and $q = 2$ in the example above. The first row of the generator matrix is a word of minimum weight 3. Thus puncturing in position 1 yields a $[6,4,2]$ code.

Extended Code

A somewhat dual operation to puncturing is extension.

Definition X.19: If C is a linear code of length n and size q^k that has minimum distance d, let

$$\hat{C} = \left\{ (x_1, \ldots, x_{n+1}) \mid (x_1, \ldots, x_n) \in C, \sum_{i=1}^{n+1} x_i = 0 \right\}$$

the *extended code*.

Clearly we have that the length of the extended code increased by 1 and the minimum weight does not decrease and cannot increase by more than 1. Over $\mathbb{F}_2$, the minimum weight of the extended code is guaranteed to be even.

Dual Codes and the Weight Enumerator

If $GH^T = 0$, transposition gives $HG^T = 0$. This means that for a linear code the roles of generator matrix and check matrix can be exchanged.

Definition X.20: Let C be a linear code with generator matrix G and check matrix H. The *dual code* $C^\perp$ is defined by generator matrix H and check matrix G.

Note that $C^\perp$ is the orthogonal complement to C with respect to the standard inner product. Its dimension is $n - k$.

We want to determine a relation between the weights of code words in C and those in $C^\perp$.

Definition X.21: For a code C let a_i be the number of code words $c \in C$ of weight $\mathrm{wt}(c) = i$. The polynomial

$$A(z) = \sum_{i=0}^{n} a_i z^i \in \mathbb{C}[z]$$

is called the *weight enumerator* of C.

Clearly $A(0) = a_0 = 1$ and $A(1) = |C| = q^k$.

Theorem X.22 (MacWilliams identity): Let C be a linear code of length n and dimension k over $\mathbb{F}_q$ whose weight enumerator is $A(z)$ and let $B(z)$ be the weight enumerator of $C^\perp$. Then

$$B(z) = q^{-k}(1 + (q-1)z)^n A\left(\frac{1-z}{1+(q-1)z}\right).$$

Proof: We give the proof only in the case $q = 2$ to avoid having to introduce linear characters (mappings from a finite field to the roots of unity), but may simply use a ±1. (The proof otherwise is identical.)

We define a function $g\colon \mathbb{F}_2^n \to \mathbb{C}$ by

$$g(u) = \sum_{v \in \mathbb{F}_2^n} (-1)^{(u,v)} z^{\mathrm{wt}\, v}.$$

(with $(\cdot,\cdot)$ the standard inner product) and thus

$$\sum_{u \in C} g(u) = \sum_{v \in \mathbb{F}_2^n} z^{\mathrm{wt}\, v} \sum_{u \in C} (-1)^{(u,v)}.$$

If $v \in \mathcal{C}^\perp$, the inner sum is over $1^0 = 1$ and thus has value $|\mathcal{C}|$. If $v \notin \mathcal{C}^\perp$, then (u, v) is a linear map with nonzero image. It thus takes as values all field elements with equal frequency. The inner sum then cancels out to zero, and we get that

$$\sum_{u \in \mathcal{C}} g(u) = |\mathcal{C}|\, B(z).$$

On the other hand we can split up

$$(-1)^{(u,v)} z^{\mathrm{wt}\, v} = \prod_{i=1}^{n} \left((-1)^{u_i v_i} z^{v_i}\right),$$

and thus have

$$\begin{aligned} g(u) &= \sum_{v=(v_1,\ldots,v_n)\in\mathbb{F}_2^n} (-1)^{(u,v)} z^{\mathrm{wt}\, v} \\ &= \prod_{i=1}^{n} \sum_{v_i \in \mathbb{F}_2} (-1)^{u_i v_i} z^{v_i} \\ &= \prod_{i=1}^{n} (1 + (-1)^{u_i} z) \\ &= (1-z)^{\mathrm{wt}(u)} (1 + (q-1)z)^{n-\mathrm{wt}(u)}. \end{aligned}$$

Therefore

$$\begin{aligned} q^k B(z) &= |\mathcal{C}|\, B(z) = \sum_{u \in C} (1-z)^{\mathrm{wt}(u)} (1 + (q-1)z)^{n-\mathrm{wt}(u)} \\ &= (1 + (q-1)z)^n \sum_{u \in C} \left(\frac{1-z}{1+(q-1)z}\right)^{\mathrm{wt}(u)} \\ &= (1 + (q-1)z)^n A\left(\frac{1-z}{1+(q-1)z}\right), \end{aligned}$$

proving the theorem. □

Code Equivalences

The basic property of the Hamming space is distance. The appropriate set of equivalence operations are thus transformations of the Hamming space, called *isometries*, that preserve distances. As distance is defined as the number of coordinates in which words differ, two kinds of isometries are:

- Permutations of the alphabet in one coordinate.
- Permutations of the coordinates.

Together they generate a group $S_q \wr S_n$, acting in the product action on $H(n,q)$.

NOTE X.23: This is the full group of isometries of Hamming space: Pick a "base" element $v = (a, a, \ldots, a)$. By considering code elements at distance 1 to v and distance 1 to each other, we can establish a permutation of coordinates. What remains are symbol permutations.

As usual, we call two codes *equivalent* if they can be mapped to each other under an isometry. The automorphism group of a code is the set of those isometries that map the set of code words back in the code.

When considering linear codes, the zero word has a special role and needs to be preserved. Furthermore we cannot permute the alphabet – $\mathbb{F}_q$ arbitrarily, but need to do so in a way that is compatible with vector space operations, that is we need to replace the set of all permutations S_q of the alphabet with scalar multiplication by nonzero elements, the group $\mathbb{F}_q^*$.

The resulting linear isometry group thus is $\mathbb{F}_q^* \wr S_n$. It elements, can be represented by *monomial matrices*, that is matrices that have exactly one nonzero entry in every row and column. In this representation the action on F_q^n is simply by matrix multiplication.

If $q = 2$, there is no nontrivial scalar multiplication and the linear isometry group becomes simply S_n.

X.5 Cyclic Codes

An important subclass of linear codes are cyclic codes: they allow for the construction of interesting codes, they connect coding theory to fundamental concepts in abstract algebra and — this is what gives them practical relevance — the existence of practical methods for decoding (which we shall not go into detail of).

DEFINITION X.24: A linear code $\mathcal{C} \subset \mathbb{F}_q^n$ is *cyclic*, if cyclic permutations of codewords are also codewords, that is $\mathcal{C}$ is (as a set) invariant under the permutation action of the cyclic group $\langle(1,2,3,\ldots,n)\rangle$.

Assume that $\gcd(n,q) = 1$ and let $R = \mathbb{F}_q[x]$ the polynomial ring and the ideal $I = (x^n - 1) \lhd R$. Then $\mathbb{F}_q^n$ is isomorphic (as a vector space) with $\mathbb{F}_q[x]/I$, the isomorphism being simply

$$(a_0, a_1, \ldots, a_{n-1}) \leftrightarrow I + a_0 + a_1x + \cdots + a_{n-1}x^{n-1}.$$

In this representation, cyclic permutation corresponds to multiplication by x, as $I + ax^{n-1} \cdot x = +ax^n = I - a(x^n - 1) + ax^n = I + a$. That means that a code $\mathcal{C}$, considered as a subset of R/I, is cyclic if and only if it is (as a set) invariant under multiplication by x.

Since $\mathcal{C}$ is a linear subspace, this is equivalent to invariance under multiplication by arbitrary polynomials, that is

PROPOSITION X.25: A cyclic code of length n is an ideal in $\mathbb{F}_q[x]/(x^n - 1)$.

We know from abstract algebra that ideals of a quotient ring R/I correspond to ideals $I \leq J \lhd R$ (as J/I). Furthermore $\mathbb{F}_q[x]$ is a principal ideal domain, that is every ideal is generated by a polynomial.

LEMMA X.26: Let R be a ring and $a, b \in R$. Then $(a) \subset (b)$ if and only if $b \mid a$ in R.

Proof: The statement $(a) \subset (b)$ is equivalent to $a \in (b)$. By definition this is if and only if there exists $r \in R$ such that $a = rb$, that is $b \mid a$. □

We thus get

THEOREM X.27: The cyclic codes of length n over $\mathbb{F}_q$ are given by the divisors of $x^n - 1$ over $\mathbb{F}_q$.

NOTE X.28: This is the reason for the condition that $\gcd(n, q) = 1$. Otherwise $q = p^a$ and $p \mid n$ but

$$x^n - 1 = \left(x^{\frac{n}{p}}\right)^p - 1 \equiv \left(x^{\frac{n}{p}} - 1\right)^p \pmod{p}$$

is a power.

If $\mathcal{C} = (I + g(x))$ for a cyclic code $\mathcal{C} \leq \mathbb{F}_q[x]/(x^n - 1)$, we call $g(x)$ the *generator polynomial* of $\mathcal{C}$ and $h(x) = (x^n - 1)/g(x)$ the *parity check polynomial* of $\mathcal{C}$.

Considered as a polynomial, we have that the code words $c \in \mathcal{C}$ are simply multiples of g, we that get that an arbitrary polynomial $f(x) \in \mathcal{C}$ if and only if $f(x)h(x) \equiv 0 \pmod{x^n - 1}$.

Example: For $n = 7$ and $q = 2$ we have that $x^7 - 1 = (x+1)(x^3+x+1)(x^3+x^2+1)$. Thus $g(x) = x^3 + x^2 + 1$ creates a cyclic code of length 7 and $2^{7-3} = 2^4 = 16$ code words, and check polynomial $h(x) = x^4 + x^3 + x^2 + 1$.

While this is a slick way of constructing codes, we have not yet said a single word about their ultimate purpose, error correction; respectively the minimum distance of cyclic codes.

The first step on that path will be to look at a different version of check matrix:

THEOREM X.29: Let $\mathcal{C} \subset \mathbb{F}_q[x]/(x^n - 1)$ be a cyclic code with generator polynomial g of degree $n - k$, and let $\alpha_1, \ldots, \alpha_{n-k}$ be the roots of g. Then

$$H = \begin{pmatrix} 1 & \alpha_1 & \alpha_1^2 & \cdots, \alpha_1^{n-1} & \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha_{n-k} & \alpha_{n-k}^2 & \cdots, \alpha_{n-k}^{n-1} & \end{pmatrix}$$

is a check matrix for $\mathcal{C}$, that is $f(x) = \sum f_i x^i \in \mathbb{F}_q[x]/(x^n - 1)$ is in the code if and only if $(f_0, \ldots, f_{n-1}) \cdot H^T = 0$.

Proof: The criterion is that for all j we have $0 = \sum f_i \alpha_j^i = f(\alpha_j)$. This is the case if and only if $g(x) \mid f(x)$, that is if $f(x) \in \mathcal{C}$. □

Note X.30: By expressing the α_i^j as (column) coefficient vectors in an $\mathbb{F}_q$ basis of a suitable field $\mathbb{F}_{q^m} \le \mathbb{F}_{q^n}$ (that is multiplying the number of rows by a factor) we can replace H with an $\mathbb{F}_q$ matrix.

We now can use Proposition X.13 on this matrix H to determine the minimum distance. We note that rows associated to roots of the same minimal polynomial do not contribute to further checking.

Lemma X.31: Suppose that $g(x) = \prod_{i=1}^m g_i(x)$ as a product of (different) irreducible polynomials over $\mathbb{F}_q$ and that (after reordering) we have $g_i(\alpha_i) = 0$ for $i = 1, \ldots, m$. Then

$$H' = \begin{pmatrix} 1 & \alpha_1 & \alpha_1^2 & \cdots & \alpha_1^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha_m & \alpha_m^2 & \cdots & \alpha_m^{n-1} \end{pmatrix}$$

(that is we only take one root for each irreducible factor) is a check matrix for C.

Proof: Let β be a root of g that is not amongst $\alpha_1, \ldots, \alpha_m$. Then (WLOG) β and $\alpha = \alpha_1$ must be roots of the same irreducible factor of g. This implies that there is a Galois automorphism σ of the field extension $\mathbb{F}_q(\alpha) = \mathbb{F}_q(\beta)$ that will map $\alpha \to \beta$. But then for any $f(x) \in \mathbb{F}_q[x]$ we have that

$$\sigma(f(\alpha)) = f(\sigma(\alpha)) = f(\beta).$$

As $\sigma(0) = 0$ the statement follows. □

Note X.32: The reader might wonder whether these check matrices H violate the theorems we have proven about rank and number of rows of these matrices. They don't, because they are not defined over $\mathbb{F}_q$ but over an extension – If α is of degree m we would need to replace the row for α by m rows, representing α (and its powers similarly) with coefficients with respect to an $\mathbb{F}_q$-basis.

Example: Let $n = 15$ and $q = 2$. We take

$$g(x) = (x^4 + x + 1) \cdot (x^4 + x^3 + x^2 + x + 1) \mid x^{15} - 1,$$

that is $\deg(g) = 8$ and it defines a code of dimension $15 - 8 = 7$. Let α be a root of $p_1(x) = x^4 + x + 1$ in $\mathbb{F}_{16}$, then $p_1(x) = (x - \alpha)(x - \alpha^2)(x - \alpha^4)(x - \alpha^8)$ (the other roots must be images of the first under the Frobenius automorphism), α is a generators of the multiplicative group of $\mathbb{F}_{16}$ (explicit calculation), and (also explicit calculation) we have that α^3 is a root of $x^4 + x^3 + x^2 + x + 1$. Thus we can take the check matrix as

$$H = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots, \alpha^{14} \\ 1 & \alpha^2 & \alpha^4 & \cdots, \alpha^{28} \\ 1 & \alpha^3 & \alpha^6 & \cdots, \alpha^{42} \\ 1 & \alpha^4 & \alpha^8 & \cdots, \alpha^{56} \end{pmatrix}$$

or – removing some rows as in the prior lemma –

$$H = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots, \alpha^{14} \\ 1 & \alpha^3 & \alpha^6 & \cdots, \alpha^{42} \end{pmatrix}.$$

Lemma X.33 below shows that any four columns of this matrix H are linearly independent, thus g defines a code of minimum distance $d \geq 5$.

The minimum bound in this example follows from the following technical lemma:

Lemma X.33: Let α be an element of multiplicative order n (also called an *primitive n-th root of unity*) and $b > 0$. Then any m columns of the $m \times n$ matrix

$$H = \begin{pmatrix} 1 & \alpha^b & \alpha^{2b} & \cdots, \alpha^{(n-1)b} & \\ 1 & \alpha^{b+1} & \alpha^{2(b+1)} & \cdots, \alpha^{(n-1)(b+1)} & \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \alpha^{b+m-1} & \alpha^{2(b+m-1)} & \cdots, \alpha^{(n-1)(b+m-1)} & \end{pmatrix}$$

are linearly independent.

Proof: Let $i_1, \dots i_m$ be the indices (starting with 0) of the columns chosen. Then these columns form the matrix

$$M = \begin{pmatrix} \alpha^{bi_i} & \alpha^{bi_2} & \alpha^{bi_3} & \cdots, \alpha^{bi_m} & \\ \alpha^{(b+1)i_1} & \alpha^{(b+1)i_2} & \alpha^{(b+1)i_3} & \cdots, \alpha^{(b+1)i_m} & \\ \vdots & \vdots & \vdots & & \vdots \\ \alpha^{(b+m-1)i_1} & \alpha^{(b+m-1)i_2} & \alpha^{(b+m-1)i_3} & \cdots, \alpha^{(b+m-1)i_m} & \end{pmatrix}$$

and we get by the usual determinant rules that

$$\det(M) = \alpha^{b(i_1+i_2+\cdots+i_m)} \det(M_0)$$

with

$$M_0 = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ \alpha^{i_1} & \alpha^{i_2} & \alpha^{i_3} & \cdots & \alpha^{i_m} \\ \vdots & \vdots & \vdots & & \vdots \\ \alpha^{(m-1)i_1} & \alpha^{(m-1)i_2} & \alpha^{(m-1)i_3} & \cdots & \alpha^{(m-1)i_m} \end{pmatrix}.$$

But M_0 is a Vandermonde matrix, so

$$\det(M_0) = \prod_{1 \leq j < k \leq m} \left(\alpha^{i_j} - \alpha^{i_k}\right) \neq 0,$$

since all powers of α will be different. □

We generalize this result to a class of codes that was discovered independently by R.C. Bose (as in previous chapters) and D.K. Ray-Chaudhuri, as well as by A. Hocquenghem and which are named after these discoverers.

DEFINITION X.34: Let q be a prime power, n be an integer and $\alpha \in \mathbb{F}_{q^e}$ be a primitive n-th root of unity (that is $n \mid q^e - 1$). A *BCH code* over $\mathbb{F}_q$ of length n and *designed distance* $2 \le d \le n$ is the cyclic code defined by the polynomial over $\mathbb{F}_q$ with roots

$$\alpha^b, \alpha^{b+1}, \ldots, \alpha^{b+d-2}.$$

The polynomial that defines such a code is simply

$$\operatorname{lcm}(m_1(x), m_2(x), \ldots, m_{b+d-2})$$

where m_i is the minimal polynomial of α^i over $\mathbb{F}_q$. In general several of these m_i will be equal or could be of degree 1.

PROPOSITION X.35: A BCH code of designed distance d has minimum distance $\ge d$.

Proof: If the code is of designed distance d, its check matrix H will include the $d-1$ rows of the matrix in Lemma X.33 as rows. Thus any $d-1$ columns of H must be independent, showing that the code has minimum distance $\ge d$. □

BCH codes are particularly easy to work with if no extension is involved, that is $n \mid q-1$:

DEFINITION X.36: A BCH code of length $n = q-1$ over $\mathbb{F}_q$ is called a *Reed Solomon code.*

Reed Solomon codes have been the work-horse of error correction. They are used, for example for error correction on CD's DVD's, Blu-Ray, QR codes, DSL networking and several wireless standards.

X.6 Perfect Codes

In general, perfect codes only exist for a select few parameter sets and thus are rarely used in practice. The codes which exist however are often mathematically significant, which is the reason for studying them.

In the case of 1-error correcting perfect codes we easily get a full classification:

PROPOSITION X.37: a) A perfect 1-error correcting code over an alphabet of prime-power order q has length $\begin{bmatrix} d \\ 1 \end{bmatrix}_q$ (for a suitable value of d).
b) A linear, perfect, 1-error correcting code is a Hamming code.

Proof: Let $\mathcal{C}$ be a perfect 1-error correcting code. By definition this means that $|\mathcal{C}| = q^n/(1+n(q-1))$, that is $1+n(q-1)$ divides q^n. As $q = p^a$ this means that $1+n(q-1) = q^d p^e$ for nonnegative d, e with $e < a$. Reduction modulo $q-1$ gives $p^e \equiv q^d p^e \equiv 1 \pmod{q-1}$, as $p^e < q$ this implies that $p^e = 1$ and thus $1+n(q-1) = q^d$, that is $n = (q^d-1)/(q-1)$.
If $\mathcal{C}$ is also linear we have that $q^n/(1+n(q-1)) = |\mathcal{C}| = q^{n-d}$. That means a check matrix H for $\mathcal{C}$ must have n columns of length d each. By X.13 these columns are

pairwise linear independent, which implies that they are representatives of the 1-dimensional subspaces of $\mathbb{F}_q^d$. By definition, $\mathcal{C}$ is a Hamming code. □

We now consider e-error correcting codes for $e > 1$ and $q = 2$.

For $e = 2$ the Hamming bound gives

$$2^{\text{dim}} = |\mathcal{C}| = 2^n / \left(\binom{n}{2} + n + 1 \right) = 2^{n+1}/(n^2 + n + 2)$$

This implies that $n^2 + n + 2^a$ for some $a \geq 0$. Setting $x = 2n + 1$ and $y = a + 2$ we get $2^y - 7 = x^2$. This is *Nagell's equation*, whose solution is a standard topic in algebraic number theory. The only solutions are

x	± 1	± 3	± 5	± 11	± 181
y	3	4	5	7	15

Since the code is to be 2-error correcting we must have that $5 \leq n = \frac{x-1}{2}$, so only $n = 5, 90$ are possible. For $n = 5$, the repetition code of length 5 is an example.

We will see below (X.40) that $n = 90$ is impossible.

Next we consider $q = 2$ and $e = 3$ and get

$$2^{\text{dim}} = |\mathcal{C}| = 2^n / \left(\binom{n}{3} + \binom{n}{2} + n + 1 \right) = 2^{n+1}/(n^2 + n + 2) = 2^n 6/(n^3 + 5n + 6).$$

Thus (with $m = n + 1$):

$$2^a 3 = n^3 + 5n + 6 = (n+1)(n^2 - n + 6) = (n+1)((n+1)^2 - 3(n+1) + 8) = m(m^2 - 3m + 8).$$

As the code is 3-error correcting we know that $n \geq 7$, i.e. $m \geq 8$. We now use unique factorization to find the possible solutions:

If $m \equiv 0 \pmod{16}$, then $m^2 - 3m + 8 \equiv 8 \pmod{16}$. But we also know that $m^2 - 3m + 8 = 2^x 3$ for some x. This means that the congruence can only be satisfied if $x \leq 3$ and $m^2 - 3m + 8 = 24$ is the only possibility. However, for $m \geq 8$ we have that $m^2 - 3m + 8 \geq 48$, contradiction.

This means that $m \not\equiv 0 \pmod{16}$, which implies that m must be a divisor of $2^3 \cdot 3$, of which only 8, 12, and 24 satisfy that $m \geq 8$. But for $m = 12$ we have that $m^2 - 3 * m + 8 = 116 = 2^2 29$. Thus $m \in \{8, 24\}$, respectively $n \in \{7, 23\}$ are the only possibilities.

If $n = 7$, then $|\mathcal{C}| = 6 \cdot 128/384 = 2$ and the code must be a repetition code. We thus consider $n = 23$ (for which a perfect code would need to be 12-dimensional:

Considering cyclic codes, we note that

$$x^{23} - 1 \equiv (x+1)(x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1)(x^{11} + x^{10} + x^6 + x^5 + x^4 + x^2 + 1) \pmod{2}$$

and that $g(x) = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1$ will yield a $23 - 11 = 12$-dimensional code[1].

Let $\alpha \in \mathbb{F}_{2^{11}}$ a root of g. Then the other roots of g are α^x for (we note as an aside, that these exponents are exactly the squares modulo 23) $x \in \{1, 2, 3, 4, 6, 8, 9, 12, 13, 16, 18\}$. As the exponent set includes $1, 2, 3, 4$, g defines a BCH code of designed distance 5. We call this code the *Golay code* $\mathcal{G}_{23}$ and study it more in Section X.7.

In particular, we shall show that the minimum weight of $\mathcal{G}_{23}$ is in fact 7, so $\mathcal{G}_{23}$ is a perfect code!

Golay also discovered a perfect code of length 11 over $\mathbb{F}_3$ that is 2-error correcting. In fact one can show that we have found all perfect codes for alphabets of prime-power order:

Theorem X.38 (Tietäväinen - van Lint): The only perfect e-error correcting codes for $e > 1$ over alphabets of prime-power size q are the repetition codes for $q = 2$, $n = 2e + 1$; as well as the two codes discovered by Golay.

Perfect codes have an intimate relation with designs.

Proposition X.39: Let $\mathcal{C}$ be a linear perfect code of length n over $\mathbb{F}_q$ that is e-error correcting. Then the supports (i.e. the indices of the nonzero entries) of the code words of minimal weight $2e + 1$ are the blocks of an $(e + 1) - (n, 2e + 1, (q - 1)^e)$ design.

Proof: Choose a set of $e + 1$ indices and let $w \in \mathbb{F}_q^n$ with support on this index set. Then there is a unique $c \in \mathcal{C}$ that is at distance $\leq e$ to w. This word must have $\operatorname{wt} c = 2e + 1$, so it is of minimal weight. The support of c must include the initially chosen $e+1$ indices, and c must agree with w on these indices. Thus different choices of w (there are $(q-1)^{e+1}$ words in $\mathbb{F}_q^n$ with the same support) imply different choices of c.

Scalar multiples of code words have the same support, so the $(q - 1)^{e+1}$ code words c defined this way define $(q - 1)^e$ index sets by their support. □

Corollary X.40: There is no perfect 2-error correcting code of length 90 over $\mathbb{F}_2$.

Proof: Let $\mathcal{C}$ a code as described. If the code is not linear we translate it in the Hamming space so that WLOG $0 \in \mathcal{C}$. As there are no nontrivial scalar multiples over $\mathbb{F}_2$, the argument of Proposition X.39 still holds for this translated code.

Thus a code as described would define a $3 - (90, 5, 1)$ design which by IX.2 would have $88/3$ blocks, an impossibility. □

[1] A similar argument will work for the other degree 11 factor and produce an equivalent code

X.7 The (extended) Golay code

(interversion, 10 fois intervertie par elle-même, sur 12 valeurs, 12 sons, 4 attaques, 5 intensités)
interversion I
interversion II

Île de feu II
Olivier Messiaen

Some years ago he had embarked [...] on a series of commentaries on Jane Austen [...] The object of the exercise [...] was not to enhance others' enjoyment and understanding of Jane Austen, still less to honour the novelist herself, but to put a definitive stop to the production of any further garbage on the subject.

Changing Places
David Lodge

Much in this section are marvelous coincidences whose properties we can observe, but that alas do not generalize to other settings. Thus the calculations should not be considered as special cases of general principles!

Recall $\mathcal{G}_{23}$ as the cyclic code over $\mathbb{F}_2$ of length 23, generated by the polynomial $g(x) = x^{11}+x^9+x^7+x^6+x^5+x+1$. We know that it has dimension 12 and minimum weight at least 5.

We extend it to a new code $\mathcal{G}_{24}$ of length 24 (called the *extended Golay code* by adding a parity bit that ensures that all words have even weight. The minimum weight of $\mathcal{G}_{24}$ thus is at least 6.

An inspection of the (in figure X.1 triangulized) generator matrix of $\mathcal{G}_{24}$ shows that $\mathcal{G}_{24}$ is self-dual (i.e. the matrix can also serve as check matrix).

The arithmetic of $\mathbb{F}_2$-vectors now gives the following surprising consequences

Lemma X.41: Let $v, w \in \mathbb{F}_2^n$. Then

$$\mathrm{wt}(v+w) = \mathrm{wt}(v) + \mathrm{wt}(w) - 2(v,w)$$

with $(\cdot,\cdot)$ being the standard (characteristic zero) inner product.

$$\begin{pmatrix}
1 & . & . & . & . & . & . & . & . & . & . & . & 1 & 1 & . & . & . & 1 & 1 & 1 & . & 1 & . & 1 \\
. & 1 & . & . & . & . & . & . & . & . & . & . & . & 1 & 1 & . & . & . & 1 & 1 & 1 & . & 1 & 1 \\
. & . & 1 & . & . & . & . & . & . & . & . & . & 1 & 1 & 1 & 1 & . & 1 & 1 & . & 1 & . & . & . \\
. & . & . & 1 & . & . & . & . & . & . & . & . & . & 1 & 1 & 1 & 1 & . & 1 & 1 & . & 1 & . & . \\
. & . & . & . & 1 & . & . & . & . & . & . & . & . & . & 1 & 1 & 1 & 1 & . & 1 & 1 & . & 1 & . \\
. & . & . & . & . & 1 & . & . & . & . & . & . & 1 & 1 & . & 1 & 1 & . & . & 1 & 1 & . & . & 1 \\
. & . & . & . & . & . & 1 & . & . & . & . & . & . & 1 & 1 & . & 1 & 1 & . & . & 1 & 1 & . & 1 \\
. & . & . & . & . & . & . & 1 & . & . & . & . & . & . & 1 & 1 & . & 1 & 1 & . & . & 1 & 1 & 1 \\
. & . & . & . & . & . & . & . & 1 & . & . & . & 1 & 1 & . & 1 & 1 & 1 & . & . & . & 1 & 1 & . \\
. & . & . & . & . & . & . & . & . & 1 & . & . & 1 & . & 1 & . & 1 & . & . & 1 & . & 1 & 1 & 1 \\
. & . & . & . & . & . & . & . & . & . & 1 & . & 1 & . & . & 1 & . & . & 1 & 1 & 1 & 1 & 1 & . \\
. & . & . & . & . & . & . & . & . & . & . & 1 & 1 & . & . & . & 1 & 1 & 1 & . & 1 & . & 1 & 1
\end{pmatrix}$$

Figure X.1: The generator matrix of the extended Golay code

Proof: The only elements of $\mathbb{F}_2$ are zero and one. The nonzero entries of $v + w$ thus are the nonzero entries in v and not in w or vice versa. Thus the weight of $v + w$ equals the sum of the weights of v and w minus twice (once to avoid double counting by PIE, once as $1 + 1 = 0$) the number of common 1-entries in v and w, which is given by (v, w). □

Proposition X.42: For any $c \in \mathcal{G}_{24}$ we have that $4 \mid \mathrm{wt}(c)$.

Proof: Every code word can be expressed as sum of a suitable set of rows of the generating matrix. We use induction on the number of summands.

For the base case, observe that the generating matrix of $\mathcal{G}_{24}$ consists of translates of g with a parity 1 appended (as $\mathrm{wt}(g) = 7$), that is all rows have weight 8 which is a multiple of 4.

For the inductive step suppose that $a, b \in \mathcal{G}_{24}$ with $4 \mid \mathrm{wt}(a), \mathrm{wt}(b)$. Then, by the previous lemma

$$\mathrm{wt}(a + b) = \mathrm{wt}(a) + \mathrm{wt}(b) - 2(a, b).$$

Since $\mathcal{G}_{24}$ is self-dual, we know that $(a, b) \equiv 0 \pmod{2}$. Thus $2(a, b) \equiv 0 \pmod{4}$, proving the theorem. □

Corollary X.43: The minimum weight of $\mathcal{G}_{24}$ is 8.

Corollary X.44: The minimum weight of $\mathcal{G}_{23}$ is 7, thus $\mathcal{G}_{23}$ is perfect.

Proposition X.45: a) The weights of the words of $\mathcal{G}_{23}$ are

wt	0	7	8	11	12	15	16	23
#	1	253	506	1288	1288	506	253	1

b) The weights of the words of $\mathcal{G}_{24}$ are

wt	0	8	12	16	24
#	1	759	2576	759	1

Proof: $\mathbb{F}_2^{23}$ has $\binom{23}{4}$ words of weight 4. These must be covered (uniquely, as the code is perfect) by the balls around the code words of minimum weight 7. To achieve distance 3, this will happen by changing 3 bits of 1 to 0, thus each code word covers exactly $\binom{7}{3}$ words of weight 4. Thus there are $\binom{23}{4}/\binom{7}{3} = 253$ words of weight 7.

The $\binom{23}{5}$ words of weight 5 are in balls around words of weight 7 (each covering $\binom{7}{2}$ words), or around words of weight 8, each covering $\binom{8}{3}$ words). Thus there are

$$\frac{\binom{23}{5} - 253\binom{7}{2}}{\binom{8}{3}} = 506$$

words of weight 8. The argument for 11, 12 is similar, the larger degrees follow from the fact that the all-1 word is in the code.
b) As the weights in $\mathcal{G}_{24}$ must be $\equiv 0 \pmod 4$, words of weight 7, 11, and 15 will increase by 1. □

By X.39, $\mathcal{G}_{23}$ defines a $4-(23,7,1)$ design and $\mathcal{G}_{24}$ a $5-(24,8,1)$ design. The latter has 759 blocks and is called the *Witt design*.

Automorphisms

We now consider the linear automorphism group of $\mathcal{G}_{24}$, considered as a subgroup of S_{24}: For this we return to the definition of $\mathcal{G}_{23}$ and label the 24 indices by $0,1\ldots,22,\infty$, that is $\mathrm{PG}(1,23)$. The cyclicity of $\mathcal{G}_{23}$ gives that the map $x \mapsto x+1 \bmod 23$ (leaving ∞ fixed) is a code automorphism. This corresponds to the permutation

$$p_1 = (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23)$$

On polynomials dividing $x^{23}-1$, the map $x \mapsto -x^{-1}$ exchanges the two factors of degree 11 (and thus $\mathcal{G}_{23}$ with the cyclic code defined by the other factor $\hat{g}(x)$ of degree 11). By definition of cyclic codes, $\frac{x^{23}-1}{g(x)} = (x-1)\hat{g}(x)$ defined the check matrix H of $\mathcal{G}_{23}$. But the check matrix $\hat{H}$ for $\mathcal{G}_{24}$ then is obtained from H (by adding a column 0 and an all-1 row).

As $\mathcal{G}_{24}$ is self-dual, this implies that the map $x \mapsto -x^{-1}$ must be an automorphism of $\mathcal{G}_{24}$, which on $0,1,\ldots,22,\infty$ (defining $1/0=\infty$) acts as

$$p_2 = (1,24)(2,23)(3,12)(4,16)(5,18)(6,10)(7,20)(8,14)(9,21)(11,17)(13,22)(15,19)$$

(e.g. 7 maps to 20 as $-1/6 \equiv 19 \pmod{23}$).

(Given these permutations, one also could verify directly that p_1 and p_2 preserve $\mathcal{G}_{24}$.)

Lemma X.46: The group $PSL_2(p)$ is generated by the automorphisms $x \mapsto x+1$ and $x \mapsto -x^{-1}$.

Proof: $SL_2(p)$ is generated by the matrices

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

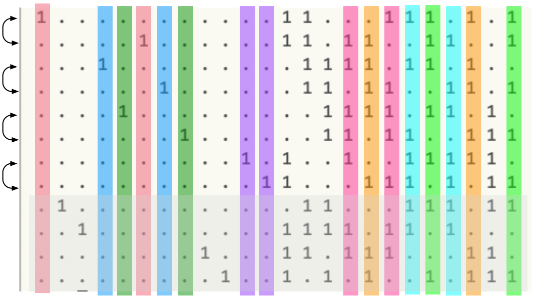
□

We thus know that the automorphism group of $\mathcal{G}_{24}$ must contain a subgroup isomorphic to $PSL_2(23)$.

In fact there are further automorphisms. Consider the permutation

$$p_3 = (1,6)(4,7)(5,8)(11,12)(16,18)(17,22)(19,21)(20,24)$$

One can verify easily, that this permutation of entries, applied to the rows of the generator matrix, will cause a swap of rows: Consider the generator matrix with rows reordered as 1, 6, 4, 7, 5, 8, 11, 12, 2, 3, 9, 10 in Figure X.2. Now the permutation



```
1 . . . . . . . . . . . 1 1 . . . 1 1 1 . 1 . 1
. . . . . 1 . . . . . . 1 1 . 1 1 . . 1 1 . . 1
. . . 1 . . . . . . . . . 1 1 1 1 . 1 1 . 1 . .
. . . . . . 1 . . . . . . 1 1 . 1 1 . . 1 1 . 1
. . . . 1 . . . . . . . . . 1 1 1 1 . 1 1 . 1 .
. . . . . . . 1 . . . . . . 1 1 . 1 1 . . 1 1 1
. . . . . . . . . . 1 . 1 . . 1 . . 1 1 1 1 1 .
. . . . . . . . . . . 1 1 . . . 1 1 1 . 1 . 1 1
. 1 . . . . . . . . . . . 1 1 . . . 1 1 1 . 1 1
. . 1 . . . . . . . . . 1 1 1 1 . 1 1 . 1 . . .
. . . . . . . . 1 . . . 1 1 . 1 1 1 . . . 1 1 .
. . . . . . . . . 1 . . 1 . 1 . 1 . . 1 . 1 1 1
```

Figure X.2: The column permutation p_3

of columns, given by p_3 will swap the first four row pairs, and fix the bottom four rows and thus also preserve the code $\mathcal{G}_{24}$.

The group $M_{24} = \langle p_1, p_2, p_3 \rangle$ generated by these permutations is in fact the full automorphism group of $\mathcal{G}_{24}$. It is called the *Mathieu group* of degree 24.

How large is it? Since $|PSL_2(23)| = 6072$, $|p1 \cdot p3| = 15$, $|p2 \cdot p1 \cdot p3| = 21$, we know that it must have order at least 212520 (which already is surprisingly large).

But using standard computer tools (such as GAP), we can calculate its order as $244823040 = 2^{10}3^3 5 \cdot 7 \cdot 11 \cdot 23$. This group is one of the sporadic simple groups. It acts quintuply transitive on the 24 points, and the underlying combinatorial structure of the Witt design can be used to investigate its structure.

We just note a few of its subgroups.

M_{23} The point stabilizer $\mathrm{Stab}_{M_{24}}(24)$ is a (sporadic) simple group of order 10200960, quadruply transitive.

M_{22} The 2-point stabilizer $\mathrm{Stab}_{M_{24}}(23,24)$ is a (sporadic) simple group of order $|M_{24}|/(23\cdot 24) = 443520$, triply transitive on the points $\{1,\dots,22\}$.

$P\Gamma L_3(4)$ The stabilizer of a 3-element set $\mathrm{Stab}_{M_{24}}(\{22,23,24\})$ acts faithfully on the points $\{1,\dots,21\}$ as a group of order $|M_{24}|/\binom{24}{3} = 120960$, isomorphic to $P\Gamma L_3(4)$. The permutation action on the 21 points stems from the action of $\mathrm{GL}(3,4)$ on the $(4^3-1)/(4-1) = 21$ nonzero vectors of $PG(3,4)$.

M_{12} The stabilizer of an 12-element set that is the support of a code word of weight 12 acts on these 12 points as a quintuply transitive (sporadic) simple group M_{12} of order $|M_{24}|/2576 = 95040$.

M_{11} The point stabilizer $\mathrm{Stab}_{M_{12}}(12)$ is a quadruply transitive group of order 7920, its is the smallest sporadic simple group.

M_{10} A point stabilizer in M_{11} is a group of order 720, that has a normal subgroup of index 2 that is isomorphic to A_6. It acts on this A_6 as an outer automorphsim that does not come from S_6.

The simple groups M_n, $n \in \{11,12,22,23,24\}$ were all originally discovered by Emile Mathieu (1835-1890) as groups of permutations, without the connection to the Golay codes. They are, apart from symmetric and alternating groups, the only k-transitive groups for $k \geq 4$.

We finally mention another connection of the Golay codes to algebra, number theory, and geometry:

Take the set Λ_{24} of vectors in $\mathbb{R}^{24}$ that have the form

$$\frac{1}{\sqrt{8}}(2c+4x) \qquad \text{and} \qquad \frac{1}{\sqrt{8}}(o+2c+4y) \qquad \text{where}$$

- o is the all-one vector
- $x,y \in \mathbb{Z}^{24}$ such that $\sum x_i \equiv 0 \pmod 2$, $\sum y_i \equiv 1 \pmod 2$.
- $c \in \mathcal{G}_{24}$ (interpreted as 0/1-vector).

The scalar factor $1/\sqrt{8}$ is only there to rescale norms. Considering the remaining vectors modulo 4, we have the Golay code scaled by 2, as well as a displaced version of it.

This set of vectors is closed under addition and subtraction, it thus is an $\mathbb{Z}$-lattice[2], called the *Leech lattice*. It is a closest sphere packing in 24-dimensional space. Its automorphism group gives rise to the sproradic simple group Co_1.

[2] One of the other usages of the word lattice in mathematics, no relation to posets

Chapter

XI Algebraic Graph Theory

We have encountered some concepts of graphs in prior chapters of this book. With automorphism groups we have seen connections between graphs and their automorphism groups. In this chapter we will look at techniques from (linear) algebra that can be used to analyze particular interesting classes of graphs.

We start with two examples of classes:

Example: Let $Z_n = \{0, \ldots, n-1\}$ the (additive) group modulo n, and $C \subset Z_n \setminus \{0\}$ such that $-C = C$. The *circulant graph* $X(Z_n, C)$ has vertices $V = Z_n$ and edges $(i, j) \in E$ if and only if $i - j \in C$.

Clearly $S_n \geq \mathrm{Aut}(X(Z_n, C)) \geq D_{2n}$.

Example: Let $v, k, i \in \mathbb{Z}_{>0}$ with $v \geq k \geq i$ and let $\Omega = \{1, \ldots, v\}$.

The *Johnson graph* $J(v, k, i)$ is the graph with vertex set $X = \{S \subset \Omega \mid |S| = k\}$, and $(S, T) \in E$ if and only if $|S \cap T| = i$.

An example is the Petersen graph $J(5, 2, 0)$ – see Exercise **??**.

Lemma XI.1: If $v \geq k \geq i$ then $J(v, k, i) \cong J(v, v-k, v-2k+i)$.

Proof: The isomorphism is given by mapping every set to its complement. □

Definition XI.2: A graph $\Gamma = (X, E)$ is called *transitive* if its automorphism group acts transitively on the vertices.

An *arc* (1-arc) in a graph is an ordered pair of adjacent vertices. A graph is *arc-transitive* if its automorphism group acts transitively on the set of arcs.

A graph is *distance transitive* if any pair of vertices u, v with distance (length of the shortest connecting path) d can be mapped under the automorphism group to any other pair of vertices of the same distance.

Clearly every distance transitive graph is arc transitive

Lemma XI.3: a) A transitive graph Γ is arc transitive if and only if for a vertex v its stabilizer acts transitively on the neighbors of v.

b) The graphs $J(v,k,i)$ are arc transitive.

Proof: a) Let (u,v) an arc and (a,b) another arc. As the graph is transitive there is $\alpha \in \mathrm{Aut}(\Gamma)$ such that $\alpha(b) = v$. Clearly a must be mapped to a neighbor w of v. But then there exists $\beta \in \mathrm{Stab}(v)$ such that $\beta(w) = u$, implying that $\alpha\beta$ maps (a,b) to (u,v).

b) The graph is clearly transitive.

Consider the vertex $\{1,\ldots,k\}$. Its stabilizer is $S_k \times S_{n-k}$. It clearly can map any two sets which intersect $\{1,\ldots,k\}$ in i vertices to each other. □

Lemma XI.4: a) The graph $J(v,k,k-1)$ is distance transitive.

b) The graph $J(2k+1,k+1,0)$ is distance transitive.

The proof is exercise **??**.

XI.1 Strongly Regular Graphs

We can interpret the concept of a transitive automorphism group as indicating that a graph "looks the same from every vertex". We now consider a condition that demands this property without enforcing transitivity.

Remember that we call a graph *regular* if every vertex has the same degree.

Definition XI.5: Let Γ be a regular graph that is neither complete, nor empty. Then Γ is called *strongly regular* with parameters (n,k,a,c) if it has n vertices, is regular of degree k, and every pair of adjacent vertices has a common neighbors and any pair of (distinct) nonadjacent vertices has c common neighbors.

Example: The cycle of length 5 is strongly regular with parameters $(5,2,0,1)$.

The Petersen Graph is strongly regular with parameters $(10,3,0,1)$.

If Γ is strongly regular with parameters (n,k,a,c) it is easily seen that the complement $\overline{\Gamma}$ is strongly regular with parameters

$$(n, n-k-1, n-2-2k+c, n-2k+a).$$

We will from now on assume that both Γ and $\overline{\Gamma}$ are connected. Since there cannot be a negative number of adjacent vertices this implies that the parameters of such a graph must satisfy

$$n \geq 2k-c+2, \qquad n \geq 2k-a.$$

Note XI.6: A strongly regular graph does *not* need to be (vertex-)transitive. The smallest examples are the so-called *Chang graphs*, three strongly regular graphs, all with parameters $(28,12,6,4)$.

A further dependency is given by:

Proposition XI.7: The parameters of a strongly regular graph satisfy that

$$k(k-a-1)=(n-k-1)c.$$

Proof: For a vertex x, consider the edges $\{y,z\}$ with y adjacent to x, but z not. There are k vertices adjacent to x, and each of them has k neighbors. One of these is x, and there are a who are also neighbor of x, so there are $k-a-1$ neighbors not adjacent to x.

On the other hand, there are $n-k-1$ vertices z not adjacent to x and different from x. For each such z there are c vertices y adjacent to x and z. □

Example: For $m\geq 4$, the Johnson graph $J(m,2,1)$ is strongly regular with parameters $n=\frac{m(m-1)}{2}$, $k=2(m-2)$, $a=m-2$, $c=4$.

Example: For a prime power q with $q\equiv 1 \pmod 4$ (ensuring that -1 is a square in $\mathbb{F}_q$), the *Paley graph* $P(q)$ has vertex set $\mathbb{F}_q$ with two vertices adjacent if their difference is a non-zero square in $\mathbb{F}_q$.

It is strongly regular with parameters $n-q$, $k=\frac{q-1}{2}$, $a=\frac{q-5}{4}$ and $c=\frac{q-1}{4}$.

Example: The *Clebsch graph* has as vertices the 16 subsets of $\{1,2,3,4,5\}$ of even size. To vertices are adjacent if their symmetric difference is of size 4. It is strongly regular with parameters $(16,5,0,2)$.

Another example class of strongly regular graphs is given by orthogonal Latin squares:

Let $\{A_i,\ldots,A_k\}$ a set of MOLS of order n. We define a graph $\Gamma=(X,E)$ with $X=\{(i,j)\mid 1\leq i,j\leq n\}$ and an edge given between (i,j) and (a,b) if one of the following conditions holds (compare Section VIII.9):

- $i=a$
- $j=b$
- For one of the latin squares A_x, its (i,j) entry and its (a,b) entry are equal.

Theorem XI.8: The graph defined by k MOLS of order n is strongly regular with parameters

$$(n^2,(n-1)(k+2),n-2+k(k+1),(k+1)(k+2)).$$

The proof is exercise **??**.

Note XI.9: IT is possible to generalize the concept of strongly regular to that of a *distance regular* graph in which the number of common neighbors of x,y depends on the distance of x and y.

XI.2 Eigenvalues

A principal tool for studying strongly regular graphs (and generalizations) is their adjacency matrix A with $A_{i,j}=1$ iff vertex i is adjacent to vertex j and 0 otherwise.

Note XI.10: A is a symmetric matrix, thus (by the spectral theorem) it has real eigenvalues and can be diagonalized by orthogonal transformations, respectively its eigenspaces for different eigenvalues are mutually orthogonal.

Lemma XI.11: If Γ is regular of degree k, then A has largest eigenvalue k with the all-one vector as associated eigenvector. k is the largest eigenvalue.

If Γ is connected, the multiplicity of k is one.

Proof: If Γ is regular, then the sum over every row of A is k, showing that the all-one vector is eigenvector for eigenvalue k.

Now let v be an eigenvector of A for eigenvalue $\lambda \geq k$. WLOG we assume that v is scaled such that its largest entry is 1. Let i be an index of this largest entry in v. Then the i-th entry e of Av (which by assumption must equal $\lambda \geq k$) is the sum of the k entries of v, at indices of the vertices adjacent to vertex i. All of these entries are ≤ 1, thus $e \leq k$ with equality only if all these entries are 1.

This shows that k is the largest eigenvalue. Furthermore, for v to be an eigenvector, its entries at all indices neighboring i must be 1. By induction this shows that the entries of v at indices of distance $j \geq 1$ from i must be 1, showing that for a connected Γ the multiplicity of k is one. □

Lemma XI.12: Let Γ be a graph with adjacency matrix A. Then the i, j entry of A^k is the number of walks (like path but we permit to walk back on the same route) of length k from i to j.

Proof: By induction over k it is sufficient to consider a product AB with $B = A^{k-1}$. The i, j entry of this product is $\sum_s a_{i,s} b_{s,j}$, which is the sum, over the neighboring vertices s of i, of the number of walks of length $k - 1$ between s and j. □

In a strongly regular graph with parameters (n, k, a, c), the number of walks of length 2 between two vertices i and j is

- k if $i = j$ (walk to neighbor and back),
- a if $i \neq j$ are adjacent (walk to common neighbor and then to other vertex),
- c if $i \neq j$ are not adjacent.

This means that

$$A^2 = kI + aA + c(J - I - A)$$

(where J is the all-one matrix). We write this in the form

$$A^2 - (a - c)A - (k - c)I = cJ$$

Note XI.13: The adjacency matrix for a strongly regular graph with parameters (v, k, a, a) satisfies – compare to Lemma IX.8 –

$$AA^T = A^2 = (k - a)I + aJ$$

and thus can be interpreted as incidence matrix of a symmetric $2-(v,k,a)$-design.

Let v be an eigenvector of A for eigenvalue λ, that is orthogonal to the all-1 vector. (By Note XI.10, this can be assumed.) Then

$$(\lambda^2-(a-c)\lambda)-(k-c))v = A^2v-(a-c)Av-(k-c)v = cJv = 0$$

because of the orthogonality of v and the columns of J. This implies that $\lambda^2-(a-c)\lambda)-(k-c)=0$.

Setting $\Delta=(a-c)^2+4(k-c)$, we thus get that

$$\lambda_\pm = \frac{(a-c)\pm\sqrt{\Delta}}{2}.$$

As $\lambda_+\cdot\lambda_- = c-k$ the assumption that $c<k$ implies that λ_+ and λ_- are nonzero with opposite sign.

Let $m_\pm$ be the dimensions of the respective eigenspaces. Then $m_++m_-=n-1$ and (as the trace of A, the sum of the eigenvalues, is zero) $m_+\lambda_++m_-\lambda_-=-k$.

We solve this system of equations (with signs corresponding) as

$$m_\pm = \mp\frac{(n-1)\lambda_\mp+k}{\lambda_+-\lambda_-}$$

and note that

$$(\lambda_+-\lambda_-)^2 = (\lambda_++\lambda_--4\lambda_+\lambda_- = (a-c)^2+4(k-c)=\Delta.$$

This yields the multiplicities

$$m_\pm = \frac{1}{2}\left((n-1)\mp\frac{2k(n-1)(a-c)}{\sqrt{\Delta}}\right),$$

which must be nonnegative integers, imposing a condition on the possible parameters.

We also note a converse result

Proposition XI.14: A connected, regular, graph Γ with exactly three distinct eigenvalues is strongly regular.

Proof: Suppose Γ is connected and regular. Then the valency (degree of each vertex) k is an eigenvalue, let λ,μ be the other two. Let A be the adjacency matrix and

$$M=\frac{1}{(k-\mu)(k-\lambda)}(A-\mu I)(A-\lambda I).$$

Then (using the orthogonality of eigenspaces), M has eigenvalues only 0 or 1 and the eigenspaces of λ and μ in the kernel.

Thus the rank of M is the multiplicity of the eigenvalue k. By Lemma XI.10 this multiplicity is one, if Γ is connected.

As the all-one vector is an eigenvector of A (and thus of M) we must have that $M = \frac{1}{n}J$. Thus J is a quadratic polynomial in A and A^2 is a linear combination of I, J, A which implies that Γ is strongly regular. □

We can interpret this situation also in different language: The matrices A, I, J generate a 3-dimensional $\mathbb{C}$-vector space $\mathfrak{A}$ that also is a ring under multiplication.

Such a set of matrices is called an *algebra*, this particular algebra $\mathfrak{A}$ is called the *Bose-Mesner algebra*.

The Krein Bounds

The eigenspaces of A clearly also are eigenspaces for the other two generators I, J of $\mathfrak{A}$ and thus for all of $\mathfrak{A}$. Call them V_1,V_+ and V_- and let E_i for $i \in \{1, +, -\}$ be the matrix projecting onto V_i (and mapping the other two eigenspaces to 0). They lie in $\mathfrak{A}$, as they can be constructed from $A - \lambda I$. Vice versa (as any operator can be written as linear combination of the projections to its eigenspaces) we have that $\mathfrak{A} = \langle E_1, E_+, E_- \rangle$ (as a vector space or as an algebra).

The E_i are *idempotent* (that is they satisfy that $x \cdot x = x$), they are *orthogonal* (that is $E_i E j = 0$ if $i \neq j$) and they decompose the identity $E_1 + E_+ + E_- = I$. (Such idempotents are part of a standard analysis of the structure of algebras.)

We now consider a different multiplication $\circ$, called the *Hadamard product* in which the i, j entry of $A \circ B$ is simply $a_{i,j} b_{i,j}$. As $0, 1$ matrices, A, I and J are idempotent under the Hadamard product, thus $\mathfrak{A}$ is closed under this product as well.

This means that

$$E_i \circ E_j = \sum_k q_{ij}^k E_k$$

But the E_i are positive definite, and the Hadamard product of positive definite matrices it positive definite as well[1].

Thus the $q_{i,j}^k$ must be positive. Using explicit expressions for the E_i in terms of A, I, J, one can conclude:

Theorem XI.15 (Krein bounds): Let Γ be strongly regular such that Γ and its complement are connected and let k, r, s be the eigenvalues of Γ. Then

$$\begin{aligned}(r+1)(k+r+2rs) &\leq (k+r)(s+1)^2,\\ (s+1)(k+s+2rs) &\leq (k+s)(r+1)^2.\end{aligned}$$

[1] This can be shown by observing that $A \circ A$ is a principal submatrix of $A \otimes A$.

Association Schemes

The concept of the Bose-Mesner algebra can be generalized to higher dimensions:

Definition XI.16: A set of square 0/1 metrics $A_0, A_1, \ldots, A_d$ is called an *Association Scheme* with d classes if

1. $A_0 = I$.
2. $A_1 + A_1 + \cdots + A_d = J$ the all-one matrix.
3. For all i there is a j such that $A_i^T = A_j$.
4. $A_iA_j = \sum_k p_{i,j}^k A_k$ for nonnegative integers $p_{i,j}^k$. The $p_{i,j}^k$ are called the *intersection numbers*, *parameters* or *structure constants* of the scheme.

Example: Let G be a transitive permutation group on $\Omega = \{1, \ldots n\}$. Then G acts on $\Omega \times \Omega$, the orbits being called *orbitals*. For each orbit we define a 0/1 matrix $A = (a_{x,y})$ such that $a_{x,y} = 1$ if and only if (x, y) is in the orbit.

As one orbit consists of pairs (i, i) we may assume that $A_0 = I$, and clearly $\sum A_i = J$. Furthermore, is (i, j) and (x, y) are in the same orbit, clearly also (j, i) and (y, x) are in the same orbit, showing that $A_i^T = A_j$.

Finally, let $A_i = (a_{x,y})$ and $A_j = (b_{x,y})$. Then the (x, y) entry of A_iA_j is $\sum_z a_{x,z}b_{z,y}$, that is the number of z's such that (x, z) is in the first orbit and (z, y) is in the second orbit, thus it is a nonnegative integer. Call this number $q_{i,j}(x, y)$.

Furthermore, if (x', y') lies in the same orbit as (x, y), that is for $g \in G$ we have that $x^g = x'$ and $y^g = y'$, we have that the z' such that (x', z') is in the first orbit and (z', y') in the second are exactly the elements of the form z^g. Thus the values of $q_{i,j}(x, y)$ are constant on the orbit of (x, y), that is we can set $p_{i,j}^k = q_{i,j}(x, y)$ for an arbitrary (x, y) in the k-th orbit.

XI.3 Moore Graphs

Definition XI.17: The *diameter* of a graph is the maximum distance between two vertices.

The *girth* of a graph is the length of the shortest closed path.

A connected graph with diameter d and girth $2d + 1$ is called a *Moore graph*.

Example: The 5-cycle and the Petersen graph both are both examples of Moore graphs.

Lemma XI.18: Let Γ be a Moore graph. Then Γ is regular.

Proof: Let v, w be two vertices at distance d and P a path connecting them. Consider a neighbor u of v that is not on P. Then v must have distance d from w, as otherwise $w - u - v - w$ would be a cycle of length $< 2d + 1$. Then there is a unique path from u to w (as otherwise the differing paths would give a cycle), it will go through one neighbor of w. Different choices of u yield (again cycle length) different neighbors of w and thus w has at least as many neighbors as v. By symmetry thus v and w must have the same degree. Since we have seen that all neighbors of v, that are not on P also have distance d from w, by the same argument they also have the same degree, as will be all neighbors of w not on P.

Now Let C be a cycle of length $2d+!$ involving a vertex v. By stepping in steps of length d along the circle the above argument shows that all vertices on C must have the same degree. If x is a vertex not on C take a path of length i from x to C. By taking $d-i$ steps along C we have that x is at distance d from a vertex on C and thus has the same degree. □

Lemma XI.19: A Moore graph Γ of diameter d and vertex degree k has $1+k((k-1)^d-1)/(k-2)$ vertices.

Proof: Vertex degree k implies (easy induction) that there are at most $k(k-1)^{i-1}$ vertices as distance $i \geq 1$ from a chosen vertex v.

Suppose that w is at distance i. Then there is one neighbor of w art distance $i-1$, and no other at distance i (as otherwise we could form a cycle using w and this neighbor). Thus $k-1$ neighbors of w must be at distance $i+1$ from v, which shows that there are at least that many neighbors.

Thus there are in total

$$1+k+k(k-1)+k(k-1)^2+\cdots+k(k-1)^{d-1}=1+k\frac{(k-1)^d-1}{k-2}$$

vertices. □

Consider a Moore graph of diameter 2 Then G has $1+k+k(k-1)=k^2+1$ vertices and girth 5. Thus two adjacent vertices have no common neighbor, and two non-adjacent vertices have exactly one common neighbor.

Thus Γ is strongly regular and its adjacency matrix A satisfies

$$A^2=kI+(J-I-A),$$

and A has eigenvalues k and

$$\lambda_\pm=\frac{-1+\sqrt{4k-3}}{2}.$$

If $4k-3$ is not a square, then $\lambda_\pm$ will be irrational and thus the multiplicities $m_\pm$ must be equal (and thus be $(n-1)/2=k^2/2$.

As A has only zeroes on the diagonal, it has trace 0. But the trace is the sum of all eigenvalues. Thus (with the irrationalities canceling out)

$$0=k+\frac{k^2}{2}\left(\frac{-1}{2}+\frac{-1}{2}\right)$$

and thus $k = 2$. A graph of valency 2 must be a cycle, thus the only possibility here is the 5-cycle.

If $4k - 3$ is a square, it must be the square of an odd number, $4k - 3 = (2s + 1)^2$. We solve for $k = s^2 + s + 1$ and the eigenvalues of A are k, s, and $-s - 1$, the latter two with multiplicities f and g such that $f + g = n - 1 = k^2$. The trace of A is $k + fs + g(-s - 1) = 0$. These are two linear equations in f and g, which gives

$$f = \frac{s(s^2 + s + 1)(s^2 + 2s + 2)}{2s + 1}.$$

This number must be an integer. We expand the numerator as

$$s^5 + 3s^4 + 5s^3 + 4s^2 + 2s = \frac{1}{32}\left((16s^4 + 40s^3 + 60s^2 + 34s + 15)(2s + 1) - 15\right).$$

We conclude that $2s + 1$ must divide 15, which implies that $s = 0, 1, 3, 7$. This corresponds to $k = 1, 3, 7, 57$, respectively $n = 2, 10, 50, 3250$.

The choice of $n = 2$ vertices is spurious, thus a Moore graph of diameter 2 and valency k can exist only for $k = 2, 3, 7, 57$ and associated $n = 2, 10, 50, 3250$.

We already have seen the 5-cycle for $k = 2$. For $k = 3$ the Petersen graph is an example. There is also an example for $k = 7$, the *Hoffman-Singleton* graph. The case $k = 57$, $n = 3250$ is open.

It has been shows that these three, possibly four, cases are the only Moore graphs of diameter 2 and that for diameter $d \geq 3$ the $2d + 1$ cycle is the only Moore graph.

Bibliography

[AU17] Maria Axenovich and Torsten Ueckerdt, *Lecture notes combinatorics*, https://www.math.kit.edu/iag6/lehre/combinatorics2017s/media/script.pdf, 2017.

[Bab16] László Babai, *Graph isomorphism in quasipolynomial time [extended abstract]*, STOC'16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, ACM, New York, 2016, pp. 684–697.

[Cam76] Peter J. Cameron, *Transitivity of permutation groups on unordered sets*, Math. Z. **148** (1976), no. 2, 127–139.

[Cam94] ______, *Combinatorics: topics, techniques, algorithms*, Cambridge University Press, Cambridge, 1994.

[CQRD11] Hannah J. Coutts, Martyn Quick, and Colva M. Roney-Dougal, *The primitive permutation groups of degree less than 4096*, Comm. Algebra **39** (2011), no. 10, 3526–3546.

[CvL91] P. J. Cameron and J. H. van Lint, *Designs, graphs, codes and their links*, London Mathematical Society Student Texts, vol. 22, Cambridge University Press, Cambridge, 1991.

[Eul15] Leonard Euler, *Decouverte d'une loi tout extraordinaire des nombres par rapport a la somme de leurs diviseurs*, Opera Omnia, I.2 (Ferdinand Rudio, ed.), Birkhäuser, 1915, pp. 241–253.

[Fel67] William Feller, *A direct proof of Stirling's formula*, Amer. Math. Monthly **74** (1967), 1223–1225.

[GKP94] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik, *Concrete mathematics*, second ed., Addison-Wesley Publishing Company, Reading, MA, 1994.

[GNW79] Curtis Greene, Albert Nijenhuis, and Herbert S. Wilf, *A probabilistic proof of a formula for the number of Young tableaux of a given shape*, Adv. in Math. **31** (1979), no. 1, 104–109.

[GR01] Chris Godsil and Gordon Royle, *Algebraic graph theory*, Graduate Texts in Mathematics, vol. 207, Springer-Verlag, New York, 2001.

[Hal86] Marshall Hall, Jr., *Combinatorial theory*, second ed., Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons, Inc., New York, 1986, A Wiley-Interscience Publication.

[Jac90] C.G.J. Jacobi, *De investigando ordine systematis aequationum differentialum vulgarium cujuscunque*, Gesammelte Werke, Fünfter Band (Karl Weierstrass, ed.), G. Reimer, Berlin, 1890, pp. 193–216.

[Knu98] Donald E. Knuth, *The art of computer programming. Vol. 3*, Addison-Wesley, Reading, MA, 1998, Sorting and searching, Second edition [of MR0445948].

[LN94] Rudolf Lidl and Harald Niederreiter, *Introduction to finite fields and their applications*, Cambridge University Press, Cambridge, 1994.

[LTS89] C. W. H. Lam, L. Thiel, and S. Swiercz, *The nonexistence of finite projective planes of order* 10, Canad. J. Math. **41** (1989), no. 6, 1117–1123.

[Rei84] Philip F. Reichmeider, *The equivalence of some combinatorial matching theorems*, Polygonal Publ. House, Washington, NJ, 1984.

[Sta99] Richard P. Stanley, *Enumerative combinatorics. Vol. 2*, Cambridge Studies in Advanced Mathematics, vol. 62, Cambridge University Press, Cambridge, 1999, With a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin.

[Sta12] ______, *Enumerative combinatorics. Volume 1*, second ed., Cambridge Studies in Advanced Mathematics, vol. 49, Cambridge University Press, Cambridge, 2012.

[vLW01] J. H. van Lint and R. M. Wilson, *A course in combinatorics*, second ed., Cambridge University Press, Cambridge, 2001.

[Wie64] Helmut Wielandt, *Finite permutation groups*, Academic Press, 1964.

[Zei84] Doron Zeilberger, *Garsia and Milne's bijective proof of the inclusion-exclusion principle*, Discrete Math. **51** (1984), no. 1, 109–110.

Index