Counting Networks

An Introduction with Specific Interest in Computational Molecular Biology

(Dietmar Cieslik, University of Greifswald)

Contents

Chapter 1

Introduction

> Nothing in biology makes sense except in the light of evolution.
>
> Theodosius Dobzhansky

> The Universe is a grand book which cannot be read until one first learns to comprehended the language and become familiar with the characters in which it is composed. It is written in the language of mathematics.
>
> Galileo Galilei

Configurations of nodes and connections occur in many applications of real world problems, they are modeled by combinatorial structures which are usually called graphs. The script attempts to describe the world of graph theory with emphasis on counting of specific kinds of graphs, and in particular trees. A specific focus will be given to its applications in biology.[1]

Graphs lend themselves as natural models of transportation as well as communication networks. They are among the most basic of all mathematical structures. Correspondingly, they have many different versions, representations and incarnations. The fact is that graph theory serves as a mathematical model for any system involving a binary relation.
Numerous challenging problems in graph theory have attracted the attention and imagination of scientists in the area of "network science", where a network will be a graph with some additional properties. In particular, we restrict the structure of the graph, label some vertices, give graphs an order, consider collection of trees, add functions on graphs, ...

Trees were first used in 1847 by Kirchhoff in his work on electrical networks. They were later redeveloped and named by Cayley in order to enumerate different isomers

[1] Networks cover a wide range in form of metabolic networks, protein-protein interactions, genetic regulatory networks, food webs and several more, compare [179].

of specific chemical molecules. Since that time, enumerative methods for counting various classes of graphs, including trees, have been developed, but are still far from completely scientific.

As it became accepted that evolution was to be understood in terms of Mendelian genetics and Darwinian natural selection, so too it became clear that this understanding could not be sought only at a qualitative level.[2] A fundamental problem is the reconstruction of species' evolutionary past, which is called the phylogeny of those species. The underlying principle of phylogeny is to try to group "living entities" according to their level of similarity. Trees are widely used to represent evolutionary relationships.[3] In biology, for example, the dominant view of the evolution of life is that all existing organisms are derived from some common ancestor and that a new species arises by the splitting of one population into two or more populations that do not cross-breed, rather than from the mixing of two populations into one. Here, the high level history of life is ideally organized and displayed as a tree. This was already seen by Darwin in 1859, [63]:

> The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of extinct species... The limbs divided into great branches, and these into lesser and lesser branches, were themselves once, when the tree was small, budding twigs; and this connexion of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups... From the first growth of the tree, many a limb and branch has decayed and dropped off, and these lost branches of various sizes may represent those whole orders, families, and genera which have now no living representatives, and which are known to us only from having been found in a fossil state... As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all a feebler branch, so by generation I belive it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications.

Darwin spoke of "descent with modification", which is the central phrase of biological evolution, it refers to a genealogical relationship of species through time. These relationships are described in a phylogenetic tree, which can therefore be thought of as a central metaphor for evolution, providing a natural and meaningful way to order data and containing an enormous amount of evolutionary information within its branches.[4]

[2]Nowak [181]: "Evolution has became a mathematical theory."

[3]Thorley and Page [185]: "The holy grail of phylogenetics is the reconstruction of the one true tree of life."

[4]Note that in Darwin's fundamental book *The origin of species* [63] there is exactly one figure,

Note, that there are several difficulties. Darwin's evolutionary tree is neither obvious, nor easy to find. In a letter to Huxley he wrote: "The time will come, I believe, though I shall not live to see it, when we shall have fairly true genealogical trees of each great kingdom of Nature." The main problem is that all known exact algorithms need exponential time; frequently more time than that taken by the evolutionary processes themselves.[5]
Moreover, we should note that the collection of all living entities is much more difficult, Doolittle [68]:

> It has been argued that the "Tree of Life" is perhaps really a "Web of Life", as mechanisms such as hybridization, recombination and swapping of genes probably play a role in evolution.

In the widest sense, a classification scheme may represent simply a convenient method for organizing a large set of data so that the retrieval of information may be made more efficiently. In this sense, classification is the beginning of all science.[6]

A classification is the formal naming of a group of individuals N.[7] The following statements are pairwise equivalent.

- $\mathcal{C}$ is a classification for N.
- $\mathcal{C}$ represents a rooted N-tree.
- $\mathcal{C}$ consists of a series of partitions for N which become finer and finer.

More and more discrete structures are used in genetics, biochemistry, evolution, agriculture, experimental design and other parts of modern biology. Here, the results are very powerful and the research frontier are perhaps more accessible than in some more traditional areas of applied mathematics.[8]

We will embed this question in the context of counting graphs in general. It is not the purpose of this script to provide a complete survey of counting methods for trees and related networks, but it should summarizes all facts which are important in biological applications. In particular, we will focus on the counting of specific classes

---

and this shows the description of the evolutionary history by a tree.
Historically, this was a new idea: The concept of species having a continuity through time was only developed in the late 17th century; higher life forms were no longer thought to transmute into different kinds during the lifetime of an individual. It took over 150 years from the development of this concept before a rooted tree was proposed by Darwin.

[5] Penny [189]: "The real evolution runs faster than the calculation can follow it." But nature
- Performs many computations in parallel; and
- Does not check all possibilities.

[6] Classifications has played a central role in other fields too. In particular, the classification of the elements in the periodic table, given by Mendeleyev 150 years ago, has had a profund impact on the understanding of the structure of atoms. Another example in astronomy is the classification of stars in the Hertzsprung-Russel plot, which has strongly affected theories of stellar evolution.

[7] Everitt [81]: "Naming is classifying."

[8] To radically simplify, in the cases, human beings and behaviour may be classified into classes named by *low*, *medium* and *high*.

of graphs, which are used in phylogeny. On the other hand, the present script will discuss several relatives which are counts specific structures in graph theory. All needful terms and definitions will be included. Often several different solutions to the same problem will be provided so that the reader has an opportunity to become acquainted with a variety of methods.

There is something to be said for regarding enumerative methods in mathematical sciences. Although Euler counted certain types of graphs, the major activity in graphical enumeration launched in the mid of 19'th century, starting with Kichhoff's matrix-tree theorem and Cayley's spanning tree enumeration. In further investigations of discrete objects this question are become more and more interest. Discrete Mathematics is devoted to the study of discrete objects, these are a finite or countable set of distinct and unconnected elements; which are separated and discontinuous. It is used whenever objects are counted, when relationships between finite or countable sets are studied, and when processes involving a finite number of steps are analyzed. A main reason for the growth of the importance of discrete mathematics is that information can be stored and manipulated by computing machines in a discrete fashion. Graph counting is a well-established subject in discrete mathematics, but it is more than only simple enumeration, it also include

- Questions whether certain objects exist. More exactly, for a problem Π with S as set of solutions

  1. Decide whether $S = \emptyset$ or not.
  2. Find a member of S.
  3. Find all members of S.
  4. Calculate $|S|$.

  There are the following implications:

  $$3 \Rightarrow \left\{ \begin{array}{c} 4 \\ 2 \end{array} \right\} \Rightarrow 1.$$

  Furthermore there is the interest to solve these questions algorithmically.

- Generally speaking, optimization is concerned with finding an object that fulfills some predetermined requirements and minimizes one or more given objective functions. This is quite common model which covers a wide variety of problems.
  Combinatorial reasoning underlies all considerations of discrete programming problems: Analysis of the speed and logical structure of problems entails combinatorial mathematics.

- Combinatorics is the "Art of Counting".
  Many of the problems can be phrased in the form "How many ways...?", "Does there exist an object such that...?" or "Can we construct... ?"

Combinatorics is a modern mathematical field, which now has rapidly increasing research activity with applications in many areas of science. More and more combinatorial structures are used in genetics, biochemistry, evolution, agriculture, experimental design and other parts of modern biology. Here, the results are very powerful and the research frontiers are perhaps more accessible than in some more traditional areas of applied mathematics.

- Knowing the number of graphs with a particular property may enable to estimate the length of an algorithmic calculation. Here we describe the border between linear, quadratic, cubic, ..., polynomial, exponential and super-exponential growing of (time-) amount to deal with graphs and to understand the complexity of algorithms.[9] Moreover, the investigations about counting graphs are interesting in view of the "borderline" between $\mathcal{P}$ and $\mathcal{NP}$.
  In particular, counting graphs and trees are helpful to see why the reconstruction of evolutionary processes are hard in the sense of complexity.

- Counting problems are closely associated with probability. Indeed, any problem of the kind "How many objects are there which ..." has the closely related form "What fraction of all objects ...", which in turn can be posed as "What is the probability that a randomly chosen object ...?" when expressed in terms of the theory of probability. In this sense Laplace defined probability.
  There is a deep interplay between graph enumeration and the theory of random graphs. Here the asymptotic behavior of the counting functions plays an important role.

- It will be give the possibility to introduce qualitative parameters in graph theory. In particular, we can exactly describe what terms "dense", "sparse", "rare" and "almost all/no" does mean.
  On the other hand, we must be carefully, what we consider: Concrete objects or structural properties. In other terms, what we count: Distinguished graphs or isomorphic classes.[10]

- We also investigate graph theory as a part of geometry and topology. In view of the description of relations and polyhedra as specific graphs it will be possible to count such objects.

The present script will reflect about all of these topics.
For readers which are interested in further facts about counting, generating and storing graphs and trees we give a list of books which continue our considerations and give several new hints and investigations. In particular:

1. Harary: Graph Theory; [121].[11]

[9] For centuries almost all mathematicians believed that any mathematical problem could be solved using an algorithm. However, this view has been questioned over the course of time as more and more problems have arisen for which no algorithmic solution has been found or for which the algorithms are too difficult to deal.

[10] For instance, for 10 vertices the number of unlabeled trees is 106, but there are 100 millions of labeled trees.

[11] Including a big list of references about counting graphs.

2. Harary, Palmer: Graphical Enumeration; [122].[12]

3. Martin: Counting: The Art of Enumerative Combinatorics; [173].

4. Stanley: Enumerative Combinatorics; vol 1 and 2, [226], [227].

We will only use methods which are present in the first classes of undergraduate studies. Further studies need more mathematics than are given here; in particular methods of higher algebra, which are beyond the scope of the present script[13], see the pioneering work by Polya [192].
A background in in elementary set theory, mathematical logic, linear algebra, probability theory and calculus is assumed. If several facts about discrete and combinatorial mathematics are not present for the reader, this book includes an appendix with the most of the important results in short reviews.[14] Additionally, we will give references for further reading.
As an textbook the present script contains several exercises, but there are vast differences in level of these questions: a) Exercises which are straightforward from the text; b) Problems which need a longer discussion; and c) Open tasks. Many enumerating problems are still unsolved.

The present script originates from lectures, seminars, and exercises given by the author at Greifswald University (Germany), the University of Bielefeld (Germany), the Massey University, Palmerston North (New Zealand) and the University of Science, Hanoi (Vietnam). It is the extended version of the book "Counting Graphs - An Introduction with specific Interest in Phylogeny" [55].

I hope the present book, as a mixture of textbook, handbook and monograph, will be balanced in the sense of understanding and research interest.

Acknowledgments.

I thank everybody which gave me helpful advice on how to write this book: C. Bandt (Greifswald), K.-E. Biebler (Greifswald), A. Dress (Shanghai), W. Fitch (Irvine), R. Graham (La Jolla), M. Haase (Greifswald), A. v.Haeseler (Wien), J. Haß (Dresden), M.D. Hendy (Palmerston North/Dunedin), E. Herrholz (Neubrandenburg), A. Kemnitz (Braunschweig), V. Liebscher (Greifswald), S. Matuszewski (Wien), R. Schimming (Potsdam) and M. Steel (Christchurch).
I thank my students S. König and C. Malsch for proof-reading the manuscript.

Almost nothing in this script is original, except perhaps by mistake. The author accepts full responsibility for any mistakes that may have occur. He is absolutely

[12] Which is the mostly devoted monograph about determining the number of graphs until today.
[13] although we will give several hints of this approach
[14] In a restricted sense the present script can be read as an introduction into Discrete Mathematics with focus in Graph Theory.

interested in all hints which decrease the number of errors, which show new facts and exercises, and which give further applications.[15]

[15] the reader can understand each mistake as a subtle form of an exercise.

Chapter 2

Sets

Set theory, founded by Cantor in the second half of the 19th century, has profoundly transformed mathematics. It is the foundation of modern mathematics.

2.1 Basics

A set is a collection of distinct objects. Usually, but not exclusive, we refer to the objects in a set as the elements. If S is a set and x is an element which belongs to S, we write $x \in S$. A set S' is a subset of a set S, written $S' \subseteq S$, if every element of S' is also an element of S. A proper subset is a subset with fewer elements than the whole set. Two sets S and S' are equal if they contain the same elements. In other words

$$S = S' \text{ if and only if } S \subseteq S' \text{ and } S' \subseteq S. \tag{2.1}$$

Two sets with no common elements are called disjoint.
The empty set $\emptyset$ containing no element. The empty set is subset of any set S:

$$\emptyset \subseteq S. \tag{2.2}$$

The proof of this observation is not so simple. Do you see why?

In general, one cannot list the elements of a, in particular infinite, set. Nor it is practical to list the elements of a very large finite set. To determine a set of either kind we specify a property P shared by all of its elements and not belonging to any element not in the set:

$$S = \{x \in \mathcal{U} : x \text{ satisfies } P\}. \tag{2.3}$$

Then S designate the set of all elements for which the property P is true. P is called the defining property. Then the logical universe of discourse defines the set by all objects which posses an attribute.

With $|S|$ we denote the number of elements of the set S. We can count the elements of S by finding a bijection (a one-to-one map) between S and $\{1, \ldots, n\}$.[1]
At first several simple facts about counting of sets: Let R and S be sets in a universe U.

$|S^c| = |U| - |S|$.

$|R \setminus S| = |R| - |R \cap S|$.

$|R \cup S| = |R| + |S| - |R \cap S|$.

$|R \triangle S| = |R \cup S| - |R \cap S| = |R| + |S| - 2|R \cap S|$.

Consider the power set of a set:

$$\mathcal{P}(S) = \{S' : S' \subseteq S\}. \tag{2.4}$$

Each subset S' of a set $S = \{x_1, \ldots, x_n\}$ can be uniquely described by a 0/1-sequence $b_1, \ldots, b_n$ of length n:

$$b_i = \begin{cases} 1 & : \quad x_i \in S' \\ 0 & : \quad \text{otherwise} \end{cases}$$

Obviously, there are 2^n such sequences. This implies that the power set $\mathcal{P}(X)$ contains more elements than the set X itself.

Observation 2.1.1 $|\mathcal{P}(S)| = 2^{|S|}$.

2.2 Infinite Sets

The concept of infinity has always fascinated philosophers and theologians, but that was avoided or met with open hostility throughout most of the history of mathematics. Only within the last two centuries mathematicians dealt with it head on and accepted infinity as a number.
How can we count the elements of an infinite set? We have to compare the sets; that means we ask for a bijective mapping between these sets.

One dogma that we have to brush aside is the statement "A part is less than the whole". This is indisputably true for finite sets, but it loses its force when we try to apply it to infinite sets. Consider the following mapping:

$$f : I\!N \to I\!N : n \mapsto 2n. \tag{2.5}$$

This sets up a one-to-one correspondence between the set of natural numbers and a proper part of this set: the set of even numbers, which was already observed by Galilei.[2] In about 1888 Dedekind introduced the concept of infinite sets by the following definition.

[1] Here we assume that S is a finite set; later we will discuss the counting of infinite sets.

[2] Here is the story of Hilbert's hotel: It is a hotel with an infinite number of rooms. All the rooms are full, but more guests are waiting outside. We amke space by the following operation: the guest occupying room 1 moves to room 2, the occupant from room 2 moves to room 4, and so on, all the way down the line, an infinite number of newcomer can be placed in the empty rooms.

An infinite set is as one that can be placed into a one-to-one correspondence with a proper subset of itself.

2.3 Transfinite numbers

At the end of the 19th century Cantor developed the idea of levels of infinity. To carry a notion of equal size of two finite or infinite sets X and Y we define that this is given if a bijective mapping from X onto Y exists. In other terms, the elements of X and Y may be paired with each other in such a way that to each element of X there corresponds one and only one element of Y and vice versa. This notation for finite sets coincides with the ordinary notation of equality of numbers. It is a straight foreward generalization for infinite sets.

Since the notion of equal size is an equivalence relation, we can associate a number, called cardinal number, with every class of equal-sized sets.[3] The cardinal numbers of infinite sets are called transfinite numbers.
To compare transfinite numbers, we define for two sets X and Y that the number of elements in X is less than or equal the number of elements in Y, written $|X| \leq |Y|$ if there is a subset Y' of Y such that X and Y' are of equal size.
The following theorem is crucial, well-known in common sense, but not simple to prove, compare [6].

Theorem 2.3.1 *Let X and Y be sets. Then it holds*

a) $|X| \leq |Y|$ *or* $|Y| \leq |X|$.

b) *(Cantor-Bernstein) If* $|X| \leq |Y|$ *and* $|Y| \leq |X|$, *then* $|X| = |Y|$.

We call sets with as many elements as the set of natural numbers countable. A countable set is the smallest of the infinite sets:

Lemma 2.3.2 *Any infinite set contains a countable set.*

[3] But note, that the set of all sets does not exists. The original paradox was given by Russel in 1901. Consider

$$R = \{x : x \text{ is a set with } x \notin x\}. \tag{2.6}$$

Then it holds

$$R \in R \quad \text{if and only if} \quad R \notin R. \tag{2.7}$$

Russel [208]:

> In terms of classes the contradiction appears even more extraordinary. A class as one may be a term of itself as many. Thus the class of all classes is a class; the class of all the terms that are not men is not a man, and so on. Do all the classes that have this property form a class? If so, it is as one a member of itself as many or not? If it is, then it is one of the classes which, as ones, are not members of themselve as many, and vice versa. Thus we must conclude again that the classes which as ones are not members of themselve as many do not form a class - or rather, that they do not form a class as one, for the argument cannot show that they do not form a class as many.

Proof. We can select a countable subset from an infinite set S in the following way: Take any element x_0 from S. Clearly, we have not exhausted the elements of S with the selection of x_0, so we can proceed to select a second element x_1. After that we select a third element x_2 and so on. We have thus extracted from S a countable subset of indexed element. □

Example 2.3.3 *The set* $\mathbf{\Gamma}$ *of integers is countable, since the function* $f : \mathbb{N} \to \mathbf{\Gamma}$ *is one-to-one and onto:*

$$f(n) = \begin{cases} -\frac{n}{2} & : \ n \text{ even} \\ \frac{n+1}{2} & : \ \text{otherwise} \end{cases}$$

It is more difficult to show that the rational numbers are also countable. Obviously this is paradoxical: Between any two rational numbers we can still find infinitely many rational numbers. So it is quite unclear how we should go about numbering them.
First we prove that $\mathbb{N}^2$ is countable. Consider the following tabulation, which is called Cantor's first diagonal principle.

$x \setminus y$	0	1	2	3	4	...
0	0	1	3	6	10	...
1	2	4	7	11	16	...
2	5	8	12	17	23	...
3	9	13	18	24	31	...
4	14	19	25	32	40	...
⋮	⋮	⋮	⋮	⋮	⋮	

That means at first we count all pairs (x, y) with $x + y = 0$, then all pairs with $x + y = 1$, then with $x + y = 2$, and so on. The pair (x, y) lies in position number x between $(0, x + y)$ and $(x + y, 0)$. Before $(0, x + y)$ we have exactly

$$1 + 2 + \ldots + (x + y) = \frac{(x + y)(x + y + 1)}{2}$$

pairs. Hence,

Theorem 2.3.4 *The function*

$$c : (x, y) \mapsto \frac{(x + y)(x + y + 1)}{2} + x, \tag{2.8}$$

called the Cantor function, is a bijective mapping from $\mathbb{N}^2$ *onto* $\mathbb{N}$*.*

The inverse functions for c are not so easy to find. For $n = c(x, y)$ we define $x = l(n)$ and $y = r(n)$, understanding as

$$x = \text{left of } n \quad \text{and} \tag{2.9}$$
$$y = \text{right of } n, \tag{2.10}$$

such that $n = (l(n), r(n))$. That means

$$2n = (x + y)(x + y + 1) + 2x.$$

In view of 2.3.4 we have

$$\begin{aligned} 8n + 1 &= 4 \cdot 2n + 1 \\ &= 4 \cdot ((x + y)(x + y + 1) + 2x) + 1 \\ &= 4 \cdot ((x + y)^2 + 3x + y) + 1 \\ &= (2x + 2y)^2 + 12x + 4y + 1, \end{aligned}$$

such that

$$8n + 1 = (2x + 2y + 1)^2 + 8x \tag{2.11}$$

and

$$8n + 1 = (2x + 2y + 3)^2 - 8y - 8. \tag{2.12}$$

Hence,

$$2x + 2y + 1 \leq \lfloor \sqrt{8n + 1} \rfloor < 2x + 2y + 3, \tag{2.13}$$

or, equivalently,

$$x + y + 1 \leq \frac{\lfloor \sqrt{8n + 1} \rfloor + 1}{2} < x + y + 2. \tag{2.14}$$

This implies

$$x + y + 1 = \left\lfloor \frac{\lfloor \sqrt{8n + 1} \rfloor + 1}{2} \right\rfloor. \tag{2.15}$$

Compared with 2.3.4 this equation gives us

Corollary 2.3.5 *The functions*

$$l(n) = x = n - \frac{1}{2} \left\lfloor \frac{\lfloor \sqrt{8n + 1} \rfloor + 1}{2} \right\rfloor \left\lfloor \frac{\lfloor \sqrt{8n + 1} \rfloor - 1}{2} \right\rfloor \tag{2.16}$$

$$r(n) = y = \left\lfloor \frac{\lfloor \sqrt{8n + 1} \rfloor + 1}{2} \right\rfloor - l(n) - 1 \tag{2.17}$$

are the inverse mappings of the Cantor function c.

As an exercise discuss all these questions for the following function from $\mathbb{N}^2$ onto $\mathbb{N}$.

$x \setminus y$	0	1	2	3	4	...
0	0	2	4	6	8	...
1	1	5	9	13	17	...
2	3	11	19	27	35	...
3	7	23	39	55	71	...
⋮	⋮	⋮	⋮	⋮	⋮	

With 2.3.4 in mind, we have several considerations. For instance, the set of all rational numbers is countable. But we count tuples of natural numbers:

Corollary 2.3.6 *For any integer $n \geq 2$ there exist a bijective mapping $c^{(n)}$ from $\mathbb{N}^n$ onto $\mathbb{N}$.*

Proof. Let c be a bijective mapping from $\mathbb{N}^2$ onto $\mathbb{N}$, compare 2.3.4. We create $c^{(n)}$ by the following recursive equations:

$$c^{(2)} = c, \tag{2.18}$$

$$c^{(n)}(x_1, \ldots, x_n) = c^{(n-1)}(c(x_1, x_2), x_3, \ldots, x_n), \tag{2.19}$$

$n = 3, 4, \ldots$.
The inverse functions l and r by using 2.3.5. For

$$c^{(n)}(x_1, \ldots, x_n) = x$$

we have

$$\begin{aligned} x_n &= r(x) \\ x_{n-1} &= r \circ l(x) \\ &\vdots \\ x_2 &= r \circ l \circ \cdots \circ l(x) \\ x_1 &= l \circ l \circ \cdots \circ l(x). \end{aligned}$$

□

2.4 Words

An alphabet A is a nonempty and finite set of distinguished letters (or symbols). Important examples of alphabets are:

- $A = \{0, 1\}$ is an alphabet which plays a central role in coding theory. Moreover, we consider a word of 0's and 1's as a description of some individual, perhaps a genetic sequence in which each entry may take on one of two possible values.
- The English language needs 26 letters: A,B,...,Y,Z, and a letter for the empty space. German needs several letters more: Ä, Ö, Ü, ß. More generally

alphabet	language		# letters
ASCII	Computer		128
Cyrillic	Russian		32
Latin	German	a,...,z,ä,ö,ü,ß	30
Latin	English	a,...,z	26
Greek	Greek	α,...,ω	24
Hebrew		$\aleph$,...	22

- $A = \{a, c, g, t\}$ is the alphabet which codes the nucleotides of a DNA molecule, where a stands for adenine, c for cytosine, g for guanine and t for thymine. A similar alphabet, namely $A = \{a, c, g, u\}$ is used for the nucleotides of RNA, where u codes for uracil.
  Derived from this alphabet there is a binary alphabet $A' = \{r, y\}$ in which r codes for a purine (a or g), and y codes for a pyrimidine (c or t).

- The amino acids commonly found in proteins are coded by the alphabet $A = \{\text{ala}, \text{arg}, \ldots, \text{val}\}$, where the letters abbreviate the amino acids alanine, arginine,...,valine. In the usual genetic code $|A| = 20$ amino acids are coded:

	One-letter code	Three-letter code	Name
1	A	ala	alanine
2	C	cys	cysteine
3	D	asp	aspartic acid
4	E	glu	glutamatic acid
5	F	phe	phenylalanine
6	G	gly	glycine
7	H	his	histidine
8	I	ile	isoleucine
9	K	lys	lysine
10	L	leu	leucine
11	M	met	methionine
12	N	asn	asparagine
13	P	pro	proline
14	Q	gln	glutamine
15	R	arg	arginine
16	S	ser	serine
17	T	thr	threonine
18	V	val	valine
19	W	trp	tryptophan
20	Y	tyr	tyrosine

A word (also called a sequence, a string) over an alphabet A is a finite sequence of letters from A. The length $|w|$ of the word w is the number of letters composing it. We additionally define an empty word λ of length 0.[4]
Note that the description of a word contains a left-to-right order of the letters. We

[4]Fitch [86] gives the following exemplary genome sizes written in its length base pairs (bp):

Domain	Organism	Size (bp)
Viruses	HIV	$9 \cdot 10^3$
Bacteria	E. coli	$4 \cdot 10^6$
Eukaryotes	mammals	$3 \cdot 10^9$

Roughly speaking, the order of genome sizes is kbp, Mbp and Gbp for Viruses, Prokarya and Eukarya, respectively.

will write $w = a_1 a_2 \dots a_n$ for a word w consisting of the letters a_1, a_2, ... a_n in this order.[5]
We say that two words $w = a_1 a_2 \dots a_n$ and $w' = b_1 b_2 \dots b_m$ over the same alphabet are equal, and we write $w = w'$, if $n = m$ and $a_i = b_i$ for all $i = 1, \dots, n$.
Let $w = a_1 a_2 \dots a_n$ and $w' = b_1 b_2 \dots b_m$ be two words over the same alphabet A. The concatenation of w and w', written ww', is the word $a_1 a_2 \dots a_n b_1 b_2 \dots b_m$ over A. Hence, $|ww'| = |w| + |w'|$. Moreover, we will write $w^k = \underbrace{w \dots w}_{k-\text{times}}$ and $w^0 = \lambda$ for each word w.
A^n denotes the set of all words over A with length exactly n. Clearly, A^n is a finite set:

Theorem 2.4.1 $|A^n| = |A|^n$.

On the other hand, we consider the set

$$A^\star = \bigcup_{n \geq 0} A^n, \tag{2.20}$$

which contains all words over the alphabet A.
Equipped with concatenation as a binary operation it satisfies the following properties:

Closure: For all $v, w \in A^\star$, $vw \in A^\star$;

Associativity: For all $u, v, w \in A^\star$, $(uv)w = u(vw)$;

Identity: For the unity λ it holds that for any $v \in A^\star$ it is $v\lambda = \lambda v = v$.

(Usually a set with such a operation is called a semigroup.)

Theorem 2.4.2 *For any alphabet A the set $A^\star$ is infinite, but countable.*[6]

Proof. To see the countableness, we give a method to count the words. First count the word λ, then the members of A itself, then the words of length 2, and so on. More precisely, let $A = \{a_1, \dots, a_n\}$, then

[5] The Central Dogma of Molecular Biology describes the interaction of these polymers:

DNA acts as a template to replicate itself;
DNA is also transcribed into RNA; and
RNA is translated into protein.

More precisely,

- Integral form: DNA makes RNA makes protein.
- Differential form: Changed DNA can make changed protein.

For instance human insulin is composed by two words (chains):

A: gly ile val glu gln cys cys thr ser ile cys ser leu tyr glu leu glu asn tyr cys asn.

B: phe val asn gln his leu cys gly ser his leu val glu ala leu tyr leu val cys gly glu arg gly phe phe tyr thr pro lys thr.

[6] Also for a one-element alphabet $A = \{|\}$: $A^\star = \{\lambda, |, ||, |||, |^4, |^5, \dots\}$.

$\mathbb{N}$	$A^\star$
0	λ
1	a_1
2	a_2
$\vdots$	$\vdots$
n	a_n
$n+1$	a_1a_1
$n+2$	a_1a_2
$\vdots$	$\vdots$
$2n$	a_1a_n
$2n+1$	a_2a_1
$2n+2$	a_2a_2
$\vdots$	$\vdots$
$3n$	a_2a_n
$3n+1$	a_3a_1
$\vdots$	$\vdots$
n^2+1	a_na_1
$\vdots$	$\vdots$
n^2+n	a_na_n
n^2+n+1	$a_1a_1a_1$
$\vdots$	$\vdots$

□

If there is an order $<$ of the letters in A, the set $A^\star$ is endowed with the following linear order $<_L$ of the words, which is called the lexicographic order: For two words $w = a_1a_2\dots a_n$ and $w' = b_1b_2\dots b_m$ we define $w <_L w'$ if

1. $n < m$ and $a_1 = b_1, \dots, a_n = b_n$; or
2. $a_1 = b_1, \dots, a_k = b_k$ for $k < n, m$ and $a_{k+1} < b_{k+1}$.

We write $w \leq_L w'$ if $w <_L w'$ or $w = w'$.[7]

All the sets we have constructed so far have been countable. This naturally leads us to ask whether all infinite sets are countable. But the situation turns out to be more complicated than that; uncountable sets exist, and of more than one cardinality. First we show, using Cantor's second diagonal principle

Theorem 2.4.3 *The set of all (infinite) binary sequences is not countable.*

[7]Note that this order is fundamentally different from the order which we used to count $A^\star$. For instance there are infinitely many words between a_1 and a_2.

Proof. Assume that there is a counting of $\{0,1\}^\infty$ given by the following double infinite array:

$\mathbb{N}$	$\{0,1\}^\infty$
0	$b_{00}, b_{01}, b_{02}, b_{03}, \ldots$
1	$b_{10}, b_{11}, b_{12}, b_{13}, \ldots$
2	$b_{20}, b_{21}, b_{22}, b_{23}, \ldots$
3	$b_{30}, b_{31}, b_{32}, b_{33}, \ldots$
$\vdots$	$\vdots$

The sequence $b_0, b_1, b_2, \ldots$ with $b_i = 1 - b_{ii}$ cannot be in this table. □

2.5 Functions

The essence of mathematics resides in its freedom.

Georg Cantor

Let X and Y are nonempty sets. $\mathcal{F}(X,Y)$ denotes the collection of all functions $f : X \to Y$.
Not hard to see:

Theorem 2.5.1 *For finite sets X and Y we have*

$$|\mathcal{F}(X,Y)| = |Y|^{|X|}. \tag{2.21}$$

2.5.1 justifies the notation $Y^X = \mathcal{F}(X,Y)$.[8]

Theorem 2.5.2 *The set $\mathcal{F}(X,Y)$ contains more elements than X whenever Y contains at least two elements.*[9]

Proof. First we show that there are as many functions in $\mathcal{F}(X,Y)$ as elements in X. Consider for each $x_0 \in X$ the function $f[x_0]$ defined by

$$f[x_0](x) = \begin{cases} y_1 & : \quad x = x_0 \\ y_2 & : \quad \text{otherwise} \end{cases}$$

[8] A function

$$f : \{0,1\}^n \to \{0,1\} \tag{2.22}$$

is called a Boolean function. Since $|\{0,1\}^n| = 2^n$ we find by 2.5.1 that the number of Boolean functions increase astronomically, namely superexponentially:

$$|\mathcal{F}(\{0,1\}^n, \{0,1\})| = 2^{2^n} \tag{2.23}$$

[9] The proof will show that this assertion is true for finite and infinite sets X.

where y_1 and y_2 are distinct elements of Y.
If $x_0 \neq x_1$ then $f[x_0] \neq f[x_1]$.
Now assume that there is a bijective mapping

$$x \in X \mapsto f[x] \in \mathcal{F}(X, Y). \tag{2.24}$$

Choose $y_x \in Y$ such that $y_x \neq f[x](x)$. The function f defined by

$$f : x \mapsto y_x, \tag{2.25}$$

cannot be one of the function $f[x]$. □

As an exercise show that the power set $\mathcal{P}(X)$ contains more elements than X. Consequently, a largest cardinal number, both finite and transfinite, does not exist, see

$$X, \mathcal{P}(X), \mathcal{P}(\mathcal{P}(X)), \mathcal{P}(\mathcal{P}(\mathcal{P}(X))), \ldots. \tag{2.26}$$

With this in mind the cardinal number of all subsets of a countable set, which is a set of size $\aleph_0$ is a bigger form of infinity. Furthermore, we have an infinite sequence of bigger and bigger infinite numbers:

$$\aleph_0, 2^{\aleph_0}, 2^{2^{\aleph_0}}, 2^{2^{2^{\aleph_0}}}, \ldots. \tag{2.27}$$

The next bigger number after $\aleph_0$ Cantor called $\aleph_1$.

$$\aleph_1 \leq 2^{\aleph_0}. \tag{2.28}$$

Cantor believed that $\aleph_1$ was identical with the size of the real numbers, which means that in (2.28) equality holds. This is Cantor's continuum hypothesis, which is equivalent to saying that there is no infinite set with a cardinality between that of the integers and the reals; in other words the number of real numbers is the next "level" of infinity above the countable sets. In 1940 Gödel showed that Cantor's guess can never be disproved from the other axioms of mathematics. In 1963 Cohen showed that it could not proved either. That means, that the continuum hypothesis is neither true nor false, but undecidable, that means independent from the other axioms of set theory.

2.6 Discrete Mathematics

Roughly speaking: Mathematics can be concerned as the essentially scientific part of any thoery. When investigating a "real world problem" we make a lot of assumptions. The logical combination of these assumptions yields hints to the solution of the problem. Mathematics gives the possibility to order and to verify scientific facts.
Discrete mathematics devoted to the study of discrete objects, these are

- a finite or countable set of distinct and unconnected elements; which are

- separated and discontinuous.

Discrete mathematics is used whenever objects are counted, when relationships between finite or countable sets are studied, and when processes involving a finite number of steps are analyzed.
To justify this view we prove the invariance of the class of countable sets under counting.

Theorem 2.6.1 *Let $S_1, S_2, \ldots$ be a countable number of finite sets, then the union $S = \bigcup_i S_i$ is finite or countable.*

Proof. We define sets $R_1, R_2, \ldots$ where R_i contains the elements of S_i which do not belong to preceding sets, that means

$$R_1 = S_1 \tag{2.29}$$
$$R_i = S_i \setminus (S_1 \cup S_2 \cup \ldots \cup S_{i-1}) \tag{2.30}$$

for $i \geq 2$. Then the R_i are disjoint and $\bigcup_i R_i = S$.
Let

$$R_i = \{b_{i1}, b_{i2}, \ldots, b_{im_i}\}. \tag{2.31}$$

If $S = \{b_{ij}\}$ is infinite, then we define a bijective function f from S onto the natural numbers by

$$f(b_{ij}) = m_1 + m_2 + \ldots + m_{i-1} + j. \tag{2.32}$$

□

Theorem 2.6.2 *A countable union of countable sets is countable.*

Proof. Let $S_1, S_2, \ldots$ be a countable number of countable sets, and suppose that $a_{i1}, a_{i2}, \ldots$ are the elements of S_i. We define sets $R_2, R_3, R_4, \ldots$ as follows:

$$R_k = \{a_{ij} : i + j = k\}. \tag{2.33}$$

Observe that each R_k is a finite set and

$$\bigcup_k R_k = \bigcup_i S_i. \tag{2.34}$$

Then we apply 2.6.1. □

Chapter 3

Selecting objects

The study of ways of choosing (= selecting) and arranging objects from a given collection and the study of other kinds of problems relating to counting the number of ways to do something are key questions in discrete mathematics.

Consider a set of n objects. How many ways are there of selecting k, $0 \leq k \leq n$, from these? We will distinguish two kinds of choosing:

- ordered or unordered;
- repetitions allowed or not.

This gives us four distinct questions.

3.1 The number of subsets

As introductory example choose two elements from $\{1, 2, 3, 4\}$, where we respect its order, but ignore repetitions:

1,2	1,3	1,4
2,1	2,3	2,4
3,1	3,2	3,4
4,1	4,2	4,3

More systematically, first recall that there are $n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$ ways to place n objects in a linear arrangement. If we select only k objects, we start with n possibilities and count down k numbers, the last one will be $n - k + 1$. Hence, we have the following theorem.

Theorem 3.1.1 *The number of subsets with k ordered elements of a set with n elements is*

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}. \tag{3.1}$$

From this we can easily derive one of the most important counting results.

Theorem 3.1.2 *The number of subsets containing k elements of a set with n elements is*

$$\frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!}. \tag{3.2}$$

Proof. In 3.1.1 we counted ordered subsets. If we want to know the number of unordered subsets, then every subset was counted exactly $k!$ times, namely all possible orderings of the elements. So we have to divide this number by $k!$ to get the assertion. □

As an example choose two elements from $\{1,2,3,4\}$:

1,2	1,3	1,4
	2,3	2,4
		3,4

The number defined in 3.1.2 is such an important quantity that there is a special notation for it:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \tag{3.3}$$

read "n choose k". These numbers are also called binomial coefficients; we will later see why.[1] In view of 3.1.1 we will write $\binom{X}{k}$ for the collection of all subsets of X with exactly k elements. This gives for the power set

$$\mathcal{P}(X) = \bigcup_{k=0}^{|X|} \binom{X}{k}. \tag{3.4}$$

3.2 Selections with Repetitions

Recall what we discussed until now for selecting k objects from a set of n:

	ordered	unordered
no repetitions	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$
repetitions allowed	n^k	?

As an example choose ordered two elements from $\{1,2,3,4\}$ where repetitions are allowed:

[1] Of course, for a calculation of a simple binomial coefficients it is not pleasant to use this formula; better:

$$\binom{n}{k} = \frac{n}{k} \cdot \frac{n-1}{k-1} \cdot \frac{n-2}{k-2} \cdots \frac{n-k+1}{1}.$$

1,1	1,2	1,3	1,4
2,1	2,2	2,3	2,4
3,1	3,2	3,3	3,4
4,1	4,2	4,3	4,4

Finally, "?" means that we have to determine the number of ways these are to choose k objects from n, where repetitions are allowed, but where order does not matter. As an example choose two elements from $\{1, 2, 3, 4\}$:

1,1	1,2	1,3	1,4
	2,2	2,3	2,4
		3,3	3,4
			4,4

More systematically,

Theorem 3.2.1 *The number of unordered choices of k from n, with repitiotions allowed is*

$$\binom{n+k-1}{k} = \binom{n+k-1}{n-1}. \tag{3.5}$$

Proof. Any choice will consist of x_1 choices of the first object, x_2 choices of the second object, and so on, where the condition $x_1 + \ldots + x_n = k$ is satisfied.
We can represent such collection $x_1, \ldots, x_n$ of integers by a binary sequence:

$$\underbrace{0, \ldots, 0}_{x_1-\text{times}}, 1, \underbrace{0, \ldots, 0}_{x_2-\text{times}}, 1, \underbrace{0, \ldots, 0}_{x_3-\text{times}}, 1, \ldots, 1, \underbrace{0, \ldots, 0}_{x_n-\text{times}} \tag{3.6}$$

In this representation there will be $n-1$ times the digit 1 and k times the digit 0, and so each sequence will be of length $n+k-1$, containing exactly k 0s. Conversely, any such sequence corresponds to a nonnegative integer solution of $x_1 + \ldots + x_n = k$. The k 0s can be in any of the $n+k-1$ positions, so the number of such sequences is $\binom{n+k-1}{k}$. □

From the first fact in the proof we get

Theorem 3.2.2 *The number of solutions for the equation $x_1 + \ldots + x_n = k$ in nonnegative integers x_i equals*

$$\binom{n+k-1}{k}. \tag{3.7}$$

3.3 The Principle of Inclusion and Exclusion

Consider an experiment with specific garden plots have to be treated with lime, potash, urea and phosphate. According to the design, 32 plots are to be treated with just one of the individual chemicals and perhaps some others, 16 plots are to be treated with a pair of chemicals and perhaps some others, eight are to be treated

with three of the chemicals and perhaps another, and four are to be treated with all four chemicals. How many plots are needed if none are to receive no treatment at all?

Suppose some event can occur in α ways and a second event can occur in β ways, and suppose both events cannot occur simultaneously. Then both events can occur in $\alpha + \beta$ ways. More generally, if E_i, $i = 1, \ldots, n$, are n events such that no two of them can occur at the same time, and that if E_i can occur in α_i ways, then one of the events can occur in $\alpha_1 + \alpha_2 + \ldots + \alpha_n$ ways. In other terms,

Observation 3.3.1 *(The addition principle)*
For a collection of pairwise disjoint sets the following holds:

$$|S_1 \cup \ldots \cup S_n| = |S_1| + \ldots + |S_n|. \tag{3.8}$$

Remember that the addition principle tells us that the size of the union of a collection of disjoint sets is the sum of the sizes of the sets. To determine the size of a union of overlapping sets, we clearly must use information about the way the sets overlap.
Consider two sets S_1 and S_2. Note that if we add $|S_1|$ and $|S_2|$, we include the size of $S_1 \cap S_2$ twice in this sum. So we get the formula

$$|S_1 \cup S_2| = |S_1| + |S_2| - |S_1 \cap S_2|. \tag{3.9}$$

Similar for three sets: In the sum $|S_1| + |S_2| + |S_3|$ an element of $S_1 \cap S_2$ is included at least twice; an element of $S_1 \cap S_2 \cap S_3$ three times. Hence,

$$\begin{aligned} |S_1 \cup S_2 \cup S_3| \quad = \quad & |S_1| + |S_2| + |S_3| \\ & -|S_1 \cap S_2| - |S_1 \cap S_3| - |S_2 \cap S_3| \\ & +|S_1 \cap S_2 \cap S_3|. \end{aligned} \tag{3.10}$$

A formula such as 3.10 is called an inclusion-exclusion formula, and we can generalize it to:

Theorem 3.3.2 *Let $S_1, \ldots, S_n$ be a collection of sets. Then*

$$\begin{aligned} |\bigcup_{i=1}^{n} S_i| \quad = \quad & \sum_{i=1}^{n} |S_i| - \sum_{1 \le i < j \le n} |S_i \cap S_j| \\ & + \sum_{1 \le i < j < k \le n} |S_i \cap S_j \cap S_k| \mp \ldots - (-1)^n |\bigcap_{i=1}^{n} S_i|. \end{aligned} \tag{3.11}$$

Corollary 3.3.3 *Let $S_1, \ldots, S_n$ be subsets of a finite set S. Then the number of elements of S that are in none of the subsets is*

$$|S| - \sum_{i=1}^{n} |S_i| + \sum_{1 \le i < j \le n} |S_i \cap S_j| - \sum_{1 \le i < j < k \le n} |S_i \cap S_j \cap S_k| \pm \ldots + (-1)^n |\bigcap_{i=1}^{n} S_i|. \tag{3.12}$$

Proof. Since

$$|S_1^c \cap S_2^c \cap \ldots \cap S_n^c| = |S| - |S_1 \cup S_2 \cup \ldots \cup S_n| \tag{3.13}$$

the equation follows by 3.3.2. □

A very useful generalization of 3.3.2 is the principle of inclusion-exclusion (PIE):

Theorem 3.3.4 *Let $\mathcal{U}$ be a set of objects, called the universe, and let $P_1, \ldots, P_r$ be a collection of properties which the elements of $\mathcal{U}$ may or may not have. Let $N(i, j, \ldots, s)$ denote the number of elements of $\mathcal{U}$ which possess the properties $P_i, P_j, \ldots, P_s$ (and possibly some others as well). Then the number of elements of $\mathcal{U}$ having none of those properties is*

$$\begin{aligned} &|U| - \sum_{i=1}^{r} N(i) + \sum_{1 \le i < j \le r} N(i,j) \\ &- \sum_{1 \le i < j < k \le r} N(i,j,k) \pm \ldots + (-1)^r N(1,2,\ldots,r). \end{aligned} \tag{3.14}$$

Proof. Consider $S_i = \{x \in U : x \text{ possess } P_i\}$. Then $N(i, \ldots, s) = |S_i \cap \ldots \cap S_s|$ holds and (3.14) is the same as 3.3.3. □

We want to emphasize an important logical point about applying this formula. To use it in a counting problem, one must select a universe and a collection of subsets in that universe such that the elements to be counted are the subset of elements in the universe that are in none of the given subsets. That is, the given subsets represent properties not satisfied by the elements being counted.
For many applications the quantity $N(i, j, \ldots, s)$ depends only from the number of properties which possess, and we get

Theorem 3.3.5 *Let $\mathcal{U}$ be a set of objects, called the universe, and let $P_1, \ldots, P_r$ be a collection of properties which the elements of $\mathcal{U}$ may or may not have. Let N_k denote the number of elements of $\mathcal{U}$ which posssess k of the properties. Define $N_0 = |\mathcal{U}|$. Then the number of elements of $\mathcal{U}$ having none of those properties is*

$$\sum_{k=0}^{r} (-1)^k \binom{r}{k} N_k. \tag{3.15}$$

3.4 Counting functions

Recall that $\mathcal{F}(X, Y)$ denotes the collection of all functions $f : X \to Y$. We have already count the cardinality of this set: $|\mathcal{F}(X, Y)| = |Y|^{|X|}$. We will generalize the observation 2.5.1 for several collections of $\mathcal{F}(X, Y)$. Consider $f \in \mathcal{F}(X, Y)$, and let $|X| = m$ and $|Y| = n$. The image of the function f is the set of elements of Y which actually arise as a value $f(x)$ for some $x \in X$:

$$\text{im} f = \{y \in Y : y = f(x) \text{ for some } x \in X\}. \tag{3.16}$$

For each function f the image is a nonempty subset of Y. How many of these functions have an image of size k?

- If f takes on precisely k values then X can be partitioned into k parts $X_1, \ldots, X_k$, where X_i consists of those elements of X which are mapped onto the ith member of $\mathrm{im}f$.
  A partition of X into k parts can be done in $S(m,k)$ ways, where $S(m,k)$ denotes the Stirling number of the second kind, which counts the number of partitioning a set of m elements into k parts.
- We have to choose the image of f in Y.
  This can be done in $\binom{n}{k}$ ways.
- We have to pair off each X_i with one of the members of $\mathrm{im}f$.
  This can be done in $k!$ ways.

Altogether, we obtain

Theorem 3.4.1 *The number of functions $f : X \to Y$ where $|X| = m$, $|Y| = n$ and $|imf| = k$ equals*

$$S(m,k) \cdot \binom{n}{k} \cdot k!. \tag{3.17}$$

Thus, since k can take any value from 1 to n, and using 2.5.1 we get

Corollary 3.4.2 *The number of functions $f : X \to Y$ where $|X| = m$ and $|Y| = n$ equals*

$$|\mathcal{F}(X,Y)| = n^m = \sum_{k=1}^{n} S(m,k) \cdot \binom{n}{k} \cdot k!. \tag{3.18}$$

Now we consider specific classes of functions.
A function $f \in \mathcal{F}(X,Y)$ is called a surjection if each element of Y belongs to the image of f. $\mathcal{F}^{\mathrm{onto}}(X,Y)$ denotes the set of all surjections from X onto Y. Of course this set is empty if $m < n$.
A surjection f is characterized by the condition $\mathrm{im}f = Y$.
A 1-1 function is called an injection, and $\mathcal{F}^{1-1}(X,Y)$ denotes the set of all injections from X into Y. Of course this set is empty if $m > n$. An injection is characterized by the condition $|\mathrm{im}f| = |X|$.
Recall that a function f is called a bijection if f is both, a surjection and an injection. $\mathcal{F}^{\mathrm{bi}}(X,Y)$ denotes the set of all bijections from X onto Y. Of course this set is empty if $m \neq n$.
Altogether, as a special case of 3.4.2 we obtain

Theorem 3.4.3 *Consider two sets X and Y with m and n elements, respectively.*

a) *Assume that $m \geq n$. The number of surjections from X onto Y equals*

$$|\mathcal{F}^{onto}(X,Y)| = n! \cdot S(m,n). \tag{3.19}$$

b) *Assume that $m \leq n$. The number of injections from X into Y equals*

$$|\mathcal{F}^{1\text{-}1}(X,Y)| = \frac{n!}{(n-m)!}. \tag{3.20}$$

c) *Assume that $m = n$. The number of bijections from X onto Y equals*

$$|\mathcal{F}^{bi}(X,Y)| = n!. \tag{3.21}$$

Let $f : X \to Y$ be a surjection, then if each element of Y occurs as the image $f(x)$ of some element $x \in X$. Let the universe be given by

$$\mathcal{U} = \{f : X \to Y = \{y_1, \ldots, y_n\}\},$$

where $|X| = m$. How many surjections are there in $\mathcal{U}$? First, it is easy to see that

$$|\mathcal{U}| = n^m. \tag{3.22}$$

Let P_i be the property: y_i is not in the image of f, $i = 1, \ldots, n$. Then, in view of 2.5.1,

$$N(i) = (n-1)^m, \tag{3.23}$$

since each of the m elements of X can be mapped onto any of the $n-1$ other elements of Y. Because this is independent of the choice of y_i we get $N_1 = (n-1)^m$. Similarly, $N_k = (n-k)^m$. Then we can apply 3.3.5:

$$\sum_{k=0}^{n}(-1)^k\binom{n}{k}N_k = n^m - n(n-1)^m + \binom{n}{2}(n-2)^m \mp \ldots + (-1)^n(n-n)^m.$$

Hence,

Theorem 3.4.4 *Consider two sets X and Y with m and n elements, respectively. The number of surjections from X onto Y equals*

$$|\mathcal{F}^{onto}(X,Y)| = \sum_{k=0}^{n-1}(-1)^k\binom{n}{k}(n-k)^m. \tag{3.24}$$

One immediate consequence of this result is a formula for the Stirling number of the second kind, see K.3.2.

Chapter 4

Networks

4.1 Graphs

We have to introduce several knowledge of graphs and networks.[1] A graph G is defined to be a pair (V, E) where

- V is a nonempty and finite set of elements, called vertices, and
- E is a finite family of elements which are unordered pairs of vertices, called edges.

The notation $e = \underline{uv}$ means that the edge e joins the vertices u and v. In this case, we say that u and v are incident to this edge and that u and v are the endvertices of e. Two vertices u and v are called adjacent in the graph G if $\underline{uv}$ is an edge of G. Different edges $e_1 = \underline{vw}$ and $e_2 = \underline{vw}$ are called multiple or parallel edges. A graph with multiple edges is called a multigraph. Any graph is also a multigraph.[2]
$N(v) = N_G(v)$ denotes the set of all vertices adjacent to the vertex v. This set of all neighbors is called the neighborhood of v.
For a vertex v of a graph G the degree $g_G(v)$ is defined as the number of edges which are incident to v. If G has no parallel edges, then the cardinality of $N(v) = N_G(v)$ is the degree of the vertex v:

$$g(v) = g_G(v) = |N_G(v)|. \tag{4.1}$$

If we sum up all the vertex degrees in a graph, we count each edge exactly twice, once from each of its endvertices. Thus,

Observation 4.1.1 *In any graph $G = (V, E)$ the equality*

$$\sum_{v \in V} g_G(v) = 2 \cdot |E| \tag{4.2}$$

[1] Since the terminology of graph theory is not standard, the reader may find some differences between terms used here and in other texts.

[2] In any case, we assume that $u \neq v$, that means we do not admit loops.

holds. Particularly, in every graph the number of vertices with odd degree is even.

A graph G is said to be a complete graph if any two vertices are adjacent. A complete graph with n vertices has exactly

$$\binom{n}{2} = \frac{n(n-1)}{2} \tag{4.3}$$

edges, and each vertex is of degree $n-1$.
Let V_1 and V_2 be two sets with n_1 and n_2 elements, respectively. The complete bipartite graph K_{n_1,n_2} is defined by

$$K_{n_1,n_2} = (V_1 \cup V_2, \{\underline{vw} : v \in V_1, w \in V_2\}). \tag{4.4}$$

The complete bipartite graph K_{n_1,n_2} has $n_1 + n_2$ vertices and $n_1 \cdot n_2$ edges.
In general a graph $G = (V, E)$ is called bipartite if it is possible to split V into subsets V_1 and V_2 such that every edge joins a vertex of V_1 to a vertex of V_2. In other terms, a graph is bipartite if and only if one can color the set of vertices such that the two endvertices of an edge have different colors. The following theorem should be an exercise for the reader.

Theorem 4.1.2 *A graph is bipartite if and only if it contains no cycle of odd length.*

Q^D denotes the D-dimensional hypercube. That is the graph whose set of vertices consists of all binary vectors from $\{0,1\}^D$, with an edge joining two vectors if and only if they differ in exactly one coordinate.[3]

Observation 4.1.3 *We may also define the hypercube Q^D inductively by letting Q^0 be a single vertex and then obtaining Q^D by taking two copies of Q^{D-1} and joining corresponding vertices.*

For practice the reader should prove the following facts.[4]

Theorem 4.1.4 *The hypercube Q^D has the following properties:*

a) *Q^D has 2^D vertices and $D \cdot 2^{D-1}$ edges;*

[3]The hypercube and its relatives play an important role in coding theory, see [217], and in the theory of molecular evolution, see [70] and [256].
As sequence data became readily available, the biological units are written in words constructed from the letters corresponding either to amino acids, which generate proteins, or to nucleotides forming DNA or RNA molecules. By comparing such words one can construct evolutionary (phylogenetic) trees showing how closeness of the words in the tree corresponds to the closeness of the unit. First, it was used by Fitch and Margoliash in their landmark paper [84] from 1967 dealing with cytochrome c sequences. As the basic idea, they construct a metric space which forms a model for the phylogeny. Further suggested by Fitch [85] in 1971, and explicitly written by Foulds et al. [91] in 1979. Unfortunately, this idea does not give a simple method. (And seems to have been rather forgotten in the field of biology after tree-building program packages became widely available.) Compare Bern and Graham [26].

[4]Hint: For the statement given in a) use 4.1.3 and solve the recurrence $f(D) = 2 \cdot f(D-1) + 2^{D-1}$.

b) *the hypercube is regular, that means each vertex in Q_D has degree D;*

c) *Q^D is bipartite.*

Let $G = (V, E)$ be a graph. Then $G' = (V', E')$ is called a subgraph of G if V' is a subset of V and E' is a subset of E such that any edge in E' joins vertices from V'. In other terms,

$$V' \subseteq V \tag{4.5}$$

and

$$E' \subseteq E \cap \binom{V'}{2}. \tag{4.6}$$

In the case $V' = V$ we call G' a spanning subgraph.

Let $W \subseteq V$ be a set of vertices, then

$$G[W] = (W, E \cap \binom{W}{2}) \tag{4.7}$$

is called the induced subgraph of W in $G = (V, E)$, that means all edges of G that connect vertices of W are also edges of $G[W]$.
When modifying a graph $G = (V, E)$, to delete an edge e means simply to remove e from E:

$$G - e = (V, E \setminus \{e\}). \tag{4.8}$$

However, to delete a vertex v, one must remove v from V and all edges that are incident to v:

$$G - v = G[V \setminus \{v\}]. \tag{4.9}$$

4.2 Multigraphs

The term graph, as the name of a system of points and lines, came from the phrase graphic notation, first introduced in chemistry by Frankland, and adopted by Crum Brown in 1884. Each atom of a chemical structure is represented by the vertex of a multigraph and each atomic bonds are represented by edges. The degree of the vertices represent the valences of the atoms. For instance

atom	abbreviation	valence
carbon	C	4
oxygen	O	3
nitrogen	N	2
hydrogen	H	1.

Thus the corresponding vertices in the associated graphs have similar degrees.
As example we consider the Kekule's structure which are the atomic structure of molecules.

- Water: $(\{H^1, H^2, O\}, \{\underline{H^1O}, \underline{H^2O}\})$.
- Amino acids: is a multigraph $G = (V, E)$ with an α-carbon atom C_α such that $G - C_\alpha$ has four components:

  (i) The amino-group $(\{H^1, H^2, N\}, \{\underline{H^1N}, \underline{H^2N}\})$.

  (ii) The carboxy-group $(\{C, O^1, O^2, H\}), \{\underline{CO^1}_1, \underline{CO^1}_2, \underline{CO^2}, \underline{C^2H}\}$.

  (iii) A single hydroxy atom H.

  (iv) The residue, which give the amino acid its name.
- A hydrocarbon is a compound formed from hydrogen atoms and carbon atoms.
- A benzene molecule C_6H_6 has double bonds for some pairs of its atoms, so it is modeled by a multigraph: Six carbon atoms forms a cycle, alternating by parallel edges; each carbon atom is adjacent with a hydrogen atom.
- A benzenoid system is a connected collection of benzene molecules in such a way that two benzene molecules are either disjoint or have one common edge. To each benzenoid system we can assign a benzenoid graph taking the vertices of benzenes as the vertices of the graph, and the sides of benzenes as its edges. This graph is planar, 2-connected, bipartite, and all internal regions are hexagons.

With help of our graph-theoretical concepts we can discuss which molecule structures can theoretically exists, and in how many ways.

4.3 Graph partitioning

The problem of graph partitioning is to divide the set of vertices into a several number of (disjoint) parts of a given size such that the number of edges connecting vertices of different parts is minimized. This number is called the cut-size.[5]
At first note that there are many ways for dividing, see C.3.1:

Observation 4.3.1 *Let $G = (V, E)$ be a graph with n vertices. Let n_i be positive integers with $n_1 + \ldots + n_k = n$, then the number of partitioning V equals*

$$\frac{n!}{n_1! \cdot n_2! \cdots n_k!}. \tag{4.10}$$

This implies that solving the problem of graph partitioning by simple enumerating is not an efficient approach. This remains true in the simplest case $k = 2$, the so-called bi-section problem. The number of edges connecting vertices of different parts is at most $n_1 \cdot n_2$ It is not hard to see that for given positive and even integers n_1 and n_2 with $n_\cdot + n_2 0 n$, the maximum cut-size is achieved if and only if

$$n_1 = n_2 = \frac{n}{2}.$$

In view of I.2.1 we obtain that the number is a very rapidly growing function:

[5] This is a first approach to the problem of classification.

Theorem 4.3.2 *The number to divide a graph of n vertices in two parts of nearly equal parts is roughly*

$$\frac{n!}{(n/2)!^2} \approx \sqrt{\frac{2}{\pi}} \frac{2^n}{\sqrt{n}}.$$

Now we go back to our optimization version. We create a simple heuristic algorithm which for te bisection problem.

Algorithm 4.3.3 *(Kernighan, Lin [150]) Let $G = (V, E)$ be a graph with n vertices. Let n_1 and n_2 be given, $n_: + n_2 = n$. Then split V in the following way:*

1. *Divide V into two parts V_1 and V_2 of the required size n_1 and n_2 in any way you like*[6]*;*
2. *For each pair v and v' of vertices whereby v lies in one part and v' in the other calculate how much the cut-size between the parts would change if interchanging v and v';*
   If no pair reduces the cut-size, then STOP;
   Otherwise, find the pair that reduces the cut size by the largest amount and swap the pair of vertices and repeat the process.

4.4 Connected graphs

A chain is a sequence $v_1, e_1, v_2, e_2, v_3, ..., v_m, e_m, v_{m+1}$ of edges and vertices of G such that the edge e_i is incident to the vertices v_i and v_{i+1} for any index $i = 1, ..., m$. A chain in which each vertex appears at most once is called a path; more exactly, the path interconnecting the vertices v_1 and v_{m+1}. Then the number m denotes the length of the path. It is important to understand that the length of a path is the number of its edges. A single vertex is a path of length 0.
A cycle is a chain with at least one edge and with the following properties: No edge appears twice in the sequence and the two endvertices of the chain are the same. A graph which does not contain a cycle is called acyclic.

I. A key notion in graph theory is that of a connected graph. It is intuitively clear what this should mean: A graph $G = (V, E)$ is called a connected graph if for any two vertices there is a path (or, equivalently, a chain) interconnecting them. Clearly,

Observation 4.4.1 *The relation "There is a path in G connecting v and v'" is an equivalence relation on $V \times V$.*

The equivalence classes of this relation divide V into subsets, which induce connected subgraphs of G. These classes are called the connected components, or briefly the components of the graph G. A component is a maximal subgraph that is connected. A connected graph has exactly one component. Of course, the number of components is an integer between 1 and the number of vertices. In any textbook of graph theory we find the following facts:

[6] Maybe randomly.

Theorem 4.4.2 *Let $G = (V, E)$ be a graph with $n = |V|$ vertices and c components. Then*

$$n - c \leq |E| \leq \frac{(n-c)(n-c+1)}{2} = \binom{n-c+1}{2}. \tag{4.11}$$

In particular, a connected graph has at least $n - 1$ and at most $\binom{n}{2}$ edges. Conversely,

Corollary 4.4.3 *Let $G = (V, E)$ be a graph with n vertices. If*

$$|E| > \binom{n-1}{2}, \tag{4.12}$$

then G is connected.

A vertex v of a graph $G = (V, E)$ is called an articulation of G if $G - v = G[V \setminus \{v\}]$ is disconnected.
An edge e of a graph $G = (V, E)$ is called a bridge if $G - e = (V, E \setminus \{e\})$ contains one component more than G.

Observation 4.4.4 *An edge e of a connected graph G is a bridge if and only if e does not lie in a cycle of G.*

As an exercise prove the following fact.

Theorem 4.4.5 *Let $G = (V, E)$ be a graph, and let $\delta(G)$ be the minimum degree in G:*

$$\delta(G) = \min\{g_G(v) : v \in V\}. \tag{4.13}$$

If

$$\delta(G) \geq \frac{|V| - 1}{2}, \tag{4.14}$$

then G is connected.

Let $G = (V, E)$ be a graph. The complement G^c of G is the graph with the same set V of vertices, such that two vertices are adjacent in G^c if and only if they are not adjacent in G: $G^c = (V, \binom{V}{2} \setminus E)$. Of course, $(G^c)^c = G$.[7]
For a graph $G = (V, E)$ and its complement $G^c = (V, E')$ it holds

$$|E| + |E'| = \binom{|V|}{2}. \tag{4.15}$$

Theorem 4.4.6 *Any graph and its complement cannot both be disconnected.*

[7] This would not be the case for multigraphs.

Proof. Let G be a disconnected graph. Suppose that v and v' are two vertices of G^c, the complement of G.
If v and v' belong to different components of G, then both are adjacent in G^c.
If v and v' belong to the same component of G, say H. Let w be a vertex of some other component, say H' of G. By definition, both v and w, and v' and w are adjacent in G^c. In either case, v is connected to v' by a path in G^c. Thus G^c is connected. □

II. A graph $G = (V, E)$ is called k-connected if for each pair of different vertices v and v' there are k pairwise vertex-disjoint paths interconnecting v and v', $k \geq 0$.

Remark 4.4.7 *A graph is k-connected if and only if k is the minimum number of vertices whose removal results in a disconnected or trivial graph.*

The proof one can find in any textbook of graph theory.

Let G be a k-connected graph with n vertices and minimum degree $\delta(G)$. Then it holds $k \leq \delta(G)$ and the number of edges is at least $kn/2$.[8]

We generalize 4.4.5 to the following result.

Theorem 4.4.9 *Let $G = (V, E)$ be a graph with n vertices. Let k be an integer with $1 \leq k \leq n - 1$. If*

$$\delta(G) \geq \frac{n + k - 2}{2}, \tag{4.16}$$

then G is k-connected.[9]

Proof. If G is the complete graph K_n, then G is k-connected.
Assume, that G is not complete, and G is not k-connected. Then there exists a set V' with $p < k$ vertices of G such that $G[V \setminus V']$ is disconnected.
Let G_1 be a component of $G[V \setminus V']$ with a minimum number of vertices. Since,

[8] In particular, 2-connected graphs play an important role.

Theorem 4.4.8 *Let G be a graph with at least three vertices. Then the following statements are pairwise equivalent:*

- *The graph G is 2-connected.*
- *G does not contain an articulation.*
- *For any two vertices of G, there is a cycle containing both.*
- *For any vertex and any edge of G, there is a cycle containing both.*
- *For any two edges of G, there is a cycle containing both.*

[9] A deep generalization of this theorem for $k = 2$, which we can find in any text book of graph theory, is given by

Theorem 4.4.10 *(Dirac) Let $G = (V, E)$ be a graph with n vertices. If for each vertex v $g(v) \geq \frac{n}{2}$ holds, then G is Hamiltonian, that means there is a cycle containing all vertices of G.*

$G[V \setminus V']$ contains $n-p$ vertices, G_1 has at most $(n-p)/2$ members. If v is a vertex of G_1, then v is adjacent only to vertices of V' or other vertices of G_1. Consequently,

$$g_G(v) \leq p + \frac{n-p}{2} - 1 = \frac{n+p-2}{2} < \frac{n+k-2}{2},$$

which contradicts the hypothesis. □

4.5 Degree sequences

A sequence $g_1, g_2, \ldots, g_n$ of nonnegative integers is called graphical if there is a graph with n vertices $v_1, v_2, \ldots, v_n$ such that $g(v_i) = g_i$ for $i = 1, \ldots, n$. Of course there is the question of whether a given sequence is graphical or not.

Firstly, in view of 4.1.1, the sum of degrees in any graph must be an even number. Secondly,

Theorem 4.5.1 *A graph contains at least one pair of vertices whose degrees are equal.*

Proof. Suppose that the graph G has n vertices. Then there appear to be n possible degree values, namely $0, 1, \ldots, n-1$. However, there cannot be both a vertex of degree 0 and a vertex of degree $n-1$. Hence the n vertices of G can realize at most $n-1$ values for the degrees. Consequently, the assertion. □

But all these conditions are not sufficient. (Example?)[10,11]

A complete answer to the problem of characterizing the graphical sequences is given by the following theorem.

Theorem 4.5.3 *(Erdös, Gallai, [78]) A non-increasing sequence*

$$S : g_1 \geq g_2 \geq \ldots \geq g_n \tag{4.19}$$

[10]Remember that we assume that there are no loops in the graph. On the other hand, allowing loops, we have the following result, which is an exercise for the reader.

Observation 4.5.2 *A sequence*

$$S : g_1, g_2, \ldots, g_n \tag{4.17}$$

of nonnegative integers is graphical if and only if the sum of S is even.

[11]When we are interested in the degree sequence of a connected graph, we have to add the conditions $g_i \geq 1$ for $i = 1, \ldots, n$ (obviously); and

$$\sum_{i=1}^{n} g_i \geq 2(n-1), \tag{4.18}$$

paying attention to 13.1.2 and 4.6.2.

of nonnegative integers is graphical if and only if

$$\sum_{i=1}^{n} g_i \equiv 0(mod\ 2), \tag{4.20}$$

and

$$\sum_{i=1}^{r} g_i \leq r(r-1) + \sum_{i=r+1}^{n} \min\{r, g_i\} \tag{4.21}$$

for every integer r with $1 \leq r \leq n-1$.

For a proof compare [25] or [121].

A much more algorithmically approach is introduced by the following theorem.

Theorem 4.5.4 *(Hakimi) A non-increasing sequence*

$$S : g_1 \geq g_2 \geq \ldots \geq g_n, \tag{4.22}$$

where $n \geq 2$ and $g_1 \geq 1$, is graphical if and only if the sequence

$$S' : g_2 - 1, g_3 - 1, \ldots, g_{g_1+1} - 1, g_{g_1+2}, \ldots, g_n \tag{4.23}$$

is graphical.

Proof. Assume that S' is graphical. This means there is a graph

$$G' = (V' = \{v_2, \ldots, v_n\}, E')$$

with $n-1$ vertices and with S' as its degree sequence:

$$g_{G'}(v_i) = \begin{cases} g_i - 1 & : \quad 2 \leq i \leq g_1 + 1 \\ g_i & : \quad g_1 + 2 \leq i \leq n. \end{cases}$$

A new graph G can be constructed in the following way: $G = (V' \cup v_1, E)$ with

$$E = E' \cup \{\underline{v_1 v_i} : 2 \leq i \leq g_1 + 1\}.$$

Then $g_G(v_i) = g_i$ for each vertex v_i of G; hence S is graphical.

Conversely, assume that S is a graphical sequence.
Among all graphs with n vertices with degree sequence S, let $G = (\{v_1, \ldots, v_n\}, E)$ be one of them such that the sum of the degrees of the vertices adjacent to v_1 is a maximum. We will verify that in G the vertex v_1 must be adjacent to g_1 vertices having degrees $g_2, \ldots, g_{g_1+1}$.
Suppose the converse fact, that means v_1 is not adjacent to vertices having degrees $g_2, \ldots, g_{g_1+1}$. Then there must exist two vertices v_j and v_k with $g_j > g_k$ such that v_1 is adjacent to v_k but not to v_j. Since $g(v_j) > g(v_k)$, there must be some vertex

v_r adjacent to v_j, but not to v_k. Removing the edges $\underline{v_1v_k}$ and $\underline{v_jv_r}$ and adding the edges $\underline{v_1v_j}$ and $\underline{v_rv_k}$ produces a graph H that has the degree sequence S, but the sum of the degrees of the vertices adjacent to v_1 is larger than that in G, a contradiction to the property of G.
Thus, the vertex v_1 must be adjacent to g_1 vertices having degrees $g_2, \ldots, g_{g_1+1}$. Then the graph $G - v_1$ has degree sequence S'; hence S' is graphical. □

Reading 4.5.4 carefully we see that S' is obtained from S by deleting the first number g_1 and subtracting 1 from exactly the next g_1 numbers. The sequence S' is shorter and some numbers are smaller. This immediately gives

Algorithm 4.5.5 *Let*

$$S : g_1, g_2, \ldots, g_n$$

be a sequence of nonnegative integers.
The following algorithm decides whether S is graphical:

1. *If S contains a number exceeding $n-1$, then S is not graphical;*
   Otherwise continue;

2. *If all integers in S are 0, then S is graphical;*
   If S contains a negative integer, then S is not graphical;
   Otherwise continue;

3. *Sort S into a non-increasing sequence S;*

4. *Delete the first number, say g, from S;*
   Subtract 1 from the next g numbers in S;
   Return to step 2.

As an exercise decide whether the following sequences are graphical:
$0, 0, 1, 3, 3, 3, 4, 4, 5, 5$;
$7, 6, 1, 0, 0, 2, 2, 2$; and
$3, 3, 1, 1$.
If the answer is "yes", find a graph with this degree sequence.

Altogether, there are algorithms which decide in linear time whether a given sequence is graphical or not. On the other hand, to describe all possible graphs with a given sequence or the number of such graphs are, in general, hard problems. Later in the present script we will answer several subquestions.

4.6 Trees and forests

A tree is defined to be a connected graph without cycles.

A vertex with degree one is called a leaf. A vertex in a tree that is not a leaf is called an internal vertex.

Theorem 4.6.1 *Each tree with more than one vertex has at least two leaves.*

Sketch of the *proof.* Consider a longest path in the tree. Its endvertices must be leaves. □

The following theorem establishes several of the most useful characterizations of a tree. Each contributes a deeper understanding of the structure of this basic type of graphs. In our further investigations we will use these equivalences permanently.

Theorem 4.6.2 *Let $G = (V, E)$ be a graph with n vertices, where $n > 1$.*[12] *Then the following properties are pairwise equivalent (and each characterizes a tree):*

- *G is connected and has no cycles.*
- *G is connected and contains exactly $n - 1$ edges.*
- *G has exactly $n - 1$ edges and has no cycles.*
- *G is maximally acyclic; that means G has no cycles, and if a new edge is added to G, exactly one cycle is created.*
- *G is minimally connected; that means G is connected, and if any edge is removed, the remaining graph is not connected.*
- *Each pair of vertices of G is connected by exactly one path.*

Proof. We will only show that a tree with n vertices has $n-1$ edges. The remaining statements should be exercises for the reader.

The proof uses induction. A tree with exactly two vertices has one edge; the result follows for $n = 2$.
Assume that the result is true for all trees with less than n vertices. Consider a tree T with n vertices. In view of 4.6.1 T contains a leaf v. Obviously $T-v$ is a tree with $n-1$ vertices, and by the induction hypothesis, with $n-2$ edges. Thus T has $n-1$ edges. □

As a consequence of our considerations, we are interested in the distribution of vertices of a given degree in a tree. Let $T = (V, E)$ be a tree with n vertices. n_i denotes the number of vertices of degree i and $\Delta = \Delta(T)$ the maximum degree in T:

$$\Delta(T) = \max\{g_T(v) : v \in V\}. \tag{4.24}$$

Then, of course,

$$n_1 + n_2 + \ldots + n_\Delta = n. \tag{4.25}$$

In view of 4.1.1 and 4.6.2, we have

$$n_1 + 2 \cdot n_2 + \ldots + \Delta \cdot n_\Delta = 2|E| = 2n - 2. \tag{4.26}$$

Subtracting this equation from two times (4.25) yields

[12] By definition a graph with one vertex and without edges is also a tree.

Theorem 4.6.3 *It holds that*

$$n_1 = 2 + \sum_{i=3}^{\Delta(T)} (i-2) \cdot n_i, \tag{4.27}$$

for any tree T, where n_i denotes the number of vertices of degree i and $\Delta(T)$ is the maximum degree in the tree.

Consequently,

a) Considering only trees without vertices of degree two, the number of internal vertices is less than the number of leaves and a binary tree has the maximum possible number of internal vertices for a given number of leaves.

b) Each tree T with more than one vertex has at least $\Delta(T)$ leaves.

Another consequence is that trees can be recognized in linear time:

Observation 4.6.4 *Trees can be generated recursively by appending repeatedly leaves starting with one vertex and vice versa leads to an elimination scheme where repeatedly leaves are deleted.*

An obvious generalization: A forest is defined as a graph whose connected components are trees. That means, in view of 4.6.2, a forest is

- an acyclic graph, or equivalently;
- a graph in which each edge is a bridge.

Observation 4.6.5 *Each forest is a bipartite graph.*

Theorem 4.6.6 *Let G be a forest with n vertices and c components. Then G contains exactly $n - c$ edges.*

Proof. Let $G_i = (V_i, E_i)$, $i = 1, \ldots, c$ be the components of G. Each component is a tree, and 4.6.2 says

$$|E_i| = |V_i| - 1.$$

Summing up all equations, the addition principle gives the assertion. □

4.7 The matrix of adjacency

Representing graphs by matrices remains important as a conceptual tool, and it helps us bring the power of linear algebra to graph theory.

Let $G = (V, E)$ be a graph and assume that the vertices are labeled, i.e. $V = \{v_1, ..., v_n\}$, that means that A is based on an ordering chosen for the vertices. Then

we define the matrix of adjacency $A(G) = (a_{ij})_{i,j=1,\ldots,n}$ by

$$a_{ij} = \begin{cases} 1 & : \text{ the vertices } v_i \text{ and } v_j \text{ are adjacent} \\ 0 & : \text{ otherwise.} \end{cases}$$

For any graph G the matrix $A(G)$ is symmetric and

$$\det A(G) = 0. \tag{4.28}$$

For a multigraph G we similarly define the matrix of adjacency by

$$a_{ij} = \text{ number of edges from } v_i \text{ to } v_j. \tag{4.29}$$

The number of elements in the matrix is n^2.[13]
The matrix of adjacency contains all information about the structure of the graph.[14]
As an example we consider the Petersen graph

$$G_{\text{petersen}} = (\{v_1, \ldots, v_5, w_1, \ldots, w_5\}, E)$$

with

$$\begin{aligned} E \quad = \quad & \{\underline{v_i v_{i+1}} : i = 1, \ldots, 4\} \cup \{\underline{v_5 v_1}\} \\ & \cup \{\underline{v_i w_i} : i = 1, \ldots, 5\} \\ & \cup \{\underline{w_1 w_3}, \underline{w_3 w_5}, \underline{w_5 w_2}, \underline{w_2 w_4}, \underline{w_4 w_1}\}. \end{aligned}$$

Or equivalently,

$$A(G_{\text{petersen}}) = \begin{pmatrix} & 1 & & & 1 & 1 & & & & \\ 1 & & 1 & & & & 1 & & & \\ & 1 & & 1 & & & & 1 & & \\ & & 1 & & 1 & & & & 1 & \\ 1 & & & 1 & & & & & & 1 \\ 1 & & & & & & & 1 & 1 & \\ & 1 & & & & & & & 1 & 1 \\ & & 1 & & & 1 & & & & 1 \\ & & & 1 & & 1 & 1 & & & \\ & & & & 1 & & 1 & 1 & & \end{pmatrix}$$

I. Let $A = (a_{ij})$ be the adjacency matrix for the graph $G = (V = \{v_1, ..., v_n\}, E)$. Then, obviously, the equation $a_{ij} = 1$ means that there is a chain of length 1 from v_i to v_j. Now consider the k-th power of A.

[13] If the number of edges is small in comparison to n^2, then many of the elements are 0. Then the amount of memory capacity is excessively big. In such a case storage by adjacency lists, where we store with each vertex v_i a list containing all vertices that are adjacent to v_i, are more useful. Compare [101].

[14] The adjacency matrix of a graph does depend on the labeling of the vertices; that is, a different labeling of the vertices may result in a different matrix, but they are closely related in that one can be obtained from the other simply by interchanging rows and columns.

Theorem 4.7.1 *Let $A = A(G) = (a_{ij})_{i,j=1,...,n}$ be the adjacency matrix for the graph $G = (V = \{v_1, ..., v_n\}, E)$. Let*

$$A^k = (a_{ij}^{(k)})_{i,j=1,...,n} \tag{4.30}$$

be the k-th power of A.
The coefficient $a_{ij}^{(k)}$ is the number of different chains of length exactly k from the vertex v_i to the vertex v_j.

Proof. Using induction over k.
For $k = 0, 1$ the assertion is obvious.
Now let $k > 1$.

$$a_{ij}^{(k)} = \sum_{l=1}^{n} a_{il}^{(k-1)} a_{lj}. \tag{4.31}$$

By the induction hypothesis $a_{il}^{(k-1)}$ is the number of different chains of length exactly $k-1$ from v_i to v_l. Whenever l and j are adjacent we can continue such chains to chains of length k. □

Hence, the graph G is connected if and only if for any pair of distinct vertices v_i and v_j there is a number $k = k(i, j)$ between 1 and $n-1$ such that $a_{ij}^{(k)} > 0$.

II. Each (connected) graph $G = (V, E)$ is a metric space: Note that the length of a path is defined as the number of its edges. A distance ρ on V is given by

$$\rho(v, v') = \text{length of a shortest path between } v \text{ and } v' \text{ in } G, \tag{4.32}$$

for two different vertices v and v', and $\rho(v, v) = 0$.
The length of a shortest path between v and v' in G can be (easily) found with an algorithm created by Dijkstra [67].[15,16] For any connected graph $G = (V, E)$ the pair (V, ρ) is a metric space, called the metric closure of G.[17]

[15] which is a consequence of the following principle:

Observation 4.7.2 *(Bellman [21]) Let $G = (V, E)$ be a graph, and let v and v' be two vertices of G. If $e = \underline{wv'}$ is the final edge of some shortest path $v, \ldots, w, v'$ from v to v', then $v, \ldots, w$ (that is the path without the edge e) is a shortest path from v to w.*

[16] The problem of a longest path is intractable, since it is $\mathcal{NP}$-complete, [97].

[17] We can find the metric closure in a simpler way by the following typical "matrix" algorithm:

Algorithm 4.7.3 *(Floyd [89]) Let $G = (V = \{v_1, \ldots, v_n\}, E)$ be a graph. The metric closure $G^f = (V, \rho)$ can be found by the following procedure:*

1. **for** $i := 1$ **to** n **do**
   for $j := 1$ **to** n **do**
   if $\underline{v_i v_j} \in E$ **then** $\rho(v_i, v_j) := 1$ **else** $\rho(v_i, v_j) := \infty$;
2. **for** $i := 1$ **to** n **do**
   for $j := 1$ **to** n **do**
   for $k := 1$ **to** n **do**
   if $\rho(v_j, v_i) + \rho(v_i, v_k) < \rho(v_j, v_k)$ **then** $\rho(v_j, v_k) := \rho(v_j, v_i) + \rho(v_i, v_k)$.

Theorem 4.7.4 *Let $G = (V = \{v_1, \ldots, v_n\}, E)$ be a connected graph, let $A = A(G)$ be its adjacency matrix and let $A^k = (a_{ij}^{(k)})_{i,j=1,\ldots,n}$, $k = 1, 2, \ldots$. Then*

$$\rho(v_i, v_j) = \min\{k : a_{ij}^{(k)} > 0\} \tag{4.33}$$

holds true for any two distinct vertices v_i and v_j.

For more properties of the metric closure see [258].

III. It is an interesting topic to investigate whether other terms in linear algebra have an graph-theoretical impact.
For instance, let $A = A(G)$ be the adjacency matrix for the graph $G = (V, E)$. Then

$$|E| = \frac{1}{2} \cdot \text{trace } A^2 \tag{4.34}$$

$$\#\text{ triangles in } G = \frac{1}{6} \cdot \text{trace } A^3 \tag{4.35}$$

Is there a meaning of the eigenvalues of $A(G)$?[18]

4.8 Planar graphs

Planarity asserts that it is possible to represent the graph in the plane in such a way that the vertices correspond to distinct points and the edges to simple Jordan curves connecting the points of its endvertices such that every two curves are either disjoint or meet only at a common endpoint. Not each graph is planar.[19]

I. An embedding of a planar graph is called a plane graph. It determines a partition of the plane into regions. Exactly one of these regions is unbounded.
The number of regions can be computed by the classical formula of Euler, which is the earliest known equation in topology:

[18] The answer is "yes", compare [29] and [61]. In particular,

Theorem 4.7.5 *Let G be a graph with n vertices, and let $A(G)$ be its adjacency matrix.*

a) *All eigenvalues are real numbers.*

b) *Every eigenvalue λ satisfies $|\lambda| \leq \Delta(G)$.*

c) *$\Delta(G)$ is an eigenvalue if and only if G is regular.*

d) *If $-\Delta(G)$ is an eigenvalue then G is regular and bipartite.*

e) *$\delta(G) \leq \lambda_{max} \leq \Delta(G)$.*

f) *If G' is a spanning subgraph of G then $\lambda_{min}(G) \leq \lambda_{min}(G') \leq \lambda_{max}(G') \leq \lambda_{max}(G)$.*

[19] As contrast each graph can be embedded into the three-dimensional space $\mathbb{R}^3$ such that no two curves which are the embeddings of the edges intersect each other outside of the vertices. This is easy to see: Arrange the vertices on a line, and consider planes through this line.

Theorem 4.8.1 *Let $G = (V, E)$ be a connected and plane graph, and let f denote the number of regions (including the single unbounded region) of an embedding of G in the plane. Then*

$$|V| - |E| + f = 2. \tag{4.36}$$

Proof. If there is a cycle, remove one edge from it. The effect is to reduce f by 1, since two regions are amalgamated into one. So the resulting graph satisfies the same equation. Repeat this process until no cycles remain. The final graph must be a tree, with

$$|V| - |E| + f = |V| - (|V| - 1) + 1 = 2,$$

paying attention 4.6.2. □

Of course, there may be several different embeddings of a planar graph in the plane.[20] 4.8.1 implies that, no matter how a connected planar graph is embedded, the number of regions is determined.

Corollary 4.8.2 *Under the assumption of 4.8.1 and assuming that $|V| \geq 3$ it holds that*

$$|E| \leq 3|V| - 6 \text{ and} \tag{4.37}$$

$$f \leq 2|V| - 4. \tag{4.38}$$

The first inequality can be strengthened to

$$|E| \leq \frac{t}{t-2}(|V| - 2), \tag{4.39}$$

where all cycles are of length at least t.

Proof. Embed G in the plane. Then there are r_j regions with j edges on the boundary. With similar arguments than for the proof of 4.1.1, we find

$$2|E| = \sum_{j \geq 3} j r_j \geq \sum_{j \geq 3} 3 r_j = 3f.$$

Therefore, $f \leq \frac{2}{3}|E|$. Together with 4.8.1 this gives the first two inequalities.
The second part of the proof remains as an exercise for the reader.[21] □

As an exercise deduce that a planar bipartite graph with n vertices has at most $2n - 4$ edges.

[20] Embedding a graph in the plane is equivalent to embedding it on the sphere. This can be seen with the aid of a stereographic projection.
The unbounded region in the plane correspond with the region on the sphere which includes the "North Pole". Consequently, each planar graph can be embedded in the plane in such a way that any one of its regions may be made the unbounded region.

[21] The corollary should not be misinterpreted to mean that if $|E| \leq 3|V| - 6$, then a connected graph is planar. Many non-planar graphs also satisfy this inequality.

Theorem 4.8.3 *In each planar graph $G = (V, E)$ there exists a vertex v with $g_G(v) \leq 5$. In other terms, $\delta(G) \leq 5$.*

Proof. Otherwise

$$2 \cdot |E| = \sum_{v \in V} g_G(v) \geq 6 \cdot |V|,$$

contradicts 4.8.2. □

Corollary 4.8.4 *In any planar graph $G = (V, E)$ the average degree is less than 6:*

$$\tilde{d}(G) = \frac{1}{|V|} \sum_{v \in v} g(v) < 6. \tag{4.40}$$

If the graph has the additional property that each vertex has at least the degree three, it is easy to prove that the following inequalities are valid for the numbers $n = |V|$ and $m = |E|$ for a planar graph $G = (V, E)$ with f regions (including the single unbounded region) in the embedding of G:

$$n \leq \frac{2}{3}m, \qquad m \leq 3n - 6, \tag{4.41}$$

$$m \leq 3f - 6, \qquad f \leq \frac{2}{3}m, \tag{4.42}$$

$$n \leq 2f - 4, \qquad f \leq 2n - 4. \tag{4.43}$$

A planar graph G is called maximal planar, if G is planar but cannot be extended to a larger planar graph by adding an edge. It is easy to see that the following properties are true for any maximal planar graph $G = (V, E)$:

a) $|E| = 3|V| - 6$, that means, in the first inequality of 4.8.2 holds equality.

b) G has at least three vertices of degree not exceeding 5.

And not so easy to see

c) G is 3-connected.

d) G can be embedded in the plane such that each (internal) region is a triangle.

II. A planar graph $G = (V, E)$ is called outer-planar if it can be embedded into the plane such that all vertices lying on the boundary of exactly one region. For practice the reader should discuss the following results. Similar to 4.8.2 we find.

Theorem 4.8.5 *It holds for any outer-planar graph $G = (V, E)$ and its embedding in the plane*

$$|E| \leq 2|V| - 3 \text{ and} \tag{4.44}$$

$$f \leq |V| - 1. \tag{4.45}$$

For practice show that for any outer-planar graph G it holds $\delta(G) \leq 3$ and $\tilde{d}(G) < 4$.

An outer-planar graph is called maximal outer-planar if no edge can be added without losing outer-planarity.

Theorem 4.8.6 *Let G be the embedding of a maximal outer-planar graph with at least $n \geq 3$ vertices, all lying on the exterior (unbounded) region. Then G has $n-2$ interior regions.*

Proof. We use induction over n.
Obviously, the theorem is true for $n = 3$.
Let G have $n+1$ vertices and f interior regions. Clearly, G must have a vertex v of degree 2. v lies on the boundary of the exterior region. In $G - v$ we reduce the number of the vertices and the number of interior regions each by exactly 1, so that $f - 1 = n - 2$. Hence, the assertion. □

The following properties are true for a maximal outer-planar graph $G = (V, E)$:

a) $|E| = 2|V| - 3$, that means, in the first inequality of 4.8.5 holds equality.

b) G has at least three vertices of degree not exceeding 3.

c) G has at least two vertices of degree 2.

d) G is 2-connected.[22]

And not so easy to see

e) G can be embedded as a triangulation of a polygon.

Theorem 4.8.7 *Consider graphs and its complements.*

a) *Every planar graph with at least nine vertices has a non-planar complement, and nine is the smallest such number.*

b) *Every outer-planar graph with at least seven vertices has a non-outer-planar complement, and seven is the smallest such number.*

Proof. These results are proved by exhaustion; the upper bounds can be created with help of (4.15) and 4.8.2 or 4.8.5, respectively. No elegant proofs are known. □

III. For disconnected graph the situation is a little bit different. Obviously, a graph is planar if and only if each component is planar.

Theorem 4.8.8 *Let $G = (V, E)$ be a plane graph with c components, and let f denote the number of regions (including the single unbounded region) of an embedding of G in the plane. Then*

$$|V| - |E| + f = c + 1. \tag{4.46}$$

The proof remains as an exercise for the reader.

[22] Can an outer-planar graph be 3-connected?

4.9 Directed graphs

For further discussions we introduce graphs whose edges are directed. This concept of directed graphs, called digraphs, is more complicated than the concept of graphs, since there are several new questions.

A digraph is a pair $G = (V, E)$ consisting of a finite set V of vertices and a set $E \subseteq V \times V$ of (ordered) pairs of vertices, which we call arcs. Hence, a digraph $G = (V, E)$ is essentially a relation over V, where arcs (v, v) are not allowed.
The terminology used in discussing digraphs is quite similar to that used for graphs. Moreover, we will understand each digraph also as a graph and we will use the graph-theoretical methods for digraphs, too.
Let $G = (V, E)$ be a digraph. For two vertices v and v' with $e = (v, v') \in E$ we say that v is the immediate ancestor of v', and v' is the immediate successor of v and v. Furthermore we say that v is adjacent to v', whereas v' is adjacent from v. The arc e is called directed from v to v'.
The indegree $g^{in}(v)$ of v is the number of immediate ancestors of v and the outdegree $g^{out}(v)$ of v is the number of immediate successors of v. Obviously,

$$g(v) = g^{in}(v) + g^{out}(v), \tag{4.47}$$

for each vertex v in a digraph. It is easy to see that

Observation 4.9.1 *For each digraph $G = (V, E)$ it holds*

$$\sum_{v \in V} g^{in}(v) = \sum_{v \in V} g^{out}(v) = |E|. \tag{4.48}$$

A directed chain is a sequence

$$v_1, (v_1, v_2), v_2, (v_2, v_3), v_3, \ldots, v_m, (v_m, v_{m+1}), v_{m+1}$$

of arcs and vertices of G. A directed chain in which each vertex occurs only once is called a directed path; more exactly, a directed path interconnecting the vertices v_1 and v_{m+1}. Then the number m denotes the length of the path. A single vertex is a directed path of length 0.
A digraph is called strongly connected if for any pair v and v' of distinct vertices there is a directed path from v to v' and also a directed path from v' to v.

Observation 4.9.2 *The relationship of being strongly connected is an equivalence relation.*

A directed cycle is a directed chain with at least one arc and with the following properties: No arc appears twice in the sequence and the two endvertices of the chain are the same.

Theorem 4.9.3 *Every acyclic digraph contains at least one vertex of outdegree 0 and at least one vertex of indegree 0.*

Proof. Let P be a directed path of maximum length in the digraph. Assume that the path connects the vertices u and v. If u is adjacent from any vertex of P, then a cycle is produced, a contradiction. If u is adjacent from a vertex w not in P, then $w, (w, u), u, P$ is a directed path whose length exceeds that of P, also a contradiction. Hence $g^{in}(u) = 0$.
Similarly, $g^{out}(v) = 0$. □

For a digraph $G = (V, E)$ we have an ancestor/successor-relation: We say that the vertex v is an ancestor for v', and the vertex v' is a successor of v if there is a directed path from v to v'.

An important question is: Given a (connected) graph; is there an assignment of directions to the edges so that the resulting digraph is strongly connected? If the answer is positive, the graph is called orientable. Remember that an edge of a graph is called a bridge if its removal disconnects the graph. Assume that the edge $\underline{vv'}$ is a bridge, and we choose the direction by (v, v'), then the vertices v and v' cannot be strongly connected in the digraph. Thus a graph with a bridge is not orientable. Surprisingly the converse fact is also true.

Theorem 4.9.4 *A connected graph without bridges is orientable.*

Proof. Let $G = (V, E)$ be a connected and bridgeless graph.
In view of 4.4.4 every edge of G is part of a cycle.
Let $C : v_1, v_2, \ldots, v_n, v_1$ be a cycle of G. Then we direct its edges as follows:

i) (v_n, v_1);

ii) (v_i, v_{i+1}) for $i = 1, \ldots, n-1$;

iii) If there exists an edge joining nonconsecutive vertices of C, then choose a direction for this edge arbitrarily.

Of course the digraph $G[C]$ is strongly connected. If every vertex of G belongs to C, then G is orientable.
Assume that there is a vertex of G not belonging to C. Since G is connected, there exists a vertex w_1 outside of C such that $\underline{w_1 v_j}$ is an edge for an index j with $1 \leq j \leq n$. Since this edge is not a bridge, we may assume that there is a cycle $C_1 : w_1, w_2 = v_j, w_3, \ldots, w_m, w_1$ of G. Then we direct the edges as follows:

i) (w_m, w_1);

ii) (w_1, w_2);

iii) (w_i, w_{i+1}) for $i = 2, \ldots, m-1$ provided that the edge $\underline{w_i w_{i+1}}$ has not already been given a direction;

iv) If there exists an edge joining two vertices of C_1 or joining a vertex of C_1 to a vertex of C, and has not yet received a direction then choose a direction for this edge arbitrarily.

It is not hard to see that $G[C \cup C_1]$ is strongly connected. If this digraph contains all the vertices of G, the desired result follows immediately. Otherwise, we continue this process. □

Let $G = (V, E)$ be a digraph and assume that the vertices are labeled, i.e. $V = \{v_1, ..., v_n\}$. Then we define the Matrix of adjacency $A(G) = (a_{ij})_{i,j=1,...,n}$ by

$$a_{ij} = \begin{cases} 1 & : \text{ there is an arc from } v_i \text{ to } v_j \\ 0 & : \text{ otherwise.} \end{cases}$$

Another kind of matrix identifies the edges with its incident vertices: Let $G = (V, E)$ be a digraph and assume that the vertices and also the edges are labeled, i.e. $V = \{v_1, ..., v_n\}$ and $E = \{e_1, ..., e_m\}$. Then we define the Matrix of incidence $I(G) = (d_{ij})_{i=1,...,n,j=1,...,m}$ by

$$d_{ij} = \begin{cases} 1 & : \quad e_j = (v_i, .) \\ -1 & : \quad e_j = (., v_i) \\ 0 & : \text{ otherwise} \end{cases}$$

Observation 4.9.5 *The sum of the entries in any column equals 0.*
Conversely, each matrix in which the entries in any column are 0 except exactly one 1 and exactly one -1, is the incidence matrix of a digraph.

As an exercise discuss the interrelation between the matrix of adjacency and of the matrix of incidence for a digraph.

4.10 Intersection graphs

Let $\mathcal{C}$ be a collection of nonempty sets in a universe. The intersection graph of $\mathcal{C}$ is obtained by representing each set in $\mathcal{C}$ by a vertex and connecting two vertices by an edge if and only if their corresponding sets intersects.
When $\mathcal{C}$ is allowed to be an arbitrary collection of sets, the class of graphs obtained as intersection graphs is the class of all graphs.

Theorem 4.10.1 *(Marczewski) Every graph is (isomorphic to) the intersection graph of some collection of sets.*

Proof. Let $G = (V, E)$ be a graph.

$$S(v) = \{\{v, w\} : \underline{vw} \in E\} \cup \{v\}. \tag{4.49}$$

It is easy to see that for different vertices v and w,

$$\underline{vw} \in E \text{ if and only if } S(v) \cap S(w) \neq \emptyset. \tag{4.50}$$

Thus G is isomorphic to the intersection graph of

$$\mathcal{C} = \{S(v) : v \in V\}. \tag{4.51}$$

□

Perhaps the most interesting applications of intersection graphs have arisen from taking special classes of sets.

The intersection graph of a collection of intervals on a linearly ordered set (like the real line) is called an interval graph.[23,24]
Given a graph $G = (V, E)$, we shall ask whether it is isomorphic to an interval graph. The characterization of interval graphs is not easy.
A graph G is called chordal if every cycle with at least four vertices has an edge (called a chord) connecting two non-consecutive vertices in the cycle.[25] Not hard to see, compare [101] or [220], is the following implication.

Lemma 4.10.3 *An interval graph is chordal.*

A graph $G = (V, E)$ is called transitive orientable if each edge can be assigned a direction in such a way that the the resulting digraph $G' = (V, F)$ satisfies the following condition: $(u, v) \in F$ and $(v, w) \in F$ imply $(u, w) \in F$. Easy to see, compare [101], is the following implication.

Lemma 4.10.4 *The complement of an interval graph is transitive orientable.*

4.10.3 and 4.10.4 provide necessary, but not sufficient, conditions for interval graphs. Put these two properties together, we get a sufficient condition:

Theorem 4.10.5 *(Gilmore, Hoffman, compare [101]) A graph G is an interval graph if and only if G is chordal and its complement G^c is transitive orientable.*

Now we go a dimension higher. Given a finite set of nonoverlapping circles in the plane.[26] Considering $\mathcal{C}$ as set of circles, the intersection graph of $\mathcal{C}$ is called a coin graph if the intersection is only touching of the circles. The following theorem gives a nice description.

Theorem 4.10.6 *(Koebe) Every planar graph is a coin graph.*

For this and related topics compare Ziegler [257].

[23]The intervals may be open closed, or half-open.

[24]Interval graphs arose from a problem of genetics as follows. On the basis of mutation data, one can tell if two subsets of the fine structure inside the gene overlap. Is this overlap information consistent with the hypothesis that the fine structure inside the gene is linear? More exactly, tests can be performed to determine if two chromosomes overlap one another, and the problem is to prove or disprove that a set of chromosomes are linked together in linear order. Construct the graph whose edges are the pairs of overlapping chromosomes; if this graph is not an interval graph, it follows that the chromosomes cannot be linked in linear order. For a detailed discussion compare [204].
Roberts [205] describes problems in social sciences, which are problems of seriation. The approach starts with overlap information, where the intervals are a possible chronological order.

[25]The concept of chordality gives a partial answer to the question when in (4.21) equality holds:

Theorem 4.10.2 *([90] and [120]) A graph G satisfies the inequality (4.21) for $\tilde{g} = \max\{i : g_i \geq i - 1\}$ with equality if and only if G itself and its complement G^c are chordal.*

[26]We do not assume that the circles are of the same size.

4.11 Further reading

We introduced several knowledge of graphs and networks. Graphs are among the most basic of all mathematical structures. Consequently, there are a lot of other aspects to consider. Further graph theoretic terminology and statements are given in most standard textbooks, for example

1. Bang-Jensen, Gutin: Digraphs; [17].
2. Bollobas: Graph Theory; [29].
3. Chartrand, Lesniak: Graphs and Digraphs; [45].
4. Diestel: Graph Theory; [66].
5. Gross, Yellen: Graph theory and its Applications; [109].
6. Lovász, Pelikán, Vestergombi: Discrete Mathematics; [164].

For a history of the theory of graphs see Aigner [3], Prömel [195], Sachs [210], and Sedlacek [219].

Chapter 5

Labeled Graphs

A graph $G = (V, E)$ is called labeled if there is a bijective mapping from V onto a set of $|V|$ distinct names in such a way as to be they are distinguishable from each other.

With most enumeration problems, counting the number of unlabeled things is harder than counting the number of labeled things. So it is with graphs. This observation holds not only for graphs, but also for trees, digraphs, relations, and so forth.

5.1 All graphs

Let us consider the problem of counting all graphs with n vertices. Such a graph has at most $\binom{n}{2} = n(n-1)/2$ edges. Hence, when we observe that each of the possible edges is either present or absent,

Theorem 5.1.1 *For enumerating graphs we obtain*

a) *The number of labeled graphs with n vertices and exactly m edges equals*

$$\binom{\binom{n}{2}}{m}. \tag{5.1}$$

b) *The number $graph(n)$ of labeled graphs with n vertices equals*

$$graph(n) = 2^{\binom{n}{2}} = \sqrt{2}^{n(n-1)}. \tag{5.2}$$

As an exercise estimate the number (5.1) from above and from below, and discuss these bounds.

To deal with graphs it is often necessary to generate these structures algorithmically. This question is closely related to the problem of counting graphs.[1] To generate

[1] And the problem of storing a graph in a computer, compare [164].

all labeled graphs is not hard: remember that a labeled graph is completely described by its adjacency matrix. There is a one-to-one correspondence between labeled graphs with n vertices and $n \times n$ symmetric binary matrices with all entries on the leading diagonal equal to 0. Hence, we have the following optimal generating technique:

Algorithm 5.1.2 *Let n be an integer greater than 1. The following procedure generates all labeled graphs with n vertices:*

1. *Determine $b := n(n-1)/2$;*
2. *Initialize $a = (0, \ldots, 0)$ in $\{0,1\}^b$;*
3. *Assuming a is the upper half of an $n \times n$ matrix A;*
   complete the matrix by setting $a_{ji} = a_{ij}$ and $a_{ii} = 0$, yielding the adjacency matrix for a graph;
   Set $a := a + 1$ (in $\{0,1\}^b$).

For more facts about generating all graphs see Nägler, Stopp [177].

$\mathcal{G}(n,m)$ denotes the set of all graphs with n vertices and m edges. Let S be a set of edges of the complete graph K_n. What is the proportion of graphs in $\mathcal{G}(n,m)$ which contains all the edges in S? Answer: Let $|S| = q$. If a graph in $\mathcal{G}(n,m)$ contains S, that uses up q of its edges, so there are $m-q$ remaining edges to place in $N-q$ positions, where $N = \binom{n}{2}$. The number of ways of doing that is

$$\binom{N-q}{m-q} = \frac{(N-q)!}{(m-q)!(N-m)!}.$$

Hence,

Theorem 5.1.3 *The proportion of graphs in $\mathcal{G}(n,m)$ containing a set of q edges is*

$$\frac{\binom{N-q}{m-q}}{|\mathcal{G}(n,m)|} = \frac{m \cdot (m-1) \cdots (m-q+1)}{N \cdot (N-1) \cdots (N-q+1)},$$

where $N = \binom{n}{2}$.

5.2 The number of bipartite graphs

Consider two finite and disjoint sets V_1 and V_2 with $n_i = |V_i|$, $i = 1, 2$. Then there are $n_1 \cdot n_2$ possible edges interconnecting the vertices of both. In view of 2.1.1 we obtain.

Theorem 5.2.1 *There are $2^{n_1 \cdot n_2}$ bipartite graphs with $n_1 + n_2$ labeled vertices.*

Let n be a positive integer, and k a number with $1 \leq k \leq n$. Selecting k elements from an n-element set creates a split (=bipartition), since are also the remaining $n-k$ elements are selected. Therefore,

Theorem 5.2.2 *The number of bipartite graphs for n vertices is at most*

$$\frac{1}{2}\sum_{k=1}^{n-1}\binom{n}{k}2^{k(n-k)}. \tag{5.3}$$

Numerically,

n	1	2	3	4	5	6	7
	0	2	12	80	720	9152	165312

Asymptotitally,

$$\begin{aligned}\frac{1}{2}\sum_{k=1}^{n-1}\binom{n}{k}2^{k(n-k)} &\leq \frac{1}{2}\sum_{k=1}^{n-1}\binom{n}{k}2^{n^2/4} = \frac{1}{2}\cdot 2^{n^2/4}\sum_{k=1}^{n-1}\binom{n}{k}\\ &< \frac{1}{2}\cdot 2^{n^2/4}\sum_{k=0}^{n}\binom{n}{k} = \frac{1}{2}\cdot 2^{n^2/4}\cdot 2^n,\end{aligned}$$

such that for a nonnegative integer n there are less than

$$2^{\frac{n^2}{4}+n-1} = \sqrt[4]{2}^{n^2+4n-4} = (1.18920\ldots)^{n^2+4n-4} \tag{5.4}$$

bipartite graphs with n labeled vertices.
In 4.6.5 we remarked that each forest is a bipartite graph. Consequently, by estimating the number of forests we will create lower bounds for the number of bipartite graphs.

5.3 Regular graphs

In further considerations we will count specific classes of graphs, and start with the following concept.
A graph $G = (V, E)$ is called regular, or more exactly regular of degree r, if each vertex has degree exactly r, $0 \leq r \leq |V| - 1$. The empty graph is 0-regular. A graph which is regular of degree 1 is called a perfect matching, of degree 2 is a collection of cycles, and of degree 3 is called a cubic graph. For practice the reader should prove for a connected r-regular graph G:

- If r is even, then G contains no bridge.
- If r is odd, then always exists such graph with a bridge.

In view of 4.1.1 we find

Theorem 5.3.1 *For a graph $G = (V, E)$ which is regular of degree r it holds*

$$r \cdot |V| = 2 \cdot |E|. \tag{5.5}$$

Consequently, if r is odd then $|V|$ must be an even number.

It is easy to see that if r and $n = |V|$ are not both odd and $0 \leq r \leq n-1$, then always exists an r-regular graph with n vertices. The reader should prove this statement and use the proof to show that the number of regular graphs increases at least exponentially.
On the other hand, the number of labeled r-regular graphs is at most n^{rn}, which is easy to see: For each vertex we must determine r adjacent vertices. There are no more than n^r possibilities, and altogether are no more than $(n^r)^n = n^{rn}$ possible ways to specify such a graph. The bound is essentially less than $2^{\binom{n}{2}}$ if

$$r < \frac{n-1}{2 \cdot \log n}.$$

We can estimate the number of regular graphs better, using 5.3.1 in (5.1).

$$\binom{\binom{n}{2}}{\frac{rn}{2}} \leq \frac{1}{e}\left(\frac{e\binom{n}{2}}{\frac{rn}{2}}\right)^{rn/2} = \frac{1}{e}\left(\frac{e(n-1)}{r}\right)^{rn/2}.$$

Hence,

Theorem 5.3.2 *The number of r-regular labeled graphs with n vertices and $m = \frac{r}{2} \cdot n$ edges is at most*

$$\binom{\binom{n}{2}}{\frac{rn}{2}} \leq \frac{1}{e}\sqrt{\frac{e(n-1)}{r}}^{rn} = \frac{1}{e}\left(\frac{2em - er}{r^2}\right)^m. \tag{5.6}$$

1- and 2-regular graphs are of minor interest, but 3-regular graphs will play a role in our further investigations.

Corollary 5.3.3 *The number of labeled cubic graphs with n vertices is less than*

$$\frac{1}{e}(n\sqrt{n})^n. \tag{5.7}$$

5.4 The number of connected graphs

Of course, the number of connected graphs is less than the number of all graphs; and we are interested in this fact more exactly.

Theorem 5.4.1 *Denote by $conn(n)$ the number of connected graphs with n labeled vertices. Then*

$$conn(n) = 2^{n(n-1)/2} - \sum_{i=1}^{n-1} \binom{n-1}{i} \cdot 2^{i(i-1)/2} \cdot conn(n-i). \tag{5.8}$$

Proof. We show that the sum

$$\sum_{i=1}^{n-1} \binom{n-1}{i} \cdot 2^{i(i-1)/2} \cdot \mathrm{conn}(n-i) = \sum_{i=1}^{n-1} \binom{n-1}{i} \cdot \mathrm{graph}(i) \cdot \mathrm{conn}(n-i) \quad (5.9)$$

is the number of disconnected graphs.

A proper component of a graph has at least one and at most $n-1$ vertices. Let i be the number of vertices outside of such a component and let $n-i$ be the number of vertices inside, $1 \leq i \leq n-1$.

In view of 5.1.1, for a fixed number i there are

$$\mathrm{graph}(i) = 2^{\binom{i}{2}} = 2^{i(i-1)/2} \quad (5.10)$$

different graphs with i vertices. We can choose

$$\binom{n-1}{i} \quad (5.11)$$

vertices. (5.10) and (5.11) together imply the assertion. □

Another way to compute $\mathrm{conn}(n)$ is given by the method of generating functions, which means that expand two series and compare the coefficients:

$$\sum_{n\geq 1} \mathrm{conn}(n) \frac{x^n}{n!} = \ln \sum_{n\geq 0} 2^{\binom{n}{2}} \frac{x^n}{n!}. \quad (5.12)$$

For a proof see [227].

Unfortunately, an explicit formula for the function *conn* is unknown, but in [122] and [177] the number of connected labeled graphs is estimated by

Number n of vertices	Number $\mathrm{conn}(n)$ of connected graphs
1	1
2	1
3	4
4	38
5	728
6	26,704
7	$1.8662\ldots \cdot 10^6$
8	$2.5154\ldots \cdot 10^8$
9	$6.6296\ldots \cdot 10^{10}$
10	$3.4487\ldots \cdot 10^{13}$

It seems that the function conn increases exponentially, and indeed this is true, which we will prove later in a more common context: 5.8.2 and 7.3.3.

5.5 Eulerian and Hamiltonian graphs

I. The origin of graph theory is the so-called "Königsberger Brückenproblem". (In English: The Königsberg bridge problem). In the town of Königsberg in what was once East Prussia, the two branches of the River Pregel converge and flow through to the Baltic Sea. Parts of the town were on an island and parts on a headland that were both joined to the outer river banks and to each other by seven bridges. The townspeople wanted to know if it was possible to take a walk that crossed each of the bridges exactly once before returning to the starting point. Euler proved in 1736 that no such walk was possible.[2]
More systematically, let us start with the following theorem of existence.

Theorem 5.5.1 *A graph G with minimum degree $\delta(G) \geq 2$ has a cycle of length at least $\delta(G) + 1$.*

Proof. Let $P = \{v_0, v_1, \ldots, v_n\}$ be a path of G of maximum length. Then v_0 is adjacent only to vertices of P, since otherwise P could be lengthen.
On the other hand, v_0 has at least $\delta(G)$ neighbors. Let v_i be the neighbor with maximum index which implies $i \geq \delta(G)$. Then $C = \{v_0, v_1, \ldots, v_i, v_0\}$ is such a searched cycle. □

Let G be a graph. A Eulerian cycle of G is defined as a cycle that uses each edge of G exactly once.[3] A graph which contains a Eulerian cycle is called a Eulerian graph. One of the oldest combinatorial problems, accredited to Euler and written in the terminology of graph theory, can be stated as follows: When does a multigraph have a Eulerian chain or a Eulerian cycle?[4] The answer is:

Theorem 5.5.2 *(Euler) A multigraph has a Eulerian cycle if and only if it is connected and all vertices have even degree.*[5]

[2] In solving the problem Euler laid the foundations for what was coming as a new type of geometry, which he called *geometris situs*; today called *topology*, with graph theory as a part. Compare [233] or [255].

[3] Note that an Eulerian cycle is not a cycle in the usual sense, since it can contain a vertex more than once. In this sense Euler's question is essentially different from Hamilton's problem which is the analogous question: When does a graph contain a cycle that contains every vertex exactly once.

[4] This question plays an important role in computational molecular biology, namely in determining an RNA sequence from its fragments, compare [109] and [205].

[5] The proof of this theorem gives a fast method to construct a Eulerian cycle explicitly:

Algorithm 5.5.3 *(Hierholzer, [132], [143], [158]) Let $G = (V, E)$ be a Eulerian (multi-) graph. Choose a vertex v_1 arbitrarily and apply the following recursive procedure Euler(G, v_1):*

1. *Set $C := v_1$; $v := v_1$;*
2. *If $g_G(v) = 0$ then goto 4. else let $w \in N_G(v)$;*
   choose one edge $e = \underline{vw}$;
3. *Set $C := C, e, w$ and $v := w$;*
   Set $E := E \setminus \{e\}$;
   goto 2.;
4. *Let $C = v_1, e_1, v_2, e_2, \ldots, v_k, e_k, v_{k+1}$;*
   For $i := 1$ to k do $C_i :=$ Euler(G, v_i);

Theorem 5.5.2 also has several consequences, all are exercises for the reader:

a) Any Eulerian graph is the union of cycles.

b) Any connected graph contains a chain that uses each edge exactly twice.

c) Each graph can be extended to an Eulerian graph by adding edges.

d) A multigraph has an (open) Eulerian chain if and only if it is connected and has exactly two vertices of odd degree.

The number of connected graphs we discussed above. Additionally,

Theorem 5.5.4 *The number of graphs with n labeled vertices each having even degree equals*

$$2^{\binom{n-1}{2}}. \tag{5.13}$$

The *proof* is not difficult, since the number itself suggests its method. We establish a one-to-one correspondence between all graphs G with $n-1$ labeled vertices and our graphs under consideration. In view of 4.1.1 G must have an even number of vertices of odd degree. Next we add add a new vertex v which is connected to each of the vertices of odd degree. This new graph has n vertices all of even degree. It is easy to see that conversely each graph with $n-1$ can be obtained from an "even" graph. □

To combine 5.5.4 and 5.4.1 in the sense of 5.5.2 is not so simple. An explicit formula is not known, but we may assume that the number grows exponentially, since for a given number n there are $(n-1)!/2$ many cycles, which are Eulerian graphs of itself.

$n=$	the number of Eulerian graphs with n vertices
3	1
4	3
5	13

On the other hand, we compute the ratio between the number given in 5.5.4 with the number of all graphs:

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

That means that graphs with only even degrees are rare, and consequently

Theorem 5.5.5 *Almost no graph is Eulerian.*[6]

5. Set $C = C_1, e_1, C_2, e_2, \ldots, C_k, e_k, v_{k+1}$.

[6]The terms "rare" and "almost all/no" will be more specified later.

But the situation is not hopeless: We can extend each graph to an Eulerian one, that means by adding new edges we create a graph with an Eulerian cycle. Let G be a graph with n vertices, m edges and c components. First add $c-1$ edges to make the graph connected. Now assume $c = 1$. Second, let W the set of all vertices with an odd degree. In view of 4.1.1 we have that $|W| = k$ is an even number. Then add for each pair of vertices from W an edge connecting both. The new graph has only vertices of even degree.[7]

II. Hamilton's problem asks whether a graph contain a cycle that contains every vertex exactly once. A graph which contains a Eulerian cycle is called a Eulerian graph.
Although it is clear that only connected graphs can be Hamiltonian, there is no simple criterion to tell us whether or not a graph is Hamiltonian as there is for Eulerian graphs. And indeed, no efficient algorithmic method is known to check whether a given graph has a Hamiltonian cycle, which is no strange since Karp [147] shows that the problem is $\mathcal{NP}$-complete.[8]
For practice the reader should decide whether the Petersen graph G_{petersen} is Hamiltonian; and shoulduse induction to prove that the hypercube Q^D contains a Hamilton cycle.[9]
The number of Hamiltonian graphs is still uncounted.

5.6 RNA secondary Structure

Proteins are not laid out simply as straight chains of amino acids. The fact that they curl and fold into complex forms plays a crucial role in determining the distinctive biological properties of each protein. The function of a protein is furthermore a direct consequence of its three-dimensional structure, shortly written by: Sequence $\Rightarrow$

[7] There mis an optimization version of Euler's problem making a given (connected) graph Eulerian by adding edges, such that the total length of the graph is minimal. This problem was introduced by the Chinese mathematician Guan [111] and later named as "The Chinese Postman Problem".

[8] To check that a graph is not Hamiltonian is often simpler. Gross, Yellen [109] describe the following rules during a construction of a Hamilton cycle:

1. If a vertex has degree 2, then both of its incident edges must be in the cycle.
2. During the construction no cycle can be formed until all vertices have been visited.
3. If during the construction two of the edges incident on a common vertex, then all other incident edges can be deleted.

[9] There is an important application of this fact in coding theory. A Gray code is a cyclic arrangement of binary sequences such that any pair of adjacent sequences differ in only one position. Example: $000 \to 010 \to 110 \to 100 \to 101 \to 111 \to 011 \to 001 \to$. This sequence corresponds to a Hamilton cycle in Q^3.

Structure ⇒ Function.[10] An understanding of the structure is essential for understanding the behavior of the molecule.

I. Following Clote, Backofen [56] we model the secondary structure of RNA molecules by a labeled graph $G = (\{v_1, \ldots, v_n\}, E)$ with the following properties:

(i) For any $i = 1 \ldots, n-1$ it holds $\underline{v_i v_{i+1}} \in E$;

(ii) For any $i = 1 \ldots, n$ there exists at most one $j \neq i+1$ for which $\underline{v_i v_j} \in E$;

(iii) If $1 \leq i < k < j \leq n$, $j \neq i+1$ $\underline{v_i v_j} \in E$ and $\underline{v_k v_l} \in E$, then $i \leq l \leq j$.

Define $s(n)$ to be the number of secondary structures and let $s(0) = 1$. Obviously, $s(2) = 1$.

Theorem 5.6.1

$$s(n+1) = s(n) + \sum_{k=0}^{n-2} s(k)s(n-k-1). \tag{5.14}$$

Proof by induction on n. For the induction step we consider two subcases:
Case 1: v_{n+1} is not paired. In this case, there are $s(n)$ possible structures.
Case 2: v_{n+1} is paired to v_j for $1 \leq j \leq n-1$. In this case the structure can independently formed on the former sequence $\{v_1, \ldots, v_{j-1}\}$ and the latter sequence $\{v_{j+1}, \ldots, v_n\}$, Thus,

$$\begin{aligned}
s(n+1) &= s(n) + \sum_{j=1}^{n-1} s(j-1)s(n-j) \\
&= s(n) + s(n-1) + \sum_{j=2}^{n-1} s(j-1)s(n-j) \\
&= s(n) + s(n-1) + \sum_{k=1}^{n-2} s(k)s(n-k-1) \\
&= s(n) + \sum_{k=0}^{n-2} s(k)s(n-k-1)
\end{aligned}$$

[10] We distinguish the following structural levels for proteins:

(i) The primary structure is the amino acid sequence.

(ii) The secondary structure is the arrangement of the amino acids in space.

(iii) The tertiary structure is the three-dimensional folding pattern, which is superimposed on the secondary structure.

(iv) The quarternary structure is the composition of two or more polypeptides.

When RNA is transcribed from the DNA template, it is single stranded. The single-stranded molecule can fold back on itself and when regions of the molecule are complementary they can become double-stranded or helical. The pairing rules for RNA sequences are the so-called Watson-Crick-rules: a pairs with u and c pairs with g.

□

5.6.1 implies

$$s(n+1) = s(n) - s(n-1) + \sum_{k=0}^{n-1} s(k)s(n-k-1). \tag{5.15}$$

The last term is M.2.1, but not the same recurrence relation, since there are different initial conditions. Additionally, there is a 1-1 correspondence between the RNA secondary structure and well well-balanced paranthesis expression. Hence,

Theorem 5.6.2 *There are exponentially many secondary structures.*

As exercise the reader should find a good lower bound.

II. After counting the secondary structure we are interested in construction the structure itself.
We add some new properties. Choose a subset of $2j$ points, $0 \leq 2j < n$. The $2j$ points are arranged as j disjoint pairs and these pairs are connected by edges such that

(iv) Consecutive points are never connected by an edge;

(v) Any two points connected by an edge must be separated by at least m points;[11]

(vi) Edges cannot intersect.

The prediction of secondary structures made by finding such structures that have the maximum number of pairs. It can be found by a dynamic programming approach.

Algorithm 5.6.3 *Let $a_1a_2\ldots a_n \in \{a,c,g,u\}^n$ be a RNA sequence. Furthermore let $m \geq 1$ be an integer and*

$$p(i,j) = \begin{cases} 1 & : \quad \{a_i, a_j\} = \{a,u\} \text{ or } = \{c,g\} \\ 0 & : \quad \text{otherwise} \end{cases}$$

Define $F(i,j) =$ maximum number of pairs of all secondary structures over $a_i \ldots a_j$. $F(1,n)$ can be found by the following procedure:

1. *$F(i,j) = 0$ whenever $j \leq i+m$;*
2. *$F(i,j) = \max\{F(i,j-1), (F(i,k-1)+F(k+1,j-1)+1)\cdot p(a_k,a_j) : 1+k+m \leq j\}$.*

The procedure can be performed in quadratic time. Unfortunately the algorithm is to simple to find the real structures and more complicated algorithms must be employed, compare [154], [247] and [248].

[11] These conditions are forced by biochemical facts. In particular in RNA $m = 3$ or $= 4$ is realistic.

5.7 The number of planar graphs

We will use our knowledge about planar graphs to determine this number. We start with the following observation: For the sum of binomial coefficients we know 2^n as upper bound. A partially better bound is given by the following formula.

Lemma 5.7.1 *Let $k \leq n$, then*

$$\sum_{m=0}^{k} \binom{n}{m} \leq \left(\frac{en}{k}\right)^k. \tag{5.16}$$

Proof. (Matoušek, Nešetřil [170]) We use C.2.2:

$$\binom{n}{0} + \binom{n}{1}x + \binom{n}{2}x^2 + \ldots + \binom{n}{n}x^n = (1+x)^n \tag{5.17}$$

for all real numbers x. In particular for $0 < x < 1$

$$\binom{n}{0} + \binom{n}{1}x + \ldots + \binom{n}{k}x^k \leq (1+x)^n. \tag{5.18}$$

Dividing this by x^k we get

$$\frac{1}{x^k}\binom{n}{0} + \frac{1}{x^{k-1}}\binom{n}{1} + \ldots + \binom{n}{k} \leq \frac{(1+x)^n}{x^k}. \tag{5.19}$$

Since $x < 1$,

$$\binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{k} \leq \frac{(1+x)^n}{x^k}. \tag{5.20}$$

Note that the left-hand side is independent of the value of x; in particular we can use $x = k/n$ which gives us

$$\frac{(1+x)^n}{x^k} = \frac{\left(1+\frac{k}{n}\right)^n}{\left(\frac{k}{n}\right)^k} \leq \left(e^{\frac{k}{n}}\right)^n \left(\frac{n}{k}\right)^k = e^k \left(\frac{n}{k}\right)^k.$$

To get this result we have to apply some calculus which shows $1 + x \leq e^x$, so that we attain to the inequality

$$\frac{\left(1+\frac{k}{n}\right)^n}{\left(\frac{k}{n}\right)^k} \leq \left(e^{\frac{k}{n}}\right)^n \left(\frac{n}{k}\right)^k.$$

□

This theorem is helpful when we count graphs whose number of edges is essentially less than $\binom{n}{2}$; in particular, if the order is less than quadratic.[12] An example: in view

[12] Hence, it was and will be of interest for us to bound the number of edges depending on the number of vertices. Theoretically, the number of edges in a graph is of quadratic order of the number of vertices, both in the worst case 4.4.2 and in the average case 5.8.3, but, on the other hand, Chung [49] remarked that that empirically most of the real-world graphs have the property that the number of edges is within a constant multiple of the number of vertices.

of 4.8.2 the number of edges in a planar graph is bounded by $3n-6$. Then in (5.1) with $k = 3n-6$

$$\text{plan } (n) \leq \sum_{m=0}^{3n-6} \binom{\binom{n}{2}}{m} \leq \left(\frac{en(n-1)}{2(3n-6)}\right)^{3n-6} \leq \left(\frac{e\cdot(n+2)}{6}\right)^{3n-6},$$

paying attention 5.7.1.

Theorem 5.7.2 *Let plan(n) be the number of all planar labeled graphs with $n \geq 4$ vertices. Then*

$$plan(n) \leq \left(\frac{e}{6}\cdot(n+2)\right)^{3n-6}. \tag{5.21}$$

Later we will prove that there are exponentially many trees. Since each tree is planar, the function plan(.) cannot be less than exponentially.
On the other hand, McDiarmid et al. [171] showed that the limit

$$\beta_0 = \lim_{n\to\infty}\left(\frac{\text{plan}(n)}{n!}\right)^{1/n} \tag{5.22}$$

exists. Hence, after some simple calculation we have

Theorem 5.7.3

$$plan(n) \approx \beta_0 \cdot \beta(n) \cdot \beta^n \cdot n!, \tag{5.23}$$

with a desired chosen subexponential function $\beta(.)$, numbers β_0 and $\beta > 1$.

Of course, the quantities β_0, β and the function $\beta(.)$ should be better estimated. Extending techniques are introduced in [23]. Gimenez and Noy [98] showed

Theorem 5.7.4 *Let plan(n) be the number of planar labeled graphs with n vertices. Then*

$$plan(n) \approx \beta_0 \cdot n^{-7/2} \cdot \beta^n \cdot n!, \tag{5.24}$$

with $\beta \approx 27.22688$.
Let plan-conn(n) be the number of connected planar labeled graphs with n vertices. Then

$$plan\text{-}conn(n) \approx \beta_1 \cdot n^{-7/2} \cdot \beta^n \cdot n!. \tag{5.25}$$

An exact and explicit formula for the number of planar graphs is not known, but for specific cases. Remember that a maximal outer-planar graph can be embedded as a triangulation of a polygon. Together with M.4.1 we obtain

Theorem 5.7.5 *The number of maximal outer-planar graphs with $n > 2$ vertices equals the $(n-2)$th Catalan number, that means*

$$C_{n-2} = \frac{1}{n-1}\binom{2n-4}{n-2}. \tag{5.26}$$

Consequently, there are at least exponentially many labeled outer-planar graphs, which is not a surprise since we will see that there is exponential number of labeled trees.

5.8 Random graphs I.

Historically, counting problems have been closely associated with probability. In this section we will see that facts from probability theory are helpful to prove results in graph theory. The goal is to give the terms "almost all/no" more substance.
Aigner, Ziegler [5]:

> If in a given set of objects the probability that an object does not have a certain P is less than 1, then there must exist an object with this property.

We will use different models.

I We will use uniform probability spaces. In other terms, our probability space will be the set $\mathcal{G}_n$ of all labeled graphs with n vertices in which each graph is equally likely. In view of 5.1.1 we observe

$$|\mathcal{G}_n| = 2^{\binom{n}{2}}, \tag{5.27}$$

such that the probability distribution for each member $G \in \mathcal{G}_n$ is given by

$$\Pr(G) = 2^{-\binom{n}{2}}. \tag{5.28}$$

Given a graph theoretic property, let p_n denote the probability that a graph picked randomly from $\mathcal{G}_n$ has this property. We use the following notation:

$$\text{If } \left\{ \begin{array}{l} \lim_{n\to\infty} p_n = 1 \\ \lim_{n\to\infty} p_n = 0 \end{array} \right\} \text{we say that } \left\{ \begin{array}{l} \text{almost all} \\ \text{almost no} \end{array} \right\} \text{graphs}$$

have the considered property.

Theorem 5.8.1 *Almost all graphs have no vertex of degree 0.*

Proof. There are

$$2^{\binom{n-1}{2}} \tag{5.29}$$

graphs containing one specific vertex as an isolated one. Hence, the number of graphs in $\mathcal{G}_n$ containing at least one isolated vertex can be bounded by $n\cdot$(5.29).
In order to prove that almost all graphs do not have any isolated vertex, we see that

$$\frac{n \cdot 2^{\binom{n-1}{2}}}{2^{\binom{n}{2}}} = \frac{n}{2^{n-1}}, \tag{5.30}$$

tends to 0, which is obvious. □

Asymptotically, the number of connected graphs is the same as the number of all graphs:

Theorem 5.8.2 *Almost all graphs are connected.*

Proof. Consider disconnected graphs in $\mathcal{G}_n$. Each such graph contains a subset of $i \leq \lfloor n/2 \rfloor$ vertices in which no vertex is adjacent to any vertex outside. There are

$$2^{\binom{i}{2}} \cdot 2^{\binom{n-i}{2}} = 2^{\binom{n}{2}-i(n-i)} \tag{5.31}$$

many graphs which are disconnected because such a set of vertices. Consequently, we can bound the number of disconnected graphs in $\mathcal{G}_n$ from above by

$$\sum_{i=1}^{\lfloor n/2 \rfloor} 2^{\binom{n}{2}-i(n-i)} \leq \sum_{i=1}^{\lfloor n/2 \rfloor} n^i 2^{-i(n-i)}. \tag{5.32}$$

For n sufficiently large, the largest summand on the right side is the first one. Therefore, the sum is bounded from above by $n^2 2^{-n}$ which tends to 0 fast. □

The theorem implies that if we randomly pick up a graph from $\mathcal{G}_n$ it will most likely be connected.

II Furthermore, we will describe two concepts which generalize our approaches. Both will imply 5.8.2. For the proofs see [45] and [194].
We can describe the the underlying probability space of $\mathcal{G}_n$ in a different way, namely where each term models the presence or absence of a particular edge and assume that each edge is presented with the probability 1/2. We write this space by $\mathcal{G}_{n,1/2}$.
The observation

$$\Pr[G \in \mathcal{G}_{n,1/2}] = 2^{-\binom{n}{2}}$$

for every fixed graph G shows that this model is indeed just another description of $\mathcal{G}_n$.

What is known about the expected number of edges in a randomly chosen graph? Linearity of expectation implies

Theorem 5.8.3 *([194]) The expected number of edges of a random graph is*

$$\frac{1}{2}\binom{n}{2} = \frac{n(n-1)}{4}. \tag{5.33}$$

Remember that each connected graph $G = (V, E)$ is a metric space (V, ρ).

$$\operatorname{diam}(G) = \max\{\rho(v, v') : v, v' \in V\} \tag{5.34}$$

defines the diameter of G.[13] In other words, the diameter is the longest distance between any two vertices in the graph. It should not be confused with the longest path in the graph. Of course,

$$\operatorname{diam}(G) \leq |V| - 1. \tag{5.35}$$

[13]For disconnected graphs this quantity is undefined, or ∞.

This implies that, using the adjacency matrix, we have to check only the powers up to $k = |V| - 1$ to decide whether a graph is connected or not.

In view of 4.4.6 and its proof the following fact is not a surprise[14] although not simple.

Theorem 5.8.4 *Almost all graphs have diameter 2.*

This theorem can be seen as the origin of the so-called "small world phenomenon": any two vertices are connected by a short path, roughly spoken in many - perhaps most - networks the typical distances between vertices are surprisingly small. Compare Chung [49], Häggström [114], Kleinberg [153] or Newman [179].

III. Remember that a graph $G = (V, E)$ is called k-connected if for each pair of different vertices v and v' there are k pairwise vertex-disjoint paths interconnecting v and v', $k \geq 0$.

Theorem 5.8.5 *For any fixed $k > 0$, almost all graphs are k-connected.*

More facts we find in Harary, Palmer [122] in view of considering asymptotics.

5.9 Random graphs II.

I. To obtain a more subtle information about graphs, we refine our probability space by introducing a parameter p with $0 < p < 1$ which describes the probability of an edge to be in a graph. In other terms, we fix not the number but the probability of the edges. The space is written by $\mathcal{G}_{n,p}$.

Observation 5.9.1 *Consider $\mathcal{G}_{n,p}$. Then each graph with m edges has the probability*

$$Pr(m) = p^m \cdot (1-p)^{\binom{n}{2}-m}. \tag{5.36}$$

The observation shows that Pr is a Bernoulli trial, which is a single experiment with two possible outcomes: success and failure. A number of independent Bernoulli trials describes just the standard binomial distribution.

Theorem 5.9.2 *Consider $\mathcal{G}_{n,p}$.*

a) *The total probability of graph with m edges equals*

$$\binom{\binom{n}{2}}{m} \cdot p^m \cdot (1-p)^{\binom{n}{2}-m}. \tag{5.37}$$

[14] But Gardner, [95], [96]:

> Most people are very surprised when they meet a stranger, especially it far from home, and discover that they have a friend in common.

b) *The expected number of edges equals*

$$p \cdot \binom{n}{2}. \tag{5.38}$$

c) *The mean degree is*

$$(n-1) \cdot p. \tag{5.39}$$

d) *The total probability of a vertex of being connected to exactly k others equals*

$$\binom{n-1}{k} \cdot p^k \cdot (1-p)^{n-1-k}. \tag{5.40}$$

That means, $\mathcal{G}_{n,p}$ has a binomial degree distribution.

Proof.
a) In view of (5.1) there are $\binom{\binom{n}{2}}{m}$ possible graphs, each selected with equal probability.
b) Well-known for the expectation of the binomial distribution, since

$$\sum_{m=0}^{\binom{n}{2}} m\Pr(m) = \binom{n}{2} p.$$

c) In view of 4.1.1 and b) we determine the mean degree in

$$\sum_{m=0}^{\binom{n}{2}} \frac{2m}{n}\Pr(m) = \frac{2}{n}\binom{n}{2} p = (n-1)p.$$

d) A given vertex is adjacent with independent probability p to each other of the $n-1$ vertices. Thus the probability of being adjacent to a particular k other vertices and not to any of the other $n-k-1$ ones is $p^k \cdot (1-p)^{n-1-k}$. There are $\binom{n-1}{k}$ ways to choose those k vertices. Hence, the assertion. □

II. For further investigations we consider the following approach. Let X be a binomially distributed random variable with parameter n and p. If the number n of experiments is large and the number k of successes is small, then a good approximation is given assuming that np, the expected number of successes, is a constant. This can be seen by the following chain of equations.

$$\begin{aligned}
&\binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{k!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \qquad \text{substituting } p = \frac{\lambda}{n} \\
&= 1 \cdot \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \\
&\to \frac{\lambda^k}{k!} \cdot e^{-\lambda},
\end{aligned}$$

provided that n increases, p approaches 0, and $\lambda = np = \text{const}$. Consequently,

$$\lim_{n\to\infty, p\to 0, np=\lambda=\text{const}} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \tag{5.41}$$

called the Poisson distribution with parameter λ. Roughly spoken, let $B(k, n, p)$ be binomial distributed with the parameters n and p, then

$$B(k, n, p) \approx P(k, \lambda), \tag{5.42}$$

where $P(k, \lambda)$ is Poisson distributed with parameter $\lambda = np$. This observation is very helpful, because, in general, the quantity $B(k, n, p)$ is hard to compute if $n \gg 1$ and $p \ll 1$.

The Poisson distribution is often used in modelling situations in biology where events occur infrequently. Consider the following example [56]: From many studies, it has become clear that the rate of amino acid substitution varies between organisms and also between protein classes. We are interested in the way how amino acid substitution rates are computed.
Let w and w' be two (homologous) polypeptides of the same length n. n_d denotes the number of differences between homologous acid sites; the probability p of an amino acid substituting occuring at a given site of either w or w' can be estimated by

$$p \approx \frac{n_d}{n}. \tag{5.43}$$

A second approximation of p can be derived by assuming that the substitution of amino acids at a given site is a Poisson process. Let X be a random variable counting the number of mutations over time t at fixed site for an polypeptide having substitution rate λ per site (and per year). Then

$$p(X = k) = \frac{(\lambda \cdot t)^k}{k!} e^{-\lambda \cdot t}. \tag{5.44}$$

Thus the probability that no substitution occurs at a given site in w is

$$p(X = 0) = e^{-\lambda \cdot t}. \tag{5.45}$$

Hence the probability that no substitution occurs at a given site in w and w' is

$$q = e^{-2 \cdot \lambda \cdot t}. \tag{5.46}$$

Since $d = 2 \cdot \lambda \cdot t$ is the total number of substitutions occuring at a fixed site, we get

$$d = 2 \cdot \lambda \cdot t = -\ln q. \tag{5.47}$$

Together with (5.43) we find the following approximation

$$d \approx -\ln\left(1 - \frac{n_d}{n}\right) \tag{5.48}$$

for the protein substitution rate.

Another example is the calculation of the degree distribution of a random graph. A given vertex in the graph is connected with independent probability p to each of the other $n-1$ vertices. Then it holds 5.9.2(d). This becomes

Corollary 5.9.3 *The total probability of a vertex of being connected to exactly k others is approximately*

$$e^{-(n-1)p} \cdot \frac{((n-1)p)^k}{k!}, \tag{5.49}$$

in the limit of large n.

5.10 Threshold functions

Now the properties of graphs depend on the concrete values of p. We use the following notation: Let Q be a property of graphs. A function $t(.)$ is called a threshold function for Q

$$\text{if } \left\{ \begin{array}{l} t(n) = o(p(n)) \\ p(n) = o(t(n)) \end{array} \right\} \text{ implies } \left\{ \begin{array}{l} \lim_{n\to\infty} \Pr[G_{n,p} \text{ has property } Q] = 1 \\ \lim_{n\to\infty} \Pr[G_{n,p} \text{ has property } Q] = 0. \end{array} \right\}$$

Note that not every property has a threshold function (but almost all of the interesting properties), and that threshold functions are not unique. (Exercises)
In [45] and [194] there are little collections of threshold functions:

graph-property Q	threshold function $t(n)$
having no isolated vertices	$\frac{\ln n}{n}$
is connected	$\frac{\ln n}{n}$
diameter 2	$\sqrt{\frac{\ln n}{n}}$
containing a path of length k	$n^{-(k+1)/k}$
containing a Hamiltonian path	$\frac{\ln n}{n}$
containing a triangle	$\frac{1}{n}$
containing a cycle	$\frac{1}{n}$
containing a K_r	$n^{-2/(r-1)}$
non-planar	$\frac{1}{n}$

Of more interest for our investigations is the following result.

Theorem 5.10.1 *(Prömel, Steger [194]) The function $t(n) = \frac{\ln n}{n}$ is the threshold function for the property for connectedness of graphs.*

In view of the last theorems graphs are "randomly dense". And indeed:

Theorem 5.10.2 *Almost all graphs are non-planar.*

In view of 5.7.2 this is not a surprise. For a complete proof see [45].

As strange fact it is known by Robinson, Warmald [206], [207], that for every integer $r \geq 3$, almost all r-regular graphs are Hamiltonian, but, surprisingly, planarity changes the picture completely, since Richmond et al. [202] show that almost all 3-connected 3-regular planar graphs are not Hamiltonian.

Random graph theory concerns many more properties of random graphs and we have just scratched the surface of this area. More facts we find in Bollobas [30], and in the algorithmic sense in Mitzenmacher et al. [175].

Chapter 6

The Number of Labeled Trees

Clearly, we have to distinguish between labeled an unlabeled trees. A tree $T = (V, E)$ with n vertices is called labeled if a bijective mapping from V onto the set $\{1, \ldots, n\}$ of integers is given.[1] On the other hand, in the case of unlabeled trees the word "different" means non-isomorphic, and each set of isomorphic trees is counted as one.

6.1 Permutations

We start with a very simple example that we have to distinguish between permutations, cycles and paths, which are as trees also linear arrangements. Let $n = 3$ then

class	permutations	(oriented) cycles	cycles	paths
	123	123 = 231 = 312	all	123 = 321
	132	132 = 321 = 213		132 = 231
	213			213 = 312
	231			
	312			
	321			
number	6	2	1	3

Well-known

Observation 6.1.1 *There are*

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 \tag{6.1}$$

permutations on n elements.

[1] Or onto another set of n distinguished names.

Proof. There are n possibilities for the first place, then $n-1$ for the second place, and so on until there is just one for the last place. So we get the assertion from the multiplication principle. □

In general,

Theorem 6.1.2 *Consider $n \geq 3$ vertices. Then*

a) *There are $(n-1)!$ labeled (oriented) cycles.*

b) *There are $(n-1)!/2$ labeled cycles.*

c) *There are $n!/2$ labeled paths.*

Proof. There are $(n-1)!$ ways to place n objects at a round table, when we count oriented cycles. And each cycle has two orientations.
Consider paths. We select two vertices as the leaves and order all other in any way.

$$\binom{n}{2} \cdot (n-2)! = \frac{n!}{2!(n-2)!} \cdot (n-2)! = \frac{n!}{2}. \tag{6.2}$$

□

6.2 Trees with a given degree sequence

We start counting with the number of different labeled trees and we will describe this number in terms of the vertex degrees.

Let $T = (V, E)$ be a tree with n vertices $v_1, ..., v_n$, and let $g_i = g(v_i)$ be the degree of each vertex v_i. Then, obviously, each of the numbers g_i is a positive integer, and, in view of 4.1.1 and 4.6.2,

$$\sum_{i=1}^{n} g_i = 2n - 2. \tag{6.3}$$

Conversely, by an induction argument, we find that this equality is also sufficient:

Lemma 6.2.1 *Let $g_1, \ldots, g_n$ be a sequence of positive integers satisfying (6.3). Then there exists a tree on n vertices with these predetermined degrees.*

Proof. Let $g_1, \ldots, g_{n+1}$ be a sequence with

$$\sum_{i=1}^{n+1} g_i = 2(n+1) - 2 = 2n. \tag{6.4}$$

Not all of the g_i can be equal 1, since otherwise

$$\sum_{i=1}^{n+1} g_i = \sum_{i=1}^{n+1} 1 = n + 1 < 2n.$$

Not all of the g_i can be greater than 1, since otherwise

$$2n = \sum_{i=1}^{n+1} g_i \geq \sum_{i=1}^{n+1} 2 = 2(n+1).$$

Hence, without loss of generality, we may assume that $g_{n+1} = 1$ and $g_n > 1$.
Define $g'_1, \ldots, g'_n$ by

$$g'_i = g_i \tag{6.5}$$

for $i = 1, \ldots, n-1$, and

$$g'_n = g_n - 1. \tag{6.6}$$

For this sequence it holds

$$\sum_{i=1}^{n} g'_i = \sum_{i=1}^{n-1} g_i + (g_n - 1) + (g_{n+1} - 1) = \sum_{i=1}^{n+1} g_i - 2 = 2n - 2.$$

By the induction assumption there is a tree $T' = (V' = \{v_1, \ldots, v_n\}, E')$ such that $g(v_i) = g'_i$. Then the tree $T = (V' \cup \{v_{n+1}\}, E' \cup \{\underline{v_n v_{n+1}}\})$ fulfills the assertion. □

In view of 3.2.2 there are $\binom{2n-3}{n-2}$ solutions of the Diophantic equation (6.3) in positive integers. Hence, the number of different trees increases exponentially, but not faster:

Theorem 6.2.2 *Let $n \geq 2$ be an integer and let $g_1, ..., g_n$ be a sequence of positive integers. When we denote by $t(n, g_1, ..., g_n)$ the number of different labeled trees $T = (\{v_1, ..., v_n\}, E)$ of n vertices with the degree sequence*

$$g_T(v_i) = g_i \tag{6.7}$$

for $i = 1, ..., n$, we have

$$t(n, g_1, ..., g_n) = \frac{(n-2)!}{\prod_{i=1}^{n}(g_i - 1)!} = \binom{n-2}{(g_1 - 1) \ldots (g_n - 1)} \tag{6.8}$$

if (6.3) holds, and

$$t(n, g_1, ..., g_n) = 0 \tag{6.9}$$

otherwise.

Proof. ([25] or [173]) In view of 6.2.1 and its proof we know that $t(n, g_1, ..., g_n) > 0$ if and only if (6.3) holds.
Without loss of generality, we may assume that

$$g_1 \geq g_2 \geq \ldots \geq g_n.$$

That means: $g_1 = \Delta(G)$ and $g_n = \delta(G) = 1$ and v_n must be a leaf.
Let C_i be the collection of all trees T with vertices $v_1, \ldots, v_n$ and degrees $g_j = g_T(v_j)$, such that the leaf v_n is adjacent to v_i. Assuming $g_i \geq 2$ we have

$$|C_i| = t(n-1, g_1, \ldots, g_{i-1}, g_i - 1, g_{i+1}, \ldots, g_{n-1}).$$

Since the collection of all trees is the union of the sets C_i for $g_i \geq 2$ we obtain, by the addition principle,

$$t(n, g_1, ..., g_n) = \sum_{g_i \geq 2} t(n-1, g_1, \ldots, g_{i-1}, g_i - 1, g_{i+1}, \ldots, g_{n-1}). \tag{6.10}$$

Now, we use induction. The theorem is true for $n = 2$. Assume that $n \geq 3$ and that the theorem is true for $n - 1$. Then

$$\begin{aligned}
& t(n, g_1, ..., g_n) \\
&= \sum_{g_i \geq 2} t(n-1, g_1, \ldots, g_{i-1}, g_i - 1, g_{i+1}, \ldots, g_{n-1}) \\
&= \sum_{g_i \geq 2} \frac{(n-3)!}{(g_1-1)! \cdots (g_{i-1}-1)!(g_i-2)!(g_{i+1}-1)! \cdots (g_{n-1}-1)!} \\
&= \sum_{g_i \geq 2} \frac{(n-3)!(g_i-1)}{(g_1-1)! \cdots (g_{n-1}-1)!} \\
&= \frac{(n-3)!}{(g_1-1)! \cdots (g_{n-1}-1)!} \cdot \sum_{g_i \geq 2} (g_i - 1) \\
&= \frac{(n-2)!}{(g_1-1)! \cdots (g_{n-1}-1)!} \\
&= \frac{(n-2)!}{(g_1-1)! \cdots (g_n-1)!},
\end{aligned}$$

where we use C.3.2. □

Summing up over all degree sequences satisfying (6.3), whereby

$$t(n) = \#\text{ labeled trees }, \tag{6.11}$$

gives

$$t(n) = \sum_{(6.3)} t(n, g_1, ..., g_n) = \sum_{(6.3)} \frac{(n-2)!}{\prod_{i=1}^{n}(g_i - 1)!} = \underbrace{(1 + \ldots + 1)}_{n-\text{times}}^{n-2} = n^{n-2},$$

by C.3.2. And we have one of the most beautiful formulas in enumerative combinatorics:

Theorem 6.2.3 *(Cayley [41]) The number of different labeled trees with n vertices equals n^{n-2}.*

This theorem shows that the number of trees grows very rapidly in the number of vertices. We will find the same fact in most of our considerations. This makes exhaustive strategies handling trees using all trees infeasible for datasets involving

more than a dozen vertices.
It should be noted that n^{n-2} is the number of distinct trees, but not the number of non-isomorphic ones. This gives a new question which we will discuss later.

Now we will give a partial answer to the following problem of reconstruction: In how many ways we can extend a subforest to a tree? For bipartitions this is easy to answer: Two trees of a split (a bipartiton) of a set in two parts with n_1 and n_2 vertices, respectively, can be extended in $n_1 \cdot n_2$ ways to a tree for all. And in general,

Theorem 6.2.4 *(Moon) Consider the complete graph $K_n = (V, E)$ with n vertices. Let $\{V_1, \ldots, V_c\}$ be a partition of V with $|V_i| = n_i$, and let $T_i = (V_i, E_i)$, $i = 1, \ldots, c$, be pairwise disjoint trees. Then the number of spanning trees of K_n that have $T_1, \ldots, T_c$ as subgraphs is*

$$n_1 \cdots n_c \cdot n^{c-2}. \tag{6.12}$$

Proof. If each set V_i were contracted to a vertex v_i, then the number of trees T with $g_T(v_i) = g_i$ is

$$\binom{c-2}{(g_1-1)\ldots(g_c-1)}.$$

To each tree T correspond exactly $n_1^{g_1} \cdots n_c^{g_c}$ different spanning trees and consequently the searched number equals

$$\sum \binom{c-2}{(g_1-1)\ldots(g_c-1)} \cdot n_1^{g_1} \cdots n_c^{g_c} = n_1^{g_1} \cdots n_c^{g_c} \cdot (n_1 + \ldots + n_c)^{c-2}.$$

The assertion follows. □

6.3 The Prüfer code

Cayley's formula yields to equally beautiful proofs drawing on a variety of combinatorial and algebraic techniques. Here and later, we will outline several of these approaches.

Prüfer [196] established a bijection between trees and sequences of $n-2$ integers between 1 and n, providing a constructive proof of Cayley's result. This bijection can then be exploited to give algorithms for systematically generating labeled trees. More precisely: The strategy of the proof is to establish a one-to-one correspondence between the labeled tree and the Prüfer code, which is a sequence of length $n-2$ of integers between 1 and n, with repetitions allowed; in other words, a member of $\{1, \ldots, n\}^{n-2}$. Algorithmically this coding is described by

Algorithm 6.3.1 *Let $T = (V = \{v_1, \ldots, v_n\}, E)$ be a labeled tree. Then the Prüfer code for T can be constructed by performing the following steps:*

1. *Initialize T to be the given tree;*

2. *For $i = 1$ to $n - 2$ do*
   Let v be the leaf with the smallest label;
   Let s_i be the label of the only neighbor of v;
   $T := T[V \setminus \{v\}]$;

3. *The code is $(s_1, \ldots, s_{n-2})$.*

We will now use the correspondence between Prüfer codes and labeled trees to generate trees. We first note that the following decoding procedure maps a given Prüfer code to a labeled tree:

Algorithm 6.3.2 *A Prüfer code P is given. Then a labeled tree $T = (V, E)$ can be constructed by performing the following steps:*

1. *Initialize the list P as the input;*

2. *Initialize the list V as $1, \ldots, n$;*

3. *Initialize T as the forest of isolated vertices on V;*

4. *For $i = 1$ to $n - 2$ do*
   Let k be the smallest number in list V that is not in list P;
   Let j be the first number in list P;
   Add an edge joining the vertices labeled k and j;
   Remove k from list V;
   Remove the first occurrence of j from list P;

5. *Add an edge joining the vertices labeled with the two remaining numbers in the list V.*

It is not hard to see (exercise) that the decoding procedure 6.3.2 is the inverse of the encoding procedure 6.3.1. Altogether this establishes again 6.2.3.[2]
In other terms, we have created a one-to-one correspondence between the set of labeled trees and the set of Prüfer codes.

Algorithm 6.3.3 *Let n be an integer with $n \geq 2$. Then the following algorithm generates all trees with n labeled vertices:*

1. *Generate, by simple counting, all Prüfer codes in $\{1, \ldots, n\}^{n-2}$;*

2. *For each code apply 6.3.2.*

This procedure consumes $n^{n-2} \cdot O(n) = O(n^{n-1})$ time, since 6.3.2 runs in linear time. Hence, it is an effective technique.

[2]For several other proofs compare [5] and [170].

6.4 Trees with given leaves

We are interest in the number of trees with n vertices of which exactly k leaves. We assume $n > 1$, such that k is a number between 2 and $n-1$.

Theorem 6.4.1 *$t_k(n)$ denotes the number of labeled trees with n vertices of which k leaves, $k = 2, \ldots, n-1$. Then*

$$\frac{k}{n} t_k(n) = (n-k)t_{k-1}(n-1) + kt_k(n-1), \tag{6.13}$$

with the initial conditions $t_1(n) = t_n(n) = 0$, except $t_2(2) = 1$.

Idea of the *proof.* Consider sets with n elements and one of these explicitly. We distinguish two cases: It is a leaf or an internal vertex. Then we use the addition principle. □

Of course, in view of 6.2.3, $\sum_{k=2}^{n-1} t_k(n) = n^{n-2}$. We discuss several specific cases.

$k = 2$: We count paths.

$$\frac{2}{n} t_2(n) = (n-2)t_1(n-1) + t_2(n-1) = 2t_2(n-1).$$

Together with $t_2(2) = 1$ this implies $t_2(n) = n!/2$, in accordance with 6.1.2.

$k = n-1$: We count stars.

$$\frac{n-1}{n} t_{n-1}(n) = t_{n-2}(n-1) + (n-1)t_{n-1}(n-1) = t_{n-2}(n-1).$$

Therefore, well-known,

$$t_{n-1}(n) = n. \tag{6.14}$$

$k = n-2$: We count double-stars.

$$t_{n-2}(n) = \frac{2n}{n-2} t_{n-3}(n-1) + nt_{n-2}(n-1) = \frac{2n}{n-2} t_{n-3}(n-1) + n(n-1) =: f(n),$$

paying attention (6.14). Solving this recurrence we get

$$\begin{aligned} f(n) &= \frac{2n}{n-2} f(n-1) + n(n-1) \\ &= \frac{2n}{n-2} \left(\frac{2(n-1)}{n-3} f(n-2) + (n-1)(n-2) \right) + n(n-1) \\ &= \frac{2^2 n(n-1)}{(n-2)(n-3)} f(n-2) + (2+1)n(n-1) \\ &= \frac{2^2 n(n-1)}{(n-2)(n-3)} \left(\frac{2(n-2)}{n-4} f(n-3) + (n-2)(n-3) \right) \end{aligned}$$

$$
\begin{aligned}
& +(2+1)n(n-1) \\
= \quad & \frac{2^3 n(n-1)}{(n-3)(n-4)} f(n-3) + (2^2+2+1)n(n-1) \\
\vdots & \\
= \quad & \frac{2^{n-4} n(n-1)}{4 \cdot 3} f(4) + (2^{n-5} + \ldots + 1)n(n-1) \\
= \quad & 2^{n-4} n(n-1) + (2^{n-4} - 1)n(n-1) \quad \text{in view of 6.1.2} \\
= \quad & (2^{n-3} - 1)n(n-1).
\end{aligned}
$$

Hence,

$$t_{n-2}(n) = (2^{n-2} - 2)\binom{n}{2}. \tag{6.15}$$

for $n \geq 4$.

We write the numbers $t_k(n)$ as a triangle:

$n \setminus k$	2	3	4	5	6	7	n^{n-2}
2	1						1
3	3						3
4	12	4					16
5	60	60	5				125
6	360	720	210	6			1,296
7	2,520	8,400	5,250	630	7		16,807
8	20,160	100,800	109,200	30,240	1,736	8	262,144

From the theorem we obtain, by simple calculation, and consequently a weak, lower bound for the number of trees with a given number of leaves.[3]

Theorem 6.4.2 *For all numbers k with $2 \leq k \leq n-1$ it holds*

$$t_k(n) \geq \frac{n!}{k!}. \tag{6.16}$$

Proof. In view of 6.4.1 we have

$$
\begin{aligned}
t_k(n) \quad & \geq \quad n t_k(n-1) \geq n(n-1) t_k(n-2) \\
& \geq \quad n(n-1)(n-2) t_k(n-3) \geq \ldots \\
& \geq \quad n(n-1) \cdots (k+2) t_k(k+1) \\
& = \quad n(n-1) \cdots (k+2)(k+1) = \frac{n!}{k!}.
\end{aligned}
$$

□

[3]Later, investigating multi-stars, we will find a better bound for specific cases.

6.5 The number of labeled forests

Observation 6.5.1 *Let $g_1, \ldots, g_n$ be a sequence of nonnegative integers, and let c be a positive integer, satisfying*

$$\sum_{i=1}^{n} g_i = 2n - 2c. \tag{6.17}$$

Then there exists a forest on n vertices with this predetermined degrees and c components.

At first, we give a result which is not only as helpful for further considerations as well on its own interest.

Lemma 6.5.2 *(Clarke) The number of different trees T with the labeled vertices $v_1, \ldots, v_n$ and with $g_T(v_1) = g$ equals*

$$\binom{n-2}{g-1} \cdot (n-1)^{n-g-1}. \tag{6.18}$$

Proof. The desired number is

$$\begin{aligned}
&\sum_{g_2,\ldots,g_n} \binom{n-2}{(g-1)(g_2-1)\ldots(g_{n-1}-1)} \\
&= \sum_{g_2,\ldots,g_n} \frac{(n-2)!}{(g-1)!(g_2-1)!\cdots(g_{n-1}-1)!} \\
&= \sum_{g_2,\ldots,g_n} \frac{(n-2)!}{(g-1)!(g_2-1)!\cdots(g_{n-1}-1)!} \cdot \frac{(n-g-1)!}{(n-g-1)!} \\
&= \frac{(n-2)!}{(g-1)!(n-g-1)!} \sum_{g_2,\ldots,g_n} \frac{(n-g-1)!}{(g_2-1)!\cdots(g_{n-1}-1)!} \\
&= \binom{n-2}{g-1} \cdot \sum_{g_2,\ldots,g_n} \binom{n-g-1}{(g_2-1)\ldots(g_n-1)} \\
&= \binom{n-2}{g-1} \cdot (\underbrace{1+\ldots+1}_{(n-1)-\text{times}})^{n-g-1} \\
&= \binom{n-2}{g-1} \cdot (n-1)^{n-g-1},
\end{aligned}$$

by C.3.2. □

We have the following consequence of 6.2.2 for forests.

Theorem 6.5.3 *(Cayley) The number of different labeled forests with n vertices and c components, where the first c vertices are in different components, equals*

$$t'(n,c) = c \cdot n^{n-c-1}. \tag{6.19}$$

Proof. Consider the set $\mathcal{C}$ of trees with the vertices $v_0, v_1, \ldots, v_n$ and $g(v_0) = c$. In view of 6.5.2 we have

$$|\mathcal{C}| = \binom{(n+1)-2}{c-1} \cdot ((n+1)-1)^{(n+1)-c-1} = \binom{n-1}{c-1} \cdot n^{n-c}.$$

On the other hand, $\mathcal{C}$ is the union of all trees with v_i adjacent to v_0, for any i. Hence,

$$|\mathcal{C}| = \binom{n}{c} \cdot t'(n,c).$$

Altogether

$$\binom{n}{c} \cdot t'(n,c) = \binom{n-1}{c-1} \cdot n^{n-c}.$$

Thus

$$t'(n,c) = \frac{c!(n-c)!}{n!} \cdot \frac{(n-1)!}{(c-1)!(n-c)!} \cdot n^{n-c} = \frac{c}{n} \cdot n^{n-c} = c \cdot n^{n-c-1}.$$

□

For fixed n the function

$$t'(n,c) = \frac{c}{n^c} \cdot n^{n-1} \tag{6.20}$$

is a decreasing sequence in c from n^{n-2} until 1. Consequently, using the formula for (finite) geometric series D.2.1, D.2.2, we find

$$\sum_{c=1}^{n} t'(n,c) = \left(\sum_{c=1}^{n} \frac{c}{n^c} \right) \cdot n^{n-1} = \frac{\frac{1}{n} - \frac{1}{n^n}}{\left(1 - \frac{1}{n}\right)^2} \cdot n^{n-1} = \frac{n^n - n}{(n-1)^2}. \tag{6.21}$$

Corollary 6.5.4 *The number of all forests with n labeled vertices is bounded from below by*

$$\sum_{c=1}^{n} t'(n,c) = \frac{n^n - n}{(n-1)^2}. \tag{6.22}$$

Numerically,

n	1	2	3	4	5	6	7
	1	2	6	28	195	1,786	22,876

Let $t(n,c)$ be the number of all forests with n labeled vertices and c components. Then, of course, $t(n,c) \geq t'(n,c)$. In particular for $n = 4$:

$c =$	$m = n - c$	$t'(n,c)$	$t(n,c)$	comment
1	3	16	16	all trees
2	2	8	15	
3	1	3	6	
4	0	1	1	the empty graph
		28	38	

Easy to find an upper bound

Theorem 6.5.5

$$t(n,c) \leq \binom{\binom{n}{2}}{n-c}. \tag{6.23}$$

For $c = n$, $n-1$ and $n-2$ equality holds.

As an exercise Bollobas [29] gives an explicit, but very complicated, formula for $t(n,c)$, implying

$$t(n,2) = \frac{1}{2}(n-1)(n+6)n^{n-4}. \tag{6.24}$$

Numerically,

$n =$	2	3	4	5	6	7	...	10
$t(n,2) =$	1	3	15	88	1080	11.557	...	288.000.000

In general, we know, in view of 4.6.6, that the number of edges in a forest is bounded by $n-1$. Then in (5.1)

$$\text{number of forests with } n \text{ vertices } \leq \sum_{m=0}^{n-1} \binom{\binom{n}{2}}{m} \leq \left(\frac{en(n-1)}{2(n-1)}\right)^{n-1},$$

paying attention 5.7.1.

Theorem 6.5.6 *The number of forests with n labeled vertices is bounded from above by*

$$\left(\frac{e}{2}\right)^{n-1} \cdot n^{n-1} = (1.3596\ldots \cdot n)^{n-1}. \tag{6.25}$$

The comparison of 6.5.4 and 6.5.6 shows a small gap, which gives a hint for the asymptotic behavior:

Theorem 6.5.7 *(Renyi [200]) The number of forests with n labeled vertices is asymptotically*

$$\sqrt{e} \cdot n^{n-2} = 1.64872\ldots \cdot n^{n-2}. \tag{6.26}$$

This is a very interesting result, since

$$\frac{\#\text{ trees with } n \text{ vertices}}{\#\text{ forests with } n \text{ vertices}} \approx \frac{1}{\sqrt{e}} = 0.6065\ldots. \tag{6.27}$$

As an exercise transform this fact into a statement in the sense of random graphs.

Chapter 7

Unlabeled Graphs

Deciding when two graphs with different specifications are structurally equivalent, that is, whether they have the same pattern of connections is an important, but difficult question. The concept of graph isomorphism lies (explicitly or implicitly) behind almost any discussion of graphs, to the extent that it can be regarded as *the* fundamental concept of graph theory.[1]

An example for the practical importance of determining whether two graphs are isomorphic is given by working with organic compounds. Such are build up large dictionaries of compounds that they have exactly analyzed. When a new compound is found, one wants to know if it is already in the dictionary. In general, many compounds have the same molecular formula but differ in their structure as graphs.[2] Consequently, one must test the new compound to see if it is isomorphic to a known compound.

7.1 Isomorphic graphs

Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are said to be isomorphic if there exists a one-to-one, onto mapping $f : V_1 \to V_2$ such that $\underline{vv'} \in E_1$ if and only if $\underline{f(v)f(v')} \in E_2$. f is called an isomorphism.

Observation 7.1.1 *Isomorphism is an equivalence relation on the collection of all graphs.*

Isomorphic graphs are structurally equivalent.

Observation 7.1.2 *Isomorphic graphs have/are*
- the same number of vertices;

[1] Solutions to the fascinating problem of determining the number of graphs starts in 1927 by a remarkable paper by Redfield [199]. But it seems forgotten for several years, when it was independently investigated by Polya [192] in 1937.

[2] For instance the molecule C_4H_{10} can be realized as butane and as isobutane.

- *the same number of edges;*
- *the same cyclomatic number;*
- *the same number of components;*
- *the same complement graphs;*
- *the same number of bridges;*
- *the same number of articulations;*
- *the same connectivity;*
- *an equal number of vertices of any given degree;*
- *for each integer k, the same number of paths of length k;*
- *for each integer k, the same number of cycles of length k;*
- *the same chromatic number;*
- *the same chromatic index;*
- *the same chromatic polynomial;*
- *a Hamiltonian cycle or not;*
- *the same metric order;*
- *an (open or closed) Eulerian line or not;*
- *both bipartite or not;*
- *both planar or not;*
- **and:** *the same number of spanning trees.*

However, each of these properties is a necessary but not a sufficient criteria for isomorphism. Is there a collection sufficient for isomorphism?[3] For practice, give two non-isomorphic graphs with
a) the same degree sequence; or
b) equal families of neighborhood, respectively.

In general it is difficult to determine whether two graphs are isomorphic.[4] Consider two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ with n vertices. There are $n!$ possible bijections between V_1 and V_2. This immediately implies the following test for graph isomorphism.

Algorithm 7.1.3 *Given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$.*
We return "yes" or "no", according to whether G_1 is isomorphic to G_2.

1. *If $|V_1| \neq |V_2|$ return* **no***;*
2. *If the ordered sequences of the degrees of the vertices of G_1 and G_2 are not equal return* **no***;*
3. *Fix an ordering for the vertices of G_1;*
   Write the adjacency matrix $A(G_1)$ with respect to that ordering;
   For each ordering (= permutation) π of the vertices of G_2 **do**
   (a) *Write the adjacency matrix $A(G_2)$ with respect to the ordering π;*

[3] This is an open, and very hard question.
[4] Can be a graph isomorphic to its complement? If yes, find all.

(b) If $A(G_1) = A(G_2)$ *return* **yes**;

4. Return **no**.

Testing each such correspondence to see whether it preserves adjacency and non-adjacency is impractical if n is large. For practice discuss the following fact.

Observation 7.1.4 *Isomorphism of graphs is usually much harder to prove than non-isomorphism.*

It is strange, but the computational complexity to verify whether two graphs are isomorphic is still unknown: No polynomially bounded algorithm is known, on the other hand it has not been proved that this problem is in $\mathcal{NPC}$. Maybe, this problem is a member of $\mathcal{NPI}$. A monograph on isomorphism detection is given in [134].

7.2 Labeled and unlabeled graphs

Remember that counting the number of unlabeled graphs tend to be harder to solve than counting the number of labeled ones. The following observation is simple to see, but gives in many cases a first estimation for the number of unlabeled graphs.

Observation 7.2.1 *Let* $\mathcal{C}_n^{labeled}$ *be a collection of labeled graphs each with* n *vertices. Similar denotes* $\mathcal{C}_n^{unlabeled}$ *the same collection without labels. Then*

$$\left|\mathcal{C}_n^{labeled}\right| \geq \left|\mathcal{C}_n^{unlabeled}\right| \geq \frac{1}{n!} \cdot \left|\mathcal{C}_n^{labeled}\right|. \tag{7.1}$$

I. A first application: In 5.7.3 we saw that there are approximately more than $\beta^n \cdot n!$ planar labeled graphs. Together with our observation this implies

Theorem 7.2.2 *There is an exponential number of planar graphs.*

The two graphs K_5 and $K_{3,3}$ are non-planar. This can easily be seen by 4.8.2 and 4.1.2. Conversely, it holds the following surprisingly fact, when we consider a weaker form of isomorphism: Two graphs G_1 and G_2 are said to be homeomorphic if they are isomorphic or if they can both be obtained from the same graph G by a sequence of subdivisions. One may think homeomorphic graphs as being isomorphic, except, possibly for vertices of degree 2. If two graphs are homeomorphic, they are either both planar or they are both non-planar. Then it is clear that a graph containing a subgraph homeomorphic to either K_5 or $K_{3,3}$ cannot be planar. The converse of this fact, however, is also true, but much more difficult to prove.

Theorem 7.2.3 *(Kuratowski) A graph is non-planar if and only if it contains a subgraph homeomorphic to either* K_5 *or* $K_{3,3}$.

The proof of this deep topological result is beyond the scope of the present script, but is given in each of the textbooks of graph theory.
In this sense there are only two non-planar graphs, but of course, there are infinitely many non-planar graphs which can be exactly named:

a) K_n is planar if and only if $n \leq 4$.

b) K_{n_1,n_2} is planar if and only if $\min\{n_1, n_2\} \leq 2$.

c) Q^D is planar if and only if $D \leq 3$.[5]

d) The Petersen graph G_{petersen} is not planar.

A (little bit surprising) consequence of Kuratowski's theorem is that we can check in polynomially bounded time whether a graph is planar. But nevertheless such naively searching for homeomorphic subgraphs K_5 or $K_{3,3}$ consumes many time. Several good planarity algorithms have been developed, until an algorithm which runs in linear time, [44], [109].[6]

Remember that a planar graph $G = (V, E)$ is called outer-planar if it can be embedded into the plane such that all vertices lying on the boundary of exactly one region. Then together with 7.2.3 we characterize outer-planar graphs.

Theorem 7.2.4 *A graph is non-outer-planar if and only if it contains a subgraph homeomorphic to either K_4 or $K_{2,3}$.*

Recall what maximal planar and maximal outer-planar mean. For practice verify (and expand) the following table for the numbers of such graphs with n vertices.

class	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	...
max. planar	1	1	2	5	14	...
max. outer-planar	1	1	3	4	...	

II. We will go a little bit further. Let G be a graph with n vertices. An isomorphism of G to itself is called an automorphism. Thus each automorphism of $G = (V, E)$ is a permutation of V which preserves adjacency.
The set $\text{Aut}(G)$ of all automorphisms forms a group.[7]
Any two bijective functions f_1 and f_2 from the set $\{1, \ldots, n\}$ to the set of vertices of G give two identical labeled copies of G if and only if there is an automorphism $\alpha \in \text{Aut}(G)$ such that $\alpha \circ f_1 = f_2$. It follows

Theorem 7.2.6 *The number of different ways to label a graph G of n vertices is*

$$\frac{n!}{|Aut(G)|}. \tag{7.2}$$

[5] Note, D is not the number of vertices, compare 4.1.4.

[6] If a graph G is both Hamiltonian and planar, then we observe that in an embedding of G the edges which are not in the Hamilton cycle H can be divided in two sets, those inside H and those outside H. This implies a very simple planarity algorithm, see [9].

[7] And vice versa:

Theorem 7.2.5 *(Frucht, [92]) For any finite group Γ there exists a graph G such that $Aut(G)$ is algebraic isomorph to Γ.*

Proof. Certainly, there exist $n!$ labellings of G using n labels without regard to which labellings are distinct. For a given one of these, each automorphism gives rise to an identical labeling. □

For example the complete graph K_n has $n!$ automorphisms, and so one labeling. Recall that there are 16 labeled trees with four vertices. We note that among these trees

$$12 = \frac{4!}{|\mathrm{Aut}(P_4)|} = \frac{24}{2}$$

are isomorphic to the path P_4 and 4 to the star $K_{1,3}$:

$$4 = \frac{4!}{|\mathrm{Aut}(K_{1,3})|} = \frac{24}{6}.$$

It is easy to see that a graph and its complement have the same automorphism group. Using Lagrange's theorem O.1.1 from group theory we find

Theorem 7.2.7 *The order $|Aut(G)|$ of a graph G with n vertices is a divisor of $n!$, and equals $n!$ if and only if G is the complete graph K_n or its complement, the empty graph K_n^c.*

In view of 7.1.4 we may assume that it is hard to determine the automorphism group of a graph. Ihringer [138] gives an algorithm.

Of course the identity is in any case an automorphism, all other are called symmetries. In the sense of random graphs we have

Theorem 7.2.8 *Almost all graphs have no symmetries.*

7.3 The number of graphs

Since every unlabeled graph G with n vertices and m edges is isomorphic to a set of at most $n!$ labeled graphs we have

Theorem 7.3.1 *The number $g(n, m)$ of non-isomorphic graphs G with n vertices and m edges is at least*

$$\frac{\binom{\binom{n}{2}}{m}}{n!} \tag{7.3}$$

The theory of random graphs is almost exclusively concerned with graphs on distinguish vertices. 7.2.6 and 7.2.8 give the possibility to extend this to unlabeled ones. And indeed in [30] it is shown that under suitable conditions the bound (7.3) is asymptotically exact, and consequently Polya already obtained $2^{\binom{n}{2}}/n!$ as the asymptotic number of graphs with n vertices.

Of course, in view of 7.2.1, the number of unlabeled graphs with n vertices is much smaller than that given by 5.1.1(b). On the other hand, by considering all $n!$ labellings, we find that the number $g(n)$ of all non-isomorphic graphs obeys $g(n) \cdot n! \geq \text{graph}(n)$. Hence,

Theorem 7.3.2 *The number of non-isomorphic graphs with n vertices is at least*

$$g(n) \geq \sqrt{2}^{n(n-1)}/n!. \tag{7.4}$$

We are interested in the order of growing of the function $g(n)$. On the one hand,

$$\log \text{graph}(n) = \log 2^{\binom{n}{2}} = \frac{n(n-1)}{2} = \frac{n^2}{2}\left(1 - \frac{1}{n}\right).$$

On the other hand,

$$\begin{aligned}
\log g(n) &\geq \log \frac{2^{\binom{n}{2}}}{n!} \\
&= \binom{n}{2} - \log n! \\
&= \frac{n^2}{2}\left(1 - \frac{1}{n}\right) - \log n! \\
&\geq \frac{n^2}{2}\left(1 - \frac{1}{n}\right) - n \log n \\
&= \frac{n^2}{2}\left(1 - \frac{1}{n} - \frac{2\log n}{n}\right).
\end{aligned}$$

In other words,

Theorem 7.3.3 *The order of non-isomorphic graphs grows not essentially slower than of all graphs.*

Unfortunately, an explicit formula for the number of non-isomorphic graphs is unknown, but for the first values we know

Number n of vertices	Number of non-isomorphic graphs
1	1
2	2
3	4
4	11
5	34
6	156
7	1,044
8	12,346
9	274,668
10	12,005,168
11	1,018,997,864
12	165,091,172,592

A more detailed list on the number of graphs depending on the two quantities, the number n of vertices and the number m of edges, where $0 \le m \le \binom{n}{2}$, is given by Harary [121] and Harary, Palmer [122]:

$m \setminus n$	1	2	3	4	5	6	7
0	1	1	1	1	1	1	1
1		1	1	1	1	1	1
2			1	2	2	2	2
3			1	3	4	5	5
4				2	6	9	10
5				1	6	15	21
6				1	6	21	41
7					4	24	65
8					2	24	97
9					1	21	131
10					1	15	148
11						9	148
12						5	131
13						2	97
14						1	65
15						1	41
16							21
17							10
18							5
19							2
20							1
21							1
sum	1	2	4	11	34	156	1044

As an extreme case, the exact value for $n = 24$ is known, asymptotically $g(24) \approx 1.9570\ldots \cdot 10^{59}$.
The list suggests several properties of the function $g(n, m)$ of graphs with n vertices and m edges. For instance the reader should discuss

a) $g(n, m) = g(n, \binom{n}{2} - m)$, which follows from the fact that two graphs are isomorphic if and only if their complements are isomorphic.

b) $g(n, m)$ forms a unimodal sequence in m for fixed n. Is this true?

c) A maximum value of $g(.,.)$ is achieved for approximately $m = \binom{n}{2}/2$. Really?

d) $g(n, m)$ is a constant sequence in n for fixed $m < \binom{n}{2}$.

Note that all these questions are simple for the function graph(n, m) of the number of labeled graphs, see C.2.3.

7.4 The number of connected graphs

In view of 5.4.1 and 5.8.2 we assume that the number of connected graphs are not essentially less than the number of all graphs. And, indeed this seems true:

Number n of vertices	Number of non-isomorphic connected graphs	ratio of connected and all graphs
1	1	1
2	1	0.5
3	2	0.5
4	6	0.5454...
5	21	0.6176...
6	112	0.7179...
7	853	0.8170...
8	11,117	0.9004...
9	261,080	0.9505...
10	11,716,571	0.9759...
11	1,006,700,565	0.9879...
12	164,059,830,476	0.9937...

In general, we introduce the cyclomatic number for a graph G with n vertices, m edges and c components by

$$\nu(G) = m - n + c. \tag{7.5}$$

If G is connected its cyclomatic number is $m - n + 1$.

Observation 7.4.1 *$\nu(G)$ is the number of chords in G with respect to any spanning forest of G.*

With these facts in mind, the following result is easy to prove.

Theorem 7.4.2 *The number of non-isomorphic connected graphs with n vertices and m edges is at most*

$$T(n) \cdot \binom{\binom{n}{2}}{m-n+1}, \tag{7.6}$$

whereby T denotes the number of trees.

Proof. Each connected graph contains a spanning tree with $n - 1$ edges. The remaining $m - n + 1$ edges are selected from $\binom{n}{2}$. □

More and better bounds, created by extensive calculations, are given by Wetuchnowski [249].

7.5 Several specific cases

I. In 5.2.1 we counted bipartite labeled graphs. An explicit formula the unlabeled case is not known. Harary, Palmer [122] give an implicit solution. For instance, we consider bipartite graphs with $2+3$ vertices and m edges and find the following numbers:

m	0	1	2	3	4	5	6	total
number	1	1	3	3	3	1	1	13

II. Remember that the complement of a graph G, denoted by G^c, has the same set of vertices as G and two vertices are adjacent in G^c if and only if they are not adjacent in G. We define G as self-complementary if G and G^c are isomorphic.
Let G be a self-complementary graph with n vertices and m edges then

$$m = \frac{1}{2}\binom{n}{2} = \frac{n(n-1)}{4}.$$

Consequently,

Observation 7.5.1 *Every self-complementary graph has $n \equiv 0$ or $1 (mod\ 4)$ vertices.*

Harary, Palmer [122] reports the following known numbers of self-complementary graphs with n vertices:

n	4	5	8	9	12	13	16	17
number	1	2	10	36	720	5,600	703,760	11,220,000

Asymptotically the number of self-complementary graphs with $4n$ or $4n+1$ vertices is $2^{2n^2-2n}/n!$.

Chapter 8

Polyhedra

Polyhedra are convex hulls of finite sets of points in the three-dimensional (Euclidean) space. They, especially the regular ones, have been fascinating people since the antiquity. Their investigation was one of the main sources of graph theory.

Remember that a graph is called planar if it can be embedded into the plane such that no two curves which are the embeddings of the edges intersect each other outside of the vertices. In such an embedding there are regions, which are parts of the plane bounded by the curves. This shows that planar graphs and polyhedra are intimately related.[1]

8.1 The f-vector

In an embedding of a planar graph there are several other entities than vertices and edges. Why did we call it a facet? When Euler was trying "his" formula 4.8.1, he was studying polyhedra, which are solid bodies bounded by polygons. Here we translate the notations

in graphs	in polyhedra	dimension
vertex	node, vertex	0
edge	edge	1
region	facet	2.

[1]There are "nice" drawings of planar graphs:

Theorem 8.0.2 *Every 3-connected planar graph can be embedded in the plane such that all edges are segments, and all bounded regions, as well as the union of all bounded regions are convex polygons.*

For a proof see [236].
For practice the reader should prove the following corollary, given by Wagner and rediscovered by Fary: Every planar graph has a straight line drawing in the plane without intersections. It is also quite easy to prove this corollary directly by induction, see [127].

In other terms, Euler found

$$\text{number of facets } + \text{ number of nodes } = \text{ number of edges } + 2, \tag{8.1}$$

written in the terms of the so-called f-vector[2] (f_0, f_1, f_2) with f_i as the number of i-dimensional faces of the polyhedron.

Observation 8.1.1 *Combinatorially equivalent polyhedra have equal f-vectors, but not vice versa.*

As example construct two polyhedra, both with the f-vector $(6, 12, 8)$, that are not combinatorially equivalent.

Theorem 8.1.2 *Euler's relation 4.8.1 is the only (linear) restriction for f-vectors of polyhedra.*

Proof. Assume that there is a relation

$$\alpha_0 f_0 + \alpha_1 f_1 + \alpha_2 f_2 = \beta. \tag{8.2}$$

Let Q be a polygon with k nodes, and let P_1 be a pyramid and P_2 be a bi-pyramid each with Q as basis. We can express the f-vectors of P_1 and P_2

$$f(P_1) = (k+1, 2k, k+1) \tag{8.3}$$

and

$$f(P_2) = (k+2, 3k, 2k). \tag{8.4}$$

Now we have

$$\alpha_0(k+1) + \alpha_1(2k) + \alpha_2(k+1) = \beta \tag{8.5}$$

and

$$\alpha_0(k+2) + \alpha_1(3k) + \alpha_2(2k) = \beta. \tag{8.6}$$

Combining (8.5) and (8.6) yields the assertion. □

Theorem 8.1.3 *(Steinitz) (f_0, f_1, f_2) is an f-vector of a polyhedron if and only if*

$$f_0 - f_1 + f_2 = 2 \tag{8.7}$$

and

$$4 \leq f_0 \leq \frac{2f_1}{3} \tag{8.8}$$

and

$$4 \leq f_2 \leq \frac{2f_1}{3}. \tag{8.9}$$

[2]Abbreviation for "face" vector.

Proof. The first condition (8.7) is Euler's relation.
Since the "smallest" polyhedron is a tetrahedron, the lower bounds on f_0 and f_2 are necessary. Consider the graph G of a polyhedron. Each vertex has degree at least three. In view of 4.1.1 we find

$$2f_1 \geq 3f_0,$$

and so (8.8). Similarly (8.9), since each region (facet) is bounded by at least three edges and each edge lies on exactly two faces.

Here is a sketch of the converse. If the triple (f_0, f_1, f_2) satisfies the conditions, then there are nonnegative integers x and y such that

$$(f_0, f_1, f_2) - x(1,3,2) - y(2,3,1) = (6,10,6) \text{ or } (6,9,5) \text{ or } (5,9,6) \text{ or } (4,6,4), \quad (8.10)$$

which can be proved by induction on f_1. Each of the four triples just listed can be realized by a polyhedron, which we will call irreducible.
Adding $(2,3,1)$ can be realized by slightly adjusting the edges incident at a vertex and replacing the vertex by a triangular facet. Adding $(1,3,2)$ can realized similarly by introducing two triangular facets in place of two adjacent edges of some facet. By iterating, one of the four irreducible polyhedra can be built up to realize (f_0, f_1, f_2). □

In view of 8.1.3 we can eliminate one parameter of an f-vector, and consequently, he following is true for polyhedra and its integer values f_0 and f_2:

Corollary 8.1.4 *$(f_0, \cdot, f_2)$ is the f-vector of a polyhedron if and only if*

$$4 \leq f_0 \leq 2f_2 - 4 \quad \text{and} \quad 4 \leq f_2 \leq 2f_0 - 4. \quad (8.11)$$

8.2 The graph of a polyhedron

In a natural sense, the vertices and edges of a polyhedron form a graph.

Theorem 8.2.1 *(Steinitz, compare [110]) A graph is isomorphic to the graph of a polyhedron if and only if it is planar and 3-connected.*

At this point, observe that the combinatorial structure of a polyhedron is completely determined by its graph. It results from Whitney's theorem [251] that the embedding of a 3-connected planar graph is unique.[3] In other terms,

Corollary 8.2.3 *Two polyhedra are combinatorial equivalent if and only if their graphs are isomorphic.*

[3] A similar statement for higher-dimensional polytopes is not known, but the following nice result: We can define the graph of a polytope in the natural way.

Theorem 8.2.2 *(Balinski [14]) The graph of a D-dimensional polytope is D-connected.*

A graph G is called polyhedral if there exists a polyhedron which graph is isomorphic to G. Equivalently, if G is planar and 3-connected.

The high connectedness of polyhedral graphs has the following consequence.

Theorem 8.2.4 *Let G be the graph of a polyhedron with n vertices. Then*

$$diam(G) \leq \frac{n+1}{3}. \tag{8.12}$$

Proof. Let l be the diameter of G and let v and v' be two vertices which achieve l. Then there is a path from v to v' of length l, and, consequently, with $l+1$ vertices. Since G is 3-connected there are two vertex-disjoint paths interconnecting v and v', each with at least $l-1$ internal vertices. Hence,

$$(l+1) + 2(l-1) = 3l - 1 \leq n.$$

The assertion follows. □

8.3 The number of polyhedra I.

How many polyhedra are there with n vertices and f facets?

Remember that for the f-vector $(6, 12, 8)$ there are two non-isomorphic polyhedra. For the number of polyhedra with small numbers $n = f_0$ of nodes and numbers f_2 of facets we know, compare [60]:

$f_0 \setminus f_2$	4	5	6	7	8	9
4	1					
5		1	1			
6		1	2	2	2	
7			2	8	11	8
8			2	11	42	74
9				8	74	296
10				5	76	633
11					38	768
12					14	558
13						219
14						50

Note the symmetry of the array. This is a consequence of the relation of planar graphs and its duals.[4]

Summing up about the table gives

[4]The Poincare duality is a two-step process running as follows: Let $G = (V, E)$ be a planar graph embedded in the plane. Then

1. Insert a vertex into the interior of each region. Let V^d be the set of all such vertices;

Number of nodes	Number of polyhedra
4	1
5	2
6	7
7	34
8	257
9	2,606
10	32,300
11	440,564

And further:

Number of nodes	Number of polyhedra
12	$6.3 \cdot 10^6$
13	$9.6 \cdot 10^7$
14	$1.5 \cdot 10^9$
$\vdots$	
18	$1.1 \cdot 10^{14}$

This suggests that the number increases exponentially, supported by 7.2.2. And indeed this is true:

Theorem 8.3.2 *(Bender [22]) The number of polyhedra with $f_0 = i + 1$ nodes and $f_2 = j + 1$ facets is asymptotically*

$$\frac{1}{972 \cdot ij(i+j)} \cdot \binom{2i}{j+3} \cdot \binom{2j}{i+3}. \tag{8.13}$$

Croft et al. [60] name several facts about the asymptotic behavior for the number of polyhedra. In particular,

Remark 8.3.3 *The number of polyhedra with $m = f_1$ edges is asymptotic to*

$$\frac{1}{486\sqrt{\pi}} \cdot \frac{4^m}{m^{7/2}}. \tag{8.14}$$

2. Through each edge $e \in E$ draw an edge joining the vertex of V^d on one side of e to the vertex on the other side. Let E^d be the set of all such edges.

Observation 8.3.1 *Let $G = (V, E)$ be a planar graph embedded in the plane; and let $G^d = (V^d, E^d)$ be one of its dual graphs. Then G^d is itself an embedding of a planar graph with*
- $|V^d| =$ number of regions of G;
- $|E^d| = |E|$;
- number of regions of $G^d = |V|$.

8.4 The number of polyhedra II.

The results given in the section before lead up to the investigate the number of polyhedra depending to the number and the size of the facets. More exactly, for which sequences $r_3, r_4, \ldots$ do there exists a polyhedron with r_j facets bounded by j edges? This is an old and (in full generality) still unsolved question. At first we give some necessary conditions.

Theorem 8.4.1 *Let n_i be the number of vertices of degree i, and let r_j be the number of facets bounded by j edges of a polyhedron. Then it holds*

$$n_3 + r_3 = 8 + \sum_{k\geq 5}(k-4)(n_k + r_k). \tag{8.15}$$

And, the pair

$$12 + \sum_{j\geq 4}(2j-6)r_j = \sum_{i\geq 3}(6-i)n_i; \tag{8.16}$$

$$12 + \sum_{j\geq 3}(j-6)r_j = \sum_{i\geq 4}(6-2i)n_i; \tag{8.17}$$

Proof. The first equation is a consequence of 4.8.1:

$$\begin{aligned}
2 &= f_0 - f_1 + f_2 \quad \text{hence} \\
0 &= 8 - 4f_0 + 4f_1 - 4f_2 \\
&= 8 + 2f_1 - 4f_0 + 2f_1 - 4f_2 \\
&= 8 + \sum_{k\geq 3} kn_k - \sum_{k\geq 3} 4n_k + \sum_{k\geq 3} kr_k - \sum_{k\geq 3} 4r_k \\
&= 8 + \sum_{k\geq 3}(k-4)(n_k + r_k) \\
&= 8 - (n_3 + r_3) + \sum_{k\geq 5}(k-4)(n_k + r_k).
\end{aligned}$$

The result follows.
In view of 4.1.1 and 4.8.1 we have

$$\sum_{i\geq 3} in_i = 2f_1 = 2(f_0 + f_2 - 2) = 2f_0 + 2f_2 - 4 = \sum_{i\geq 3} 2n_i + \sum_{j\geq 3} 2r_j - 4,$$

which implies

$$\sum_{i\geq 3}(i-2)n_i + 4 = \sum_{j\geq 3} 2r_j. \tag{8.18}$$

Similarly,

$$\sum_{j\geq 3}(j-2)r_j + 4 = \sum_{i\geq 3} 2n_i. \tag{8.19}$$

We determine $2\times$ (8.19) $-$ (8.18):

$$2\sum_{i\geq 3} 2n_i - \sum_{i\geq 3}(i-2)n_i - 4 = 2\sum_{j\geq 3}(j-2)r_j + 8 - \sum_{j\geq 3} 2r_j,$$

which immediately implies the assertion. The other equation of the pair follows similar by interchanging the role of r_j and n_i. □

The following theorem provides a partial converse of 8.4.1.

Theorem 8.4.2 *(Eberhard [72]) Suppose that the finite sequences $r_3, r_4, r_5, \ldots$ and $n_3, n_4, n_5, \ldots$ of nonnegative integers satisfy the equations*

- *(8.15),*
- *(8.16) and*
- $\sum_{i\geq 3, i\neq 4} in_i \equiv 0 (mod\ 2)$.

Then there exists a polyhedron with n_i vertices of degree i, and r_j regions bounded by j edges for all numbers $i, j \neq 4$.

For a further discussion see [110].[5]

8.5 Regular polyhedra

We will discuss several regular conditions for polyhedra.

I. A polyhedron is called regular if there exist integers $r, s \geq 3$ such that each vertex has r faces, or equivalently r edges, incident at it, and each face has s edges on its boundary. The tetrahedron with $(r, s) = (3, 3)$; and the cube are such bodies. Regular polyhedra are also known as Platonic solids; which already played an important role in the science by the ancient Greeks.[6] They knew, proven by Euclid, that there are exactly five of such solids, which we will prove in the next theorem.

Theorem 8.5.1 *Suppose that a regular polyhedron has each vertex of degree r and each face bounded by exactly s edges. Then*

$$(r, s) = (3, 3)\ or\ (3, 4)\ or\ (4, 3)\ or\ (3, 5)\ or\ (5, 3). \tag{8.20}$$

Proof. We have $f_0 - f_1 + f_2 = 2$ by 4.8.1. Additionally, with similar arguments than for the proof of 4.1.1, we find the following facts.

$$2f_1 = rf_0, \tag{8.21}$$

[5] And visit Sloane's encyclopedia of integer sequence in the www.

[6] We discuss polyhedra in the sense of graph theory. In the sense of metric geometry all faces of a Platonic solid must be congruent and regular polygons.
Remember: Graph theory is a part of topology.

and

$$2f_1 = sf_2. \tag{8.22}$$

Hence,

$$\left(\frac{2}{r} - 1 + \frac{2}{s}\right) f_1 = 2,$$

and, equivalently

$$\frac{1}{r} + \frac{1}{s} = \frac{1}{2} + \frac{1}{f_1}. \tag{8.23}$$

In other words, the values r and s determine the quantities f_i, $i = 0, 1, 2$. In particular $\frac{1}{r} + \frac{1}{s} > \frac{1}{2}$ gives $r = 3$ or $s = 3$. Then (8.23) arises the five possibilities. □

For each possible pair (r, s) we can find the values of f_0, f_1, f_2 from (8.21), (8.22) and (8.23). We tabulate these values, and give the name of the corresponding Platonic solid, the first result in mathematics about counting.[7]

r	s	f_0	f_1	f_2	Name
3	3	4	6	4	tetrahedron
3	4	8	12	6	cube
4	3	6	12	8	octahedron
3	5	20	30	12	dodecahedron
5	3	12	30	20	icosahedron

Corollary 8.5.2 *A polyhedron whose all facets are hexagons cannot exist.*

II. As well as the regular polyhedra just discussed, there exist the semi-regular polyhedra known as the Archimedean solids, which have more than one type of facets.[8] As an important example we will discuss the polyhedron which is made up of pentagons and hexagons, with each vertex-degree three. We have

$$3f_0 = 2f_1 = 5r_5 + 6r_6. \tag{8.24}$$

Then 4.8.1 implies

$$12 = 6f_0 - 6f_1 + 6r_5 + 6r_6 = 4f_1 - 6f_1 + 2f_1 + r_5 = r_5.$$

Consequently,

Theorem 8.5.3 *A polyhedron which is made up of pentagons and hexagons, with each vertex-degree three, must contain exactly 12 pentagonal faces.*

[7] The fact that there are exactly five Platonic solids play an important role in philosophy of the ancient Greeks until Kepler, see [254].

[8] For practice discuss the number of all such solids. How many different polyhedra do you find? Hint: Compare [12].
Note, that there are classes of infinitely many semiregular polyhedra. For example consider the convex hull of two parallel regular n-gons.

In other terms,

$$f_2 = 12 + r_6. \tag{8.25}$$

Moreover, both (8.15) and (8.16) give

$$f_0 = n = 20 + 2r_6, \tag{8.26}$$

and by 4.1.1 we have

$$f_1 = 30 + 3r_6 \tag{8.27}$$

for the number n of vertices and the number f_1 of edges depending on the number r_6 of hexagons.
The case with zero hexagons corresponds to a dodecahedron. The case with 20 hexagons corresponds to the pattern of a soccer ball.[9]

8.6 Simplicial and simple polyhedra

We introduce classes of polyhedra, which are defined by a "non-degeneracy" condition, which makes these polyhedra much easier to handle than polyhedra in general. Examples are discussed in the section before.

Actually, the conditions will be dual, such that there is no formal reason to prefer one of the classes. And moreover, counting the member of one class is enumerating the member of the other one.

I. A polyhedron is called simplicial if each facet is a triangle (a simplex).
In a simplicial polyhedron each facet is incident with exactly three edges. Hence, we have $3f_2 = 2f_1$. Together with 4.8.1 this obtain

Theorem 8.6.1 *The family of f-vectors of simplicial polyhedra depends only from one parameter. In particular:*

$$(f_0, 3f_0 - 6, 2f_0 - 4) \quad \textit{with } f_0 \geq 4. \tag{8.28}$$

Remember that we find these equations in considering maximal planar graphs. Conversely, every maximal planar graph with at least four vertices is 3-connected, which is not easy to see.

Theorem 8.6.2 *A graph G is a graph of a simplicial polyhedron if and only if G is maximal planar.*

In view of these both theorems we can determine the number of simplicial polyhedra for small values $n = f_0$, [110]:

[9] These structures play a very important role in the chemistry of carbons. The molecular structure is that of 60 carbon atoms situated at the vertices of a truncated icosahedron. Also known as a fullerene named after the architect R.M.Fuller.

$n =$	Number of simplicial polyhedra
4	1
5	1
6	2
7	5
8	14
9	50
10	233
11	1,249
12	7,595
13	49,566
14	339,722
$\vdots$	
20	$\approx 5.8 \cdot 10^{10}$
21	$\approx 4.5 \cdot 10^{11}$

II. A polyhedron is called simple if each vertex is incident to precisely 3 edges. In other terms, the graph of a simple polyhedron is 3-regular.

a) A polyhedron is simple if and only if its dual is simplicial.

b) The dual graph of the graph of a simple polyhedron is maximal planar.

c) A polyhedron is simple if and only if each vertex is contained in exactly three facets.

For more properties of simple polyhedra see [33] and [110].
This observation shows that by investigation of simplicial polyhedra we also consider simple ones.

Theorem 8.6.3 *Let r_j be the number of facets bounded by j edges of a simple polyhedron. Then*

$$3r_3 + 2r_4 + r_5 = 12 + \sum_{j \geq 7} (j-6) r_j. \tag{8.29}$$

Proof. In view of (8.17) and $n_i = 0$ for $i \neq 3$ we find

$$12 + \sum_{j \geq 3} (j-6) r_j = 0.$$

□

On of the interesting features of the theorem is that it does not contain an information about the number r_6 of hexagons.

Theorem 8.6.4 *(Eberhard [72], [110]) Suppose that the finite sequence $r_3, r_4, r_5, r_7, \ldots$ of nonnegative integers which satisfy the equation (8.29). Then there exists a simple polyhedron with r_j regions bounded by j edges for all $j \neq 6$.*

Note that there are examples of two sequences $r_3, r_4, r_5, \ldots$ which differ only in the value r_6, one may be realized a simple polyhedron and the other not. On the other hand, if a sequence realized a simple polyhedron for some r_6, then for infinitely many values of r_6. (8.29) implies $12 \leq 3r_3 + 2r_4 + r_5$. Consider the equality case, then we find the following solutions in nonnegative integers and minimal values of r_6:

r_3	r_4	r_5	minimal r_6
4	0	0	0
3	1	1	3
3	0	3	1
2	3	0	0
2	2	2	0
2	1	4	1
2	0	6	0
1	4	1	2
1	3	3	0
1	2	5	1
1	1	7	2
1	0	9	3
0	6	0	0
0	5	2	0
0	4	4	0
0	3	6	0
0	2	8	0
0	1	10	2
0	0	12	0

The given values each realizes a polyhedron. (Exercise.)

III. We consider polyhedra having only vertices of degree 4. (8.15) implies

Theorem 8.6.5 *Let r_j be the number of facets bounded by j edges of a polyhedron having only vertices of degree 4. Then*

$$r_3 = 8 + \sum_{k \geq 5} (k-4) r_k. \tag{8.30}$$

The converse fact for sequences $r_3, r_5, r_6, \ldots$ is also true, [110].

Chapter 9

The Number of Unlabeled Trees

Now we may estimate the number of non-isomorphic trees.

Recall that the problem to decide if graphs are isomorphic is not simple. On the other hand, for trees the isomorphic problem is easy: there is a quadratic time algorithm which decides whether two trees are isomorphic; see [240]. But this does not mean that it is easy to count the number of such trees.

9.1 Upper and lower bounds

Let $T(n)$ be the number of non-isomorphic trees with n vertices. By considering all $n!$ labellings, we have $n! \cdot T(n) \geq n^{n-2}$. Hence, and by using Stirling's inequality (I.3),

Theorem 9.1.1 *Let $T(n)$ be the number of non-isomorphic trees with n vertices. Then*

$$T(n) \geq \frac{n^{n-2}}{n!} \geq \frac{e^n}{en^{5/2}}. \tag{9.1}$$

In particular, the number of non-isomorphic, that means unlabeled, trees grows exponentially.
On the other hand,

Theorem 9.1.2 *Let $T(n)$ be the number of non-isomorphic trees with n vertices. Then*

$$T(n) \leq C_{n-1}, \tag{9.2}$$

where C_{n-1} denotes the $(n-1)$th Catalan number.

Proof. We will describe a non-optimal technique, involving drawing a tree in the plane: Let $n > 1$ be an integer. A tree code w (with respect to $n-1$) is a sequence in $\{0,1\}^{2(n-1)}$ with the following properties:

1. In each prefix of w the number of 1 is at least the number of 0;
   In particular, the first letter in w must be 1;

2. The number of 1 in w equals the number of 0;
   In particular, the last letter in w must be 0.

Algorithm 9.1.3 *Let w be a tree code with respect to $n-1$. Then draw a tree by the following algorithm:*

1. *Set a vertex as the origin;*

2. *Read w letter by letter and*
   if you see a 1 then draw a new edge to a new vertex;
   if you see a 0 then move back by one edge toward the origin.

Thus the tree is described by its tree code. Hence, after generating all tree codes, we can generate all unlabeled trees with n vertices.
We know by M.1.3 that the number of tree codes is the Catalan number, which gives an upper bound for the number of non-isomorphic trees. □

Note that the tree code is far from optimal; every unlabeled tree has many different codes. For instance all the codes 11010010, 10110100, 11101000, 10101100, 11011000 and 11100100 generate the same tree.

It holds,

$$T(n) \leq C_{n-1} = \frac{1}{n}\binom{2n-2}{n-1} \approx \frac{4^n}{\sqrt{2\pi \cdot n^3}}. \tag{9.3}$$

Consequently, we have that $T(\cdot)$ is bounded from above by an exponential function. All together we expect that

$$T(n) = \Theta\left(\frac{a^n}{f(n)}\right) \tag{9.4}$$

with $e \leq a \leq 4$ and a function $f(n)$ which is bounded by a low degree polynomial. And indeed, according to a difficult result of Pólya, compare [122] or [249], the number of unlabeled trees is asymptotically completely determined:

Theorem 9.1.4 *Let $T(n)$ be the number of non-isomorphic trees with n vertices. Then*

$$T(n) \approx \frac{c \cdot a^n}{n^{5/2}}, \tag{9.5}$$

where

$$a = 2.9557\ldots \text{ and } c = 0.5349\ldots. \tag{9.6}$$

Note that, using the Stirling approximation I.2.2,

$$\frac{n^{n-2}}{n!} \approx \frac{n^n}{n^2} \cdot \frac{e^n}{\sqrt{2\pi n} \cdot n^n} = \frac{e^n}{\sqrt{2\pi} \cdot n^{5/2}}. \tag{9.7}$$

In other terms, the approximation ratio is not far from (9.5), since

$$\sqrt{2\pi} = 2.5066\ldots \text{ and } e = 2.7183\ldots. \tag{9.8}$$

In view of 7.2.8 this is not a surprise, because almost all graphs have nontrivial automorphism groups.

9.2 Generating functions

Now we introduce a key tool for enumeration. The generating function of a sequence $a_0, a_1, a_2, \ldots$ is defined to be

$$f(x) = \sum_{i=0}^{\infty} a_i x^i. \tag{9.9}$$

(Sometimes we will use the index starting at 1.)

The concept of generating functions is very powerful, it rests upon its ability not only to solve the kinds of problems we have considered so far, but also to help us in new situations where additional restrictions could be involved. This is not a surprise since generating functions are formal power series, which are well-understood in the sense of higher algebra and calculus.

As an example, in view of C.2.2, we immediately see

Observation 9.2.1 *$(1+x)^n$ is the generating function for the sequence*

$$\binom{n}{0}, \binom{n}{1}, \binom{n}{2}, \ldots, \binom{n}{n}, 0, 0, \ldots \tag{9.10}$$

of binomial coefficients.

Let $T(n)$ be the number of trees with n vertices. No simple formula for $T(.)$ exists (better: is known), although $T(n)$ is the coefficient of x^n in a desired chosen generating function

$$T(x) = \sum_{n=1}^{\infty} T(n) x^n. \tag{9.11}$$

Otter [184] finds such a generating function for $T(.)$.[1] It shows that the number of non-isomorphic trees at first is very small, but then increases rapidly:

[1] The proof used counting rooted trees, these are trees where one vertex (the root) distinguished from the others. Let $T'(n)$ be the number of rooted trees with n vertices. Then

$$T(x) = T'(x) - \frac{1}{2}(T'^2(x) - T'(x^2)). \tag{9.12}$$

See [121] and our discussion in later sections.

Number n of vertices	Number $T(n)$ of trees
1	1
2	1
3	1
4	2
5	3
6	6
7	11
8	23
9	47
10	106
11	235
12	551
13	1,301
14	3,159
15	7,741
$\vdots$	$\vdots$
23	14,828,074
24	39,299,897
25	$1.0464 \cdot 10^8$
26	$2.7979 \cdot 10^8$

For these numbers and other facts about the counting of trees see Carter et al. [37], Hendy et al. [131] and Riordan [203].

We have an essential difference between the growing of the number of labeled and unlabeled trees. In terms of $p(.)$ in $2^{p(n)}$:

Labeled trees	$p(n) = O(n \log n)$
unlabeled trees	$p(n) = O(n - \log n)$.

In numerical terms:

Number n of vertices	$t(n)$	$T(n)$
10	10^8	106
20	$2.6 \cdot 10^{23}$	823,065
30	$2.3 \cdot 10^{41}$	$1.4 \cdot 10^{10}$

An interesting question: What is the number of non-isomorphic trees with a trivial automorphism group. A partial answer give Harary and Prins [123]:

Number n of vertices	Number of non-isomorphic asymmetric trees
1	1
2	0
3	0
4	0
5	0
6	0
7	1
8	1
9	3
10	6
11	15
12	29

Chapter 10

Digraphs

A digraph or directed graph is a pair $G = (V, E)$ consisting of a finite set V of vertices and a set $E \subseteq V \times V$ of (ordered) pairs of vertices, which we call arcs. Hence, a digraph $G = (V, E)$ is essentially a relation over V. In other terms, enumerating digraphs is counting relations.

10.1 The number of labeled digraphs

Let us consider the problem of counting all labeled digraphs with n vertices. There are the following possibilities for two different vertices v and v':

i) There is no arc.

ii) There is the arc (v, v').

iii) There is the arc (v', v).

iv) There are the arcs (v, v') and (v', v).

A digraph has at most $\binom{n}{2} = n(n-1)/2$ (undirected) pairs between vertices. Hence, when we observe that each of the four possible situations is either present or absent, we have

Theorem 10.1.1 *There are*

$$4^{\binom{n}{2}} = 2^{n^2-n} \tag{10.1}$$

digraphs with n labeled vertices.

Remember that a digraph is called connected if on ignoring all orientations on the arcs, the resulting multigraph is connected. Consequently, connected digraphs can counted in terms of connected graphs.
As an example we consider labeled transitive digraphs, which plays an important role in many application. An explicit formula is not known, but in Harary, Palmer [122] we find the following list for small values n of vertices:

n	number of transitive digraphs
1	1
2	4
3	29
4	355
5	6,942
6	209,527
7	9,535,241
8	642,779,354

For practice the reader should discuss the following observation.

Observation 10.1.2 *The number of labeled transitive digraphs for n vertices equals the number of finite topologies on n labeled points.*

10.2 Relations

Let $\mathcal{U}$ be a universe. A subset of $\mathcal{U}^2$ is called a relation over $\mathcal{U}$. In contrast to digraphs in relations loops, these are arcs (v, v), are allowed. Easy to see

Theorem 10.2.1 *There are*

$$2^{n^2} \tag{10.2}$$

relations over n elements.

I. Several specific kinds of relations play important roles.
A relation R over a universe $\mathcal{U}$ is said to

- reflexive: For all $v \in \mathcal{U}$ it holds that $(v, v) \in R$;
- symmetric: If $(v, v') \in R$ then $(v', v) \in R$;
- antisymmetric: If $(v, v') \in R$ and $(v', v) \in R$ then $v = v'$;
- transitive: For any three elements v, v' and w, if $(v, v') \in R$ and $(v', w) \in R$ then $(v, w) \in R$.

Theorem 10.2.2 *There are*

$$2^{n^2-n} \tag{10.3}$$

reflexive relations over n elements.

Proof. We have n^2 pairs, whereby n be of the form (v, v). Consider the other $n^2 - n$ pairs as we construct a reflexive relation we either include or exclude each these pairs. □

To count the symmetric relations over $\mathcal{U} = \{v_1, \ldots, v_n\}$, we split $\mathcal{U} = X \cup Y$ whereby $X = \{(v_i, v_i) : i = 1, \ldots, n\}$ and $Y = \{(v_i, v_j) : i, j = 1, \ldots, n, i \neq j\}$. Then

$$|Y| = |\mathcal{U} \times \mathcal{U}| - |X| = n^2 - n.$$

The set Y can be divided into $(n^2 - n)/2$ parts of the form $\{(v_i, v_j), (v_j, v_i)\}$.[1] In constructing a symmetric relation we have for the X all choices of exclusion or inclusion a pair and for Y two choices for each part. Then the number of symmetric relations equals

$$2^n \cdot 2^{(n^2-n)/2} = 2^{(n^2+n)/2}.$$

Consequently

Theorem 10.2.3 *There are*

$$2^{(n^2+n)/2} = 2^{\binom{n+1}{2}} \tag{10.4}$$

symmetric relations over n elements.

For practice the reader should prove

Corollary 10.2.4 *There are*

$$2^{(n^2-n)/2} = 2^{\binom{n}{2}} \tag{10.5}$$

relations over n elements that are both reflexive and symmetric.

Note that "not symmetric" is not synonymous with "antisymmetric" (example as exercise).

Theorem 10.2.5 *There are*

$$2^n \cdot 3^{(n^2-n)/2} \tag{10.6}$$

antisymmetric relations over n elements.

Proof. We make two observations as we try to construct an antisymmetric relation over a set $\mathcal{U}$:

i) Each pair $(v, v) \in \mathcal{U} \times \mathcal{U}$ can be either included or excluded with no concern about whether or not the relation is antisymmetric.

ii) For a pair (v, v') with $v \neq v'$, we must consider three alternatives: include (v, v'), include (v', v), or include neither (v, v') nor (v', v) in the relation.

Then there are the number of antisymmetric relations equals $2^n \cdot 3^{\binom{n}{2}}$. □

II. There is no known general formula for the number of transitive relations, but of relations which are simultaneously reflexive, symmetric and transitive, called an equivalence relation if it is reflexive, symmetric and transitive. The following observation reveals that equivalence relations and partitions are intimately related.

[1] Note that $n^2 - n = n(n-1)$ is in any case an even number.

Observation 10.2.6 *There is a one-to-one correspondence between the set of equivalence relations and the collection of partitions.*

This is the first step in discussing classifications, which are hierarchies of partitions, and of great relevance in biology.[2] As an immediate consequence of 10.2.6 we find

Theorem 10.2.7 *There are $B(n)$ equivalence relations over n elements, where $B(n)$ denotes the nth Bell number.*

There is a formula which gives the exact value of $B(.)$, see K.3.2 and (K.7).

$n =$	$B(n) =$
1	1
2	2
3	5
4	15
5	52
6	203
7	877
8	4,140
9	21,147
10	115,975
11	678,570
12	4,213,597
13	27,644,437
14	190,899,322

The number of equivalence relations grows exponentially:

Theorem 10.2.8

$$(n-1)\cdot 2^{n-2} \leq B(n) \leq n!. \tag{10.7}$$

Proof. In view of K.4.2 we have for all $n \geq 1$,

$$B(n) = \sum_{k=0}^{n-1} \binom{n-1}{k} B(k) \geq \sum_{k=0}^{n-1} k \binom{n-1}{k} = (n-1)2^{n-2}.$$

[2] In the book *The System of Nature* Linnaeus introduced a system still in use today. He gave every species two Latinized names; the first for the group it belongs to, the genus; and the second for the particular organism itself. Each group is a partition of the set of all species. For example

group \ species	human	fruit fly
Domain	Eukarya	Eukarya
Kingdom	Animalia	Animalia
Phylum	Chordata	Arthropoda
Class	Mammalia	Insecta
Order	Primata	Diptera
Family	Hominidae	Drosophilidae
Genus	Homo	Drosophila
Species	*sapiens*	*melanogaster*

Paying attention C.2.6.
The second inequality is given in K.4.1 □

As an exercise improve the lower bound to $\frac{1}{2} \cdot (3^n + 1) \leq B(n)$. What kind of graphs the Stirling number of the second kind do count?

III. A relation $\preceq$ in $\mathcal{U}$ is called a partial order if it is reflexive, antisymmetric and transitive. The pair $(\mathcal{U}, \preceq)$ is named a poset.
If, in addition, for every pair $v, v' \in \mathcal{U}$ either (v, v') or (v', v), then the relation is said to be a linear order.[3]
What we are really asking for here is whether we can take the partial order $\preceq$ and find a linear order $\leq$ for which

$$\preceq \subseteq \leq . \tag{10.8}$$

The answer is yes, and uses the following idea, known as topological sorting:

Algorithm 10.2.9 *Given a poset $(\mathcal{U}, \preceq)$. Then a linear order $\leq$ satisfying (10.8) can be found by the following algorithm.*

1. *Find a minimal element;*
2. *Put it on top of anything already in the linear ordering;*
3. *Remove it from the poset;*
4. *Repeat the process until the poset is empty.*

In other terms, each partial order can be embedded in a linear order.
A linear orders for a universe $\mathcal{U}$ is a permutation of the elements. Hence,

Theorem 10.2.10 *The number of linear orders for a universe $\mathcal{U}$ of n elements equals $n!$.*

There is no simple general formula for the number of posets known, but must be faster growing than for linear orders, see [226]. Consequently,

Corollary 10.2.11 *The number of posets increases exponentially in the number of elements.*

10.3 Sperner's theorem

I. Consider the power set of a set X: $\mathcal{P}(X) = \{S : S \subseteq X\}$. in 2.1.1 we saw that This implies that the power set contains more elements than the set X itself.[4]
Our consideration gives us a way to enumerate methodically the subsets, beginning with the empty set $\emptyset$, and then adding each successive element of S to a copy of each of

[3] Sometimes called a total order.
[4] Note, that the power set of the empty set $\emptyset$ is not empty, since it contains the $\emptyset$.

all the previously listed subsets. Let us illustrating it on the subsets of $X = \{a, b, c\}$. We look at the elements one by one, and write down a "1" if the element occurs in the subset and "0" if it does not. Thus each subset corresponds with a binary sequence of length 3. Moreover, such sequences remind us of the binary representation of integers.

subset	binary seqence	integer
$\emptyset$	000	0
$\{c\}$	001	1
$\{b\}$	010	2
$\{b, c\}$	011	3
$\{a\}$	100	4
$\{a, c\}$	101	5
$\{a, b\}$	110	6
$S = \{a, b, c\}$	111	7

In other terms, the subsets correspond to the numbers $0, \ldots, 7.$, and similarly, we count the power set of a set of n elements by $0, \ldots, 2^n - 1$.

II. The subsets of a set X, when they are ordered by set inclusion, form a partial ordered set. A linear order of subsets is called a chain.

Theorem 10.3.1 *Let X be a finite set of n elements. A chain contains at most n members, and there are at most $n!$ chains in $\mathcal{P}(X)$.*

Proof. Consider chains of subsets

$$\emptyset = C_0 \subset C_1 \subset C_2 \subset \ldots \subset C_n = X, \tag{10.9}$$

where $|C_i| = i$ for $i = 0, \ldots, n = |X|$.
There are $n!$ many chains, since we obtain a chain by adding one of the elements of X after the other.[5] □

III. Let $\mathcal{F}$ be a family of subsets of a set X. $\mathcal{F}$ is called an antichain if no member of $\mathcal{F}$ contains another member of $\mathcal{F}$.

Theorem 10.3.2 *(Sperner [225]) Let X be a finite set of n elements. Let $\mathcal{F} \subseteq (\mathcal{P}(X), \subseteq)$ be an antichain. Then*

$$|\mathcal{F}| \leq \binom{n}{\lfloor n/2 \rfloor}. \tag{10.10}$$

Proof. There are many proofs of this theorem; one due to Lubell [165] is probably the most elegant.
In view of 10.3.1, there are $n!$ many chains
Let $\mathcal{F}$ be an arbitrary antichain. For a set $A \in \mathcal{F}$ we ask how many chains does A

[5] Or, in other terms, since a chain is a linear order we use 10.2.10.

contain. To get from $\emptyset$ to A have to add the elements of A one by one, and then to pass from A to X we have to add the remaining elements. Thus if A contains k elements, then by considering all these pairs of chains linked together there are precisely

$$k! \cdot (n-k)! \tag{10.11}$$

such chains.
It is important to note that no chain can pass through two different sets A and A' of $\mathcal{F}$, since $\mathcal{F}$ is an antichain.
Let f_k be the number sets in $\mathcal{F}$ with exactly k elements. The number of chains passing through some member of $\mathcal{F}$ cannot exceed the number $n!$ of all chains. In other terms,

$$\sum_{k=0}^{n} f_k \cdot k! \cdot (n-k)! \leq n!. \tag{10.12}$$

Equivalently,

$$\sum_{k=0}^{n} \frac{f_k}{\binom{n}{k}} \leq 1. \tag{10.13}$$

Paying attention C.2.3, replacing the dominator by the largest binomial coefficient we obtain

$$\frac{1}{\binom{n}{\lfloor n/2 \rfloor}} \sum_{k=0}^{n} f_k \leq 1. \tag{10.14}$$

This complete the proof, since $|\mathcal{F}| = \sum_{k=0}^{n} f_k$. □

In view of Stirling's inequalities I.2.1, we can describe the asymptotic behavior of set families. In particular, doing so gives us

$$\binom{n}{\frac{n}{2}} \approx \sqrt{\frac{2}{\pi}} \frac{2^n}{\sqrt{n}}.$$

Thus the number of subsets of size $n/2$ grows exponentially with n; in fact it is the fraction $\sqrt{2/(\pi n)}$ of the total number of subsets of an n element set. In terms of the exponent $p(n)$ for $2^{p(n)}$:

family	$p(n) =$
power set	n
chain	$O(\log n)$
antichain	$O(n - 0.5 \cdot \log n)$

These numbers suggests an intimately relation between chains and antichains.[6]

[6] And indeed this is true, not only for the power set.

Theorem 10.3.3 *(Dilworth, see [28]) If $(X, \leq)$ is a poset and if the largest antichain(s) have k elements, then there are a family of k chains whose union is X.*

Theorem 10.3.4 *(Erdös, Ko, Rado [79], [5]) The largest size of an intersecting family in which all sets have the same size k is $\binom{n-1}{k-1}$ when $n \geq 2k$*

IV. A similar question, solved by another method: Let $\mathcal{F}$ be a finite family of subsets of a set X.[7] We assume

$$\mathcal{F} = \{X_1, \ldots, X_m\} \tag{10.15}$$

with

$$|X_i| = k \quad \text{and} \quad |X_i \cap X_j| = t$$

for all i and $i \neq j$, and given numbers k and t.
What is the maximal possible number for m?
Consider the characteristic function $x_i = f_{X_i}$ for X_i. The functions x_i can be regarded as essential vectors. We find for the (standard) inner product of such vectors

$$(x_i, x_i) = k \quad \text{and} \quad (x_i, x_j) = t$$

for all i and $i \neq j$.
When we consider the Gram-matrix

$$G = \begin{pmatrix} (x_1,x_1) & (x_1,x_2) & \ldots & (x_1,x_m) \\ (x_2,x_1) & (x_2,x_2) & \ldots & (x_2,x_m) \\ \ldots & \ldots & & \ldots \\ (x_m,x_1) & (x_m,x_2) & \ldots & (x_m,x_m) \end{pmatrix} = \begin{pmatrix} k & t & \ldots & t \\ t & k & \ldots & t \\ \ldots & \ldots & & \ldots \\ t & t & \ldots & k \end{pmatrix}. \tag{10.16}$$

For $t < k$ it is easy to see that $\det G \neq 0$. Consequently, $x_1, \ldots, x_m$ are linearly independent. In particular, $m \leq |X|$. Hence,

Theorem 10.3.5 *Let $\mathcal{F}$ be a finite family of k-element subsets of a set X with*

$$|Y \cap Z| = t < \min\{k, |X|\} \tag{10.17}$$

for some integer t. Then

$$|\mathcal{F}| \leq |X|. \tag{10.18}$$

For a broader description of Sperner's theory see Engel, Gronau [75].

10.4 Tournaments

A tournament G is a digraph in which for every two distinct vertices v and v', either (v, v') or (v', v) is an arc of G. In other terms, a tournament is a digraph obtained when directions are assigned to each edge of a complete graph. In particular, for any vertex v in a tournament with n vertices

$$g^{in}(v) + g^{out}(v) = n - 1. \tag{10.19}$$

[7] A pair $(X, \mathcal{F})$ is sometimes called a hypergraph. The "hyperedges" are elements of $\mathcal{F}$ and in general not 2-elementary. Two vertices X_i and X_j are adjacent if there exists an edge $e \in \mathcal{F}$ such that $X_i, X_j \in e$.

Theorem 10.4.1 *There are* $2^{\binom{n}{2}}$ *tournaments with* n *vertices.*

For the *proof* note that there are two choices for which way to put a direction of an edge. □

We observe that the number of tournaments is the precisely the number of all (labeled) graphs, see 5.1.1(b). The correspondence between these two classes of graphs is indicated by the order in which they appear: Each tournament corresponds to that labeled graph in which the vertices with labels i and j are adjacent if and only if $i < j$ and the arc from i to j is present in the tournament.

A tournament is a model for a game in which every two players engage in a match; and a match cannot end in a tie.

Theorem 10.4.2 *Let* v *be a vertex of maximum outdegree in a tournament* G*. Then the minimum number of arcs from* v *to any other vertex in* G *equals 1 or 2.*

Proof. Assume that $G = (V, E)$ has n vertices, $g^{out}(v) = k$, and that v is adjacent to the vertices $v_1, \ldots, v_k$. Then

$$\begin{aligned} (v, v_i) \in E \quad &\text{for} \quad i = 1, \ldots, k; \quad \text{and} \\ (u_j, v) \in E \quad &\text{for} \quad j = 1, \ldots, n-k-1. \end{aligned}$$

We show that there is a directed path from v to each of the u_j with two arcs.
If each vertex u_j, $j = 1, \ldots, n-k-1$, is adjacent from some vertex v_i, then this creates such path. Otherwise, suppose now that some vertex u_k is adjacent from no vertex v_i. Then u_k is adjacent to each v_i. Moreover, u_k is also adjacent to v. Hence,

$$g^{out}(u_k) = k + 1 > g^{out}(v),$$

which is a contradiction. □

10.4.2 gives the following interpretation for a game: We define a winner w as a player with the most victories.[8] The theorem says that a winner w has been defeated only by players that were themselve defeated by a player which had been beaten by w.

Let $G = (V, E)$ be a digraph. A Hamiltonian cycle (path) is a directed cycle (path) that contains all vertices of G. The problem is to decide whether or not G has a Hamilton cycle; if so then G is called a Hamiltonian digraph.

Theorem 10.4.3 *Every tournament has a Hamiltonian path.*

Proof. Assume on the contrary, that $G = (V, E)$ is a tournament that does not have a Hamiltonian path. Then $|V| > 3$. Let $P : v_1, \ldots, v_n$ be a longest directed path in G. Since P is not a Hamiltonian path there is a vertex v of G that does not belong

[8] Maybe there is more than one winner.

to P.
If $(v, v_1) \in E$ then $v, v_1, \ldots, v_n$ is a path whose length exceeds that of P, which is a contradiction to the choice of P. Hence, $(v_1, v) \in E$; similarly, $(v, v_n) \in E$.
Since G is a tournament we know that for any vertex v_i, $1 \leq i \leq n$ either $(v, v_i) \in E$ or $(v_i, v) \in E$. We already know that $(v_1, v), (v, v_n) \in E$, such that there must be an integer j, $1 \leq j < n$ such that $(v_j, v), (v, v_{j+1}) \in E$. However,

$$P' : v_1, \ldots, v_j, v, v_{j+1}, \ldots, v_n$$

is a path whose length exceeds that of P. Since we cannot avoid this contradiction, the theorem is proved. □

For practice the reader should discuss 10.4.3 in terms of games. Does this path run from a winner to a loser?

Corollary 10.4.4 *Every transitive tournament contains exactly one Hamiltonian path.*

Consider a tournament with n vertices. A Hamiltonian path has exactly $n-1$ arcs. Hence, the number of tournaments containing a given Hamiltonian path is $2^{\binom{n}{2}-n+1}$. On the other hand, the number of Hamiltonian paths possible in such a tournament is $n!$. With each occurring in $2^{\binom{n}{2}-n+1}$ tournaments, the average number per tournament is

$$\frac{2^{\binom{n}{2}-n+1} \cdot n!}{2^{\binom{n}{2}}} = 2^{-n+1} \cdot n!.$$

Hence,

Theorem 10.4.5 *At least one tournament on n vertices contains at least*

$$\frac{n!}{2^{n-1}} \geq 2e \left(\frac{n}{2e}\right)^n \tag{10.20}$$

Hamiltonian paths.

Contrary, while 10.4.3 shows that each tournament contains a Hamiltonian path, and paying attention 10.4.5 many more, not every tournament is Hamiltonian, that means not every tournament has a Hamiltonian cycle. Of course,

Observation 10.4.6 *Every Hamiltonian tournament is strongly connected.*

The converse statement is also true. We give an even stronger result, but without a proof. One can find it in [44].

Theorem 10.4.7 *Every vertex of a strongly connected tournament with n vertices is contained in a cycle of length k for every k with $3 \leq k \leq n$.*

10.5 The shortest superstring problem

The problem of Hamiltonian paths is strongly connected with the shortest common superstring problem (SCS):
Given: A set of strings (= sequences = words) over the same alphabet.
Find: A shortest sequence that contains each of the given sequences as a subsequence.
It plays a very important role in DNA sequencing. In fact, in DNA sequencing is routinely done by sequencing large numbers of relatively short fragments and then finding a SCS. Compare [31].

Let $S = \{w_1, \ldots, w_n\} \subseteq A^\star$ be a set of strings over the alphabet A.
Throughout the discussion we assume that no string in S is a substring of any other string in S. Any such substring can be efficiently detected (How?) and consequently removed. After that the problem has the same solution as before, which means that a SCS for the remaining strings is also an SCS of the original set.
For two strings $w, w' \in A^\star$ we define the string

$$\text{Merge}(w, w') = xyz, \tag{10.21}$$

where y is a suffix of w; y is a prefix of w'; and $|y|$ is maximal. The string y is called the overlap of w and w', and written by $y = \text{overlap}(w, w')$.
$\text{prefix}(w, w')$ is the prefix of w ending at the start of the overlap. Of course,

$$|\text{prefix}(w, w')| = |w| - |\text{overlap}(w, w')| \tag{10.22}$$

Note that the function $|\text{prefix}(.,.)|$ satisfies the triangle inequality, but it is not a metric, since the symmetry fails.

Let π be a permutation of $\{1, \ldots, n\}$, then

$$w[\pi] = \text{prefix}(w_{\pi(1)}, w_{\pi(2)})\text{prefix}(w_{\pi(2)}, w_{\pi(3)}) \ldots \text{prefix}(w_{\pi(n-1)}, w_{\pi(n)})w_{\pi(n)}. \tag{10.23}$$

is the concatenation of the non-overlapping prefixes of the pairs of adjacent strings, followed by the full string of the last index.

Theorem 10.5.1 *For a set $S = \{w_1, \ldots, w_n\} \subseteq A^\star$ of strings and the permutation π the string $w[\pi]$ is a superstring of S with length*

$$|w[\pi]| = \sum_{i=1}^{n-1} |\mathit{prefix}(w_{\pi(i-1)}, w_{\pi(i)})| + |w_{\pi(n)}|. \tag{10.24}$$

That is, looking for an SCS is the search for a permutation π such that $|w[\pi]| = \min$.
For S we define the distance-graph $G = (V, E)$ by the following definitions:

a) $V = S$;

b) $E = V \times V$, this means that G includes loops;

c) There is a length-function $c : E \to I\!N$ with

$$c(w, w') = \begin{cases} |\text{prefix}(w, w')| & : \quad w \neq w' \\ |w| & : \quad w = w' \end{cases}$$

Then, looking for an SCS is the search for a minimal tour through the distance graph.

In view of the fact that the shortest superstring problem is $\mathcal{NP}$-complete, [241], we are interested in an approximation strategy. We use a greedy strategy: In the overlap-graph we sequentially chose the longest edge that does not form a cycle with already chosen edges.
For S we define the overlap-graph $G = (V, E)$ by the following definitions:

a) $V = S$;

b) $E = V \times V \setminus \{(w, w) : w \in V\}$;

c) There is a length-function $c : E \to I\!N$ with

$$c(w, w') = |\text{overlap}(w, w')|$$

for $w \neq w'$.

(10.22) says that we can derive the overlap-graph from the distance-graph.

Algorithm 10.5.2 *Let $S = \{w_1, \ldots, w_n\} \subseteq A^\star$ be a set of strings.*
While $|S| > 1$ do

1. *Find $w_i, w_j \in S$, $i \neq j$, such that*

$$|overlap(w_i, w_j)| = \max\{|overlap(w, w')| : w, w' \in S\};$$

2. *$w = Merge(w_i, w_j)$;*
3. *$S := S \setminus \{w_i, w_j\} \cup \{w\}$.*

The remaining string is the searched superstring.

10.6 The number of unlabeled digraphs

0. Two digraphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are said to be isomorphic if there exists a one-to-one, onto mapping $f : V_1 \to V_2$ such that $(v, v') \in E_1$ if and only if $(f(v), f(v')) \in E_2$. Of course,

Observation 10.6.1 *Isomorphism is an equivalence relation on the family of all digraphs.*

I. Similar to 7.2.1 the number of unlabeled digraphs increases at least exponentially. Numerical better, the following table due to Oberschelp [182] gives the number of digraphs with n vertices.[9]

Number n of vertices	Number of digraphs	number of connected digraphs
1	1	1
2	3	2
3	16	13
4	218	199
5	9,608	9,364
6	1,540,944	1,530,843
7	882,033,440	880,471,142
8	1,793,359,192,848	1,792,473,955,306

II. In 10.4.1 we find a simple formula for the number of labeled tournaments. Of course, counting unlabeld tournaments are not so easy, [122]:

Number n of vertices	Number of tournaments
1	1
2	1
3	2
4	4
5	12
6	56
7	456
8	6,880
9	191,536
$\vdots$	$\vdots$
12	$1.54108 \cdot 10^{11}$

III. For practice the reader should find the number of (unlabeled) posets for small values of elements:

Number n of elements	Number of posets
1	1
2	2
3	5
4	16
5	?

More information in [226].

[9] Note that in the underlying graph multiple edges are allowed, since (v, v') and (v', v) can be arcs.

Chapter 11

Phylogenetic Networks

The underlying principle of phylogeny is to try to group "living entities" according to their level of similarity. In biology for example, such trees ("phylogenies") typically represent the evolutionary history of a collection of extant species or the line of descent of some gene. No two members of a species are exactly the same - each member has slight modifications from its parents. As environmental conditions change, nature will favour that branch of a species with some particular modification; as time goes on another mutation of the basic stock will become dominant. In this way, all species are continually evolving. This evolution occurs in a number of ways at the same time: some species die out and some become new species in their own right. This forces a "tree-like" structure. This was already seen by Darwin [63].[1]

The present book is devoted to the question: How fast does the number of possible phylogenetic trees grow as function of the size of the given number of taxa? We will give partial answers.

11.1 Phylogenetic trees

Phylogenetic trees summarize the history of life according the theory of evolution.

I. In a phylogenetic tree each vertex describes a species and the leaves represent the species that exist today. In particular, in phylogenetics we search for a tree interconnecting a set N of "living entities" (species, genes, sequences, words - roughly speaking: **N**ames[2]). Such a partially labeled tree (semi-labeled tree) is usually called an N-tree, which means:

- The tree has exactly $|N|$ leaves, each labeled by a different element of N;

[1] Darwin wrote it under "I think" in his notebook, see Engels [76]. A picture of this paper can be found in the book *Phylogenetics* [220].

[2] which explains the term "phylogenetic"

- All internal vertices are unlabeled;
- The degree of each internal vertex is at least 3;
- Sometimes we accept an exception, namely that exactly one internal vertex is marked, and is permitted to have degree 2. Then this vertex is called the root of the tree, and such a tree is called a rooted N-tree.

In view of the application of N-trees in phylogeny, these trees are also usually called phylogenetic trees, namely that in biology, phylogenetic trees are used as "evolutionary" trees or cladograms. In other terms: Starting with a set of known present-day objects a phylogenetic tree may be constructed by first assigning each object a leaf of the tree and then assigning ancestral and unknown objects to the internal nodes.

Furthermore, if N is a set of objects to be classified, then an N-tree can also be viewed as a type of hierarchical classification of N. Here the central idea is that the classification of organisms reflects its evolutionary history.[3]
A classification is the formal naming of a group of individuals. In the sense of set theory a classification $\mathcal{C}$ of a set N of individuals is given by a collection of subsets of N satisfying

- $\emptyset \notin \mathcal{C}$;
- $N \in \mathcal{C}$;
- $\{v\} \in \mathcal{C}$ for any $v \in N$; and
- For any two members N' and $N"$ of $\mathcal{C}$ it holds that

$$N' \cap N" \in \{N', N", \emptyset\}. \tag{11.1}$$

In other words, any two sets in $\mathcal{C}$ are disjoint or one is contained in the other.

A member of a classification is called a class or a cluster of N.
Extending 10.2.6,

[3]In the widest sense, a classification scheme may represent simply a convenient method for organizing a large set of data so that the retrieval of information may be made more efficiently. In this sense, classification is the beginning of all science. Consequently, phylogenetic analysis does not only be a question in biology. As an example in linguistics we give a very partial representation of branches of Indo-European language family.

Indo-European	Germanic	German English Danish
	Slavic	Russian Polish
	Indo-Iranian	Persian Hindi

For a classification of languages see Comrie et al. [57].

Observation 11.1.1 *(Hendy et al. [131]) There is a one-to-one correspondence between the collection of classifications for a set N and the collection of (rooted) N-trees.*

In other words, classifications for a set N and rooted N-trees contain essentially the same information: The following statements are pairwise equivalent.

- $\mathcal{C}$ is a classification for N.
- $\mathcal{C}$ represents a (rooted) N-tree.
- $\mathcal{C}$ consists of a series of partitions for N which become finer and finer.

For example a first classification of the life on earth is given by

$$\begin{aligned} \text{organisms} \quad = \quad & \{\{\text{prokaryota}\} = \{\text{archea}\} \cup \{\text{bacteria}\}, \\ & \{\text{eukaryota}\} = \{\text{protista}\} \cup \{\text{plantae}\} \cup \{\text{fungi}\} \cup \\ & \{\text{animalia}\}\}. \end{aligned}$$

Roughly speaking, for the description of classifications, and related trees, we have the following relationship:

Level		
In taxonomy	OTU = operational taxonomic unit	HTU = hypothetical taxonomic unit
Species/genes	extant	extinct
Placement in time	existing unit	ancestor
Classification	individuals	class
Vertex in the tree	leaf	internal vertex

For more facts see Page und Holmes [185]. An elementary introduction is given by Cieslik [54].[4] A broad discussion on the basis of molecular evolution we find by Graur, Li [107]. A very short introduction by Dossing et al. [69].

II. From now on we will investigate different classes of trees in parallel:

- Graphs and digraphs.
- Rooted and unrooted trees.
- Labeled, unlabeled and semi-labeled graphs and digraphs.

[4] In a first view, it seems impossible to describe the "Great Darwin Tree" since the diversity of the living world is staggering: more than two million existing species of plants and animals have been named and described; many more - both existing and past - remain to be discovered. But tolweb.org attempts to describe the "Tree of Life".
Biodiversity is twofold:

- The presence of numerous species on Earth; and
- The polymorphism within each species.

For using evolutionary history for describing the biodiversity see [99] or [163].

A helpful description of binary trees with labeled leaves is given by the following procedure: Let $T = (V, E)$ be a binary N-tree for $N = \{v_1, \ldots, v_n\}$.

1. If $n = 2$, then write T as $(v_1 v_2)$; otherwise,

2. Let v_i and v_j be two leaves of T which are adjacent to the same vertex v. Then
   - Delete the leaves v_i and v_j, and its incident edges;
   - Replace the vertex v by $(v_i v_j)$, which is now a leaf;
   - Consider the new tree with $n-1$ leaves and repeat the procedure.

Clearly, this procedure gives a simple written description of the tree, called the "bracket" or Newick format. But note that it is not unique, for example for the one N-tree for $n = 3$ we have the descriptions $((v_1 v_2)v_3)$ and $(v_1(v_2 v_3))$ and $((v_1 v_3)v_2)$.

11.2 Semi-labeled trees

A tree in which each vertex has degree one or three is called a binary tree.[5] Such trees play an important role in the theory of evolution, since it is assumed that a phylogenetic tree is a "bifurcation" tree. This follows from the assumption that evolution is driven by bifurcation events.[6]
To count the number more precisely, we start with the following observation, which is easy to verify.

Observation 11.2.1 *A binary tree with n leaves has exactly $n-2$ internal vertices. In particular, the tree has an even number of vertices, namely $2n-2$, and $2n-3$ edges.*

With this in mind we have the following result.

Theorem 11.2.2 *The following formulas are true for the number of binary trees:*

a) *The number of binary trees with n labeled leaves and $n-2$ labeled internal vertices is*

$$\frac{(2n-4)!}{2^{n-2}}. \tag{11.2}$$

b) *(Cavalli-Sforza, Edwards [38]) The number of binary trees with n labeled leaves and $n-2$ unlabeled internal vertices (i.e. binary N-trees having $|N| = n$) is*

$$(2n-5)!! = 1 \cdot 3 \cdot 5 \cdots (2n-5) = \Omega\left(\left(\frac{2n}{3}\right)^{n-2}\right). \tag{11.3}$$

[5] This definition is a little bit strange, but later we will define an order on such trees, and our condition means that each internal vertex has exactly two "children".

[6] In practice phylogenetic trees are allowed to be multifurcating when the bifurcations are sufficiently close together or the exact order of two or more bifurcations cannot be determined unambiguously with the available data.

Proof. Any tree has exactly n leaves and, in view of 11.2.1, exactly $n-2$ vertices of degree 3. Then 6.2.2 counts the number of trees by

$$\binom{(2n-2)-2}{\underbrace{(1-1)\dots(1-1)}_{n-\text{times}}\underbrace{(3-1)\dots(3-1)}_{(n-2)-\text{times}}} = \frac{(2n-4)!}{2^{n-2}}.$$

If the $n-2$ internal vertices are unlabeled, then this number must be divided by $(n-2)!$, thus

$$\begin{aligned}\frac{(2n-4)!}{2^{n-2}(n-2)!} &= \frac{(2n-4)\cdot(2n-5)\cdot(2n-6)\cdot(2n-7)\cdots 4\cdot 3\cdot 2\cdot 1}{2(n-2)\cdot 2(n-3)\cdots 2\cdot 2\cdot 2\cdot 1}\\ &= (2n-5)(2n-7)(2n-9)\cdots 5\cdot 3\cdot 1\\ &= (2n-5)!!.\end{aligned}$$

□

Independently, there is a much simpler proof of (b) by induction, which also gives a procedure to create all such trees: If $n=3$, then there is clearly one such tree. By adding a new leaf to any of the edges in this tree we obtain all three possible bifurcating trees on 4 leaves. In general, we can obtain every tree by adjoining a new leaf and edge to some tree T' on the first $n-1$ leaves, in exactly one way. T' has $2(n-1)-3 = 2n-5$ edges. It follows 11.2.2(b) in another form:

Theorem 11.2.3 *There are* $(2n-5)!!$ *different bifurcating phylogenetic trees on* n *taxa.*

The function $(2n-5)!!$ grows very rapidly with n. That means, the number of possible trees increases rather dramatically as the number of taxa increases. In the following tabular, the result is applied to phylogenetic trees, Hall [119]:

Number of taxa	Number of binary trees	Comment
3	1	
4	3	
5	15	
6	105	
7	945	
8	10,395	
9	135,135	
10	2,027,025	
11	34,459,425	
12	654,729,075	
22	$\approx 3\cdot 10^{23}$	Almost a mole of trees
50	$\approx 3\cdot 10^{74}$	More trees than the number of atoms in the universe
100	$\approx 2\cdot 10^{182}$	out of any range

11.3 Multi-stars

Specific "topologies" of trees play an important role in the phylogenetic analysis.

I. A star is defined as a tree in which exact one internal vertex.[7] Of course, there are n labeled stars with n vertices and exactly one unlabeled star.

II. A double-star is a tree with exact two internal vertices. Easy to see.

Observation 11.3.1 *The both internal vertices of a double-star must be adjacent.*

In other terms, a double tree with n vertices has $n-2$ leaves where $k \geq 1$ adjacent to one and $l = n-2-k \geq 1$ adjacent to the other internal vertex.
Let v and v' be the both centers. The internal edge $\underline{vv'}$ divides the set of all $n-2$ leaves in two disjoint sets. In view of 12.1.1 there are $2^{n-3}-1$ of such splits.
This is also true vice versa: Assume that v and v' are unlabeled internal vertices. We can select $\binom{n}{2}$ such pairs, and get two possible correspondences.
If all vertices are unlabeled we count all bipartitions of the integer $n-2$, see L.2.1.

Theorem 11.3.2 *Consider double-stars.*

a) *There are*

$$\binom{n}{2} \cdot (2^{n-2}-2) \tag{11.4}$$

double-stars for n labeled vertices.[8]

b) *There are*

$$2^{n-3}-1 \tag{11.5}$$

double-stars for $n-2$ labeled leaves and 2 unlabeled internal vertices.

c) *There are*

$$\begin{cases} \frac{n-2}{2} & : \quad \textit{if } n \textit{ even} \\ \frac{n-3}{2} & : \quad \textit{otherwise} \end{cases} \tag{11.6}$$

double-stars with n unlabeled vertices.

Now we distinguish double trees for given numbers $n \geq 4$, k and l. In view of 11.3.1 there is exactly one such tree in the unlabeled case.

$n=$	$(k,l)=$	Number of labeled trees	Number of partially labeled trees
4	(1,1)	12	1
5	(1,2)	60	3
6	(1,3)	(Exercise)	4
6	(2,2)	(Exercise)	3
7	(1,4)		4
7	(2,3)		11

[7] Or equivalently as $K_{1,n-1}$.
[8] Remember that we found this result already in (6.15).

Very important in phylogeny are the partially labeled double stars with four leaves splited in $k = l = 2$, namely ((12)(34)), ((13)(24)) and ((14)(23)).
The so-called quartet puzzling method, originated by Strimmer, von Haeseler [231], use these trees as input, then combine the quartet trees into an N-tree, which tries to respect to the neighbor relation of all quartet trees. Repeat these steps many times and output the majority consensus tree.

III. It would be an interest topic for further investigations to discuss these questions for multi-stars, which are graphs (!) with the following properties: The internal vertices induce a complete graph and each internal vertex is adjacent to at least one leaf. Considering n vertices and k leaves, the number $n - k$ of internal vertices must be satisfies $n - k \leq k$. Hence,

$$n - 1 \geq k \geq \left\lceil \frac{n}{2} \right\rceil . \tag{11.7}$$

Theorem 11.3.3 *Let $t^{(m)}(n, k)$ be the number of multi-stars with n vertices, which k are leaves, satisfying (11.7). Then*

$$t^{(m)}(n, k) = \frac{n!}{k!} \cdot S(k, n - k), \tag{11.8}$$

where $S(.,.)$ denotes the Stirling number of the second kind.

Proof. To count $t^{(m)}$

(i) We have to choose k vertices as leaves. This can be done in $\binom{n}{k}$ ways.

(ii) The set of all leaves has to partition into $n - k$ parts. This we count by the Stirling numbers of the second kind in $S(k, n - k)$ ways.

(iii) We have to pair off each internal vertex with one of the parts. This can be done in $(n - k)!$ ways.

Altogether we obtain

$$\begin{aligned} t^{(m)}(n, k) &= \binom{n}{k} \cdot S(k, n - k) \cdot (n - k)! \\ &= \frac{n!}{k!(n - k)!} \cdot S(k, n - k) \cdot (n - k)! \\ &= \frac{n!}{k!} \cdot S(k, n - k). \end{aligned}$$

□

We write the numbers $t^{(m)}(n, k)$ as a triangle:

$n \setminus k$	2	3	4	5	6	7	sum
3	3						3
4	12	4					16
5		60	5				65
6		120	210	6			336
7			1260	630	7		1897
8			1680	8400	1736	8	11824

Combining 11.3.3 and 6.2.3 we find

Theorem 11.3.4 *Let $t_k(n)$ be the number of trees with n (labeled) vertices, which k are leaves, satisfying (11.7). Then*

$$t_k(n) \geq \frac{n!}{k!} \cdot S(k, n-k) \cdot (n-k)^{n-k-2}, \tag{11.9}$$

where $S(.,.)$ denotes the Stirling number of the second kind.

(Why does not equality hold?)

More generally, a multi-star $K^{(m)}(g_1, \ldots, g_m)$ is a graph such that the m vertices form a K_m, and additionally each vertex is adjacent with $g_i \geq 1$ pendants[9], $i = 1, \ldots, m$. Easy to see, $K^{(m)}(g_1, \ldots, g_m)$ has $m + \sum_{i=1}^{m} g_i$ vertices, and $\binom{m}{2} + \sum_{i=1}^{m} g_i$ edges. And for counting we have to use the knowledge about the partition of integers. See [180].

IV. Another generalization are the caterpillars. These are trees with the property that after removing all leaves and its incident edges there is only a path. The number of caterpillars are determined by Lifschitz [162]. In particular

Remark 11.3.5 *(Harary, Schwenk [124]) The number of unlabeled caterpillars with n vertices is*

$$2^{n-4} + 2^{\lfloor (n-4)/2 \rfloor}. \tag{11.10}$$

11.4 The structure of rooted trees

The most important point in a phylogenetic tree is its root. In a rooted tree exactly one distinguished vertex is marked as the root. This tree has a natural orientation from ancestors to descendants, and the root as the common ancestor of the leaves. In this sense, a universal ancestor must exists.[10]

I. Rooted trees are representations for evolutionary relationships: For a rooted N-tree T we view the edges as being directed away from the root, and then regard T

[9] These are vertices of degree 1, in trees called leaves.

[10] Unrooted trees neither make assumptions nor require knowledge about common ancestors.

as describing the evolution of the set N of given "names" from a common (hypothetical) ancestral name; the other internal vertices of T correspond to further ancestral names. The distinction between rooted and unrooted trees is important, because many methods for reconstructing phylogenetic trees generated unrooted ones[11], and we need more information for rooting an unrooted tree.[12]

The root is placed at this position to indicate that

- it corresponds to the (theoretical) last universal common ancestor of everything in the tree;
- gives directionality to evolution within the tree; and
- it identifies which groups of vertices are "true", given that the root does not lie within a group.

The question is: On which edge should the root be placed? There are three popular ways to find this position:

1. On the longest edge[13].
2. In the middle of the longest path between two leaves.
3. An "outgroup" can be added to the set of given points. Then the root is placed at the bifurcation between outgroup and the main group.

II. Remember that in a tree any two vertices are connected by exactly one path. Hence, a unique path leads from the root to any other vertex of a rooted tree. Let w be the root and v be an arbitrary vertex in a rooted tree $T = (V, E)$. The length of the path from w to v is called the level of v:

$$\text{level}(v) = \rho(w, v). \tag{11.11}$$

The depth of the tree itself is defined by

$$\text{depth}(T) = \max\{\text{level}(v) : v \in V\}. \tag{11.12}$$

If T is a rooted tree, then it is customary to draw T with root w at the top, at level 0. The vertices adjacent to w are placed on level 1. Any vertex adjacent to a vertex of level 1 is at level 2, and so on. In general, every vertex at level $i > 0$ is adjacent to exactly one vertex at level $i - 1$. The definition of the depth gives

Observation 11.4.1 *Let T be a rooted tree. Then*

$$\mathit{diam}(T) \leq 2 \cdot \mathit{depth}(T). \tag{11.13}$$

[11] Such trees are also biologically relevant since they are typically what tree reconstruction methods generate.

[12] Rooting a tree has a strong relationship to the molecular clock; but especially, proteins evolve at different rates, making it difficult to relate the (evolutionary) distance to the historical time.

[13] This approach of course requires that there is a length-function for the graph.

III. We may consider a rooted tree $T = (V, E)$ as a digraph if we direct the edges $\underline{vv'} \in E$ from v to v' if and only if $\text{level}(v') = \text{level}(v) + 1$. Then $g^{in}(w) = 0$ characterizes the root, and $g^{out}(v) = 0$ characterizes the leaves of T.
In this sense we have an ancestor/successor-relation for the vertices of a rooted tree. In particular, the root is the common ancestor of all vertices of the tree. In other words, a rooted tree has a vertex identified as the root from which ultimately all other vertices descend.
For a rooted tree $T = (V, E)$ a natural partial order $\leq_T$ on the set V of vertices is obtained by setting $v \leq_T v'$ if

- The path from the root of T to v' includes v; or, equivalently,
- v' is the successor of v, and v is the ancestor of v'.

Observation 11.4.2 *Let $T = (V, E)$ be a rooted tree, and $v, v' \in V$. Then $v \leq_T v'$ implies level$(v) \leq$ level(v'), but not vice versa.*

IV. Let $T = (V, E)$ be a rooted N-tree and let N' be a subset of N. We will refer to the unique vertex v of T that is the greatest lower bound of N' under the order $\leq_T$ as the last universal common ancestor (LUCA) of N' in T. That means

1. v is an ancestor for each vertex in N'; and
2. $\text{level}(v) = \max\{\text{level}(v') : v' \text{ ancestor for each member in } N'\}$.

It is extremely simple to see, but has deep consequences in biology,[14] that the following fact holds true.

Observation 11.4.3 *Let $T = (V, E)$ be a rooted N-tree, and $N' \subseteq N$. Then LUCA for N' in T exists.*

This is a central tenet of modern evolutionary biology: All "living things" trace back to a single common ancestor.[15,16]
Humans and other mammals are descended from shrew-like creatures that lived more than 150 Mya (million years ago); mammals, birds, reptiles and fish share as ancestors

[14] And in its application in medicine, for instance when we discussed which animal is the "mixing vessel" for a specific virus disease, see [113].

[15] It has become clear that the basic metabolic processes of all living cells are very similar. A number of identical mechanisms, structures, and metabolic pathways are found in all living entities so far observed. In particular

- All cells utilize phosphates, particularly adenosine triphosphate (ATP), for energy transfer.
- The metabolic reactions are catalyzed largely by proteins.
- Proteins are manufactured in the cell by a complete coding process. The sequence of amino acids in each protein is determined by the sequence of nucleotides in its gene, "written" as a DNA.
- The universal genetic code.

[16] Note that this proposition does not assert that life arose just once, but that all starting points except one became extinct.

aquatic worms that lived 600 Mya; all plants and animals are derived from bacteria-like organisms that originated more than 3000 Mya. If we go back far enough, humans, frogs, bacteria and slime moulds share a common ancestor.[17]
Finding the LUCA for a set of species, or a set of populations, or a collection of genes is a very difficult task.[18] To find LUCA for all species is discussed in [68], [187], and [250]. Eigen [73] found that the LUCA for genes is an RNA-molecule with the following properties: 3.5 - 4 Gya and 76 bp. A complete discussion of this subject is given by [190].[19]

11.5 The number of rooted trees

We are interested to count rooted trees, and, in view of 11.1.1 simultaneously the number of classifications.

Remember that in order to find the number of trees it was necessary to investigate rooted trees. Let $T_r(n)$ be the number of (without the root unlabeled) rooted trees with n vertices. No simple formula for $T_r(.)$ is known, but $T(n)$ is the coefficient of x^n in a desired chosen generating function

$$T'(x) = \sum_{n=1}^{\infty} T_r(n)x^n. \tag{11.14}$$

Halder and Heise [118] prove

Theorem 11.5.1 *The generating function T' satisfies*

$$T'(x) = x \cdot exp\left(\sum_{n=1}^{\infty} \frac{T'(x^n)}{n}\right). \tag{11.15}$$

[17] In 11.4.1 we see that for two existing species with a common ancestor in d Mya, the "evolution distance" is $2 \cdot d$ Mya.

[18] In particular, find this entity for all humans!
A centaur is not a common ancestor of human and horse.

[19] Darwin claimed that the African apes are mans closest relatives, and suggested that evolutionary origins of man were to be found in Africa. In other words, the commonly held view was that humans were phylogenetically distinct from the great apes (chimpanzees, gorillas and orang-utans), being placed in different taxonomic families, and that this split occurred at least 15 Mya. These conclusions were based on fossils.
In 1967, Sarich and Wilson [214] measured the extent of immunological cross-reaction in the protein serum albumin between various primates. The results were striking: humans, chimpanzees and gorillas were genetically equidistant and clearly distinct from the orang-utan.
The breakthrough for understanding came with a publication in *Nature* in 1987 [36] by Wilson and two of his students, Cann and Stoneking, entitled "Mitochondrial DNA and human evolution". They used mother-only genes, known technically as mitochondrial DNA. Wilson and his colleagues examined the mother-only genes in 134 individuals from around the world. They found remarkable similarities as well as differences in all the samples. The centrepiece of the article was a diagram which bears a superficial resemblance to a tree. It contains a hypothetical common female ancestor of all extant humans, called Eve, or in more scientific terms Mitochondrial Eve.
For more facts about this question compare [13], [16] and [185].

Consequently, in numerical terms, see [203]:

Number n of vertices	Number $T_r(n)$ of rooted trees
1	1
2	1
3	2
4	4
5	9
6	20
7	48
8	115
9	286
10	719
11	1,842
12	4,766
13	12,486
14	32,973
15	87,811
$\vdots$	$\vdots$
23	268,282,855
24	743,774,984
25	$2.0671 \cdot 10^9$
26	$5.7596 \cdot 10^9$

Another approach to count rooted trees is given by the following considerations. Let $T_r(n, g)$ be the number of rooted trees in which the root has degree g.

Theorem 11.5.2

$$T_r(n) = \sum_{g=1}^{\infty} T_r(n, g) = \sum_{g=1}^{n-1} T_r(n, g), \tag{11.16}$$

and

$$T_r(n, 1) = T_r(n - 1). \tag{11.17}$$

$n \setminus g$	1	2	3	4	5
2	1				
3	1	1			
4	2	1	1		
5	4	3	1	1	
6	9	6	3	1	1

Consider a rooted tree with the root w, whereby $g(w) = g$ and

$$1 \leq g \leq n - 1. \tag{11.18}$$

Then $T - w$ is a forest, where we may assume that each tree is itself a rooted tree. g_i denotes the number of such roots which represents a rooted tree of i vertices. The sum of the degrees of the roots of each tree equals g, and furthermore the sum of vertices must be $n - 1$. Hence,

$$g_1 + \ldots + g_{n-1} = g \tag{11.19}$$

and

$$g_1 + 2 \cdot g_2 + \ldots + (n-1) \cdot g_{n-1} = n - 1. \tag{11.20}$$

For example, for $n = 5$ we have $g = 1, \ldots, 4$ and $(g_1 m \ldots, g_4) = (0, 0, 0, 1)$ or $= (1, 0, 1, 0)$ or $= (0, 2, 0, 0)$ or $= (2, 1, 0, 0)$ or $= (4, 0, 0, 0)$.
To compute $T_r(n, g)$ we have to determine all possibilities for $T_r(i)$.

Theorem 11.5.3 *(Flachsmeyer [87]) The number of rooted trees can be recursively computed by*

$$T_r(n) = \sum \binom{T_r(1) + g_1 - 1}{g_1} \cdots \binom{T_r(n-1) + g_{n-1} - 1}{g_{n-1}}, \tag{11.21}$$

where the sum runs over all integers which satisfy (11.18), (11.19) and (11.20).

11.6 The number of rooted binary trees

A tree T is called a rooted binary tree if for its vertices

$$g_T(v) = \begin{cases} 1 & : \text{ if } v \text{ is a leaf} \\ 2 & : \text{ if } v \text{ is the root} \\ 3 & : \text{ otherwise} \end{cases}$$

holds. In other words, we create a rooted binary tree from a binary tree by choosing an edge and placing the root there.[20]

Observation 11.6.1 *A rooted binary tree with n leaves has exactly $n - 1$ internal vertices.*

And,

Observation 11.6.2 *Let T be a rooted binary tree of depth d. Then T has at least $d + 1$ and at most 2^d leaves.*

Conversely, the depth of such a tree with n leaves lies between $\Omega(\log n)$ and $O(n)$.

For each of the labeled trees we have n rooted trees, because any of the n vertices can be made a root. Hence, as a consequence of Cayley's tree formula we find[21]:

[20] Remember that this procedure is called rooting a tree.
[21] Similar discussed by Harper [125] for problems in paleontology.

Theorem 11.6.3 *The number of different rooted labeled trees with n vertices equals* n^{n-1}.

For semi-labeled trees we have together with 11.2.2 the following result.

Theorem 11.6.4 *The number of rooted binary trees with n labeled leaves and unlabeled internal vertices (i.e. rooted binary N-trees having* $|N| = n$*) is*

$$(2n-3)!! = 1 \cdot 3 \cdot 5 \cdots (2n-3) = \Omega\left(\left(\frac{2n}{3}\right)^{n-1}\right). \tag{11.22}$$

Moreover, in view of 11.2.2(b), the number of rooted binary trees with n labeled leaves and unlabeled internal vertices grows $2n-3$ times faster than the number of binary trees with the same kind of vertices. And, in 11.2.3 we saw that the function $(2n-5)!!$ grows very rapidly with n.[22] Hall's tabular [119]:

Number of taxa	Number of unrooted trees	Number of rooted trees
3	1	(not defined)
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425
11	34,459,425	654,729,075
12	654,729,075	13,749,310,575
⋮		
20	$2.2 \cdot 10^{20}$	$8.2 \cdot 10^{21}$

The assumption that only one of these trees correctly represents the true evolutionary relationship among the taxa, it is usually very difficult to identify the true phylogenetic tree when the number of taxa is large.

11.7 The shape of phylogenetic trees

Now we go back from semi-labeled to unlabeled trees. Consider a N-tree T and ignore the labels (of the leaves), we obtain a tree shape $\mathcal{T}(T)$.
This term introduce an equivalence relation of phylogenetic trees: The N-trees T_1 and T_2 are called shape equivalent if $\mathcal{T}(T_1)$ is isomorphic to $\mathcal{T}(T_2)$.

[22] This was one of the pessimistic view by Graham and Foulds [103], that it will be unlikely that minimal phylogenies for realistic number of "living entities" can be constructed in reasonable computational time. Today we are a little bit more optimistic. In particular by applying PAUP, which stands for "Phylogenetic analysis using parsimony"; see Hall [119] and Swofford [232].

This concept extends naturally to the rooted trees and binary trees.

Obviously, there are two questions:

1. Given a positive integer n. Determine the number of tree shapes for phylogenetic trees on n leaves.

2. For a given tree shape $\mathcal{T}$, count the number of phylogenetic trees on a given label set with shape $\mathcal{T}$.

The first question we discussed together with the considerations about unlabeled trees. Now we consider the second. We follow an approach of Semple and Steel [220] to give an outline.
We need some facts from the theory of groups: An action of a group Γ on a set X is a map $X \times \Gamma \to X$. Two elements x and x' are equivalent if there is an element π of Γ such that $\pi x = x'$, or equivalently $\mathrm{orb}(x) = \mathrm{orb}(x')$, when we define

$$\mathrm{orb}(x) = \{\pi(x) : \pi \in \Gamma\}, \tag{11.23}$$

We denote by $[x]$ the equivalence class of x under this relation.

We find Burnsides lemma O.3.5 in the following form:

Lemma 11.7.1 *Let Γ be a finite group acting on a finite set X and let $x \in X$. Then*

$$|[x]| = \frac{|\Gamma|}{|stab(x)|}, \tag{11.24}$$

where

$$stab(x) = \{\pi \in \Gamma : \pi(x) = x\}. \tag{11.25}$$

In our context, X is the collection of all rooted phylogenetic trees with n labeled vertices, while Γ is the group of all $n!$ permutations.
We consider for an element $\pi \in \Gamma$ and an element $\mathcal{T} \in X$, the action which maps $\mathcal{T}$ to the phylogenetic tree obtained from $\mathcal{T}$ by permuting the labels according to π.

An internal vertex v of a rooted binary tree shape $\mathcal{T}$ is called a symmetry vertex if the two maximal rooted subtrees that lie below v have the same shape. Let $s(\mathcal{T})$ denote the number of symmetry vertices of $\mathcal{T}$. Combining all our results we obtain the following theorem.

Theorem 11.7.2 *The number of rooted binary phylogenetic trees T with n leaves and of shape $\mathcal{T}$ is*

$$n! \cdot 2^{-s(\mathcal{T})}. \tag{11.26}$$

11.8 Generalized binary trees

A generalized binary tree consists of a root and its left and right subtree, which are themselves smaller generalized binary trees.
Our recursive view of such trees makes mathematical induction the method of choice for proving many important facts about binary trees. Typically, (Exercise)

Observation 11.8.1 *Consider a generalized binary tree T.*

- *If T is of depth d, then its left and right subtree both have a depth less than or equal to $d-1$, and equality holds for at least one of them.*
- *If T is of depth d, then it has at most $2^{d+1}-1$ vertices.*

Now, we derive a formula for the number b_n of different generalized binary trees on n vertices.
First, we will establish a recursive formula for b_n. The recursion can be expressed more conveniently by artificially defining $b_0 = 1$. Since the only binary tree with one vertex is a root without subtrees; it follows that $b_1 = 1$. For $n > 1$, a generalized binary tree T on n vertices has a left subtree with j vertices and a right subtree with $n-j-1$ vertices, for some j between 0 and $n-1$. Pairing the b_j possible left subtrees with the b_{n-j-1} possible right subtrees, the multiplication principle gives

$$b_j \cdot b_{n-j-1}$$

different combinations of left and right subtrees for T. Hence,

$$b_n = b_0 b_{n-1} + b_1 b_{n-2} + \ldots + b_{n-1} b_0. \tag{11.27}$$

This recurrence equation is well-known: M.2.1. Consequently, in view of M.3.1:

Theorem 11.8.2 *The number b_n of different generalized binary trees on n vertices is the Catalan number:*

$$b_n = C_n = \frac{1}{n+1}\binom{2n}{n}. \tag{11.28}$$

11.9 Genealogical trees

Now we consider specific trees which play a role in social models of parent-child relations.[23] Schimming [215] discussed so-called genealogical trees, which are rooted trees with linearly ordered leaves.
More exactly, we translate the notations

[23] Curiously, one of the first mathematical papers about phylogenetic trees created by Buneman [34] dealt not with biology but rather with reconstructing the copying history of manuscripts. An example, Mink [174]:

> The same data as used for creating the new printed *Editio Critica Maior* of the New Testament, commencing with Catholic Letters, allows a genealogical analysis of the witness. The objective is to establish a comprehensive theory of the structure of the tradition. Because the tradition of the New Testament is highly contaminated this

in the tree	in the social model
vertex	person
root	progenitor
internal vertex	person with an offspring
leaf	person without an offspring
adjacency	parent-child relation
level	generation

But we further assume that there is a social order in each generation: It shall be hereditary, that means pass over from the parents to their children[24] and can be written more formally as: If v is higher than w, then every child v' from v is higher than every child w' of w.

$C(n,k)$ denotes the number of non-isomorphic genealogical trees with $n+1$ vertices and k leaves.
As an example determine some simple cases and show

$$\begin{aligned} C(n,1) = C(n,n) &= 1 \quad \text{and} \\ C(n,2) = C(n,n-1) &= \binom{n}{2}. \end{aligned}$$

Theorem 11.9.1 *(Schimming [215]) For the number $C(n,k)$ of genealogical trees it holds*

$$C(n,k) = \frac{1}{n}\binom{n}{k}\binom{n}{k-1}. \tag{11.29}$$

By simple calculation:

$$C(n,k) = \frac{1}{k}\binom{n-1}{k-1}\binom{n}{k-1}. \tag{11.30}$$

The equations created the following triangle:

theory has to handle the problem of contamination, and also the problem of accidental rise of variants, and must be able to be verified at any passage of the text. Where there are variants, the witnesses have a relation that can be described by a local stemma of the different readings. These local stemmata allow or restrict relations among witness in a global stemma, which must be in harmony with the total of the local stemmata. In the first phase, local stemmata were established only at places where the development of the variants is very clear. The coherencies within each attestation were analysed.... Then the local stemmata must be revised in the light of the total of the genealogical data included in them. Now an analysis of genealogical coherence is possible and may help to find local stemmata for passages unsolved so far. Finally, the global stemma (or stemmata) mirroring all the relations of the local stemmata will be established by combining optimal substemmata, each containing a witness and its immediate ancestor, to produce the simplest possible tree.

For an application to the *Canterbury Tales Project* see [64].

[24] for example by the order of births or by a system of privileges

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

The triangle suggests the following equation, which is indeed true $C(n,k) = C(n, n+1-k)$. Of more interest is that the numbers $C(n,k)$, $k = 1, \ldots, n$ form a "partition" of the nth Catalan number.

Corollary 11.9.2 *(Schimming [215])*

$$\sum_{k=1}^{n} C(n,k) = C_n, \tag{11.31}$$

where C_n denotes the nth Catalan number.

11.10 Multifurcating trees

Until now, in our studies we assume that the evolutionary process is usually employs bifurcation trees, in which each ancestral taxon splits into two descendent taxa. Multifurcation generalizes bifurcation, when we assume that the tree is not necessarily a binary one. There are two possible interpretations for such an approach:

a) Either it represents the true sequence of events, whereby an ancestral taxon gave rise to three or more descendent taxa simultaneously;

b) Or it represents an instance in which the exact order cannot be determined unambiguously with the available data.

Although multifurcating trees perhaps model biological reality better, mostly, constructing binary phylogenetic trees is considered. Reasons:

- The construction of multifurcation trees is much more difficult.
- In evolution multifurcating events are rare.

We will discuss several aspects of multifurcating trees: structure and counting.

I. Let $m \geq 2$ be an integer. A rooted tree in which every vertex has m or fewer successors is called an m-ary tree.

Lemma 11.10.1 *An m-ary tree has at most m^k vertices at level k.*

Proof. We use induction over k.
The statement is trivially true for $k = 1$.
Assume as induction hypothesis that there are m^l vertices at level l, for some $l \geq 1$. Since each of these vertices has at most m successors, there are at most $m \cdot m^l = m^{l+1}$ successors at the next level. □

In view of this lemma an m-ary tree of depth d has at most m^d leaves.[25] Conversely, the depth of such a tree with n leaves lies between $\Omega(\log n)$ and $O(n)$.
About the total number of vertices we derive another inequality.

Theorem 11.10.2 *Let T be an m-ary tree of depth d with n vertices. Then*

$$d + 1 \leq n \leq \frac{m^{d+1} - 1}{m - 1}. \tag{11.32}$$

Proof. Let n_k be the number of vertices at level k. In view of 11.10.1

$$1 \leq n_k \leq m^k.$$

Thus

$$d + 1 = \sum_{k=0}^{d} 1 \leq \sum_{k=0}^{d} n_k = n,$$

and

$$n = \sum_{k=0}^{d} n_k \leq \sum_{k=0}^{d} m^k = \frac{m^{d+1} - 1}{m - 1}.$$

□

II. In 11.6.4 we gave the number of bifurcating (binary) trees. Now we are interested in methods for enumerating the number of trees when multifurcations are allowed. Felsenstein [83] created the following approach, using the observation that if a tree contains a multifurcation, it has fewer than $n - 1$ internal vertices.
Let m-tree(n, k) be the number of distinct (multifurcating) trees having n labeled leaves and k unlabeled internal vertices. We will extend our method for proving 11.2.3 to compute m-tree(n, k) from m-tree$(n - 1, k)$. There will be a one-to-one correspondence between the ways adding a new leaf and the desired tree.
Clearly,

$$\begin{aligned} \text{m-tree}(1, 0) &= 1 \quad \text{and} \\ \text{m-tree}(1, i) &= 0 \quad \text{otherwise.} \end{aligned}$$

Assume that we know all values of m-tree$(n - 1, i)$.
If we add the leaf v_n to a tree and obtain a tree with k internal vertices, this could happen in two ways:

[25] But maybe no more than one leaf.

1. We could take a tree with $n-1$ leaves and k internal vertices. For each of the trees of this sort there are then k possibilities at which the new leaf could be added.

2. We could take a tree with $n-1$ leaves and $k-1$ internal vertices. Each tree of this sort has $n+k-2$ internal edges. Then we place a new internal vertex in the midst of one of its edges, and have a new leaf arise from the new vertex.

Consequently,

$$\text{m-tree}(n,k) = \begin{cases} k \cdot \text{m-tree}(n-1,k) + (n+k-2) \cdot \text{m-tree}(n-1,k-1) & : \quad k > 1 \\ \text{m-tree}(n-1,1) & : \quad k = 1 \end{cases}$$

In view of 11.6.1 m-tree$(n, n-1)$ is the number of bifurcating rooted trees.

Theorem 11.10.3 *$m-tree(n,k)$ is the number of distinct (multifurcating) trees having n labeled leaves and k unlabeled internal vertices. Furthermore, there are*

$$\sum_{k=1}^{n-1} m\text{-}tree(n,k) \tag{11.33}$$

distinct multifurcating trees having n labeled leaves and unlabeled internal vertices.

III. Further generalizations, in particular for trees in which ancestors and/or leaves are partly labeled, are discussed in Felsenstein [83].

11.11 Phylogenetic forests

Let N be a finite set (of names). We generalize phylogenetic trees by the following definition: A phylogenetic forest G on N is a union of trees $T_i = (V_i, E_i)$, $i = 1 \ldots, c$, such that

$$N_i = N \cap V_i$$

form a partition of N, that means T_i is an $N \cap V_i$-tree.
If each of the trees T_i is rooted and binary, then G is called a rooted binary phylogenetic forest. Surprisingly, there is an explicit formula for the number of such trees which generalizes 11.2.2(b), 11.2.3 and 11.6.4 (exercise).

Theorem 11.11.1 *For every pair of positive integers n and c with $1 \leq c \leq n$, the number of rooted binary phylogenetic forests with n labeled vertices and consisting of c trees (components) equals*

$$\frac{(2n-c-1)!}{(n-c)! \cdot (c-1)! \cdot 2^{n-c}}. \tag{11.34}$$

For a proof see [220].

Chapter 12

Collections of Trees

> If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included such an arrangement would be the only possible one.
>
> Charles Darwin

So far we have assumed that the evolutionary relationships among sequences[1] are best represented by a tree. However, the actual evolutionary history may be not be in particular tree-like, in which case analyses that assume a tree may be seriously misleading. There are limitations in always forcing data onto a standard phylogenetic tree. Processes such as parallel mutation, hybridization, recombination, gene conversion and lateral gene transfer violate a tree-based evolutionary model. In other words, if we look for "The Great Darwin Tree", we will find a network with several cycles. But the number of (elementary) cycles will be small in relation to the number of vertices.[2]

12.1 Splits and trees

Let N be a finite set. A split for N is a bipartition, that is a partition of N into two non-empty sets. That means $N = S \cup S^c$ with $S \cap S^c = \emptyset$.

Theorem 12.1.1 *The number of splits of a set of n elements is $2^{n-1} - 1$.*[3]

Proof. If we choose k elements for N, $0 < k < n$, then we also choose $n - k$ elements for S^c. Hence, each selecting of k elements creates a split. We can do it in

[1] and genes, species, organisms,...

[2] This idea of a symbiotic evolution was first given by Margulis [168], compare also [68].

[3] In terms of the Stirling numbers of the second kind: $S(n, 2) = 2^{n-1} - 1$.

$\binom{n}{k}$ ways. But we find each split twice, namely by choosing k, and by choosing $n-k$. Consequently, the number of splits is

$$\begin{aligned}\frac{1}{2}\sum_{k=1}^{n-1}\binom{n}{k} &= \frac{1}{2}\left(\sum_{k=0}^{n}\binom{n}{k}-\binom{n}{0}-\binom{n}{n}\right)\\ &= \frac{1}{2}(2^n-1-1)\\ &= 2^{n-1}-1.\end{aligned}$$

□

Let $T=(V,E)$ be an N-tree, $|N|\geq 3$. For an edge e of T the graph $G-e$ has exactly two components $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$, and creates a split $\mathcal{S}(e)=\{N_1,N_2\}$ of the set N of leaves by setting

$$N_1 = V_1\cap N \text{ and} \tag{12.1}$$
$$N_2 = V_2\cap N = N\setminus N_1. \tag{12.2}$$

This means in particular that for a split $\mathcal{S}(e)=\{N_1,N_2\}$, each path from a vertex in N_1 to a vertex in N_2 contains the edge e. The collection

$$\mathcal{S}(T)=\{S(e):e\in E\} \tag{12.3}$$

denotes the family of all splits of N induced by the tree T.
For instance, consider the set $N=\{a,b,c,d,e\}$. Coming from the (binary) N-tree $(((ab)c)(de))$ we have the split family

$$\begin{aligned}\mathcal{S} = &\{\{a,bcde\},\{b,acde\},\{c,abde\},\{d,abce\},\{e,abcd\},\\ &\{ab,cde\},\{abc,de\}\}.\end{aligned} \tag{12.4}$$

The following result provides a fundamental equivalence between N-trees and a certain type of collection of splits of N. A pair $\{N_1,N_2\}$ and $\{M_1,M_2\}$ of splits for N is called compatible if at least one of the sets $N_1\cap M_1$, $N_1\cap M_2$, $N_2\cap M_1$ and $N_2\cap M_2$ is the empty set. Then we have the following central theorem:

Theorem 12.1.2 *(Buneman [34]) Let $\mathcal{S}$ be a collection of splits for the set N. Then there is an N-tree T such that $\mathcal{S}=\mathcal{S}(T)$ if and only if the splits in $\mathcal{S}$ are pairwise compatible. Moreover, if such a tree exists, then, up to isomorphism, T is unique.*

For a proof compare [15].

A natural way to generate splits of a set N is to arrange the elements of N on a circle in the plane and then to draw lines to divide the points into two subsets.

Theorem 12.1.3 *Let N be a set of n elements, $n\geq 3$, and let π be a cyclic permutation on N. Then the number of binary N-trees for which π is a circular ordering equals the Catalan number C_n.*

Sketch of the *proof.* On the one hand, we know by that the number of cyclic permutations of N is $(n-1)!$, see 6.1.2.
On the other hand, the number of choices of N-trees is determined in 11.2.2.
Equating these two counts gives the desired result.[4] □

12.2 Reconstructing trees

O. Each N-tree is a metric space. Let S be a split on the tree $T = (V, E)$. Then define for two leaves v and w for T:

$$d(v,w) = \begin{cases} 1 & : \quad S \text{ separates } v \text{ and } w \\ 0 & : \quad \text{otherwise} \end{cases}$$

This creates a metric on V as

$$\rho_T(v,w) = \sum_{\text{split } S} d(v,w). \tag{12.5}$$

Theorem 12.2.1 *Let v and w be leaves of a N-tree T. Then*

$$\rho_T(v,w) = \text{ length of the path from } v \text{ to } w.$$

The *proof* follows from the fact that length of a path is the number of its edges. □

I. The general idea is the following: Let $N = \{v_1, \ldots, v_n\}$ be a set of individuals (OTU's). We assume that N is embedded in a metric space (X, ρ), such that we represent the distances between the members of N by a symmetric distance matrix

$$D = (d_{ij}) = (\rho(v_i, v_j)), \tag{12.6}$$

where $i, j = 1, \ldots, n = |N|$.
We would like to build a phylogenetic tree for N. If we fix an N-tree $T = (V, E)$ we obtain a metric ρ_T. The broad aim of distance methods is to determine a (or all) tree(s) T for which ρ_T is as close to ρ as possible. Then we said that D is "tree-like".
Consider the following example. For $n = 3$ we only have one N-tree T with one internal vertex w. Looking for the edge-lengths $l_i = \rho_T(v_i, w)$, $i = 1, 2, 3$, such that $\rho = \rho_T$ means solving the following system

$$\begin{aligned} l_1 + l_2 &= d_{12} \\ l_1 + l_3 &= d_{13} \\ l_2 + l_3 &= d_{23}, \end{aligned}$$

[4] The values of the Catalan numbers represent the number of ways to cut a polygon: Examines a given convex polygon of $n \geq 3$ sides. Euler counted the number of ways the interior of the polygon to subdivide into triangles by drawing diagonals that do not intersect. Let c_n be this number.

$$c_{n+1} = c_2 c_n + c_3 c_{n-1} + \ldots + c_{n-1} c_3 + c_n c_2.$$

Then use M.2.1.

which is given by

$$\begin{aligned} l_1 &= \frac{1}{2}(d_{12} + d_{13} - d_{23}) \\ l_2 &= \frac{1}{2}(d_{12} + d_{23} - d_{13}) \\ l_3 &= \frac{1}{2}(d_{13} + d_{23} - d_{12}). \end{aligned}$$

(Note that the values on the right-hand side are non-negative due to the triangle inequality.) Hence, we have a unique tree which reflects the phylogeny with respect to given distances.

II. For specific classes of distances we have some results more. A collection of distances $D = (d_{ij}) = (\rho(v_i, v_j))$ for $i, j = 1, \ldots, n = |N|$ is called additive if there is an N-tree T such that

$$d_{ij} = \rho_T(v_i, v_j). \tag{12.7}$$

We saw that for N each distances are additive. In general, this is not true for $n \geq 4$. We characterize additive distances by the following statements.

Theorem 12.2.2 *(Buneman [35]) A collection $(d_{ij})_{i,j=1,\ldots,n}$ of distances is additive if and only if for every set of four distinct numbers it holds the so-called four-point condition, that means for $1 \leq i, j, k, l \leq n$ two of the three sums $d_{ij} + d_{kl}$, $d_{ik} + d_{jl}$ and $d_{il} + d_{jk}$ coincide and are greater than or equal to the third one:*

$$d_{ij} + d_{kl} \leq \max\{d_{ik} + d_{jl}, d_{il} + d_{jk}\}. \tag{12.8}$$

In biology, the so-called molecular clock hypothesis states that the mutation rate is constant over all edges of the tree. That implies that all leaves have the same distance to the root. We call such a tree an ultrametric ones.

Theorem 12.2.3 *A collection $(d_{ij})_{i,j=1,\ldots,n}$ of distances is ultrametric if and only if for every tripel of distinct numbers it holds the so-called three-point condition:*

$$d_{ij} \leq \max\{d_{ik}, d_{kj}\}. \tag{12.9}$$

Effective methods to construct the trees which reflect 12.7 are given in the references at the end of the present chapter.[5]

[5]Since we are looking for binary trees these approaches use pair group methods (PGM) which are known from cluster analysis, [81]:

Algorithm 12.2.4 *Let $N_1 = \{v_1\}, \ldots, N_n = \{v_n\}$ be a family of sets each containing a single element. Then do*

1. *Find the nearest pair of distinct sets, say N_i and N_j;*
2. *Merge N_i and N_j to form N';*
   Compute a new distance, or similarity from N' to each of the other sets;
   Decrement the number of sets by one;
3. *If the number of sets equal one then STOP, else go to 1.*

12.3 Fitch's algorithm

For a given semi-labeled tree, finding the labels of the internal vertices such that the total number of changing is minimized is the parsimony problem.[6]
A well-known method to compute a tree in a sequence space is a dynamic programming algorithm for finding the location of the internal vertices in a given N-tree:

Algorithm 12.3.1 *(Fitch [85]) Let N be a set of n sequences in a sequence space (A^d, ρ_H): $N = \{v_k = v_{k,1}, \ldots, v_{k,d} : k = 1, \ldots, n\}$, and let a binary N-tree $T = (V, E)$ be given. Then do:*

1. *For each position $i = 1, \ldots, d$ do*
   1. Label each leaf v_k with $\{v_{k,i}\}$;
   $L_i := 0$;
   2. Until all vertices are labeled do
   Find an unmarked vertex which is adjacent to two marked vertices with the labels N_1 and N_2;
   Label the unmarked vertices with
   (a) $N_1 \cap N_2$ if $N_1 \cap N_2 \neq \emptyset$; otherwise
   (b) $N_1 \cup N_2$ and $L_i := L_i + 1$;
2. *$L(T) := \sum_{i=1}^{d} L_i$.*

The correctness of Fitch's algorithm is proven by Hartigan [126]. In particular, it is shown that the final answer is independent of the vertices chosen when moving through the tree.
The algorithm computes the length of the tree. Since a binary N-tree has $2n-2$ vertices, it uses $O(n)$ time for each position and hence $O(d \cdot n)$ time to find the length. Hence, the Fitch algorithm uses linear time to find the length of a given binary N-tree On the other hand, there are exponentially many binary trees. Consequently, applying 12.3.1 in a sequence space uses in exponential time.
After applying 12.3.1 we have labels for all the internal vertices in the tree. However,

[6] Remember, that we have to choose a tree among a collection of many trees. The **Principle of Maximum Parsimony** involves the identification of a combinatorial structure that requires the smallest number of evolutionary changes. It is often said that this principle abides by Ockham's razor. Note, that we do not use this principle in a simple sense; Cavalli-Sforza [39]:

> ... it does not necessarily follow that a method of tree reconstruction minimizing the number of mutations is the best or uses all the information contained in the sequences. The minimization of the number of mutation is intuitively attractive because we know that mutations are rare. There may be some confusion, however, between the advantage of minimizing the number of mutations and sometimes invoked parallel of Ockham's razor ..., which was developed in the context of mediaval theology. The extrapolation of Ockham's razor to the number of mutations in an evolutionary tree is hardly convincing.

That means that in this case minimizing the number of assumptions does not mean to minimize the steps of an evolution, it means that among all possible network structures we search one which satisfy only one, namely the condition of length minimizing. What other condition can be more natural in a metric space? For more facts what does Ockham's razor in network design really mean see [26], [53] and [133].

some marks have more than one letter and hence are ambiguous. There are several methods for choosing which one of the possible states yields the most parsimonious reconstruction; the simplest one is : Go back up the tree assigning to any internal vertex that is ambiguous the intersection of its label with that of its immediate ancestor.

However, as the number of possible trees increases rapidly with the number of given sequences, it is virtually impossible to employ an exhaustive search when the number of given sequences is not small. Fortunately, there exist short-cut algorithms for identifying all shortest trees that do not require exhaustive enumeration, and work for larger sets of sequences. One such algorithm is the branch-and-bound method by Hendy and Penny [130], described briefly below:

1. Guess a "good tree" T_0 using a heuristic[7]);
   $L_0 := L(T_0)$;
   Let X be the set of all binary N-trees;

2. (Iteration:)
   1. Partition X into a small number of subsets $X_1, X_2, \ldots, X_k$;
   2. For $i := 1, \ldots, k$ do
   - Find a length $L(X_i)$ such that $L(T) \geq L(X_i)$ for all $T \in X_i$;
   - If $L(X_i) \leq L_0$ then iterate (return to 1. with $X = X_i$).

12.3.1 can be extended to find the location of Steiner points in phylogenetic spaces $(A^\star, \rho_L)$. And indeed, Sankoff [211] and Cieslik et al. [50] give a dynamic programming algorithm for tree alignment. They merges the high-dimensional version of the dynamic programming algorithm for pairwise alignment with the Fitch algorithm.

12.4 Consensus trees

A consensus tree summarizes information common to two or more trees. In other words:

- A phylogenetic tree summarizes phylogenetic information;
- A consensus tree summarizes the information in a set of trees.
  Here, we have two additional observations: a) We can combine heterogeneous data, and b) We can find hidden phylogenetic information.

For instance,

- One may want to treat different phylogenetic trees as different estimations of the same underlying true evolutionary tree. Then a consensus tree represents the evolutionary history on which the different trees agree.

[7]This tree is expected to have short length.

- Cavalli-Sforza [40] compares the species tree and the tree of languages for human populations. This gives many hints for the prehistoric development of mankind.

- In general, molecular systematics shows that the phylogenies of genes does not match those of the organisms, due lineage sorting, hybridization, recombination and other events. See v.Haeseler and Liebers [115] or Page and Holmes [185].

- Consensus trees are helpful to taxonomists, see Semple and Steel [220].

The methods differ in what aspect of tree information they use, and how frequently that information must be shared among the trees to be included in the consensus. The most commonly used are the strict consensus and the majority-rule consensus trees.
Suppose that $T_1, \ldots, T_m$ are N-trees. Each of the trees has the same leaves, namely the members of N. We are interested in a consensus N-tree T described by one of the following methods.

I. The strict consensus tree includes only those splits that occur in all the trees. That means

$$\mathcal{S}(T) = \bigcap_{i=1}^{m} \mathcal{S}(T_i). \tag{12.10}$$

II. We can relax the requirement that a split of T occurs in all trees, and instead retain those splits occuring in a majority of the trees.

Algorithm 12.4.1 *For each of the N-trees $T_1, \ldots, T_m$, mark the vertices inductively as follows:*

1. *Mark the leaf v with $\{v\}$;*
2. *If the vertices $v_1, \ldots, v_r$ have been marked with $N_1, \ldots, N_r$ and v is the common ancestor of $v_1, \ldots, v_r$, then mark v with $N_1 \cup \ldots \cup N_r$.*

The consensus tree T consists of exactly those vertices whose mark occurs in more than half of the T_i.

For more facts about consensus trees see Margush, Morris [169].

III. As a generalization and a weaker form of **II.** we define consensus trees by the following methods: Let $\mathcal{S}$ be a collection of splits and let $S \in \mathcal{S}$. Let $n(S)$ be the number of N-trees in a family $\mathcal{F} \subseteq \mathcal{T}_n$ that induces S. For all $0 \leq q < 1$, let

$$\mathcal{S}(q) = \left\{ S \in \mathcal{S} : \frac{n(S)}{|\mathcal{F}|} > q \right\}. \tag{12.11}$$

As an exercise prove the following

Theorem 12.4.2 *If $q \geq 0.5$, then $\mathcal{S}(q)$ is the set of splits of N induced by a (unique) N-tree.*[8]

[8] For $q < 0.5$ the definition of $\mathcal{S}(q)$ makes no sense.

12.5 The metric spaces of all trees

Given two or more phylogenetic trees computed from different gene families or related taxa, the problem of comparing these trees arises.

I. It is often helpful to have a measure of distance between two phylogenetic trees. More precisely, to begin with, let $\mathcal{T}_n$ denote the set of all N-trees with $n = |N|$.

$\mathcal{T}_1$ and $\mathcal{T}_2$ each contain exactly one tree. $\mathcal{T}_3$ contains one, and $\mathcal{T}_4$ four trees. A tree in $\mathcal{T}_n$ has at least one and at most $n-2$ internal vertices, and consequently, at least n and most $2n-3$ edges. Using 6.2.3 we find that

$$|\mathcal{T}_n| = \sum_{k=1}^{n-2} \frac{(n+k)^{n+k-2}}{k!} \leq \sum_{k=1}^{n-2} \frac{(2n)^{n+k-2}}{k!} = (2n)^{n-2} \cdot \sum_{k=1}^{n-2} \frac{(2n)^k}{k!} \leq (2n)^{n-2} \cdot e^{2n}.$$

On the other hand,

$$|\mathcal{T}_n| \geq (2n-5)!! = \frac{(2n-4)!}{(n-2)! \cdot 2^{n-2}} \geq \left(\frac{2}{e}\right)^{n-2} \cdot (n-2)^{n-3}.$$

In other words, $\mathcal{T}_n$ will be a finite set with a number of elements which is exponential in n, but not more or less. More facts are given in [131].

We are interested in creating a metric between the trees in $\mathcal{T}_n$ which reflects the "difference" between the trees in the sense of different phylogeny. A commonly used measure of dissimilarity between two N-trees is Penny and Hendy's [188], [131] method based on tree partitioning. It uses the binary operation $\triangle$ which is the symmetric difference between sets, defined as

$$\begin{aligned} \mathcal{S}_1 \triangle \mathcal{S}_2 &= (\mathcal{S}_1 \setminus \mathcal{S}_2) \cup (\mathcal{S}_2 \setminus \mathcal{S}_1) && (12.12) \\ &= \mathcal{S}_1 \cup \mathcal{S}_2 \setminus \mathcal{S}_1 \cap \mathcal{S}_2 && (12.13) \end{aligned}$$

for sets $\mathcal{S}_1, \mathcal{S}_2$.[9] Then

$$\begin{aligned} |\mathcal{S}_1 \triangle \mathcal{S}_2| &= |\mathcal{S}_1 \cup \mathcal{S}_2 \setminus \mathcal{S}_1 \cap \mathcal{S}_2| \\ &= |\mathcal{S}_1 \cup \mathcal{S}_2| - |\mathcal{S}_1 \cap \mathcal{S}_2| \\ &= |\mathcal{S}_1| + |\mathcal{S}_2| - 2 \cdot |\mathcal{S}_1 \cap \mathcal{S}_2|. \end{aligned}$$

In R.2.1, we saw that $\rho(S_1, S_2) = |\mathcal{S}_1 \triangle \mathcal{S}_2|$ is a metric. Then we define the metric ρ between trees as follows: Let T_1, T_2 be two trees in $\mathcal{T}_n$, $n \geq 3$, with the induced split collections $\mathcal{S}(T_1), \mathcal{S}(T_2)$, respectively.

$$\rho_S(T_1, T_2) = |\mathcal{S}(T_1) \triangle \mathcal{S}(T_2)| \tag{12.14}$$

is a distance between T_1 and T_2, which is called the split metric.

[9] Let $\mathcal{S}_1, \ldots, \mathcal{S}_k$ be a family of subsets of $\mathcal{U}$. An element of $\mathcal{U}$ is a member of $\mathcal{S}_1 \triangle \mathcal{S}_2 \triangle \ldots \triangle \mathcal{S}_k$ if and only if it is contained in an odd number of the $\mathcal{S}_i$'s. In particular, the symmetric difference of a set with itself is empty.

Observation 12.5.1 *$(\mathcal{T}_n, \rho_S)$ is a metric space.*

The proof for the following description for the split metric should be an exercise for the reader.

Remark 12.5.2 *Let $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ be two N-trees. Then*

$$\rho_S(T_1, T_2) = |E_1| + |E_2| - \#\text{ splits induced by both } T_1 \text{ and } T_2. \tag{12.15}$$

Note that it is algorithmically easy, i.e. achievable in polynomial time, to compute the distance between two trees in $(\mathcal{T}_n, \rho_S)$.

Theorem 12.5.3 *(Robinson and Foulds) Let T and T' be N-trees. Then $\rho_S(T, T')$ is equal to the smallest k for which there is a sequence $T = T_0, T_1, \ldots, T_k = T'$, and for all $i = 1, \ldots, k$ the tree T_i is obtained from T_{i-1} by either contraction or expanding an edge T_i.*

For a proof see [220].

We call an edge of T an internal edge if it connects two internal vertices. However, all splits comprising a leaf on one hand and the rest of the tree on the other are not "phylogenetically informative" in the sense that all possible N-trees will contain those splits. Using an internal edge implies for a split that $|N_1|, |N_2| \geq 2$. Since each tree in $\mathcal{T}_n$, $n \geq 3$, contains at most $n - 3$ internal edges, we observe the following:

Theorem 12.5.4 *It holds for any two trees T_1 and T_2 in $(\mathcal{T}_n, \rho_S)$*

$$\begin{aligned} \rho_S(T_1, T_2) &\leq \#\text{internal edges in } T_1 + \#\text{internal edges in } T_2 && (12.16)\\ &\leq 2n - 6. && (12.17) \end{aligned}$$

In particular, the diameter of the metric space $(\mathcal{T}_n, \rho_S)$ equals $2n-6$ (Exercise).[10] A "strange" metric space: many elements and small diameter. More exactly: exponential size and linear expansion. (Do you know some other spaces with this property?)

II. $\mathcal{T}$ denotes the set of all trees: $\mathcal{T} = \bigcup_{n \geq 1} \mathcal{T}_n$. By 2.6.1 $\mathcal{T}$ is a infinite, but countable set.
The edit distance ρ, between two trees of not necessarily equal size is the minimal number of "edit operations" required to change one tree into the other, where an edit operation are deletion of a leaf from a tree, insertion of a new leaf into a tree, and substituting a vertex in a tree for a vertex in another tree. Then $(\mathcal{T}, \rho)$ becomes a metric space.

[10] More facts about the geometry of the space of phylogenetic trees can be found by Billera, Holmes and Vogtmann [27].

12.6 Further reading

For further reading about phylogenetic networks, its application in the theory of evolution, construction methods and related topics see:

1. Böckenhauer, Bongartz: Algoritmische Grundlagen der Bioinformatik; [31].
2. Cieslik: Shortest Connectivity - An Introduction with Applications in Phylogeny; [52].
3. Clote, Backofen: Computational Molecular Biology; [56].
4. Gusfield: Algorithms on Strings, Trees, and Sequences; [112].
5. Hall: Phylogenetic Trees Made Easy; [119].
6. Huson et al.: Phylogenetic Networks; [136].
7. Knoop, Müller: Gene und Stammbäume; [156].
8. Page, Holmes: Molecular Evolution: A Phylogenetic Approach; [185].
9. Semple, Steel: Phylogenetics; [220].
10. Waterman: Introduction to Computational Biology; [248].

An annotated bibliography in Computational Molecular Biology is presented by Vingron et al. [242].

Chapter 13

Spanning Trees

Network organization is just about universal in the real world, as evidenced by systems as different as the World Wide Web, glass fiber cable interconnections, metabolic networks, protein interactions, and financial transactions. Such universality is related to the very general nature of network structures. Compare [49], [179], and [253].
The mathematical description of networks is given by graph theory; and an core investigating networks are the spanning trees. See [253].

13.1 The number of spanning trees

Let $G = (V, E)$ be a graph. A subgraph $G' = (V, E')$ is called a spanning tree of G if G' is a tree. If G' is a spanning tree of G, then G itself must be connected. Conversely, if $G = (V, E)$ is a connected graph, then G contains a subgraph $G' = (V, E')$ minimal with respect to the property that G' is connected. The graph G' is a spanning tree of G. Hence,

Theorem 13.1.1 *A graph is connected if and only if it contains a spanning tree.*

This simple result has a lot of important consequences: First, in view of our investigations about random graphs, we see that almost all graphs contain a spanning tree. Furthermore,

Observation 13.1.2 *A connected graph with n vertices contains at least $n-1$ edges.*

And

Observation 13.1.3 *Each spanning tree of a connected graph G contains all bridges of G.*

In some situations it is necessary to be able to generate a complete list of all the spanning trees of a graph. This may be the case when, for example, the "best" tree needs to be chosen, but the criterion used for deciding what tree is the "best"

is very complex. Hence, we are interested in the number of spanning trees for a graph.

Generalizing and applying 6.2.3 we have

Theorem 13.1.4 *Consider (connected) graphs G with n vertices and m edges.*

(a) *(Kelmans [149]) The number of spanning trees of G is at most*

$$\frac{1}{n}\left(\frac{2m}{n-1}\right)^{n-1} \leq n^{n-2}. \tag{13.1}$$

(b) *(Cayley [41]) If the graph is complete, that is $2m = n(n-1)$, equality holds, i.e. the number of spanning trees of K_n is exactly*

$$n^{n-2}. \tag{13.2}$$

The quantity $t(G)$ denotes the number of spanning trees for a graph G. The following facts are easy to see.

Observation 13.1.5 *Let $G = (V, E)$ be a graph. Then*

a) *If G is a tree, then $t(G) = 1$.*

b) *If G is disconnected, then $t(G) = 0$.*

c) $t(C_n) = n$.

d) $t(K_n) = n^{n-2}$.

e) *For any spanning subgraph $G' = (V, E')$, $E' \subseteq E$, of G, it holds $t(G') \leq t(G)$.*

f) *If G is a connected graph and G' a proper spanning subgraph, then $t(G') < t(G)$.*

The present chapter deduces several approaches to calculate the quantity $t(G)$ for several classes of graphs G.

13.2 The density of graphs

I. Let G be a graph with n vertices and m edges. the maximum number of edges is $\binom{n}{2}$. Then the we define the density of G by the fraction of these edges that are actually present:

$$\text{density}(G) = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)}.$$

The quantity $d = \text{density}\ (G)$ is a real number between 0 and 1, whereby $d = 0$ characterizes the empty and $d = 1$ the complete graph. The density of a k-connected graph with n vertices is at least $k/(n-1)$.[1]

[1] The expected density of a graph is 1/2.

The density of a graph G has a strong connection to the average degree $\tilde{d}(G)$, since in view of 4.1.1:

$$\text{density}(G) = \frac{2m}{n(n-1)} = \frac{1}{n(n-1)} \sum_{v \in V} g(v) = \frac{1}{n-1} \tilde{d}(G).$$

Consequently,

Lemma 13.2.1 *Let G be a graph with minimum degree δ, average degree $\tilde{d}$ and density d. Then $(n-1) \cdot d = \tilde{d} \geq \delta$ and for sufficiently large n it holds $nd \approx \tilde{d}$.*

Summarizing our further results we have for specific nonempty graphs:

	minimum degree $\delta \leq$	average degree $\tilde{d} \leq$	density $d \approx$
forests	1	$2 - \frac{1}{n}$	$\frac{2}{n}$
outer-planar graphs	3	$4 - \frac{6}{n}$	$\frac{4}{n}$
planar graphs	5	$6 - \frac{12}{n}$	$\frac{6}{n}$

It is a nice exercise to show that a connected graph with average degree greater than two has at least two cycles.
For practice discuss for planar graphs G and its dual graphs G^d the interrelation between $t(G)$ and $t(G^d)$.
Let G be a graph with $n \geq 3$ vertices and density greater than $1 - 2/n$. then in view of 4.4.3 it holds $t(g) > 1$. Conversely,

$$\text{If density}(G) \left\{ \begin{array}{l} > 1 - \frac{2}{n} \\ < \frac{2}{n} \end{array} \right\} \text{ then } G \text{ is } \left\{ \begin{array}{l} \text{connected} \\ \text{disconnected} \end{array} \right\}$$

II. Let G be connected, then, in view of 13.1.2, $m \geq n - 1$, and furthermore, paying attention 13.1.4:

Theorem 13.2.2 *Let G be a graph with n vertices, density d and average degree $\tilde{d}$. If G is connected then $d \geq 2/n$ and G contains a spanning tree, but at most*

$$n^{n-2} \cdot d^{n-1} = d \cdot (dn)^{n-2} = \frac{1}{n} \cdot \left(\frac{n}{n-1} \tilde{d} \right)^{n-1}. \tag{13.3}$$

In other terms, the number of spanning trees is asymptotically bounded by $\tilde{d}^{n-1}/n$.

The density of planar graphs is bounded by $6/n$. Therefore, a planar graph with n vertices contains at most $6^n/(6n)$ spanning trees. On the other hand, there are planar graphs with an exponential number of spanning trees: For an even number n consider a rotor R_n with $n/2$ triangle wings. Then R_n is an outer-planar graph with $n+1$ vertices and $t(R_n) = 3^{n/2} = \sqrt{3}^n$.

A graph is dense if m is large compared to n and sparse otherwise. More exactly, a graph is called dense if the density tends to a constant as $n \to \infty$.[2] A graph is said sparse if the density tends to zero.
We may expect that dense graphs has more spanning trees than sparse.

III. Consider graphs and its complements.
The density of a self-complementary graph equals 1/2. Consequently, there are at most $\frac{1}{n} \cdot \left(\frac{n}{2}\right)^{n-1}$ spanning trees in such a graph with n vertices.

Lemma 13.2.3 *Let G be a graph and G^c its complement. Then*

$$\textit{density}(G^c) = 1 - \textit{density}(G). \tag{13.4}$$

The *proof* follows from (4.15). □

Theorem 13.2.4 *Let G be a graph with n vertices and let G^c be its complement. Then*

$$t(G) + t(G^c) \leq n^{n-2}. \tag{13.5}$$

Equality holds if and only if G is the complete or the empty graph.

Proof. Let d and d' be the density of G and G^c, respectively, then

$$\begin{aligned} t(G) + t(G^c) &\leq n^{n-2} \cdot d^{n-1} + n^{n-2} \cdot d'^{n-1} \quad \text{by 13.2.2} \\ &= n^{n-2} \cdot \left(d^{n-1} + d'^{n-1}\right) \\ &\leq n^{n-2} \cdot (d + d')^{n-1} \\ &= n^{n-2} \quad \text{in view of (13.4).} \end{aligned}$$

The discussion of equality uses that d or d' must be 0. □

As an example consider the cycle C_5, whereby C_5^c is itself C_5. Then $t(C_5)+t(C_5) = 5 + 5 = 10$.
In view of 4.4.6 we know $t(G) + t(G^c) \geq 1$. This bound cannot be given better: For $n \geq 3$ consider the star G with $n - 1$ leaves. G^c is disconnected. Hence $t(G) = 1$ and $t(G^c) = 0$.

13.3 Polyhedral graphs

Recall that polyhedral graphs are planar and 3-connected, implying that the number of edges is a number between $3n/2$ and $3n - 6$. Simple calculations give

[2] In such a graph the fraction of non-zero elements in the adjacency matrix remains constant as the graph becomes large.

Lemma 13.3.1 *Let G be a polyhedral graph with n vertices. Then*

$$\frac{3}{n} < \mathit{density}(G) < \frac{6}{n}. \tag{13.6}$$

We may expect that polyhedral graphs have many spanning trees, but of which structure? A Hamiltonian path is a spanning tree with maximum degree two. Not each polyhedral graph contains such a path (Example as an exercise). Thus the the following theorem is related in a natural way.

Theorem 13.3.2 *(Barnette [18]) Every polyhedral graph has a spanning tree of maximum degree 3.*

13.4 Generating spanning trees

How we can find spanning trees algorithmically?

I. It is not a difficult task to obtain one (arbitrary) spanning tree. 4.6.2 gives a technique to generate one of a graph $G = (V, E)$:

1. Start with the empty graph $T = (V, \emptyset)$;
2. Sequentially choose an edge that does not form a circle with already chosen edges;
3. Stop when all vertices are connected, that is when $|V| - 1$ edges have been chosen.

A generalization to optimization problems are given in S.1.2 and T.2.1.

II. In general, methods to generate all spanning trees use the following approach: Let $G = (V, E)$ be a connected graph and let $T_1 = (V, E_1)$, $T_2 = (V, E_2)$ be two spanning trees of G. Then

$$\rho(T_1, T_2) = |E_1 \setminus E_2| \tag{13.7}$$

defines a distance. Question: Is ρ a metric?

If the $\rho(T_1, T_2) = 1$, that means

$$(E_1 \setminus E_2) \cup (E_2 \setminus E_1) = E_1 \cup E_2 \setminus E_1 \cap E_2 = \{e_1, e_2\},$$

where $e_i \in E_i$, $i = 1, 2$, then T_2 could be derived from T_1 by removing e_1 and introducing e_2. Such a transformation is called an elementary tree transformation.

Theorem 13.4.1 *(Christofides [48]) If T_0 and T_k are spanning trees of a graph with $\rho(T_0, T_k) = k$, then T_k can be obtained from T_0 by a sequence of k elementary tree transformations.*

Kapoor and Ramesh [145] present an algorithm for enumerating all spanning trees of a graph G having complexity $O(t(G)+n+m)$, where $t(G)$ is the number of trees.

III. Sometimes it is of interest to find spanning trees with specific properties. In general, this will be not a simple question. Consider a graph $G=(V,E)$ with n vertices, then the following problems are $\mathcal{NP}$-complete, [97], [253].[3]

- For a given integer g, is there a spanning tree for G in which no vertex has degree larger than g?[4]
- Given a sequence
$$\{g_1,...,g_n\}\subseteq\{1,2,\ldots\}\cup\{\infty\} \tag{13.8}$$
of positive integers. Is there a spanning tree such that no vertex v_i has a degree greater than g_i, $i=1,...,n$?
- For a given positive integer L, decide if there is a spanning tree T with
$$\sum_{v,v'\in V}\rho_T(v,v')\leq L?$$
- For a given integer k, is there a spanning tree for G in which k or more are leaves?[5]
- For a given tree T with n vertices, does G contain a spanning tree isomorphic to T.

On the other hand, there are several problems which can be solved in polynomially bounded time:

- Find a spanning tree of G with minimum diameter, [253].[6]
- Find a spanning tree in which a specific vertex has a degree bounded by a given positive integer, [93].

[3]For practice read all these problems with care and discuss whether for specific parameters they are easier to solve.

[4]For $g=2$ we obtain that the problem of a Hamiltonian path, which is a spanning path, is $\mathcal{NP}$-complete.

[5]Kleitman, West [155] give a partial answer:

Theorem 13.4.2 *Let $l(n,b)$ be the maximal number k such that each connected graph G with n vertices and $\delta(G)\geq b$ contains a spanning tree with at least b leaves. Then*
$$l(n,b)\leq n-3\frac{n}{\lceil b-1\rceil}+2, \tag{13.9}$$
and, conversely

a) $l(n,3)\geq n/4+2$,

b) $l(n,4)\geq(2n+8)/5$, *and asymptotically*

c) $l(n,b)\geq n\cdot(1-(c\ln b)/b)$, *where c is a desired chosen constant.*

[6]We will discuss this question below in its own section.

13.5 A recursive procedure

Another method to count the number of spanning trees of a graph is given by the following recursive procedure: Let $G = (V, E)$ be a graph and let e be an edge of G. $G - e$ denotes the graph after deleting the edge e, and $G \downarrow e$ denotes the contraction of G on e, that is the graph obtained from G by deleting e and then amalgamating its endvertices, where parallel edges may be produced. For practice consider the contracting $K_{3,3} \downarrow e$.

Theorem 13.5.1 *(Zykov; compare [109] or [258]) Let $G = (V, E)$ be a graph and denote the number of its spanning trees by $t(G)$. Then*

$$t(G) = t(G - e) + t(G \downarrow e), \tag{13.10}$$

where $e \in E$.

Proof. The number of spanning trees of G that do not contain e is $t(G - e)$ since each of them is also a spanning tree of $G - e$, and vice versa. On the other hand, the number of spanning trees that contain e is $t(G \downarrow e)$ because each of them corresponds to a spanning tree of $G \downarrow e$. □

13.5.1, together with 13.1.5 a), b) and c) as initial steps, creates a recursion to compute the number of spanning trees.[7]

As an example we consider fans, those are graphs G_n on the vertices $v_0, v_1, \ldots, v_n$ with $2n - 1$ edges defined as follows: v_0 is adjacent to each of the other vertices; and v_i is adjacent to v_{i+1} for $i = 1, \ldots, n-1$. $\mathrm{fan}(n)$ denotes the number of spanning trees for a fan with $n + 1$ vertices. Lets look at some small cases: $\mathrm{fan}(1) = 1$, $\mathrm{fan}(2) = 3$ and $\mathrm{fan}(3) = 8$.
To apply 13.5.1, we assume that $v_1, \ldots, v_n$ forms a path in this order. Then for $e = \underline{v_0 v_n}$ we

$$t(G - e) = \mathrm{fan}(n - 1)$$

and

$$t(G \downarrow e) = h_{n-1},$$

whereby h_n denotes the number of spanning trees for a "derived" fan with a multiple edge between v_0 and v_n. Simple to see by 13.5.1

$$h_{n-1} = \mathrm{fan}(n - 1) + h_{n-2}.$$

Altogether, and repeatedly applying 13.5.1

$$\mathrm{fan}(n) \quad = \quad \mathrm{fan}(n - 1) + \mathrm{fan}(n - 1) + h_{n-2}$$

[7] The running time of the algorithm is an exponential function in the number of edges, hence an exponential function in the square of the number of vertices, and so the algorithm is impractical for large and dense graphs.

$$
\begin{aligned}
&= \operatorname{fan}(n-1) + \operatorname{fan}(n-1) + \operatorname{fan}(n-2) + h_{n-3} \\
&\vdots \\
&= \operatorname{fan}(n-1) + \sum_{i=1}^{n-1} \operatorname{fan}(i).
\end{aligned}
$$

This is a recurrence that goes back through all previous values. We use a trick to compute $\operatorname{fan}(n)$ by a simpler one.

$$
\begin{aligned}
\operatorname{fan}(n) - \operatorname{fan}(n-1) &= \operatorname{fan}(n-1) + \sum_{i=1}^{n-1} \operatorname{fan}(i) - \left(\operatorname{fan}(n-2) + \sum_{i=1}^{n-2} \operatorname{fan}(i) \right) \\
&= 2 \cdot \operatorname{fan}(n-1) - \operatorname{fan}(n-2).
\end{aligned}
$$

Hence

$$
\operatorname{fan}(n) = 3 \cdot \operatorname{fan}(n-1) - \operatorname{fan}(n-2). \tag{13.11}
$$

n	1	2	3	4	5	6	7	8
$\operatorname{fan}(n)$	1	3	8	21	55	144	377	987

This sequence seems known, and indeed as an exercise prove

Theorem 13.5.2 *The number fan(n) of fan graphs on $n+1$ vertices satisfies (13.11) and it holds*

$$
\mathit{fan}(n) = f_{2n-1}, \tag{13.12}
$$

where f_n denotes the nth Fibonacci number.

A wheel W_n consists of a cycle with n vertices and an additional vertex that is adjacent to all of the vertices in the cycle. The formula for the number of spanning trees of the wheel with n spokes is difficult to derive, so we present the formula with the idea of the proof.

Theorem 13.5.3 *Let W_n, $n \geq 3$, be a wheel. Then*

$$
t(W_n) = \left(\frac{3+\sqrt{5}}{2} \right)^n + \left(\frac{3-\sqrt{5}}{2} \right)^n - 2. \tag{13.13}
$$

Sketch of the *proof.* The number $t(W_n)$ satisfies the recurrence relation

$$
t(W_n) = 3 \cdot t(W_{n-1}) - 3 \cdot t(W_{n-3}) + t(W_{n-4}). \tag{13.14}
$$

□

The formula is a little bit strange, since it handles with irrational numbers, although in any case $T(W_n)$ is an integer.

Since $W_3 = K_4$, the reader is invited to check the formula for $n = 3$.
We consider fan graphs and wheels. A fan graph is an interval graph, a wheel not; a fan graph is outer-planar, a wheel is only planar. An interesting observation we obtain when comparing wheels and fan graphs, which are differ in exactly one edge.[8]

n	1	2	3	4	5	6	7	8
$t(R_n)$	-	2	-	9	-	81	-	279
fan(n)	1	3	8	21	55	144	377	987
$t(W_n)$	1	3	16	45	121	318	831	

13.6 The matrix-tree theorem

We associate the following matrix $M(G) = (m_{ij})_{i,j=1,...,n}$ of admittance to a graph $G = (V, E)$ with $V = \{v_1, ..., v_n\}$:

$$m_{ij} = \begin{cases} g_G(i) & : \text{ if } i = j \\ -1 & : \text{ if the vertices } v_i \text{ and } v_j \text{ are adjacent} \\ 0 & : \text{ otherwise.} \end{cases}$$

That is

$$M(G) = \text{diag}(g_G(v_1), ..., g_G(v_n)) - A(G),$$

where $A(G)$ denotes the matrix of adjacency and $\text{diag}(g_G(v_1), ..., g_G(v_n))$ is the matrix which has the degrees of the graph on the diagonal and all other elements equal zero. It holds, of course:

$$\det M(G) = 0, \tag{13.15}$$

but surprisingly,

Theorem 13.6.1 *(Kirchhoff) Let G be a graph with the labeled vertices $1, ..., n$. Then the number of spanning trees of G is the determinant of the matrix obtained from the matrix of admittance $M(G)$ by deleting the i'th row and the i'th column for some i between 1 and n.*[9]

Proof. There is a very elegant proof using the Binet-Cauchy theorem of matrix-algebra: Let M_1 be an $r \times s$-matrix and M_2 be an $s \times r$-matrix, then $M_1 \cdot M_2$ is a quadratic $r \times r$-matrix such that $\det M_1 \cdot M_2$ equals the sum of the products of determinants of the $r \times r$-submatrices, where we take the same indices for the columns of M_1 and the rows of M_2. Compare [5].
Here, we will give another proof, created by Hutschenreuther, compare [209], which only uses very simple properties of matrices.
Let $A = (a_{ij})$ be an $n \times n$ matrix. We define the following operations for A:

- A_i is the matrix obtained from A by deleting the i'th row and the i'th column;

[8] In view of 13.1.5 it holds $t(G - e) \leq t(G)$. Is there constant c such that $t(G) \leq c \cdot t(G - e)$?
[9] In particular, the value of this determinant is independent of the choice of the number i.

- $A_j^\star$ is the matrix obtained from A by setting $a_{jj} = 0$ and $a_{lj} = 0$ for $l \neq j$; and
- A^k is the matrix obtained from A by setting $a_{kk} = a_{kk} - 1$.

Then the following facts are true:

$$\det A_i = \det A_i^\star. \tag{13.16}$$

For $k \neq i$ it holds

$$(A_i)^k = (A^k)_i, \tag{13.17}$$

shortly written as A_i^k.
For $k \neq i$ we write $A_{ik} = (A_i)_k$.
For $k \neq i$ it holds

$$\det A_i = \det(A_i)^k + \det(A_i)_k^\star. \tag{13.18}$$

This can be seen by the fact that $(A_i)^k$ and $(A_i)_k^\star$ only differs in the kth column; the sum of both columns equals the kth column of A_i.
Altogether,

$$\begin{aligned} \det A_i &= \det(A_i)^k + \det(A_i)_k^\star \quad \text{by (13.18)} \\ &= \det A_i^k + \det(A_i)_k^\star \quad \text{by (13.17)} \\ &= \det A_i^k + \det(A_i)_k \quad \text{by (13.16)} \\ &= \det A_i^k + \det A_{ik}, \end{aligned}$$

which gives

$$\det A_i = \det A_i^k + \det A_{ik}. \tag{13.19}$$

Now we complete the proof of the theorem by induction over the number n of vertices and the number m of edges.

The theorem is true for an edgeless graph, since such graph has no spanning tree and its admittance matrix is the zero matrix.

The theorem is true for a graph with two vertices, since the number of its spanning trees equals m, and the admittance matrix is given by

$$M = \begin{pmatrix} m & -1 \\ -1 & m \end{pmatrix}. \tag{13.20}$$

Now, assume that the theorem is true for all graphs with less than n vertices and m edges.

Let $G = (V, E)$ with $|V| = n$ and $|E| = m$ be given. Consider a vertex $v \in V$.

Case 1: v is an isolated vertex.
Then $t(G) = 0$, and $\det M_i = 0$, when i denotes the index of v.

Case 2: There is an edge e incident with v.
i and k denote the subscript of v and its neighbor. Then, in view of the induction assumption,

$$\begin{aligned} t(G \downarrow e) &= \det M(G \downarrow e)_k \\ &= \det(M_i)_k \\ &= \det M_{ik}. \end{aligned} \tag{13.21}$$

On the other hand, also using the induction assumption,

$$\begin{aligned} t(G - e) &= \det M(G - e)_i \\ &= \det(M_i)^k \\ &= \det M_i^k \quad \text{by (13.17).} \end{aligned} \tag{13.22}$$

Altogether,

$$\begin{aligned} t(G) &= t(G \downarrow e) + t(G - e) \quad \text{in view of 13.5.1} \\ &= \det M_{ik} + \det M_i^k \quad \text{by (13.21) and (13.22)} \\ &= \det M_i \quad \text{by (13.19).} \end{aligned}$$

This is the assertion. □

Given a graph G. The number of spanning trees grows exponentially in the number of vertices. But the theorem shows that it is possible to find the quantity $t(G)$ in polynomially bounded time. This is one of the few enumeration problems which has such property.

Suppose that the eigenvalues of $M(G)$ are equal to $\lambda_1, \ldots, \lambda_n$. Since $\det M(G) = 0$, we assume that $\lambda_n = 0$.[10] In view of

$$\det(M(G) - xI) = -x\Pi_{j=1}^{n-1}(\lambda_j - x),$$

we see that for M', which is the matrix $M(G)$ with the i'th row and column removed, it holds

$$(-1)^{i+j+1} \cdot n \cdot \det M' = \lambda_1 \cdots \lambda_{n-1}$$

and find an equivalent statement to 13.6.1.

Theorem 13.6.2 *(Stanley [227]) Let G be a graph with the n labeled vertices and the matrix of admittance $M(G)$. If $M(G)$ has the eigenvalues $\lambda_1, \ldots, \lambda_n$, with $\lambda_n = 0$, then*

$$t(G) = \frac{1}{n} \cdot \lambda_1 \cdots \lambda_{n-1}. \tag{13.23}$$

[10]Note: The geometric multiplicity $g(\lambda)$ of an eigenvalue λ is the dimension of its eigenspace. The arithmetic multiplicity $a(\lambda)$ is the multiplicity of λ as the zero of the characteristic polynomial. In any case $a(\lambda) \leq g(\lambda)$, but for symmetric matrices equality holds.

13.7 Applications

We give several applications of the matrix-tree theorem, showing the power of this approach.

I. As exercise prove 13.1.4(b), namely $t(K_n) = n^{n-2}$ again, by investigating

$$M(K_n) = \begin{pmatrix} n-1 & -1 & \dots & -1 \\ -1 & n-1 & \dots & -1 \\ \dots & \dots & & \dots \\ -1 & -1 & \dots & n-1 \end{pmatrix}. \tag{13.24}$$

We can do it on two ways:
1. applying 13.6.1 with calculations of the determinant of $M(G)$ with the i-th row and column removed; and
2. applying 13.6.2 showing that $M(G)$ has the eigenvalues n ($(n-1)$-times) and 0 (once).

II. Similar, considering the complete bipartite graph

$$M(K_{n_1,n_2}) = \begin{pmatrix} n_1 & 0 & \dots & 0 & -1 & -1 & \dots & -1 \\ 0 & n_1 & \dots & 0 & -1 & -1 & \dots & -1 \\ \dots & \dots & & \dots & \dots & \dots & & \dots \\ 0 & 0 & \dots & n_1 & -1 & -1 & \dots & -1 \\ -1 & -1 & \dots & -1 & n_2 & 0 & \dots & 0 \\ -1 & -1 & \dots & -1 & 0 & n_2 & \dots & 0 \\ \dots & \dots & & \dots & \dots & \dots & & \dots \\ -1 & -1 & \dots & -1 & 0 & 0 & \dots & n_2 \end{pmatrix} \tag{13.25}$$

we find the following result.

Theorem 13.7.1 *Let K_{n_1,n_2} be the complete bipartite graph with $n_1 + n_2$ vertices. Then*

$$t(K_{n_1,n_2}) = n_2^{n_1-1} \cdot n_1^{n_2-1}. \tag{13.26}$$

We are interested in specifications and generalizations of this theorem.

A consequence of 13.7.1 is simple to prove.

Corollary 13.7.2 *Let $K_{2,n}$ be the complete bipartite graph with $2+n$ vertices. Then*

$$t(K_{2,n}) = n \cdot 2^{n-1}. \tag{13.27}$$

Proof. The graph $K_{2,n}$ is a bipartite graph with two blue vertices v and w, and n red vertices. In any spanning of $K_{2,n}$, the unique path between v and w has length two. We can now count, there are n ways to choose the internal vertex, and for each of the remaining $n-1$ red vertices, there are two choices, either to be adjacent to v

or w. This gives us a total of $n2^{n-1}$ different spanning trees. □

As an exercise prove that the number of non-isomorphic spanning trees is essentially less:

Theorem 13.7.3 *The number of non-isomorphic spanning trees in $K_{2,n}$ is*

$$\left\lfloor \frac{n+1}{2} \right\rfloor. \tag{13.28}$$

This is a complete other situation than for K_n. There are n^{n-2} different spanning trees, and also an exponential number of non-isomorphic ones.

On the other hand, the complete multi-partite graph $K_{n_1,\ldots,n_r}$ denotes the graph $G = (V, E)$ which is defined by a partition of V in subsets V_i with $|V_i| = n_1$, $i = 1, \ldots, r$, such that all edges connecting only distinct V_i and V_j:

$$E = \bigcup_{i \neq j} \{\underline{vw} : v \in V_i, w \in V_j\}.$$

Theorem 13.7.4 *(Onadera [183]) For the number of spanning trees of the complete multi-partite graph it holds*

$$t(K_{n_1,\ldots,n_r}) = n^{r-2} \cdot \Pi_{i=1}^{r} (n - n_i)^{n_i - 1}, \tag{13.29}$$

where $n = |V| = \sum_{i=1}^{r} n_i$.

III. What is the probability for a given edge to be a member in a spanning tree in the complete graph K_n?
First we know $t(K_n) = n^{n-2}$. With help of 13.6.1 we find

$$t(K_n - e) = (n-2)n^{n-3},$$

where e denotes a specific edge. Consequently, the number α of spanning trees of K_n which contain e equals

$$\alpha = t(K_n) - t(K_n - e) = n^{n-2} - (n-2)n^{n-3} = 2n^{n-3}.$$

Then α / n^{n-2} is the probability that e is in a spanning tree of K_n:

Theorem 13.7.5 *The probability for a given edge in the complete graph of n vertices to be a member in a spanning tree equals $2/n$.*

An extension of the calculation for $t(K_n) - t(K_n - e)$ is given in

Theorem 13.7.6 *(Weinberg) Let $E \subseteq \binom{V}{2}$, $|V| = n$ be a set of m pairwise disjoint edges, that means in particular $m \leq n - 1$. Then the number of different trees on V that do not contain any edge in E equals*

$$n^{n-2} \cdot \left(1 - \frac{2}{n}\right)^m. \tag{13.30}$$

For a proof and further generalizations see [25].
Another property of a randomly chosen tree is given by the following considerations: Let T be a spanning tree of the K_n, then the diameter of T is at least 2, at most $n-1$ and in the average case

Theorem 13.7.7 *(Renyi, Szekeres [201]) (In the model $\mathcal{G}_{n,p}$ it holds:) The diameter of a randomly chosen spanning tree of K_n is of order $\sqrt{n}$.*

13.8 Cubes, Grids, Ladders

I. We consider hypercubes Q^D.
Q^3 has 8 vertices and the average degree 3. In view of 13.2.2 we have $t(Q^3) \leq 696$. Actually, $t(Q^3) = 384$. This is a consequence of the following theorem.

Theorem 13.8.1 *The D-dimensional hypercube has*

$$t(Q^D) = \prod_{W \subseteq \{1,\ldots,D\}, |W| \geq 2} 2|W| = 2^{2^D - D - 1} \cdot \prod_{k=1}^{D} k^{\binom{D}{k}} \tag{13.31}$$

spanning trees.

The *proof* uses the matrix-tree theorem in the form 13.6.2 and several deep results about the eigenvalues of matrices, see [227]. □

The formula gives an superexponential growing[11]:

$D =$	$n =$	$t(Q^D) =$
1	2	1
2	4	4
3	8	384
4	16	42,467,328
5	32	256,494,072,527,585,280

II. The set

$$L_{p,q} = \{1, \ldots, p\} \times \{1, \ldots, q\} \tag{13.32}$$

is called an $p \times q$-grid. As an exercise the reader should prove that $L_{p,q}$ is Hamiltonian if and only if $p \cdot q$ is an even number.
For practice determine $t(L_{p,q})$ for some small values of p and q. For instance, $t(L_{2,3}) = 15$ and $t(L_{3,3}) = 192$.

III. $L_{1,q}$ is a path, such that $t(L_{1,q}) = 1$. $L_{2,q}$ denotes a ladder. Using 13.5.1 we find

$$t(L_{2,q}) = 4 \cdot t(L_{2,q-1}) - t(L_{2,q-2}), \tag{13.33}$$

[11] in the dimension of the hypercube, but not in the number of its vertices

which solution is

$$t(L_{2,q}) = \frac{1}{2\sqrt{3}}\left((2+\sqrt{3})^q - (2-\sqrt{3})^q\right). \tag{13.34}$$

Numerically,

$q =$	1	2	3	4	5	6	7	8
$t(L_{2,q}) =$	1	4	15	56	209			

The circular ladder $CL_{2,q}$ is ladder in which the ends are joined. More exactly, it consits of two concentric cycles with q vertices each in which every pair of corresponding vertices is joined by an edge. For practice the reader should discuss that $t(CL_{2,3}) = 75$ and $t(CL_{2,4}) = 384$.

Related questions about the number of spanning trees in graphs are discussed in [100], [127] and [180].

13.9 Spanning Tree Numbers

13.1.5 shows that for each nonnegative integer t, except 2, there exists a graph G with $t(G) = t$.[12] For each positive integer n we define

$$\Upsilon(n) = \{t : \text{there is a graph } G \text{ with } n \text{ vertices and } t(G) = t\}. \tag{13.35}$$

We have for $n > 2$:

a) $0, 1, 3, \ldots, n \in \Upsilon(n)$.

b) $n^{n-2} \in \Upsilon(n)$.

c) $(n-2)n^{n-3} \in \Upsilon(n)$.

d) $\Upsilon(n) \subset \Upsilon(n+1)$.

a) and b) are obvious; c) and d) remains as an exercise for the reader.[13] Assume that the numbers in $\Upsilon(n)$ are ordered: $t_0 < t_1 < t_2 < \ldots < t_r$. Then Sedlacek [218] reported:

$$\begin{aligned}
t_r &= n^{n-2} \\
t_{r-1} &= (n-2)\cdot n^{n-3} \\
t_{r-2} &= (n-2)^2\cdot n^{n-4} \\
t_{r-3} &= (n-1)(n-3)\cdot n^{n-4} \\
t_{r-4} &= (n-2)^3\cdot n^{n-5} \\
t_{r-5} &= (n-1)(n-2)(n-3)\cdot n^{n-5}
\end{aligned}$$

[12] Why $t = 2$ is impossible?

[13] Hint: for c) compare the proof of 13.7.5.

$$
\begin{aligned}
t_{r-6} &= (n-2)(n^2-4n+2)\cdot n^{n-5} \\
t_{r-7} &= (n-3)^2\cdot n^{n-4} \\
t_{r-8} &= (n-1)^2(n-4)\cdot n^{n-5} \\
t_{r-9} &= (n-2)^4\cdot n^{n-6}
\end{aligned}
$$

Of course, there are non-isomorphic graphs with the same number of spanning trees. Let $g(n)$ be the number of non-isomorphic graphs with n vertices. Then $|\Upsilon(n)| \leq g(n)$.

$n=$	members of $\Upsilon(n)$	$\lvert\Upsilon(n)\rvert =$	$g(n) =$
1	1	1	1
2	0,1	2	2
3	and 3	3	4
4	and 4,8,16	6	11
5	and 5,9,11,12,20,21,24,40,45,75,125	17	34

13.10 Arboricity

One of the most common question in graph theory deals with decompositions of a graph into various subgraphs possessing some prescribed property. Here we decompose a graph in a union of simpler ones, where the union of two graphs $G = (V, E)$ and $G' = (V, E')$ is defined by $G \cup G' = (V, E \cup E')$.

I. Any graph G can be expressed as the union of spanning forests. A natural problem is to determine the minimum number of edge-disjoint spanning forests into which G can be decomposed. This number is called the arboricity of the graph, written as $\mu(G)$.[14]

Assume that G has n vertices and m edges. The maximum number of edges in a spanning forest is $n-1$. Consequently, the minimum number of spanning forests which composed G is at least $m/(n-1)$. Since the arboricity is an integer, we have

Lemma 13.10.1

$$\mu(G) \geq \left\lceil \frac{m}{n-1} \right\rceil . \tag{13.36}$$

In terms of edges,

$$\mu(G) \geq \left\lceil \frac{m}{m-\nu(G)} \right\rceil , \tag{13.37}$$

where $\nu(G)$ denotes the cyclomatic number of G, compare 7.4.1.
Let G' be a subgraph of G then $\mu(G') \leq \mu(G)$. Together with 13.10.1 this is the basic of the (nontrivial) proof of

[14] This is a new quantity measured the density of a graph. We already defined: Diameter, minimum degree, connectedness, and density itself. It is an interesting question to discuss the interrelation between these numbers.

Theorem 13.10.2 *(Nash-Williams [178]) Let G be a graph with n vertices and let $m(k)$ be the maximum number of edges in any subgraph of G having k vertices. Then*

$$\mu(G) = \max_k \left\lceil \frac{m(k)}{n-1} \right\rceil . \tag{13.38}$$

As an exercise prove the following facts for the arboricity of the complete graph and complete bipartite graph.

Corollary 13.10.3

$$\begin{aligned} \mu(K_n) &= \left\lceil \frac{n}{2} \right\rceil . \\ \mu(K_{n_1,n_2}) &= \left\lceil \frac{n_1 \cdot n_2}{n_1 + n_2 - 1} \right\rceil . \end{aligned}$$

In a specific case the decomposition is extremely simple:

Theorem 13.10.4 *(Beineke [20]) Assume that n is a even number. Then K_n can be decomposed into $n/2$ spanning paths.*

Is there a nice description for the arboricity of planar graphs?[15]

II. The arboricity of a graph G was the minimum number of forests whose union covers G. A star forest is a forest in which each component is a star. The star arboricity $\mu^\star(G)$ of G is the minimum number of star forests whose union covers G.

It follows from the definition that $\mu(G) \leq \mu^\star(G)$. Since every tree can be decomposed into two star forests (Exercise), we obtain

Theorem 13.10.6 $\mu^\star(G) \leq 2 \cdot \mu(G)$ *for every graph G.*

Kurek [160] shows that equality is possible, that is, for any natural k there exists a graph $G = G(k)$ with $\mu(G) = k$ and $\mu^\star(G) = 2k$.

[15] Conversely, when we are interested in the smallest number of planar graphs into which a graph G can be decomposed. This number is called the thickness of the graph, written as $\theta(G)$.
Exact in the same way as 13.10.1, we prove

Theorem 13.10.5

$$\theta(G) \geq \left\lceil \frac{m}{3n-6} \right\rceil . \tag{13.39}$$

A formula for the thickness of the complete graph is not easy to find: On the one hand, using 13.10.5, we can derive that

$$\theta(K_n) \geq \left\lceil \frac{n(n-1)}{6(n-2)} \right\rceil = \left\lfloor \frac{n(n-1)+6(n-2)-2}{6(n-2)} \right\rfloor = \left\lfloor \frac{(n+7)(n-2)}{6(n-2)} \right\rfloor = \left\lfloor \frac{n+7}{6} \right\rfloor . \tag{13.40}$$

On the other hand, it is hard to show that in almost all cases equality holds, compare [127]:

$$\theta(K_n) = \begin{cases} \left\lfloor \frac{n+7}{6} \right\rfloor & : \quad n \neq 9, 10 \\ 3 & : \quad n = 9, 10 \end{cases}$$

Chapter 14

Coloring of graphs

Suppose that the vertices of a graph represent different kinds of chemicals in some manufactoring process. For each pair of these chemicals that might explode if combined, there is an edge between the corresponding vertices. Label each chemical with a color such that those which can be combined without exploding have the same color.[1]

14.1 Vertex coloring

To color the vertices of a graph G is to assign a color to each vertex in such a way that no two adjacent vertices have the same color. In general, but not exclusively, we will refer to the colors by natural numbers $1, 2, \ldots$.
We can reformulate coloring in another way. Let $G = (V, E)$ be a graph colored with k colors. Consider the sets $V_i = \{v \in V : \text{color of } v = i\}$, $i = 1, \ldots, k$. These sets forms a partition of V, coming from the equivalence relation: $v \sim v'$ if and only if v not adjacent to v'.[2]

14.2 The number of colored and labeled graphs

Let $G = (V, E)$ be a labeled graph colored with k colors. Two such graphs are isomorphic if there is a bijective mapping between the sets of vertices which preserves not only adjacency but also the colors. In other terms, we also count the number

[1] One of the origin of graph theory is the four-color problem, which was a long-standing problem dates back to 1852 when Guthrie tried to color a map of the countries of England, and become aware of the fact that although the question seems to be very simple and finding the answer is very hard. In coloring a "geographical" map it is customary to give different colors to any two countries that have a part of their boundary in common. Guthrie attempted to prove that the countries (regions) of any (!) map could be colored with four colors. This was repeatedly misproven, but finally proved correctly in 1976 with a combination of graph theory and sophisticated computing.

[2] The sets V_i are also called independent

of color-classes. The number $c(n,k)$ denotes the number of k-colored labeled graphs with n vertices. Note

Observation 14.2.1 *The number of k-colored labeled graphs with n vertices in which the colors have fixed identities is $k!c(n,k)$.*

Remember that the the collection of V_i, $i = 1, ldots, k$ forms a partition of V. Let $n_i = |V_i| \geq 1$, then

$$\sum_{i=1}^{k} n_i = n. \tag{14.1}$$

Conversely, each solution of 14.1 determines a partition with k parts of a set of n vertices. In view of C.3.1 the number of ways that the labels can be selected for the vertices is the multinomial coefficient

$$\binom{n}{n_1 n_2 \dots n_k}.$$

Obviously, there are

$$\binom{n}{2} - \sum_{i=1}^{k} \binom{n_i}{2} = \frac{1}{2} \cdot \left(n^2 - \sum_{i=1}^{k} n_i^2 \right) \tag{14.2}$$

pairs of vertices of different colors. Since each of these pairs may or may not be adjacent we use 2.1.1 to obtain for the number of graphs with n_i vertices of color i precisely

$$\binom{n}{n_1 n_2 \dots n_k} 2^{(n^2 - \sum_{i=1}^{k} n_i^2)/2}. \tag{14.3}$$

Summing over all solutions of (14.1) and paying attention 14.2.1 we obtain

Theorem 14.2.2 *(Read [198]*

$$c(n,k) = \frac{1}{k!} \sum_{(14.1)} \binom{n}{n_1 \dots n_k} 2^{(n^2 - \sum_{i=1}^{k} n_i^2)/2}. \tag{14.4}$$

The equation in 14.2.2 creates the following triangle:

$n \setminus k$	1	2	3	4	5	6	7
1	1						
2	1	2					
3	1	12	8				
4	1	80	192	64			
5	1	720	5,120	5120	1024		
6	1	9,152	192,000	450,560	24,576	32,768	
7	1	165,312	10,938,368	59,197,120	64,225,280	22,020,096	2,097,152

14.2.2 also suggests a recursive formula for $c(n,k)$ (Exercise).
Consider the second column. Here we find the numbers which was deduced as upper bounds for the number of bipartite graphs in 5.2.2. This is not a surprise, since the proofs are similar.

Corollary 14.2.3 *the coefficient of x^m in*

$$\frac{1}{k!}\sum_{(14.1)}\binom{n}{n_1\dots n_k}(1+x)^{(n^2-\sum_{i=1}^{k}n_i^2)/2}. \tag{14.5}$$

is the number of k-colored labeld graphs with n vertices and m edges.

14.3 The chromatic number

I. The chromatic number $\chi(G)$ of a graph G is the smallest value of k for which the vertices of G can be colored by k colors. This is not a simple question. For instance determine the chromatic number of the Petersen graph G_{petersen}. As an exercise show the following facts:

a) Let G be a graph with the components $G_1,\dots,G_m$. Then
$\chi(G)=\max\{\chi(G_1),\dots,\chi(G_m)\}$.

b) $\chi(K_n)=n$.

c) A graph G is bipartite if and only if $\chi(G)=2$.

d) It holds $\chi(T)=2$ for any tree T.

e) Let C_n be a cycle of length n, then

$$\chi(C_n)=\begin{cases}2 & : \ n \text{ even}\\ 3 & : \ \text{otherwise}\end{cases}$$

f) Let K_r, the complete graph with r vertices, be a subgraph of G. Then $\chi(G)\geq r$.

It seems natural that there should be a good bound of the chromatic number in terms of the size of the largest complete subgraph. The following theorem shows that this approach fails.

Theorem 14.3.1 *(Mycielsky, see [5]) For any positive integer r, there exists a triangle-free graph G, with $\chi(G)=r$. And furthermore G does not contain any K_s for $s>3$.*

The only vertex coloring problem that can be nicely characterized is concerned with two-colorable graphs, called bichromatic. Such graphs can be divided into two sets such that all edges are between a vertex in one set and a vertex in the other set. We called such graphs bipartite. Consequently, in view of 4.1.2 we have,

Theorem 14.3.2 *A graph can be two-colored if and only if it does not contain a cycle of odd length.*

In other words, bipartite and bichromatic are equivalent terms.

II. We derive several interrelations between the chromatic number and the size of a graph. By definition, $\chi(G) \leq |V|$. That

$$|E| \geq \binom{\chi(G)}{2}, \tag{14.6}$$

or equivalently,

$$\chi(G) \leq \frac{1}{2} + \sqrt{2|E| + \frac{1}{4}}, \tag{14.7}$$

holds true is obviously. But the next inequality not.

Theorem 14.3.3 *For any graph $G = (V, E)$*

$$\chi(G) \geq \frac{|V|^2}{|V|^2 - 2|E|}. \tag{14.8}$$

Proof. Let $\chi = \chi(G)$.

$$V_i = \{v \in V : v \text{ colored with } i\} \tag{14.9}$$

denote the sets of vertices colored with the same color. Let $n_i = |V_i|$ for $i = 1, \ldots, \chi$. Consider the adjacency matrix $A(G)$ of G. Let

$$\begin{aligned} N_0 &= \text{number of 0 entries in } A(G) \text{ and} \\ N_1 &= \text{number of 1 entries in } A(G). \end{aligned}$$

a) Each set V_i induces a submatrix of $A(G)$ the entries of which are all 0. Thus

$$N_0 \geq n_1^2 + \ldots + n_\chi^2. \tag{14.10}$$

Applying the Cauchy-Schwarz inequality (G.2) to $(n_1, \ldots, n_\chi)$ and $(1, \ldots, 1)$ gives

$$n_1^2 + \ldots + n_\chi^2 \geq \frac{(n_1 + \ldots + n_\chi)^2}{\chi} = \frac{|V|^2}{\chi}. \tag{14.11}$$

b) In view of 4.1.1

$$N_1 = 2|E|. \tag{14.12}$$

Thus, the total number of entries in $A(G)$ satisfies

$$|V|^2 = N_0 + N_1 \geq 2|E| + \frac{|V|^2}{\chi}.$$

This gives the required formula. □

III. We estimate the chromatic number of a graph and its complement.

Theorem 14.3.4 *Let G be a graph with n vertices. Then*

$$\begin{aligned} 2\sqrt{n} &\leq \chi(G)+\chi(G^c) && (14.13)\\ \chi(G)+\chi(G^c) &\leq n+1 && (14.14)\\ n &\leq \chi(G)\cdot\chi(G^c) \quad \textit{and} && (14.15)\\ \chi(G)\cdot\chi(G^c) &\leq \left(\frac{n+1}{2}\right)^2. && (14.16) \end{aligned}$$

Proof. Let $F : V \to \{1,\ldots,\chi(G)\}$ be a coloring of $G = (V, E)$.

$$V_i = \{v \in V : F(v) = i\}, \tag{14.17}$$

for $i = 1,\ldots,\chi(G)$. Let $n_i = |V_i|$. Then $n_i > 0$, and

$$n = |V| = \sum_{i=1}^{\chi(G)} n_i \leq \sum_{i=1}^{\chi(G)} \max_{j=1,\ldots,\chi(G)} n_j = \chi(G)\cdot \max_{j=1,\ldots,\chi(G)} n_j.$$

Hence,

$$\chi(G) \geq \frac{n}{\max_{j=1,\ldots,\chi(G)} n_j}. \tag{14.18}$$

If $v, v' \in V_i$, then they have the same color. Consequently they cannot be adjacent in G, and must be adjacent in G^c. In other terms, K_{n_i} is a subgraph of G^c. Thus $\chi(G^c) \geq n_i$ for all $i = 1,\ldots,\chi(G)$, which implies

$$\chi(G^c) \geq \max_{j=1,\ldots,\chi(G)} n_j. \tag{14.19}$$

Together with (14.18) we get the inequality (14.15).
We state, as a consequence of G.3.2, that

$$(\chi(G)+\chi(G^c))^2 \geq 4\chi(G)\chi(G^c). \tag{14.20}$$

In view of (14.15) we have $(\chi(G)+\chi(G^c))^2 \geq 4n$, which gives (14.13).
To show the two other inequalities we use induction over the number n of vertices. For $n = 2$ the inequalities are obvious. Let G be a graph with n vertices, and let v be a vertex of G. It is not hard to see that

$$\chi(G-v)+1 \geq \chi(G) \quad \text{and} \quad \chi((G-v)^c)+1 \geq \chi(G^c).$$

We distinguish two cases:
1. In one of the inequalities there is "<".

$$\begin{aligned} \chi(G)+\chi(G^c) &\leq \chi(G-v)+1+\chi((G-v)^c)+1-1\\ &\leq \chi(G-v)+\chi((G-v)^c)+1\\ &\leq (n-1)+1+1 \quad \text{using the induction assumption}\\ &= n+1. \end{aligned}$$

2. Both inequalities are equalities.
The following must be true considering the vertex v:

$$\begin{aligned} g_G(v) &\geq \chi(G-v), \quad \text{and} && (14.21)\\ g_{G^c}(v) &\geq \chi((G-v)^c), && (14.22) \end{aligned}$$

otherwise we can color G and G^c with fewer than $\chi(G)$ and $\chi(G^c)$ colors, respectively.
Altogether,

$$\begin{aligned} \chi(G)+\chi(G^c) &= \chi(G-v)+1+\chi((G-v)^c)+1\\ &\leq g_G(v)+1+g_{G^c}(v)+1\\ &= (n-1)+1+1=n+1. \end{aligned}$$

In both cases we obtain the third desired inequality (14.14).
(14.14) and (14.20) give the last of the desired inequality. □

14.4 Spanning trees of colored graphs

The chromatic number can be understand as a measure of density. In this sense we consider a graph G with n vertices, m edges and chromatic number χ.
On one hand, (14.6) implies

$$d=\frac{2m}{n(n-1)}\geq\frac{\chi(\chi-1)}{n(n-1)}$$

for the density d of G.
On the other hand, (14.8) is equivalent to

$$\frac{2m}{n^2}\leq 1-\frac{1}{\chi}. \tag{14.23}$$

Additionally,

$$\frac{2m}{n^2}=\frac{2m}{n(n-1)}\cdot\frac{n-1}{n}=d\cdot\frac{n-1}{n}.$$

Altogether,

Lemma 14.4.1 *Let G be a graph with n vertices and of chromatic number χ. Then*

$$\frac{\chi(\chi-1)}{n(n-1)}\leq \textit{density}(G)\leq\frac{n}{n-1}\cdot\frac{\chi-1}{\chi}. \tag{14.24}$$

The chromatic number is a global parameter of a graph, as well as the average degree. For practice discuss the interrelation of both quantities.

We have

$$\frac{n-1}{n} \leq \frac{\chi}{\chi-1} \tag{14.25}$$

$(1+1/n)^{n+1}$ is a monotone decreasing sequence such that for $\chi \leq n$

$$\left(\frac{\chi}{\chi-1}\right)^{\chi} \geq \left(\frac{n}{n-1}\right)^{n}. \tag{14.26}$$

14.4.1 in 13.2.2 gives for the number of spanning trees

$$\begin{aligned}
t(G) &\leq n^{n-2} \cdot \left(\frac{n}{n-1} \cdot \frac{\chi-1}{\chi}\right)^{n-1} \\
&= n^{n-2} \cdot \left(\frac{n}{n-1}\right)^{n} \cdot \frac{n-1}{n} \cdot \left(\frac{\chi-1}{\chi}\right)^{n-1} \\
&\leq n^{n-2} \cdot \left(\frac{\chi}{\chi-1}\right)^{\chi} \cdot \frac{n-1}{n} \cdot \left(\frac{\chi-1}{\chi}\right)^{n-1} \quad \text{by (14.26)} \\
&\leq n^{n-2} \cdot \left(\frac{\chi}{\chi-1}\right)^{\chi} \cdot \frac{\chi-1}{\chi} \cdot \left(\frac{\chi-1}{\chi}\right)^{n-1} \quad \text{by (14.25)} \\
&= n^{n-2} \cdot \left(\frac{\chi-1}{\chi}\right)^{n-\chi}.
\end{aligned}$$

Hence,

Theorem 14.4.2 *Let G be a graph with n vertices and of chromatic number χ. Then*

$$t(G) \leq n^{n-2} \cdot \left(1 - \frac{1}{\chi}\right)^{n-\chi}. \tag{14.27}$$

Since bichromatic and bipartite are equivalent terms, this theorem implies

Corollary 14.4.3 *Let G be a bipartite graph with n vertices.*

$$t(G) \leq \left(\frac{n}{2}\right)^{n-2}. \tag{14.28}$$

14.5 Algorithms for coloring

It is easy to decide whether a graph is bicromatic using a naive coloring procedure. But this is the only simple case: The Graph 3-colorability problem is $\mathcal{NP}$-complete, see Stockmeyer [229].

I. An exact algorithm is given by the following theorem.

Theorem 14.5.1 *(Zykov) Let $G = (V, E)$ be a graph with two non-adjacent vertices v and w. Then*

$$\chi(G) = \min\{\chi(G + \underline{vw}), \chi(G \downarrow \underline{vw})\}. \tag{14.29}$$

Proof. $\leq$: Each coloring of $G + e$ is also a coloring of G. Each coloring of $G \downarrow e$ induces also a coloring of G given that v and w get the same color which ther also the color of the vertex representing v and w.
$\geq$: Let $F : V \to \{1, \ldots, \chi(G)\}$ be an optimal coloring of G. If $F(v) \neq F(w)$ then is F also a coloring of $G + e$. Otherwise, $F(v) = F(w)$, and F is a coloring of $G \downarrow e$. □

As an exercise use this theorem to create an algorithm for finding the chromatic number.[3]

II. There is no easy way of finding the chromatic number of a graph. Often, but not in general, a good upper bound is given by the following considerations.

Lemma 14.5.2

$$\chi(G) \leq \Delta(G) + 1, \tag{14.30}$$

where $\Delta(G)$ is the maximum degree of a vertex in G.

The *proof* is given by induction.
Brook characterized the equality in 14.5.2 completely: If the graph G is not an odd cycle or a complete graph, then $\chi(G) \leq \Delta(G)$, where $\Delta(G)$ is the maximum degree of a vertex in G.
We will omit the proof, but we will give the following greedy algorithm which creates a vertex coloring of G with at most $\Delta(G) + 1$ colors.

Algorithm 14.5.3 *Let $G = (V, E)$ be a graph. Consider the following algorithm:*

1. *List the vertices in some order: $V = \{v_1, v_2, \ldots, v_n\}$;*
2. *Assign color 1 to v_1;*
3. *for $i := 2$ to n do:*
   assign to v_i the color j as small as possible which has not yet been used to color a vertex adjacent to v_i.

This algorithm is far from being optimal, it depends essentially on the way the vertices are ordered. As an exercise find graphs G for which $\chi(G)$ and $\Delta(G)$ differ significantly.

[3] Another approach was created by Christophides using independent sets:
for $k := 1$ **to** n **do**
for all $W \in \binom{V}{k}$ **do**

$$\chi(G[W]) := 1 + \min\{\chi(G[W \setminus S]) : S \text{ an independent set with } |S| = \beta(G[W])\}.$$

where $\beta(G)$ denotes the independence number of G.

14.6 Chromatic polynomials

Two colorings of a labeled graph G are considered different if they assign different colors to the same vertex in G. The chromatic polynomial $p(G,t)$ of G is the number of different colorings of G that use t or fewer colors.
If $t < \chi(G)$, then $p(G,t) = 0$. In fact, the smallest positive integer t such that $p(G,t) > 0$ is the chromatic number:

$$\chi(G) = \min\{t : p(G,t) > 0\}. \tag{14.31}$$

The following lemma is easy to see.

Lemma 14.6.1 *Let G be a graph with the isolated vertex v. Then*

$$p(G,t) = t \cdot p(G-v,t). \tag{14.32}$$

Consider the complete graph K_n. If we color this graph with $t \geq n$ colors then there are t choices for the color of the first vertex. For each such choice, there are $t-1$ choices for the second vertex. Since a third vertex is adjacent to both the first and second vertex, we have $t-2$ choices for the color of this vertex. And so on. For the chromatic polynomial of K_n^c we use 14.6.1.

Theorem 14.6.2 *Let n and t be positive integers. Then*

$$p(K_n,t) = t(t-1)\cdots(t-n+1). \tag{14.33}$$

And

$$p(K_n^c,t) = t^n. \tag{14.34}$$

The following theorem provides an expression for $p(G,t)$ for graphs G in general.

Theorem 14.6.3 *Let G be a graph and $p(G,t)$ the chromatic polynomial of G. Then*

$$p(G,t) = p(G-e,t) - p(G \downarrow e,t), \tag{14.35}$$

where $e \in E$.

Proof. Let $e = \underline{vv'} \in E$. Consider the coloring $F : V \to \{1,2,\ldots,t\}$ of $G-e$.
If $F(v) \neq F(v')$ there is also a coloring of G.
Otherwise, if $F(v) = F(v')$, we get a coloring of $G \downarrow e$. Hence,

$$p(G-e,t) = p(G,t) + p(G \downarrow e,t).$$

□

14.6.3 suggests a way of finding the chromatic polynomial $p(G,t)$ of a graph G: From G we construct two graphs. One with the same number of vertices and one more edge, and the other with one vertex less than G. We can continue this procedure provided no graph produced in this process is complete or empty. Applying 14.6.2 we find a justification for the name "polynomial".

Theorem 14.6.4 *A graph G with n vertices is a tree if and only if*

$$p(G,t) = t(t-1)^{n-1}. \tag{14.36}$$

It is not known in general, however, what properties graphs G and G' must possess for $p(G,t) = p(G',t)$. 14.6.4 shows that non-isomorphic graphs may have the same chromatic polynomial, since there are non-isomorphic trees with the same number of vertices.

14.7 Edge coloring

Let $G = (V,E)$ be a graph. An edge-coloring of a graph is a mapping from E into a finite set, the colors, such that for each vertex all incident edges have different colors. The chromatic index $\chi'(G)$ of a graph G is the smallest value of k for which the edges of G can be colored by k colors.

Let $G = (V,E)$ be a graph with maximum degree $\Delta(G)$, then obviously $\chi'(G) \geq \Delta(G)$, but surprisingly

Theorem 14.7.1 *(Vizing) Let $G = (V,E)$ be a graph with maximum degree $\Delta(G)$, then $\chi'(G) = \Delta(G)$ or $\chi'(G) = \Delta(G) + 1$.*

We omit the proof. You can find one in [109]. Consequently, there is a polynomially bounded algorithm which finds an edge-coloring of a graph G with $\Delta(G)+1$ colors. But surprisingly, to decide whether $\chi'(G) = \Delta(G)$ is $\mathcal{NP}$-complete, [135].
For practice the reader should discuss the following facts:

- Let K_n be the complete graph with n vertices, then

$$\chi'(K_n) = \begin{cases} n & : \ n \text{ odd} \\ n-1 & : \ n \text{ even} \end{cases}$$

- $\chi'(G_{\text{petersen}}) = 4$.
- What can we say about the chromatic index of a tree?

Theorem 14.7.2 *(König compare [9]) $\chi'(G) = \Delta(G)$ for all bipartite graphs G.*

14.8 The four-color problem

Recall that a graph $G = (V,E)$ is called planar if it can be embedded into the plane such that two curves which are the embeddings of the edges intersect only at the vertices. More precisely, planarity asserts that it is possible to represent the graph in the plane in such a way that the vertices correspond to distinct points and the edges to simple Jordan curves connecting the points of its endvertices such that every two curves are either disjoint or meet only at a common endpoint.

We use the term plane graph to refer to a planar description of a planar graph. A plane graph determines a partition of the plane into regions.

The Four Color Problem: Can the regions of a plane graph be colored with four colors so that adjacent regions are colored differently?[4]

The Poincare duality construction transforms this question into the problem of deciding whether it is possible to color the vertices of every planar graph with four colors so that no two adjacent vertices are assigned the same color.

Observation 14.8.1 *The Four Color Problem: The chromatic number of each planar graph is less or equal 4.*

Now we prove the famous Five Color Theorem.

Theorem 14.8.2 *(Heawood) Every planar graph can be 5-colored.*

Proof. We only need to consider connected planar graphs.
We prove the theorem by induction on the number of vertices. Let $G = (V, E)$ be a planar and connected graph with n vertices.
Trivially for $n \leq 5$ G can be 5-colored.
Now we assume that all connected planar graphs with $n-1$ vertices are 5-colorable. In view of 4.8.3 there is a vertex $v \in V$ with degree at most 5. Deleting v from G we obtain a graph G' with $n-1$ vertices which by assumption can be 5-colored. Now we reconnect v to G' and try to properly color v.
If $g_G(v) \leq 4$, then we can assign v a color different from the colors of its neighbors. The same approach works if the degree equals 5, but the neighbors only needs 4 colors. Thus it remains to consider the case where v is of degree 5 and have the neighbors $v_1, \ldots, v_5$ with i being the color of v_i. We may assume that $v_1, \ldots, v_5$ are arranged in clockwise order around v.
Let i and j be two different colors. Then define

$$G_{i,j} = G'[\{v' \in V \setminus \{v\} : \text{ the color of } v' = i \text{ or } j\}]. \tag{14.37}$$

Note, that in any case $G_{i,j}$ is a subgraph of G.
First consider $G_{1,3}$. Suppose that there is no path in this graph from v_1 to v_3. Then we change the color of v_1 and all other vertices in its component from 1 to 3 or vice versa. This interchange will not affect v_3. Hence, the color 1 is "free" to color v. On the other hand, if v_1 and v_3 are in the same component of $G_{1,3}$, a path from v_1 to v_3 together with the edges $\underline{vv_1}$ and $\underline{vv_3}$ forms a cycle in G which blocks the possibility of any path from v_2 to v_4 in $G_{2,4}$. Thus, we can perform a 2-4 interchange in the component which contains v_2, and v can be properly colored with 2. This completes the induction step. □

[4]The embedding of a K_4 show that fewer than four colors are not sufficient in general.

There is an nice interrelation between edge and region colorings. Let G be an embedding of a connected, 3-regular, planar and bridgeless graph in the plane. Such a plane graph is called a cubic map . We can define a cubic map equivalently as an embedding of a 2-connected 3-regular planar graph.
Cubic maps are not rare. As an exercise construct an exponential number of such graphs.

Theorem 14.8.3 *(Tait) Let G be a cubic map. Then the edges of G are colorable with three colors if and only if the regions of G are colorable with four colors.*

Idea of the *proof.* We use the colors $0, a_1, a_2$ and a_3 which can be added in the sense of the Klein group.
Let $G = (V, E)$ be a cubic map with the set

$$R = \{\mathcal{R}_1, \ldots, \mathcal{R}_{2-|V|+|E|}\},$$

of regions.
Let $F : R \to \{0, a_1, a_2, a_3\}$ be a coloring of the regions. We define a coloring $F' : E \to \{a_1, a_2, a_3\}$ of the edges by: Let e be an edge on the common boundary of the regions $\mathcal{R}_i$ and $\mathcal{R}_j$, then

$$F'(e) = F(\mathcal{R}_i) + F(\mathcal{R}_j).$$

Conversely, let $F' : E \to \{a_1, a_2, a_3\}$ be a coloring of the edges. We find a coloring $F : R \to \{0, a_1, a_2, a_3\}$ of the regions by:

$$F(\mathcal{R}_1) = 0.$$

Let $\mathcal{R}_i$, $i = 2, \ldots, 2 - |V| + |E|$ be a region. Let $\mathcal{C}$ be a curve from an inner point of $\mathcal{R}_1$ to an inner point of $\mathcal{R}_i$ which avoids all vertices of G. Then

$$F(\mathcal{R}_i) = \sum_{\mathcal{C} \text{ crosses } e} F'(e).$$

□

Consequently, the Four-Color-Conjecture is true if and only if each 2-connected 3-regular graph has chromatic index 3.

Theorem 14.8.4 *The Four-Color-Conjecture is true if and only if each bridge-less planar graph can factored in (three) perfect matchings.*

In 1977 Appel, Haken [10] show that every planar graph can be 4-colored.[5]

[5]The proof was controversial at that time because it made extensive use of a computer to check the large number of special cases. it was the first proof of a major mathematical result generated in this fashion. Can a human being check its correctness?

Chapter 15

Graphs Inside

15.1 Subgraph isomorphism

A very important generalization of graph isomorphism is known as subgraph isomorphism, and it is to determine whether a graph is isomorphic to a subgraph of another graph. More formally, a subgraph isomorphism of a graph $G_1 = (V_1, E_1)$ into a graph $G_2 = (V_2, E_2)$ is an injective mapping $f : V_1 \to V_2$ such that

$$\underline{vv'} \in E_1 \text{ then } \underline{f(v)f(v')} \in E_2.$$

Remark 15.1.1 *The problem of determining whether or not a graph is isomorphic to a subgraph of another graph is $\mathcal{NP}$-complete.*

Compare [240].

Most practical applications of the problem need finding all subgraphs of a given graph which are isomorphic to another given graph.

Theorem 15.1.2 *Assume that $p \leq n$. There are*

a)

$$\binom{n}{p} = \frac{n!}{p!(n-p)!} \tag{15.1}$$

different subgraph of the complete graph K_p into the complete graph K_n.

b)

$$\frac{n!}{(n-p)!} \tag{15.2}$$

different subgraph isomorphisms of the complete graph K_p into the complete graph K_n.

Proof. Since the graphs are complete, any injection from the set of vertices of K_p into the set of vertices of K_n, will be a subgraph isomorphism. There are

$$\binom{n}{p} = \frac{n!}{p! \cdot (n-p)!}$$

different injections; and the complete graph on p vertices induced by each of which has $p!$ different automorphisms. □

Generalizing the second statement of the former theorem we find

Theorem 15.1.3 *Let G be a graph with p vertices given, and let $p \leq n$. The number of subgraphs of K_n isomorphic to G equals*

$$\frac{n \cdot (n-1) \cdots (n-p+1)}{|Aut(G)|}. \tag{15.3}$$

Proof. There are $\binom{n}{p}$ ways to choose the p vertices for a copy of G. In view of 7.2.6 for each labeling of G we have $p!/|\mathrm{Aut}(G)|$ ways to place it on these vertices. Then

$$\binom{n}{p} \cdot \frac{p!}{|\mathrm{Aut}(G)|} = \frac{n!}{p!(n-p)!} \cdot \frac{p!}{|\mathrm{Aut}(G)|} = \frac{n \cdot (n-1) \cdots (n-p+1)}{|\mathrm{Aut}(G)|}. \tag{15.4}$$

□

15.2 Trees inside

Let k be a natural with $k \leq 2m/n$, then a graph with n vertices and m edges contains a path with k vertices. (Exercise.)
All graphs of sufficiently large minimum degree contain all trees of certain orders.

Theorem 15.2.1 *Let T be a tree with n vertices, and let G be a graph with $\delta(G) \geq n-1$, where $\delta(G)$ denotes the minimum degree in G. Then T is isomorphic to some subgraph of G.*

Proof. (Chartrand, Lesniak, [45]) We proceed by induction on n.
The result is obviously true for $n = 1$ and $n = 2$.
Assume that for any tree T' with $n-1$ vertices, where $n \geq 3$, and any graph G' with $\delta(G') \geq n-2$, that T' is isomorphic to a subgraph of G'.
Now let T and G be given as above. Let v be a leaf of T and w be the neighbor of v. Then the graph $T - v$ is a tree with $n-1$ vertices. Since

$$\delta(G) \geq n-1 > n-2,$$

it follows by the induction hypothesis that $T - v$ is isomorphic to a subgraph T_1 of G. Let w_1 be the vertex of T_1 that corresponds to the vertex w under an isomorphism. Since $g_G(w_1) \geq n-1$ and the fact that T_1 has $n-2$ vertices different from

w_1, necessarily w_1 is adjacent to a vertex u of G that does not belong to T_1. Then the subgraph $T_1 + \underline{w_1 u}$ of G is isomorphic to T. □

Hartsfield and Ringel [127] count the number of trees inside the complete graph for specific cases:

Theorem 15.2.2 *There are*

$$\frac{n!}{2(n-k-1)!} \tag{15.5}$$

many subgraphs isomorphic to the path of length k in the complete graph with n vertices for $k < n$.

Proof. A path P_k of length k has $k+1$ vertices. Clearly we have to assume $k < n$. Now we count as follows: If we begin at some vertex, there are $n-1$ choices for the next vertex in the path and $n-2$ choices for the vertex following that, until we reach the $(k+1)$th vertex. This is a total of $n(n-1)(n-2)\cdots(n-k)$. But we have counted everything twice. So we divide by 2, and get the assertion. □

Theorem 15.2.3 *There are*

$$n \cdot \binom{n-1}{3} \tag{15.6}$$

many subgraphs isomorphic to $K_{1,3}$ in the complete graph with n vertices.

Proof. First we have $\binom{n}{4}$ choices of four vertices from n vertices. For every choice we have four possibilities to form a $K_{1,3}$ as a subgraph.
By simple calculation:

$$4 \cdot \binom{n}{4} = 4\frac{n!}{4!(n-4)!} = \frac{n(n-1)!}{3!(n-4)!} = n \cdot \binom{n-1}{3}.$$

□

15.3 Complete graphs inside

Maybe the "Great Darwin Tree" is not a tree, but a graph with some subgraphs which are created "clusters". Consequently, we are interested in the converse question as in the section before.

.

Let $G = (V, E)$ be a graph. Usually, a complete subgraph of G is called a clique; more exactly a k-clique when the subgraph contains exactly k vertices.[1]

[1]The complement of a clique is an independent set; that means no two vertices are adjacent in G. As an example we wish to design an aquarium, which is a collection of tanks, for a big collection of species of fishes. Several of these animals can be placed in the same tank, but not if one is the

The obvious way to find a k-clique would be subject all $\binom{|V|}{k}$ subsets of V with cardinality k and test of whether they fulfill the requirement. The catch is that k can be depend from the size $|V|$.

Observation 15.3.1 *Consider cliques in a graph.*

a) *The maximum clique problem is $\mathcal{NP}$-complete.*[2]

b) *The maximum clique problem is solvable in polynomial time for graphs obeying a fixed degree bound.*[3]

Let $G = (V, E)$ be a graph with n vertices. For any positive integer k the quantity $c_k(n)$ denotes the number of complete subgraphs of G with k vertices.
Of course,

$$c_1(n) = n = |V|, \tag{15.7}$$

and

$$c_2(n) = |E|. \tag{15.8}$$

Theorem 15.3.2 *(Ahlswed et al. [1]) Let $G = (V, E)$ be a graph with n vertices and m edges. Then for $k \geq 2$ it holds*

$$c_k(n) \geq \frac{2(k-1)m - (k-2)n^2}{kn} \cdot c_{k-1}(n). \tag{15.9}$$

The *proof* uses induction over k starting with (15.7) and (15.8). □

By simple calculation after 15.3.2 we obtain

Theorem 15.3.3 *Let $G = (V, E)$ be a graph with n vertices and m edges. If for some positive number ϵ*

$$m \geq \left(\frac{k-1}{2k} + \epsilon\right) \cdot n^2, \tag{15.10}$$

then

$$c_k(n) \geq \epsilon^{k-1} n^k. \tag{15.11}$$

Remeber that we define the density of graph $G = (V, E)$ as the quantity $|E|/\binom{|V|}{2}$. Then 15.3.3 is satisfied if density$(G) > (1 - 1/k + 2\epsilon)$.

---

predator of one other. If two species cannot be placed in the same tank, we say they cannot cohabit. What is the minimum number of tanks that are required? And how can the fishes distributed among the tanks?
We can model this question by a graph in which each vertex corresponds to a species of fish, and two vertices are adjacent if and only if the corresponding species cannot cohabit. The vertices for each species of fish in a tank must be independent. The species of fish that residue in the same tank of the aquarium correspond to an independent set of vertices.

[2] See [97].

[3] In particular this holds true for planar graphs, see 4.8.3.

15.4 Counting perfect matchings

Let $G = (V, E)$ be a graph. A subset $E' \subseteq E$ is called a matching if no two edges in E' have a common endvertex. Then, of course,

$$|E'| \leq \frac{|V|}{2}. \tag{15.12}$$

A matching E' of G is called perfect if each vertex is incident with exactly one edge in E', that means in (15.12) equality holds.[4]
Of course, not each graph contains a perfect matching. At least such a graph must have an even number of vertices.

I. How many perfect matchings are in the complete graph K_{2n}?
We shall denote this number by $f(2n)$. Of course $f(2) = 1$, as an exercise show that $f(4) = 3$.
Now, consider the graph K_{2n} and fix a vertex v. There are $2n-1$ choices for an edge in a perfect matching containing v. Once we have chosen such an edge, we can now disregard the two endvertices of the edge and consider how many perfect matchings can be made from the remaining $2n-2$ vertices. Thus

$$f(2n) = (2n-1)f(2n-2). \tag{15.13}$$

This gives immediately by repeated application

$$\begin{aligned} f(2n) &= (2n-1) \cdot f(2n-2) \\ &= (2n-1) \cdot (2n-3) \cdot f(2n-4) \\ &\vdots \\ &= (2n-1) \cdot (2n-3) \cdot (2n-5) \cdots 3 \cdot 1 \\ &= \frac{(2n-1)!}{(2n-2) \cdot (2n-4) \cdots 4 \cdot 2} \\ &= \frac{(2n-1)!}{2^{n-1} \cdot (n-1)!}. \end{aligned}$$

We proved

Theorem 15.4.1 *There are*

$$f(2n) = \frac{(2n-1)!}{2^{n-1} \cdot (n-1)!} = (2n-1)!! \tag{15.14}$$

perfect matchings in K_{2n}.

[4] In literature such a subgraph is sometimes called a 1-factor.

II. Another problem is to count the perfect matchings in the complete bipartite graph $K_{n,n}$.[5] This is the question of the number of bijections from an n-element set into itself. Hence,

Theorem 15.4.3 *There are $n!$ perfect matchings in $K_{n,n}$.*

The general question of the number of perfect matchings in bipartite graphs is much more difficult. The reader should prove

Corollary 15.4.4 *Let $G = (V, E)$ be an r-regular, $1 \leq r \leq n$, bipartite graph with $n + n$ vertices. Then G contains at least $r!$ distinct perfect matchings.*

The problem for bipartite graphs in general is discussed in [3] and [143].[6] Also the following sharper result.

Remark 15.4.5 *Let $G = (V, E)$ be an r-regular, $1 \leq r \leq n$, bipartite graph with $n + n$ vertices. Then G contains at least*

$$n! \cdot \frac{r^n}{n^n} \geq e^{3/4} \cdot \sqrt{n} \cdot \left(\frac{r}{e}\right)^n \tag{15.16}$$

distinct perfect matchings.

Now we consider the following specification: How many perfect matchings are in the graph $\tilde{K}_{n,n}$ which is the complete bipartite graph minus a perfect matching, that is $r = n - 1$? From 15.4.5 we get only a lower bound, but we can find an exact answer. By changing the indexes the problem is the question of all derangements of n objects. Denoting the number of derangements of $\{1, \ldots, n\}$ by D_n, we can give an exact result.

Theorem 15.4.6 *Let $\tilde{K}_{n,n}$ be the complete bipartite graph with $n+n$ vertices without a perfect matching. Then there are*

$$D_n = n! \cdot \sum_{k=0}^{n} (-1)^k \frac{1}{k!} \approx \frac{n!}{e} \tag{15.17}$$

perfect matchings in $\tilde{K}_{n,n}$.

Al-Knaifes, Sachs [7] give an algebraic approach to count the number of perfect matchings in arbitrary graphs.

---

[5] This is a generalizations the so-called marriage problem: Given a set of women, each of whom known a subset of men, under what conditions can each of the women marry a men whom she knows? More formally: If a bipartite graph $G = (V_0 \cup V_1, E)$ has a matching that saturates all the vertices in V_0, then we say that V_0 can be matched into V_1.

Theorem 15.4.2 *(König, Hall) In a bipartite graph $G = (V_0 \cup V_1, E)$, V_0 can be matched into V_1 if and only if*

$$|N(W)| \geq |W| \tag{15.15}$$

for all $W \subseteq V_0$.

[6] And Petersen showed that every bridgeless 3-regular graph contains a perfect matching, see [66]. More exactly, it is the sum of a perfect matching and a collection of cycles, see [191].

15.5 Systems of distinct representatives

Closely related to the marriage problem is to find a set of distinct representatives for a collection of subsets. We need to pick one element from each subset without using any element twice. More formally, let S be a set and let $\{S_1, \ldots, S_m\}$ be a collection of subsets of S. The set $\{r_1, \ldots, r_m\}$ is called a system of distinct representatives, abbreviated SDR, for $\{S_1, \ldots, S_m\}$ if

- $r_i \in S_i$ for $i = 1, \ldots, m$; and
- $r_i \neq r_j$ for $i \neq j$.

Assume that an SDR exists for $\{S_1, \ldots, S_m\} \subseteq S$. If we select k of the sets from among $S_1, \ldots, S_m$, then these sets are represented by k elements of S. Consequently, we have for each subcollection $S_{i_1}, \ldots, S_{i_k}$

$$|\bigcup_{j=i_1}^{i_k}| \geq k \tag{15.18}$$

for all $k = 1, \ldots, m$.
Thus the inequalities (15.18) are obviously necessary for the existence of an SDR. The following theorem shows that the condition is also sufficient.

Theorem 15.5.1 *Let S be a set of n elements and let $\{S_1, \ldots, S_m\}$ be a collection of subsets of S. Assume that $m \leq n$. Assume moreover, that each S_i contains at least $t \geq 1$ elements. An SDR for $\{S_1, \ldots, S_m\}$ exists if and only if the inequalities (15.18) hold.*

Sketch of the *proof.* Consider the bipartite graph

$$G = (S \cup \{S_1, \ldots, S_m\}, \{\underline{S_i x_j} : x_j \in S_i\}),$$

□

Theorem 15.5.2 *Assume the conditions of 15.5.1. Then there are at least $t!$ of SDR's if $t \leq m$ and $\frac{t!}{(t-m)!}$ otherwise.*

A proof can found in [167].

15.6 Alignments, pairwise

> Einstein said: "God does not play dice." He was right. God plays scrabble.
>
> Philip Gold

Similar to the problem of perfect matchings is to compute the number of alignments. Sequence alignment is the identification of residue-residue correspondences. It is **the** basic tool of bioinformatics.

Any assignment of correspondences that preserves the order of the residues within sequences is an alignment; gaps may be introduced.

I. Given two sequences w and w' over the same alphabet, an alignment of w and w' is a partial mapping from letters in w to w', or vice versa, which preserves the left-to-right ordering. Such an alignment can be represented by a diagram (a matrix) with aligned letters above each other, and unaligned letters placed opposite gaps. An alignment can be viewed as a way to extend the sequences to be of the same length using gaps or "dummy symbols".
For instance consider the two words $w = ac^2g^2t^2$ and $w' = agct$. The following arrays are all alignments for w and w':

```
a c c g g t t
a g c t - - -

a c c g g t t - - -
- - - - - - a g c t
```

and

```
a c c g g t t
a - - g c t -
```

where "-" denotes a "dummy" symbol.
In other words, we are looking for a diagram such that

- The elongated sequences are of the same length;
- There is no position for which the elongated sequences both have a dummy (i.e. we do not use pairs of dummies).

Consider two words $w = a_1a_2 \dots a_n$ and $w' = b_1b_2 \dots b_m$. To count alignments is to identify aligned pairs (a_i, b_j) and simply to choose subwords of w and w' to align. We do not count the pairs of $(a-, -b)$ and $(-a, b-)$ as distinct. There must be k aligned pairs, where $0 \leq k \leq \min\{n, m\}$. There are $\binom{n}{k}$ ways to choose a's and $\binom{m}{k}$ ways to choose b's, so there are $\binom{n}{k} \cdot \binom{m}{k}$ alignments with k aligned pairs. Altogether there are

$$\sum_{k \geq 0} \binom{n}{k} \binom{m}{k} = \binom{n+m}{n} \tag{15.19}$$

alignments, compare C.2.5.[7]

[7] We do not count the pairs of $(a-, -b)$ and $(-a, b-)$ as distinct. Otherwise, the number $f(n, m)$ of such alignments for two sequences of n and m letters fulfills the equality

$$f(n, m) = f(n-1, m) + f(n-1, m-1) + f(n, m-1), \tag{15.20}$$

Theorem 15.6.1 *There are*

$$\binom{n+m}{n} = \binom{n+m}{m} \tag{15.22}$$

alignments of two words with n and m letters, respectively.
In particular, if both words have the same length n there are

$$\binom{2n}{n} \approx \frac{4^n}{\sqrt{\pi n}} \tag{15.23}$$

alignments.

More about the combinatorics of alignments can be found in Waterman [247].

II. In the biological context the equality of words makes no sense, since mutations do not allow identical sequences in reality. On the other hand, in biomolecular sequences, high sequence similarity usually implies significant functional and structural similarity.[8]
Given an alignment between two sequences, we assign a score to it as follows: Each column of the alignment will receive a certain value depending on its contents and the total score for the alignment will be the sum of the values assigned to its columns.
The similarity $\mathrm{sim}(w, w')$, between two sequences $w, w' \in A^\star$ according to a scoring system is the maximum of the scores running over all alignments of w and w'. Here, a scoring system (p, g) is given by

- A symmetric function $p : A \times A \to \mathbb{R}$, and
- A non-positive real number g.

The array of p is called the (substitution) score matrix. The value $p(a, b)$ scores pairs of aligned letters a and b. The penalty g is used to penalize gaps. In general, we assume that $p(a, a) > 0$, for $a \in A$, and $g < 0$.[9]
The distance $\rho(w, w')$, between two sequences $w, w' \in A^\star$ according to a cost measure is the minimum of the costs running over all series of operations transforming w into

which does not have a nice explicit description. But it can be shown that

$$f(n, n) \approx (1 + \sqrt{2})^{2n+1} \cdot \sqrt{n}, \tag{15.21}$$

see [246].

[8] But note that the converse is, in general, not true. And in reality, for applications in biology it is sometimes necessary to take into account several other properties of the macro-molecules to measure their similarity, for instance structure, expression and pathway similarity, compare [144].

[9] In general, in a biological context a scoring matrix p is a table of values that describe the probability of a residue (amino acid or base) pair occuring in an alignment. The approach is good, if the score matrix produces good alignments.
The PAM (Point Accepted Mutation) series of score matrices are frequently used for protein alignments [8] and [65]. Each entry in a PAM matrix gives the logarithm of the ratio of the frequency at which a pair of residues is observed in pairwise comparisons of homologous proteins to the frequency expected due to chance alone. Amino acids that regularly replace each other have a positive score, while amino acids that rarely replace each other have a negative score.

w'. The function ρ is a pseudo-metric in $A^\star$.[10]
The interrelation between both approaches are given in the following theorem.

Theorem 15.6.2 *(Smith, Waterman, Fitch [223], Waterman [246]) A metric ρ and the corresponding similarity sim there holds the following interrelation: Let w and w' be sequences (words) over A. Then*

$$\rho(w, w') + sim(w, w') = c \cdot (|w| + |w'|). \tag{15.24}$$

Idea of the *proof.* Let w and w', and let α be an alignment between w and w'. We define a series σ of operations transforming w into w' by dividing α into columns corresponding to the operations in a natural way:

- matches and mismatches in the alignment correspond to substitutions in the transformation;
- gaps in the alignment corresponds to indels (= insertions and deletions) in the transformation.

□

But we should note what Gusfield [112] wrote:

> Although an alignment and an edit transcript are mathematically equivalent, from a modeling standpoint, an edit transcript is quite different from an alignment. An edit transcript emphasizes the putative *mutational events* (point mutations in the model so far) that transform one string to another, whereas an alignment only displays a relationship between two strings. The distinction is one of *process* versus *product*. Different evolutionary models are formalized via different permitted string operations, and yet these can result in the same alignment. So an alignment alone blurs the mathematical model. This is often a pedantic point but proves helpful in some discussions of evolutionary modeling.

III. How can we find the similarity of or the distance between two words? Clearly, the consideration of all possible alignments does not make sense, since there are too many; see 15.6.1. But, observe that we cannot change the order of the letters in the words. This fact suggests that a dynamic programming approach will be useful. A dynamic programming algorithm finds the solution by first breaking the original problem into smaller subproblems and then solving all these subproblems, storing each intermediate solution in a table along with a score, and finally choosing the sequence of solutions that yields the highest score.[11] The goal is to maximize the total score for the alignment.

Algorithm 15.6.3 *Let $w = a[1]a[2]\dots a[m]$ and $w' = b[1]b[2]\dots b[n]$ be two sequences in $A^\star$, equipped with a scoring system (p, q). Then, we find the similarity $sim(w, w') = sim[m, n]$ by the following procedure.*

[10] The metric space $(A^\star, \rho)$ is a discrete one, that means each bounded set is a finite one.
[11] We used this approach by 4.7.2 in finding shortest paths.

1. **for** $i := 0$ **to** m **do**
   $sim[i, 0] := i \cdot g$;

2. **for** $j := 0$ **to** n **do**
   $sim[0, j] := j \cdot g$;

3. **for** $i := 1$ **to** m **do**
   for $j := 1$ **to** n **do**
   $sim[i, j] := \max\{sim[i-1, j] + g, sim[i-1, j-1] + p[i, j], sim[i, j-1] + g\}$

In other terms, we determine each 2×2 submatrix by the following scheme:

$$\begin{array}{ccc} \mathrm{sim}[i-1, j-1] & & \mathrm{sim}[i-1, j] \\ & \searrow & \downarrow \\ \mathrm{sim}[i, j-1] & \rightarrow & \mathrm{sim}[i, j] \end{array}$$

The algorithm runs in quadratic time:

Observation 15.6.4 *Let w and w' be two words over the same alphabet A. Then the quantities $sim(w, w')$ and $\rho(w, w')$ can be determined in $O(|w| \cdot |w'|)$ time.*[12]

15.7 Alignments, multiple

Remember that the key question in phylogeny is the reconstruction of the evolutionary tree based on contemporary data. Often these data may come from a multiple alignment. which is a natural generalization of the alignment of two sequences.[13] That means that we insert gap characters (called dummies) into, or at either end of, each of the sequences to produce a new collection of elongated sequences that obeys these rules:

(i) All elongated sequences have the same length, l;

[12]Note that 15.6.3 is relatively fast but still too slow for most practical work, where the length of the sequences and the number of sequences to be compared are very large. This comes from the following often used question: You already have a particular protein or nucleic acid sequence that you are interested in and you need to find other sequences that are related to it.
There are heuristic methods which are more efficiently for "similarity-searching" an entry in a collection of sequences. In particular, the well-known BLAST method runs in linear, that is $O(|w| + |w'|)$, time, compare [221]. Usually, BLAST use a scoring system with:

$$\begin{array}{rcl} \text{match} & = & 1 \\ \text{mismatch} & = & -3 \\ \text{gap setting} & = & -5 \\ \text{gap extension} & = & -2 \end{array}$$

[13]Phylogenetic trees and networks which are constructed in view of protein or DNA sequences are general used aligned sequences, and so the first step in an evolutionary study is often to build such a scheme.
As a nice example from linguistics we compare the word for SCHOOL in different languages:

(ii) There is no position at which all the elongated sequences have a dummy.

Then the sequences are arrayed in a matrix of n rows and l columns, where

$$\max_{i=1,\ldots,n} |w_i| \leq l \leq \sum_{i=1}^{n} |w_i|. \tag{15.25}$$

Consequently,

Observation 15.7.1 *There are only finitely many multiple alignments for a collection of sequences.*

Although the notation of a multiple alignment is easily extended from two to many sequences, the score or the cost of a multiple alignment is not easily generalized. There is no function that has been universally accepted for multiple alignment as distance or similarity has been for pairwise alignment. Compare Chan et al. [42], and Wang, Jiang [245].

As an example a cost measure is a function $f : (A \cup \{-\})^n \to I\!R_{\geq 0}$, which satisfies the following conditions:

(i) f is non-negative: $f(a_1, \ldots, a_n) \geq 0$;

(ii) $f(a, \ldots, a) = 0$, for each $a \in A$;
$f(-, \ldots, -)$ is not defined;

(iii) $f(a_1, \ldots, a_n) > 0$ if $a_i = -$ holds for at least one index i;

(iv) f is symmetric:

$$f(a_{\pi(1)}, \ldots, a_{\pi(n)}) = f(a_1, \ldots, a_n) \tag{15.26}$$

holds true for any permutation π.

For a broader discussion of the relationship between multiple alignment and phylogeny construction, compare Vingron [243].

A natural and simple way to combine alignments is the following: Let A be an alphabet and let $C = \{w_1, \ldots, w_n\}$ be a collection of n sequences over A, each of length l. Given the $n(n-1)/2$ pairwise alignments of the members of C, their alignment graph $G(C) = (V, E)$ is constructed as follows:

Language								
German	-	S	C	H	U	-	L	E
English	-	S	C	H	O	O	L	-
French	E	-	C	-	O	-	L	E
Italian	-	S	C	-	U	O	L	A
Consensus, MR	-	S	C	H or -	O or U	O or -	L	E
Consensus, restricted MR	E	S	C	H	O or U	O	L	E

MR abbreviates "majority rule".

(i) There is a vertex in $G(C)$ for each position in each sequence; and

(ii) For each pair of aligned positions in each of the pairwise alignments, there is an edge between the corresponding vertices.

Observation 15.7.2 *Let $C = \{w_1, \ldots, w_n\}$ be a collection of n sequences over an alphabet A, each of length l. Then for the alignment graph $G(C) = (V, E)$,*

$$|V| = nl \quad \textit{and} \tag{15.27}$$

$$|E| \leq \frac{ln(n-1)}{2}. \tag{15.28}$$

Now a multiple alignment for C is given by a maximum clique in the alignment graph.

15.8 The center of a graph

I. Let $G = (V, E)$ be a graph. The eccentricity $e(v)$ of the vertex v of G is the distance from v to a vertex furthest away from v:

$$e(v) = \max\{\rho(v, w) : w \in V\}. \tag{15.29}$$

The radius is defined as

$$\mathrm{rad}(G) = \min\{e(v) : v \in V\}, \tag{15.30}$$

and the diameter we have as

$$\mathrm{diam}(G) = \max\{e(v) : v \in V\}. \tag{15.31}$$

The diameter is a monotone function in the following sense: Let $G = (V, E)$ be a connected graph. If $G' = (V, E')$ is a (connected) subgraph of G, $E' \subseteq E$, then $\mathrm{diam}(G) \leq \mathrm{diam}(G')$.

The hypercube Q^D is a metric space with a strange property: Depending on the quantity D (the dimension), on one hand, it is a "big" space, since it contains exponentially many points and superexponentially many spanning trees; on the other hand, it is a "small" space, since it has a linear diameter:

$$\mathrm{diam}(Q^D) = D. \tag{15.32}$$

For some deep consequences of this observation for molecular evolution see [74].
In [213] we find some facts about the diameter of randomly chosen subgraphs of the hypercube.

Theorem 15.8.1 *Let G be a graph. Then*

$$\mathit{rad}(G) \leq \mathit{diam}(G) \leq 2 \cdot \mathit{rad}(G). \tag{15.33}$$

Proof. The left inequality follows directly from the definition.
To verify the right inequality let $v, w \in V$ such that $\rho(v, w) = \mathrm{diam}(G)$, and let u be a vertex with $e(u) = \mathrm{rad}(G)$. Then

$$\mathrm{diam}(G) = \rho(v, w) \leq \rho(v, u) + \rho(u, w) \leq \mathrm{rad}(G) + \mathrm{rad}(G) = 2 \cdot \mathrm{rad}(G).$$

□

Consider a tree T. Here, any pair of vertices has a unique path. For this reason, the diameter and the radius are more related:

$$2 \cdot \mathrm{rad}(T) - 1 \leq \mathrm{diam}(T) \leq 2 \cdot \mathrm{rad}(T), \tag{15.34}$$

and two paths which are realized the diameter cannot be disjoint. Whilst the to find the diameter of a graph consumes cubic time, the diameter (and the center) of tree can be computed in linear time (Exercise). A tree has diameter 2 if and only if it is a star.

The center of a graph G is the subgraph induced by the vertices whose eccentricity equals the radius. We have a surprising fact:

Theorem 15.8.2 *Every graph is the center of some connected graph.*

Proof. Let $G = (V, E)$ be a given graph. We construct a new graph $G' = (V', E')$ in the following way:

$$\begin{aligned} V' &= V \cup \{v_1, v_2, w_1, w_2\}; \\ E' &= E \cup \{\underline{v_1 v} : v \in V\} \cup \{\underline{v_2 v} : v \in V\} \cup \{\underline{v_1 w_1}, \underline{v_2 w_2}\}, \end{aligned}$$

where v_1, v_2, w_1, w_2 are new vertices. Then for the eccentricities related to G':

$$\begin{aligned} e(v) &= 2 \quad \text{for all } v \in V; \\ e(v_1) = e(v_2) &= 3; \quad \text{and} \\ e(w_1) = e(w_2) &= 4. \end{aligned}$$

Hence, the radius of G' equals 2, and the center of G' is G. □

On the other hand, not every tree is a center of some tree. The proof is left to the reader.

Theorem 15.8.3 *The center of a tree is one vertex or two adjacent vertices including the edge incident to both of them.*

II. Now we attack the minimum diameter spanning tree problem.
Recall that the of a network is given as a) the maximum eccentricity; and b) the longest distance between any two vertices.
For a network $G = (V, E, f)$ the minimum diameter spanning tree (MDST) is a spanning tree of minimum diameter among all possible spanning trees.
We shall have a look at some properties of collections of paths in a tree helpful for computing an MDST in a network.

Lemma 15.8.4 a) *Consider the three paths interconnecting three vertices in a tree, then these paths intersect at a vertex.*

b) *Two paths in a tree, each of the length of the diameter, cannot be disjoint.*

c) *Let $\mathcal{P}$ be a set of at least two paths in a tree where the paths intersect each other. Then all paths in $\mathcal{P}$ share a common vertex.*

Proof.
a) Otherwise there exist a cycle.
b) Suppose that the paths are given between v_1 and v_2, and between v_3 and v_4, respectively. Let u_1 be an intersection point of $\{v_1, v_2, v_3\}$ and u_2 an intersection point of $\{v_1, v_3, v_4\}$.

$$\rho(v_1, v_3) + \rho(v_2, v_4) = \rho(v_1, v_2) + \rho(v_3, v_4) + 2\rho(u_1, u_2) > 2\rho(v_1, v_2),$$

since $\rho(v_1, v_2) = \rho(v_3, v_4)$ is the diameter, and $\rho(u_1, u_2) > 0$.
It implies that the path from v_1 to v_3 or the path from v_2 to v_4 is longer than the diameter, which is a contradiction. c) Otherwise there exist a cycle. □

From 15.8.4 we immediately get the following result.

Theorem 15.8.5 *All paths in a tree, each of the length of the diameter, share at least one common vertex.*

15.9 The metric orders

Let $G = (V, E)$ be a connected graph with n vertices.[14] The metric $\rho(v, v')$ denotes the length of a shortest path between the vertices v and v'.
Let k be a positive integer. A k-order for G is a permutation π of $\{1, \ldots, n\}$ such that

$$\rho(v_{\pi(i)}, v_{\pi(i+1)}) \leq k \tag{15.35}$$

for $i = 1, \ldots, n-1$, and

$$\rho(v_{\pi(n)}, v_{\pi(1)}) \leq k. \tag{15.36}$$

Obviously, each connected graph has a k-order, but for which value of k.[15] It seems that this quantity lies in a wide range. But this is not true.

Lemma 15.9.1 *(Karaganis [146]) Let $T = (V, E)$ be a tree with n vertices, and let v, v' be two vertices. Then there is a order $v = v_1, v_2, \ldots, v_{n-1}, v_n = v'$ of the vertices such that*

$$\rho(v_i, v_{i+1}) \leq 3 \tag{15.37}$$

for $i = 1, \ldots, n-1$.

[14] For disconnected graphs the following considerations are without sense.
[15] Note, that a 1-order is a Hamilton cycle.

Proof. We use induction over n.
The lemma is true for $n = 2, 3$. Now we assume that it is true for all trees with less than n vertices.
Let $v = v_1, v_2, \ldots, v_{r-1}, v_r = v'$ be the path interconnecting v with v'.

$$\begin{aligned} G_1 &= G - \underline{v_1 v_2}, \\ G_i &= G - \underline{v_{i-1} v_i} - \underline{v_i v_{i+1}}, \text{ for } i = 2, \ldots, r-1 \\ G_r &= G - \underline{v_{r-1} v_r} \end{aligned}$$

$G_1 \cup \ldots \cup G_r$ is a forest, where the tree G_i contains the vertex v_i. In view of the induction hypothesis for each $i = 1, \ldots, r$ there is a order $v_i = v_1^i, v_2^i, \ldots, v_{n_i}^i$ in G_i such that

$$\rho(v_j^i, v_{j+1}^i) \leq 3 \text{ for } j = 1, \ldots, n_i - 1 \tag{15.38}$$

$$\rho(v_i, v_{n_i}^i) = 1, \tag{15.39}$$

where n_i denotes the number of vertices in G_i.
Then we construct the desired order by

$$\begin{aligned} &v = v_1 = v_1^1, \ldots, v_{n_1}^1, v_2 = v_1^2, \ldots, v_{n_2}^2, \ldots, v_{r-1} = v_1^{r-1}, \ldots, v_{n_{r-1}}^{r-1}, \\ &v_{n_r}^r, v_{n_r-1}^r, \ldots, v_1^r = v_r = v'. \end{aligned} \tag{15.40}$$

□

Consequently, we have the following surprising result.

Theorem 15.9.2 *(Sekanina) Each connected graph has a 3-order.*

Proof. First use a spanning tree of the graph, then apply 15.9.1 for two adjacent vertices. □

As an exercise construct a graph without 2-order.[16]
Now we give the metric order another form. Let $G = (V, E)$ be a graph. For a positive integer k we define the kth power G^k of G as a graph with the same set of vertices and that the different vertices v and v' are adjacent if

$$\rho_G(v, v') \leq k.$$

Then 15.9.2 says that for a connected graph G the "cube" G^3 is Hamiltonian. And we have a very deep and surprising result.

Theorem 15.9.3 *(Fleischner [88]) Let G be a 2-connected graph. Then G^2 is Hamiltonian.*

Of course, it is of interest to extend this result.[17] Further comments we find in [244].

[16] Hint: In view of 15.9.1 it is sufficient to look for a tree.

[17] The conjecture that a 3-connected graph is Hamiltonian is not true. This can be seen in considering the complete bipartite graph $K_{k,(k+1)}$. The graph is k-connected, but not Hamiltonian, since,

15.10 Forbidden subgraphs

Remember that we are interested in estimating the number of edges for specific classes of graphs.

I. We start with the following result.

Theorem 15.10.1 *(Mantel [172], [127]) Let $G = (V, E)$ be a graph with n vertices and without triangles. Then*

$$|E| \leq \left\lfloor \frac{n^2}{4} \right\rfloor. \tag{15.41}$$

Equality holds if and only if G is a bipartite graph with $G = K_{n_1,n_2}$, whereby $n_1 = \lfloor \frac{n}{2} \rfloor$ and $n_2 = \lceil \frac{n}{2} \rceil$.

Proof. Let v_1 be a vertex with maximum degree g and let $v_n, \ldots, v_{n-g+1}$ be its neighbors. Since there is no triangle in G the neighbors are only incident with $v_1, \ldots, v_{n-g}$. Hence

$$|E| \leq g(v_1) + \ldots + g(v_{n-g}) \leq (n-g) \cdot g.$$

The discussion of equality follows by considering of the function $f(g) = (n-g)g$. □

II. 15.10.1 is the first instance of a problem in "Extremal Graph Theory", which means for a given graph H to find the maximum number of edges that a graph with n vertices can have without containing the "forbidden" subgraph H.[18]

Theorem 15.10.2 *(Turan, [237]) The largest graph $G = (V, E)$ with n vertices that contains no subgraph isomorphic to K_{r+1} is the graph $K_{n_1,\ldots,n_r}$ which is defined by a partition of V in subsets V_i with $|V_i| = n_1$, $i = 1, \ldots, r$, and $|n_i - n_j| \leq 1$ such that $|E|$ consists of all edges connecting distinct V_i and V_j:*

$$E = \bigcup_{i \neq j} \{\underline{vw} : v \in V_i, w \in V_j\}.$$

To maximize the number of edges one chooses the parts V_i to have as equal size as possible, that means

$$\left\lfloor \frac{n}{r} \right\rfloor \leq |V_i| \leq \left\lceil \frac{n}{r} \right\rceil. \tag{15.42}$$

In particular, if r divides n, then we may choose $n_i = n/r$ for all i, obtaining

$$\binom{r}{2} \left(\frac{n}{r}\right)^2 = \frac{r(r-1)}{2} \frac{n^2}{r^2} = \frac{n^2}{2} \left(1 - \frac{1}{r}\right).$$

That means

---

in view of 4.1.2, each cycle has even length and the graph has an odd number of vertices.
Planarity changes the picture completely. There are 3-connected planar graphs which are not Hamiltonian, for instance the so-called Herschel- or the Grinberg graph, see [45]. On the other hand, Tutte [239] shows that every 4-connected planar graph is Hamiltonian.

[18] As an exercise prove that if a graph G with n vertices has at least $\binom{n-1}{2} + 1$ edges, then G has a Hamiltonian path.

Theorem 15.10.3 *For the largest graph $G = (V, E)$ with n vertices that contains no subgraph isomorphic to K_{r+1} it holds*

$$|E| \leq \frac{n^2}{2}\left(1 - \frac{1}{r}\right). \tag{15.43}$$

For further comments compare [5].

III. Now we will write the problem in terms of coloring graphs.
Remember that the chromatic number $\chi(G)$ of a graph G is the smallest value of k for which the vertices of G can be colored by k colors.

Theorem 15.10.4 *The largest graph $G = (V, E)$ with n vertices and chromatic number r is the graph $K_{n_1,\ldots,n_r}$.*

The conclusion is the same as of 15.10.2, but the assumption is stronger. The assumption of 15.10.4 implies the assumption of 15.10.2, but not conversely: No graph with chromatic number r contains a subgraph isomorphic to K_{r+1}, which is of chromatic number $r + 1$, but, for instance, a cycle of length five does not contain a K_3 and still has chromatic number 3.

Remark 15.10.5 *([80]) Let H be a fixed graph of chromatic number r. Then the number of graphs with n vertices and not containing H as a subgraph is*

$$2^{\binom{n}{2}(1-\frac{1}{r-1}+o(1))}. \tag{15.44}$$

IV. Another generalization of 15.10.1 is

Theorem 15.10.6 *(Reiman) Let $G = (V, E)$ be a graph not containing a 4-cycle. Then*

$$|E| \leq \frac{n}{4}(1 + \sqrt{4n - 3}) \tag{15.45}$$

where $n = |V|$.

Proof. S is the set of pairs $(u, \{v, w\})$, $v \neq w$, where u is adjacent to v and w. We count S in two ways: Summing over u, we find

$$|S| = \sum_{u \in V} \binom{g(u)}{2}.$$

On the other hand, since 4-cycles are forbidden, implies that v and w have at most one common neighbor,

$$|S| \leq \binom{n}{2}.$$

Altogether, and rearranging, we conclude

$$\sum_{u \in V} g(u)^2 \leq n(n-1) + \sum_{u \in V} g(u). \tag{15.46}$$

We apply the Cauchy-Schwarz inequality (G.2) to the two vectors $(g(u_1), \ldots, g(u_n))^T$ and $(1, \ldots, 1)^T$, obtaining

$$\left(\sum_{u \in V} g(u) \right)^2 \leq n \sum_{u \in V} g(u)^2, \tag{15.47}$$

and hence by (15.46)

$$\left(\sum_{u \in V} g(u) \right)^2 \leq n^2(n-1) + n \sum_{u \in V} g(u). \tag{15.48}$$

In view of 4.1.1 we find

$$4|E|^2 \leq n^2(n-1) + 2n|E|$$

or, equivalently,

$$|E|^2 - \frac{n}{2}|E| - \frac{n^2(n-1)}{4} \leq 0. \tag{15.49}$$

Solving this quadratic inequality we obtain the result. □

Chapter 16

Ramsey Theory

> Every "irregular" structure, if it is large enough, contains a "regular" substructure of some given size.[1]

In mathematics one sometimes finds that an almost obvious idea, when applied in a rather subtle manner, is the key needed to solve troublesome problems. One of such is the pigeonhole principle[2]: If m pigeons occupy n pigeonholes and $m > n$, then at least one pigeonhole has two or more pigeons in it.[3]
In this sense, we introduce the main topic of the so-called Extremal Graph Theory. The core of this theory starts with the following result, introduced by Ramsey in 1930 [197], which gave the subject its name.

Let $G = (V, E)$ be a graph. An edge-coloring of a graph is a mapping from E into a set of two elements, the colors: "red" and "blue".
Suppose that the graph is the complete graph K_r with r vertices. We are interested in complete subgraphs whose edges all have the same color, called a monochromatic complete subgraph; and define the Ramsay number $R(p, q)$ as the smallest integer r such that for any 2-coloring of K_r there exist a monochromatic red K_p or a monochromatic blue K_q.
Another interpretation of $R(p, q)$ arises from the following observation. Given a coloring of K_r, one can view the red edges as a graph on the r vertices and the blue

[1] Roughly spoken in biological context: If the universe is big (and chaotic) enough, life must be in it.

[2] This is the Anglo-American notation; continental the Schubfach principle.

[3] To find deeper consequences of the pigeonhole principle, we have to describe it more exactly.

- If n objects are put into m boxes and $n > m$, then at least one box contains two (or more) objects.
- The strong version: If n objects are put into m boxes and $n > m$, then some box must contain at least $\lceil \frac{n}{m} \rceil$ objects.
- The infinite version: If an infinite number of objects are put into a finite number of boxes, then some box must contain an infinite number of objects.

edges as the complement of that graph. This yields the following fact.

Observation 16.0.7 *$R(p,q)$ is the smallest integer r such that if G is any graph with r vertices, then either G contains a K_p or G^c contains a K_q.*

16.1 Ramsey's theorem

The existence of Ramsey's numbers is not immediately obvious.
First some simple statements.

Observation 16.1.1 *For the Ramsey numbers we have*

a) *For all positive integers p and q: $R(p,q) = R(q,p)$.*

b) *$R(1,q) = 1$.*

c) *$R(2,q) = q$ for $q \geq 2$.*

Therefore the first nontrivial Ramsey number is $R(3,3)$. It is a nice exercise to see that $R(3,3) \leq 6$.[4] On the other hand, this number cannot be less than 6, since K_5 is the union of two cycles, one red and one blue. Hence,

Observation 16.1.2 *For any graph G with six vertices G or G^c contains a triangle: $R(3,3) = 6$.*

In order to show that $R(p,q) = r$, we must verify:

1. For every coloring of the edges of a K_r there exist a monochromatic red K_p or a monochromatic blue K_q.
2. There exists a coloring of the K_{r-1} such that no p vertices are pairwise adjacent by red edges and no q vertices are pairwise adjacent by blue edges.

Using only the first step gives upper bounds for the Ramsey numbers.

Theorem 16.1.3 *The Ramsey numbers $R(p,q)$ exist for all integers $p,q \geq 1$ and satisfy*

$$R(p,q) \leq R(p-1,q) + R(p,q-1). \tag{16.1}$$

Proof. We assume that $R(p-1,q)$ and $R(p,q-1)$ exist. Let $r = R(p-1,q) + R(p,q-1)$ and consider a K_r which has been colored. We now show that there must exist a red K_p or blue K_q.
Fixing a vertex v of K_r we consider the $r-1$ incident edges. Let α represent the number of red edges incident to v and β the blue edges, respectively. Since

[4] Hint: In any group of six people, at least three must be mutual friends or at least three must be mutual strangers. Consider and myself as one of the people and put the other five people in $m = 2$ boxes: **Box 1:** my friends, and **Box 2:** stranger to me.
Using the pigeonhole principle in one box there are $\lceil \frac{5}{3} \rceil = 3$ people.

$r-1 = R(p-1,q) + R(p,q-1) - 1$, it follows from the pigeonhole principle that there are at least $R(p-1,q)$ red edges or at least $R(p,q-1)$ blue edges. Without loss of generality we may then assume $\alpha \geq R(p-1,q)$. Attached to these α edges is a subgraph $K_{R(p-1,q)}$ of K_r. This subgraph must contain a red K_{p-1} or a blue K_q. If the subgraph contains a blue K_q, then we are done. If not, then the vertices of the red K_{p-1} together with the vertex v yield a red K_p.
The argument is similar if $\beta \geq R(p,q-1)$. □

Corollary 16.1.4

$$R(p,q) \leq \binom{p+q-2}{p-1}. \tag{16.2}$$

Proof. We proceed 16.1.3 by induction.
Let $n = p+q$ and assume the result is true for $n-1$. Then, by using (C.12)

$$R(p,q) \leq R(p-1,q) + R(p,q-1) \leq \binom{p+q-3}{p-2} + \binom{p+q-3}{p-1} = \binom{p+q-2}{p-1},$$

□

16.2 Known Ramsey numbers

Only few of Ramsey numbers $R(p,q)$ are exactly known. Since there is no unified methodology for evaluating Ramsey numbers, the difficulties in obtaining these numbers are formidable.

Lemma 16.2.1 *(Greenwood, Gleason)* $R(3,4) \leq 9$.

Proof. Consider the complete graph K_9, colored by red and blue. For each vertex v let $\alpha(v)$ represent the number of red edges incident with v, and let $\beta(v)$ be the number of blue edges incident with v. Then of course $\alpha(v) + \beta(v) = 8$ for each vertex v.
Suppose that $\alpha(v) = 3$ for each vertex v of the K_9. Then the degree sum of the subgraph composed of the red edges would be $3 \cdot 9 = 27$, which contradicts 4.1.1. Hence, there exists a vertex v with $\alpha(w) \neq 3$.
We consider the following two cases.
Case 1: $\alpha(w) \geq 4$.
Let $w_1, \ldots, w_4$ be four vertices such that $\underline{ww_i}$ is red, and let K_4 denote the complete graph determined by these vertices. If an edge of K_4 is red, say $\underline{w_iw_j}$, then there is a red triangle $\{w, w_i, w_j\}$. If this is not true, then every edge must be blue, and we have a blue K_4 as desired.
Case 2: $\alpha(w) \leq 2$.
It holds $\beta(w) \geq 6$. Let $w_1, \ldots, w_6$ be six vertices such that $\underline{ww_i}$ is blue, and let K_6 denote the complete graph determined by these vertices. Applying 16.1.2 to K_6

we either have a red triangle (in which case we are done) or a blue triangle, say $\{w_i, w_j, w_k\}$, then the complete graph for $\{w, w_i, w_j, w_k\}$ provides the desired blue K_4. □

Lemma 16.2.2 *(Greenwood, Gleason)* $R(3,5) \geq 13$.

Altogether,

Theorem 16.2.3 *(Greenwood, Gleason)*

$$R(3,4) = 9 \tag{16.3}$$
$$R(3,5) = 14 \tag{16.4}$$

Proof. From 16.1.3 we obtain $R(3,5) \leq R(2,5) + R(3,4)$, which implies, in view of 16.1.1, 16.2.1 and 16.2.2, $14 \leq R(3,5) \leq 5 + 9 = 14$. □

The following Ramsey numbers (for $p, q \geq 3$) and only these are exactly known, see [45], [252]:

q/p	3	4	5	6	7	8	9
3	6	9	14	18	23	28	36
4	9	18	25				

16.3 Asymptotics

In view of (C.12) and a trivial upper bound for the sum of the binomial coefficients we obtain

$$R(p,p) \leq \binom{2p-2}{p-1} = \binom{2p-3}{p-1} + \binom{2p-3}{p-2} \leq 2^{2p-3}.$$

Consequently

Theorem 16.3.1

$$R(p,p) \leq \frac{4^p}{8}. \tag{16.5}$$

Now we are interested in a lower bound for $R(p,p)$. First, we give an argument which is typical of probabilistic methods.

Lemma 16.3.2 *If* $\binom{r}{p} 2^{1-\binom{p}{2}} < 1$ *then* $R(p,p) > r$.

Proof. Let V be a set consisting of r elements (vertices), and let $S \subseteq V$ with $|S| = p$.
There are $2^{\binom{r}{2}}$ ways to color K_r and there are $2^{\binom{r}{2}-\binom{p}{2}+1}$ colorings of K_r for which the K_p for S is monochromatic. Since S can be chosen in $\binom{r}{p}$ ways there are at most

$$\binom{r}{p} 2^{\binom{r}{2}-\binom{p}{2}+1} \tag{16.6}$$

colorings which yield a monochromatic K_p.
Under the hypothesis that $2^{\binom{r}{2}}$ is greater than (16.6) there must exist a coloring of K_r which has no monochromatic K_p. Hence $R(p,p) > r$. □

The lemma 16.3.2 is the starting point for several facts about Ramsey numbers created by Erdös.

Theorem 16.3.3 *(Erdös , compare [252])*

$$R(p,p) \geq p \cdot 2^{p/2} \left(\frac{1}{e\sqrt{2}} + o(1) \right). \tag{16.7}$$

Proof. Let $r = R(p,p)$. In view of 16.3.2 and C.2.4 we have

$$2^{\binom{p}{2}-1} \leq \binom{r}{p} \leq \frac{r^p}{p!} \leq \frac{1}{e} \left(\frac{er}{p} \right)^p \leq \left(\frac{er}{p} \right)^p.$$

By taking logarithms we obtain

$$\left(\binom{p}{2} - 1 \right) \log 2 \leq p(\log(er) - \log p), \tag{16.8}$$

which gives the assertion. □

The best lower bounds, given by using the theorem, are listed below.

$p =$	3	4	5	6	7	8	9	10	11	12	13	14	15
$R(p,p) \geq$	3	6	11	17	27	42	65	100	152	231	349	527	792

16.4 Generalized Ramsey numbers

The Ramsey numbers introduced in the preceding sections are often called the classical Ramsey numbers. This subject has expanded greatly and in many directions, creating a "Ramsey theory", compare Graham et. al. [106].[5] We expand our considerations about the Ramsey numbers in direction of graphs.[6]

[5] In an extremely abstract sense:

Theorem 16.4.1 *(Ramsey) Let p, t, n be positive integers. Then there exists a positive integer r with the following property: If X is any set with at least r elements and*

$$\binom{X}{p} = S_1 \cup \ldots \cup S_t \tag{16.9}$$

is any partition of the set of all p-element subsets of X, then there exists a subset Y of X with at least n elements such that

$$\binom{Y}{p} \subseteq S_i \tag{16.10}$$

for one index i.

For a proof see [141].

[6] One of the origin of Ramsey theory was of geometric nature:

Let H and H' be two graphs. The Ramsey number $R(H, H')$ is the least positive integer n such that if G is any graph with n vertices, then H is a subgraph of G or H' is a subgraph of G^c. Obviously, this Ramsey number may also be described in terms of coloring the edges of K_n. Thus,

Observation 16.4.3 *There are the following interrelations of the Ramsey number and the Ramsey number of graphs.*

a) *For all positive integers p and q it holds*

$$R(K_p, K_q) = R(p, q). \tag{16.12}$$

b) *For any two graphs H and H' with p and q vertices, respectively,*

$$R(H, H') \leq R(p, q). \tag{16.13}$$

This observation and 16.1.3 imply

Theorem 16.4.4 *The Ramsey numbers $R(H, H')$ exist for all graphs H and H'.*

In general the determination of the Ramsey numbers of graphs is more difficult than for "simple" ones; only for some graphs we can present a complete answer. For example, but without proof:

Theorem 16.4.5 *(Chvatal) Let T be a tree with p vertices, and let q be a positive integer. Then*

$$R(T, K_q) = 1 + (p - 1)(q - 1). \tag{16.14}$$

As exercise determine the following Ramsey numbers: $R(P_3, P_3)$, $R(K_{1,3}, K_{1,3})$ and $R(P_3, K_{1,3})$.

Remark 16.4.2 *(Erdös, Szekeres, [77]) There is a minimum function $f(.)$ such that any set of $f(n)$ points in the plane in general position contains the nodes of a convex n-gon.*
The best bounds known for $f(.)$ are:

$$2^{n-2} + 1 \leq f(n) \leq \binom{2n-4}{n-2} + 1. \tag{16.11}$$

Chapter 17

Markov processes

A Markov chain describes a stochastic process in which the future state can be predicted from its present state as accurately as if its entire earlier history was known.

17.1 Transitions

Let $\mathcal{S}$ be a finite set of states. Without loss of generality we assume that the states are named by numbers:

$$\mathcal{S} = \{1, 2, \ldots, n\}. \tag{17.1}$$

We consider diagrams between states, where the transition from state i to state j occurs with given probability α_{ij}, altogether written in a transition matrix

$$A = (\alpha_{ij})_{i,j=1,\ldots,n}. \tag{17.2}$$

Of course, a transition matrix has the properties $\alpha_{ij} \geq 0$ for any $i, j = 1, \ldots, n$, and

$$\sum_{j=1}^{n} \alpha_{ij} = 1 \tag{17.3}$$

for any $i = 1, \ldots, n$.
Under these conditions, such a matrix is sometimes called stochastic. The pair $(\mathcal{S}, A)$ is called a Markov process.
Consider $A^2 = (\alpha_{ij}^{(2)})_{i,j=1,\ldots,n}$. Then

$$\sum_{j=1}^{n} \alpha_{ij}^{(2)} = \sum_{j=1}^{n} \sum_{k=1}^{n} \alpha_{ik}\alpha_{kj} = \sum_{k=1}^{n} \alpha_{ik} \sum_{j=1}^{n} \alpha_{kj} = \sum_{k=1}^{n} \alpha_{ik} = 1,$$

since A is a stochastic matrix. Therefore A^2 is stochastic, too.

Observation 17.1.1 *The power of any stochastic matrix is also stochastic.*

In view of (17.3) we have $\alpha_{ij} \leq 1$, for any $i, j = 1, \ldots, n$, so that we can really speak about a probability. Moreover

$$\begin{aligned} \alpha_{ij}^{(2)} &= \sum_{k=1}^{n} \alpha_{ik}\alpha_{kj} \\ &= \sum_{k=1}^{n} \text{probability for a transition from state } i \text{ to state } k \\ &\quad \star \text{ probability for a transition from state } k \text{ to state } j. \end{aligned}$$

Hence $\alpha_{ij}^{(2)}$ is the probability for a transition from state i to state j in two steps. By induction we obtain

Theorem 17.1.2 *(Chapman, Kolmogorov) Let $(\mathcal{S}, A)$ be a Markov process. Let*

$$A^t = (\alpha_{ij}^{(t)})_{i,j=1,\ldots,n} \tag{17.4}$$

be the tth power of A. Then $\alpha_{ij}^{(t)}$ is the probability for a transition from state i to state j in t steps.

A stochastic matrix A is called double-stochastic if A^T is also stochastic. A symmetric stochastic matrix is double stochastic, but not vice versa.

For example consider a two-state process with the matrix

$$A = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}, \tag{17.5}$$

representing the probabilities for a particular period of time (e.g. 1 million years). If this trend was to continue for 100 periods (= 100 million years), what would the corresponding transition matrix be? Answer: A^{100}. Obviously, this matrix is not easy to compute, but we can use the following trick: Consider the matrix

$$T = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \tag{17.6}$$

which is a regular matrix, called a Hadamard matrix, and

$$T^{-1} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{pmatrix}. \tag{17.7}$$

Then it holds

$$T^{-1}AT = \begin{pmatrix} 1 & 0 \\ 0 & 0.98 \end{pmatrix} = B. \tag{17.8}$$

Simple to see that $A^t = TB^tT^{-1}$, see (17.21), and B^{100} is easy to calculate:

$$B^{100} = \begin{pmatrix} 1 & 0 \\ 0 & 0.1326 \end{pmatrix}.$$

Hence,

$$A^{100} = \begin{pmatrix} 0.5663 & 0.4337 \\ 0.4337 & 0.5663 \end{pmatrix}. \tag{17.9}$$

17.2 Two-states processes

As a specific case consider $(\{1,2\}, A)$ with

$$A = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}, \tag{17.10}$$

where $0 \leq p, q \leq 1$.
For $p = q = 0$ we have $\lim_{t\to\infty} A^t = E$; for $p = q = 1$ the quantity $\lim_{t\to\infty} A^t$ does not exist.
Now we assume $0 < p, q < 1$. We have

$$A = E + \begin{pmatrix} -p & p \\ q & -q \end{pmatrix} = E + B. \tag{17.11}$$

It is easy to see that $B^2 = -(p+q) \cdot B$, such that

$$B^i = (-1)^{i-1}(p+q)^{i-1} \cdot B, \tag{17.12}$$

for all $i \geq 2$. Then[1]

$$\begin{aligned}
A^t &= (E+B)^t = \sum_{i=0}^{t} \binom{t}{i} B^i \quad \text{by C.2.2} \\
&= E + \sum_{i=1}^{t} \binom{t}{i} B^i \\
&= E + \sum_{i=1}^{t} \binom{t}{i} (-1)^{i-1}(p+q)^{i-1} B \quad \text{by (17.12)} \\
&= E - \frac{1}{p+q} \sum_{i=1}^{t} \binom{t}{i} (-1)^{i}(p+q)^{i} B \\
&= E + \frac{1}{p+q} B - \frac{1}{p+q} \sum_{i=0}^{t} \binom{t}{i} (-1)^{i}(p+q)^{i} B \\
&= E + \frac{1}{p+q} B - \frac{1}{p+q} \sum_{i=0}^{t} \binom{t}{i} (-p-q)^{i} B \\
&= E + \frac{1}{p+q} B - \frac{(1-p-q)^t}{p+q} B \quad \text{by C.2.2.}
\end{aligned}$$

Theorem 17.2.1 *Let*

$$A = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \tag{17.13}$$

[1] In view of the fact that in general the multiplication of matrices is not commutative, but for the unit matrix it is true.

be a stochastic matrix. Then

$$A^t = E + \frac{1-(1-p-q)^t}{p+q}(A-E). \tag{17.14}$$

Now, we discuss the convergence behavior.

Case 1: $p = q = 0$. Then $A = E$, and of course $A^t = E$.

Case 2: $p = q = 1$. Then

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

and $A^2 = A^4 = A^6 = \ldots = E$, but $A^3 = A^5 = A^7 = \ldots = A$, such that a limit does not exists.

Case 3: $0 < p + q < 1$. Since $\lim_{t\to\infty}(1-p-q)^t = 0$ we get

$$\lim_{t\to\infty} A^t = E + \frac{1}{p+q}B. \tag{17.15}$$

Theorem 17.2.2 *Consider a two-state Markov process with* $0 < p + q < 1$. *Then*

$$\lim_{t\to\infty} \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}^t = \begin{pmatrix} \frac{q}{p+q} & \frac{p}{p+q} \\ \frac{q}{p+q} & \frac{p}{p+q} \end{pmatrix}. \tag{17.16}$$

Corollary 17.2.3 *Consider a two-state Markov process with a double-stochastic matrix A. Then*

$$\lim_{t\to\infty} A^t = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \tag{17.17}$$

17.3 The convergence behaviour

We are interested in discussing the convergence behaviour of stochastic matrices.

Theorem 17.3.1 *Let A be a stochastic matrix. Assume that*

$$C = \lim_{t\to\infty} A^t \tag{17.18}$$

exists. Then

a) *C is a stochastic matrix.*

b) *C is a projection which commutes with A, that means*

$$C^2 = C = CA = AC. \tag{17.19}$$

Proof. The first fact is an immediately consequence of 17.1.1.

$$C = \lim_{t\to\infty} A^{2t} = \lim_{t\to\infty} A^t \lim_{t\to\infty} A^t = C^2,$$

$$C = \lim_{t\to\infty} A^{t+1} = \lim_{t\to\infty} A^t A = CA,$$

and

$$C = \lim_{t\to\infty} A^{1+t} = \lim_{t\to\infty} AA^t = AC.$$

□

We may restrict ourselves to similar matrices.

Theorem 17.3.2 *Let A be a matrix. Let B be similar to A, where $B = T^{-1}AT$ with a regular matrix T. Then*

$$\lim_{t\to\infty} B^t = T^{-1} \lim_{t\to\infty} A^t T. \tag{17.20}$$

Proof.

$$B^t = (T^{-1}AT)^t = \underbrace{T^{-1}AT\cdots T^{-1}AT}_{t-\text{times}} = T^{-1}A^tT. \tag{17.21}$$

□

Matrix diagonalization enables us to easily calculate any positive power of the matrix.

Theorem 17.3.3 *Let A be an $n\times n$ matrix with n linearly independent real eigenvectors, and corresponding eigenvalues $\lambda_1,\ldots,\lambda_n$. Let T be the transformation matrix such that*

$$T^{-1}AT = diag(\lambda_1,\ldots,\lambda_n).$$

Then

$$A^t = T\cdot diag(\lambda_1^t,\ldots,\lambda_n^t)\cdot T^{-1}. \tag{17.22}$$

As an exercise prove the following fact.[2]

Theorem 17.3.4 *Let A be an double-stochastic $n\times n$ matrix with 1 as a simple eigenvalue. Then $\lim_{t\to\infty} A^t = \frac{1}{n}E$.*

[2]Hint: First prove that the power of a double-stochastic matrix is also double stochastic.

17.4 Once again: Two-states processes

Consider again our example for the two-state case $A \in \mathcal{M}_{2,2}$:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}. \tag{17.23}$$

By simple calculation, we find the characteristic polynomial

$$\begin{aligned} p_A(\lambda) &= \lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} && (17.24) \\ &= \lambda^2 - \text{trace } A \cdot \lambda + \det A. && (17.25) \end{aligned}$$

In view of $a_{12} = 1 - a_{11}$ and $a_{21} = 1 - a_{22}$ we find the characteristic polynomial in

$$p_A(\lambda) = \lambda^2 - \text{trace } A \cdot \lambda + \text{trace } A - 1. \tag{17.26}$$

The roots are easy to find:

Lemma 17.4.1 *A 2×2 stochastic matrix A has the eigenvalues*

$$\begin{aligned} \lambda_1 &= 1 \quad \textit{and} \\ \lambda_2 &= \textit{trace } A - 1. \end{aligned}$$

Now, we create a triangular matrix similar to A: Consider again the Hadamard matrix

$$T = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \tag{17.27}$$

Then it holds

$$T^{-1}AT = \begin{pmatrix} 1 & a_{11} - a_{22} \\ 0 & \text{trace } A - 1. \end{pmatrix} \tag{17.28}$$

Theorem 17.4.2 *Each 2×2 stochastic matrix A is similar to*

$$\begin{pmatrix} 1 & 0 \\ 0 & \textit{trace } A - 1. \end{pmatrix} \tag{17.29}$$

17.5 Continuous Markov processes

Evolutionary models describe the substitution processes in DNA, RNA and amino acid sequences through time. For simplicity, we will concentrate on DNA sequences, that means the corresponding matrices of the transition probabilities are given by

$$P(t) = \begin{pmatrix} p_{aa}(t) & p_{ac}(t) & p_{ag}(t) & p_{at}(t) \\ p_{ca}(t) & p_{cc}(t) & p_{cg}(t) & p_{ct}(t) \\ p_{ga}(t) & p_{gc}(t) & p_{gg}(t) & p_{gt}(t) \\ p_{ta}(t) & p_{tc}(t) & p_{tg}(t) & p_{tt}(t) \end{pmatrix}, \tag{17.30}$$

where $p_{xx}(.)$ really means the probability that the nucleotide $x \in \{a, c, g, t\}$ is not substituted.

Observation 17.5.1 *For each time parameter $t \geq 0$ the matrix $P(t)$ is double-stochastic.*

While modeling we assume that $P(t)$ gives the probability of all possible states changes in time t. We get a continuous-time Markov process. Then we find 17.1.2 in the following theorem:

Theorem 17.5.2 *Let $P(t)$ be the matrix for the transition probabilities.*

$$P(t+t') = P(t) \cdot P(t'). \tag{17.31}$$

Now we assume that such continuous-time Markov processes are differentiable at every $t \geq 0$. For $h > 0$ it then follows, in view of 17.5.2

$$\frac{P(t+h) - P(t)}{h} = \frac{P(t)P(h) - P(t)}{h} = \frac{P(t)(P(h) - E)}{h} = P(t) \cdot \frac{P(h) - P(0)}{h}.$$

When $h \to 0$ this identity implies

$$P'(t) = P(t) \cdot P'(0). \tag{17.32}$$

This differential equation has the following solution.

Theorem 17.5.3 *Under the assumptions given above the matrix $P(t)$ has the form*

$$P(t) = e^{tQ}, \tag{17.33}$$

where Q is some (fixed) matrix.

Recall, that for a square matrix A we define the exponential matrix e^A by the sum of the following series:

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!}.$$

The matrix Q is called the matrix of instantaneous change or the rate matrix. It has the following important properties:

a) It holds the "inverse" identity $Q = P'(0)$.

b) $P(t)$ is the unique solution to $P'(t) = P(t) \cdot Q$, subject to $P(0) = E$.

c) The elements in each row of $Q = (q_{ij})$ sum up to 0. Furthermore, $q_{ij} \geq 0$ for $i \neq j$ and $q_{ii} < 0$ for all i. In particular, $\det Q = 0$.

Remark 17.5.4 *A matrix Q is a rate matrix if and only if the matrix $P(t) = e^{tQ}$ is a stochastic matrix for every t.*

By varying the matrix Q one obtains several models:
The Jukes-Cantor model is the oldest model and assumes that the probabilities to find a nucleotide site are equal for all four possible states and for all time t. The matrix of instantaneous change is given by

$$Q = \frac{1}{4}\begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}, \tag{17.34}$$

where α is a positive real number, called the evolutionary rate.
We will calculate the corresponding matrix $P(t)$. First, by induction, it is easy to see that

$$Q^n = (-\alpha)^{n-1}Q \tag{17.35}$$

is true for all integers $n \geq 1$. Now we find $P(t) = \exp(tQ)$ by the following calculations.

$$\begin{aligned} P(t) &= \sum_{n=0}^{\infty} \frac{t^n Q^n}{n!} \\ &= E + \sum_{n=1}^{\infty} \frac{t^n Q^n}{n!} \\ &= E + \left(\sum_{n=1}^{\infty} \frac{t^n (-\alpha)^{n-1}}{n!} \right) Q \quad \text{in view of (17.35)} \\ &= E - \frac{1}{\alpha}\left(\sum_{n=1}^{\infty} \frac{(-t\alpha)^n}{n!} \right) Q \\ &= E - \frac{1}{\alpha}\left(e^{-t\alpha} - 1\right) Q. \end{aligned}$$

This implies

Theorem 17.5.5 *The transition matrix in the Jukes-Cantor model equals*

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-t\alpha} & : \quad i = j \\ \frac{1}{4} - \frac{1}{4}e^{-t\alpha} & : \quad i \neq j \end{cases}$$

$(i, j \in \{a, c, g, t\})$

In other models the entries of the matrices are influenced by the nucleotide composition. The Kimura model models a certain difference between two types of nucleotide substitutions: Purines into pyrimidines or vice versa; and purines into purines or pyrimidines into pyrimidines. It is given by

$$Q = \frac{1}{4}\begin{pmatrix} -(2\beta+1)\alpha & \beta\alpha & \alpha & \beta\alpha \\ \beta\alpha & -(2\beta+1)\alpha & \beta\alpha & \alpha \\ \alpha & \beta\alpha & -(2\beta+1)\alpha & \beta\alpha \\ \beta\alpha & \alpha & \beta\alpha & -(2\beta+1)\alpha \end{pmatrix}, \tag{17.36}$$

with two parameters $\alpha, \beta > 0$.

Theorem 17.5.6 *The transition matrix in the Kimura model equals*

$$p_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{1}{4}e^{-t\alpha\beta} - \frac{1}{2}e^{-t\alpha(\beta+1)/2} & : \quad (i,j) = (a,g),(g,a),(c,t) \text{ or } (t,c) \\ \frac{1}{4} - \frac{1}{4}e^{-t\alpha\beta} & : \quad (i,j) = (a,c),(c,a),(a,t),(t,a), \\ & : \quad (c,g),(g,c),(g,t) \text{ or } (t,g) \\ \frac{1}{4} + \frac{1}{4}e^{-t\alpha\beta} + \frac{1}{2}e^{-t\alpha(\beta+1)/2} & : \quad \text{otherwise} \end{cases}$$

For more information, and other models, compare [139].

17.6 A Moran process

... is a specific birth-death process.
Consider a population of fixed size n. There are two types of individuals, A and B. In any time step, a random individual is chosen for reproduction and a random individual is chosen for elimination. They reproduce at the same rate, but assume that A has fitness r while B has fitness 1. If $r > 1$ then the selection favors A; if $r < 1$ favors B; and if $r = 1$ we have neutral drift.
The Moran process is defined on the state space $i = 0, \ldots, n$. The probability that A is chosen for reproduction is given by $ri/ri + n - i$; hence, the probability that B is chosen for reproduction is given by $n - i/ri + n - i$. On the other hand, fitness does not act on dead, which means that the probability that A is chosen for elimination is i/n; and for B is $(n - i)/n$. For the transition matrix, we obtain

$$\begin{aligned} p_{ii-1} &= \frac{n-i}{ri+n-1} \cdot \frac{i}{n} \\ p_{ii} &= 1 - p_{ii-1} - p_{ii+1} \\ p_{ii+1} &= \frac{ri}{ri+n-1} \cdot \frac{n-i}{n}, \end{aligned}$$

all other elements are zero.
Therefore, solving this system, the probability of being absorbed in the state n when starting in state i is given by

$$x_i = \frac{1 - \frac{1}{r^i}}{1 - \frac{1}{r^n}}. \tag{17.37}$$

The fixation probability of a single A individual in a population of $n-1$ B individuals is

$$p_A = x_1 = \frac{1 - \frac{1}{r}}{1 - \frac{1}{r^n}}. \tag{17.38}$$

For more information see [181].

Appendix A

Orders of Growing

A.1 The Landau symbols

Often we will use the phrase "on the order of" to express lower and upper bounds. For this purpose we introduce specific notations, called Landau symbols: Let f and g be functions from the positive integers into the real numbers. Then:

1. The function $g(n)$ is said to be of order at least $f(n)$, denoted $\Omega(f(n))$, if there are positive constants c and n_0 such that $g(n) \geq c \cdot f(n)$ for all $n \geq n_0$.

2. The function $g(n)$ is said to be of order at most $f(n)$, denoted $O(f(n))$, and often read "big oh", if there are positive constants c and n_0 such that $g(n) \leq c \cdot f(n)$ for all $n \geq n_0$.

3. The function $g(n)$ is said to be of order $f(n)$, denoted $\Theta(f(n))$, if $g(n) = \Omega(f(n))$ and $g(n) = O(f(n))$. That is, $f(n)$ and $g(n)$ both grow at the same rate; only the multiplicative constants may be different.

This notation allows us to concentrate on the dominating term in an expression describing a lower or upper bound and to ignore any multiplicative constants order notations.
Note that it is not an equation in the usual sense. It has to be read from left to right.

Example A.1.1 *If $p(n)$ is a polynomial of degree k that means*

$$p(n) = a_k x^k + a_{k-1} x^{k-1} + \ldots + a_1 x + a_0. \tag{A.1}$$

Then

$$p(n) = O(n^k). \tag{A.2}$$

Proof.

$$\begin{aligned} |p(n)| &= |a_k x^k + a_{k-1}x^{k-1} + \ldots + a_1 x + a_0| \\ &\leq |a_k x^k| + |a_{k-1}x^{k-1}| + \ldots + |a_1 x| + |a_0| \\ &= |a_k|x^k + |a_{k-1}|x^{k-1} + \ldots + |a_1|x + |a_0| \\ &\leq (|a_k| + |a_{k-1}| + \ldots + |a_1| + |a_0|)x^k \\ &= ax^k. \end{aligned}$$

□

In particular, we say that the function f is polynomially bounded if there is a positive integer k such that $f(n) = O(n^k)$.
It is not hard to see that the "Order"-notations have the following properties:

- $g(n) = O(f(n))$ if and only if $f(n) = \Omega(g(n))$.
- $f(n) = \Theta(g(n))$ if and only if $g(n) = \Theta(f(n))$.
- The relation represented by "O" is transitive.
- For the logarithmic order $O(\log n)$ the base is irrelevant since $\log_b n = \log_a n \cdot \log_b a$.
- Exponential functions grow faster than polynomial functions: $n^k = O(b^n)$ for all $k > 0$ and $b > 1$. Conversely, logarithmic functions grow more slowly than polynomial functions.

Observation A.1.2 *The big oh notation has the following hierarchy of increasing orders:*

$$c, \log\log n, \log n, n, n \cdot \log n, n^2, n^3, c^n, n!, n^n. \tag{A.3}$$

For our purpose we will use the following "classes of order", which are defined in terms of the input size n:

Order	Name of the "class"	Remark
$O(1)$	constant	the function is bounded
$O(\log n)$	logarithmic	the base is irrelevant
$O(n)$	linear	
$O(n \log n)$	log-linear	the base is irrelevant
$O(n^2)$	quadratic	
$O(n^3)$	cubic	
$\vdots$		
$O(n^k)$	polynomial	k is a fixed positive integer

Mention that the previous table shows the "slow growing" orders, this table the "fast growing" ones:

Order	Name of the "class"	Remark
$O(c^n)$	exponential	$c > 1$ is a fixed positive real number
$\vdots$		
$O(n!)$	factorial	$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$
$\vdots$		
$\Omega(2^{2^n})$	superexponential	

A.2 Approximations

Often we have no exact formula for counting the number of combinatorial objects of some kind, but we can describe its asymptotic behavior. Then we use the following notation: Let f and g be functions from the positive integers into the real numbers, then

1. The function $g(n)$ is said to be growing faster than $f(n)$, denoted $f(n) = o(g(n))$, and read "small oh", if
$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0. \tag{A.4}$$

2. The function $g(n)$ is said to be approximately $f(n)$, denoted $f(n) \approx g(n)$, if
$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1. \tag{A.5}$$

It is easy to see that the relation represented by "o" is transitive, and we have the same increasing sequence as in A.1.2.

Theorem A.2.1 *Let f and g be two functions.*

$$\textit{If } f(n) \approx g(n), \textit{ then } f(n) = \Theta(g(n)).$$

Proof. $f(n) \approx g(n)$ means that $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$. It follows that there is some number n_0 beyond which the ratio is always between $1/2$ and 2. Thus, $f(n) \leq 2g(n)$ for all $n \geq n_0$, which implies that $f(n) = O(g(n))$; and $f(n) \geq (1/2)g(n)$ for all $n \geq n_0$, which implies that $f(n) = \Omega(g(n))$. Together, $f(n) = \Theta(g(n))$. □

As exercise show that the converse statement is not true.

A broader discussion on the growth of functions can be found in the book by Aigner [4].

Appendix B

Designs

B.1 Incidence Structures

Let X and Y be sets and let I be a correspondence from X to Y. Here, we will call the tripel (X, Y, I) an incidence structure.
We write xIy to denote the fact that $(x, y) \in I$, saying that x is incident with y. Often the elements of X are called the points and the elements of Y the blocks of the incidence structure.
Specific examples of incidence structures are

- Geometry: X is the set of points, and Y the family of lines. xIy means that the point x lies on the line y.
- Graph theory: X is any finite set of vertices, and Y is a finite family of unordered pairs of vertices, called edges. The incidence $e = \underline{uv}$ means that the edge e joins the vertices u and v.

We define the incidence matrix as the $b \times v$ matrix $I = (a_{ij})$ with

$$a_{ij} = \begin{cases} 1 & : \text{ the } i\text{th block contains the } j\text{element} \\ 0 & : \text{ otherwise} \end{cases}$$

That means that I is based on the order in which the blocks, and the elements are taken. However, it turns out that the important properties of I do not depend on the particular orders chosen.

B.2 The double counting principle

The double counting principle compares the number $r(x)$ of blocks which are incident with the point x:

$$r(x) = |\{(x, y) : y \in Y, xIy\}|, \tag{B.1}$$

and the number $r(y)$ of points incident with the block y.

Theorem B.2.1 *Let (X, Y, I) be a (finite) incidence structure. Then*

$$\sum_{x\in X} r(x) = |I| = \sum_{y\in Y} r(y). \tag{B.2}$$

Proof. For $x \in X$ we consider $S_x = \{(x, y) : y \in Y, xIy\}$. Then $C = \{S_x : x \in X\}$ is a partition of I, where some of the S_x may be empty, but then $r(x) = 0$. Then by the addition principle

$$\sum_{x\in X} r(x) = \sum_{x\in X} |S_x| = |I|.$$

Similarly, we get the other equation. □

Consider the following example from number theory. Let $\tau(n)$ be the number of divisors of the natural number n. We are interested in the average number $\tilde{\tau}(n)$ of divisors. At first glance, this seems hopeless. For primes we have $\tau = 2$, while for powers of 2 we obtain $\tau(2^k) = k + 1$. So, τ is a widly jumping function. But this is not true for $\tilde{\tau}$.
Let $X, Y = \{1, \ldots, 8\}$ and let I be the divisibility relation, that means $xIy = x|y$. We write the incidence in the following matrix:

$$a_{xy} = \begin{cases} 1 & : \quad x|y \\ & : \quad \text{otherwise} \end{cases}$$

which gives

$X \setminus Y$	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	1
2		1		1		1		1
3			1			1		
4				1				1
5					1			
6						1		
7							1	
8								1

Obviously, $\tau(y)$ is the number of 1's in the column y.
To use B.2.1 consider such matrix for n. We must also count the number of 1's in the rows. This number is $\left\lfloor \frac{n}{x} \right\rfloor$. Consequently,

$$\tilde{\tau}(n) = \frac{1}{n} \cdot \sum_{y=1}^{n} \tau(y) = \frac{1}{n} \cdot \sum_{x=1}^{n} \left\lfloor \frac{n}{x} \right\rfloor \approx \frac{1}{n} \cdot \sum_{x=1}^{n} \frac{n}{x} = \sum_{x=1}^{n} \frac{1}{x},$$

which is called the nth Harmonic number. Later we will prove that the Harmonic number is approximately the logarithm.

Theorem B.2.2 *The average number of divisors of an integer is asymptotic $\tilde{\tau}(n) \approx \ln n$.*

B.3 Balanced incomplete block designs

Let V be a set with v elements. A collection $\{B_1, \ldots, B_b\}$ of subsets of V is called a balanced incomplete block design (BIBD), or (v, b, r, k, λ)-design, if the following conditions are satisfied:

a) For each index i the subset B_i contains exactly k elements, where k is a fixed constant and $k < v$;[1]

b) Each element $x \in V$ is in exactly r of the subsets B_i;

c) Every pair $x, y \in V$ appear together in exactly λ of the subsets B_i.

For example

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

is a BIBD on seven elements with each pair in exactly one block.
The subsets of $\{1, \ldots, 6\}$ have the property that each subset has three elements and each pair of elements occurs in two of the subsets:

$$\begin{array}{lllll} \{1,2,3\} & \{1,2,4\} & \{1,3,5\} & \{1,4,6\} & \{1,5,6\} \\ \{2,3,6\} & \{2,4,5\} & \{2,5,6\} & \{3,4,5\} & \{3,4,6\} \end{array}$$

This example was given by Yates in 1936 in construction of agricultural experiments.

Let (v, b, r, k, λ)-design be given. The parameters b, v, r, k and λ are not all independent.

Theorem B.3.1 *In a (v, b, r, k, λ)-design it holds $bk = vr$.*

Proof. Each block contains k elements. Therefore the total numbers of elements in the b blocks is bk. Each of the v elements occurs in r blocks, so the total number of elements in the b blocks is vr. The assertion follows. □

Theorem B.3.2 *In a (v, b, r, k, λ)-design it holds $\lambda(v-1) = r(k-1)$.*

Proof. Let x be an arbitrary element. Then x appears in r different blocks, whereby it appears with $k-1$ other of the $v-1$ elements. The total number of elements different from x that appear in the r blocks containing x is $r(k-1)$.
Each of the $v-1$ elements different from x must occur with x in λ blocks, so the total number of elements different from x that appear in the r blocks containing x is $\lambda(v-1)$. □

We assume by definition that $k < v$. This and several other inequalities are summarizes in the following result.

[1] That $k < v$ means "incomplete"; otherwise "complete" has the meaning that each block is V.

Theorem B.3.3 *In a (v, b, r, k, λ)-design it holds a) $k < v$, b) $r < b$ and c) $\lambda < r$.*

Proof. a) is the definition for "incompletness".
b) is a direct consequence from (a) and B.3.1.
c) By B.3.2 we have $\lambda(v-1) = r(k-1)$. Therefore, in view of (a),

$$\frac{\lambda}{r} = \frac{k-1}{v-1} < 1,$$

□

It should noted that for arbitrary given values v, b, r, k and λ, there need not exist a (v, b, r, k, λ)-design. Moreover, the above conditions are necessary, but they are not sufficient for the existence of a BIBD. For example it can be shown that there is no BIBD with the parameters $v = 15$, $b = 21$, $r = 7$, $k = 5$ and $\lambda = 2$ even though all of the conditions are satisfied.
An interesting special type is given when we restrict to $k = 2$ and $\lambda = 1$. Then by B.3.1 and B.3.2 we obtain $2b = rv$ and $r = v - 1$. Together $b = \binom{v}{2}$. Assuming that we have v vertices, and consider the blocks as the edges, we have the complete graph K_v.

B.4 The Fisher inequality

Theorem B.4.1 *Let I be an incidence matrix of a (v, b, r, k, λ)-design. Then*

$$I^T \cdot I = (r - \lambda)E + \lambda J, \tag{B.3}$$

where E is the $v \times v$ unit matrix and J the $v \times v$ matrix with every entry equal 1.

A very important consequence is the fact, that a design cannot contain fewer blocks than elements.

Theorem B.4.2 *(Fisher) In a (v, b, r, k, λ)-design it holds $v \leq b$.*

Proof. Let I be the incidence matrix for the design. Then

$$\begin{aligned}
\det(I^T I) &= \det \begin{pmatrix} r & \lambda & \lambda & \dots & \lambda \\ \lambda & r & \lambda & \dots & \lambda \\ \lambda & \lambda & r & \dots & \lambda \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda & \lambda & \lambda & \dots & r \end{pmatrix} \quad \text{in view of } B.4.1 \\
&= \det \begin{pmatrix} r & \lambda & \lambda & \dots & \lambda \\ \lambda - r & r - \lambda & 0 & \dots & 0 \\ \lambda - r & 0 & r - \lambda & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda - r & 0 & 0 & \dots & r - \lambda \end{pmatrix}
\end{aligned}$$

$$
\begin{aligned}
&= \det \begin{pmatrix} r+(v-1)\lambda & \lambda & \lambda & \dots & \lambda \\ 0 & r-\lambda & 0 & \dots & 0 \\ 0 & 0 & r-\lambda & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & r-\lambda \end{pmatrix} \\
&= (r+(v-1)\lambda)(r-\lambda)^{v-1} \\
&= rk(r-\lambda)^{v-1} \quad \text{in view of } B.3.2.
\end{aligned}
$$

Now $k < v$ by B.3.3, such that in view of B.3.1 $r > \lambda$. Consequently,

$$\det(I^T I) \neq 0.$$

In other terms the rank of the $v \times v$ matrix $I^T I$ must be v.

$$v = \text{rank}(I^T I) \leq \text{rank} I \leq \text{number of rows of } I = v.$$

□

As exercise show that no BIBD exists for $v = 25$, $k = 10$ and $r = 3$.

Appendix C

Polynomic Approaches

C.1 Factorials and double factorials

There are

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 \tag{C.1}$$

ways to place n objects in a row.
It is convenient to define $0! = 1$. Obviously, the function ! satisfies the following recurrence relation:

$$\begin{aligned} n! &= n \cdot (n-1)! \quad \text{for } n \geq 1; & \text{(C.2)} \\ 0! &= 1. & \text{(C.3)} \end{aligned}$$

The double factorial !! can be defined recursively by

$$\begin{aligned} n!! &= n \cdot (n-2)!! \quad \text{for } n \geq 1; & \text{(C.4)} \\ 0!! = (-1)!! &= 1. & \text{(C.5)} \end{aligned}$$

A continuous extension of the factorial function is the gamma function Γ, defined by

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x}\, dx, \tag{C.6}$$

for $k > 0$.

Theorem C.1.1 *For positive integers n*

$$\Gamma(n) = (n-1)!. \tag{C.7}$$

Proof.

$$\begin{aligned} \Gamma(n) &= \int_0^\infty x^{n-1} e^{-x}\, dx \\ &= -e^{-x} x^{n-1} |_0^\infty + (n-1) \int_0^\infty x^{n-2} e^{-x}\, dx \\ &= (n-1)\Gamma(n-1), \end{aligned}$$

since

$$\lim_{x\to\infty} e^{-x}x^{n-1} = 0.$$

In view of $\Gamma(1) = 1$ the proof is complete. □

C.2 Binomial coefficients

We call the numbers

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

the binomial coefficients. Now we will see why.

Theorem C.2.1 *(The binomial theorem) For any real numbers x and y and a nonnegative integer n*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k}y^k \tag{C.8}$$

holds.

Proof.

$$(x+y)^n = \underbrace{(x+y)\cdots(x+y)}_{n-\text{times}}. \tag{C.9}$$

So the coefficient of the term $x^{n-k}y^k$ is the number of ways of getting $x^{n-k}y^k$ when the n brackets are multiplied out. Each term in the expansion is the product of one term from each bracket; so $x^{n-k}y^k$ is obtained as many times as we can choose y from k brackets and x from the remaining $n-k$ brackets. But this can be done just in $\binom{n}{k}$ ways. □

Corollary C.2.2 *For any real number y and nonnegative integers n*

$$(1+y)^n = \sum_{k=0}^{n} \binom{n}{k} y^k \tag{C.10}$$

holds.

This corollary is the origin of

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n; \text{ and } \sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0. \tag{C.11}$$

It holds the following recursive relation:

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}. \tag{C.12}$$

The proof should be given by the reader.

Another property one observes that along any row, the entries increase until the middle, and then decrease. To verify this observation, we want to compare two consecutive entries:

$$\binom{n}{k} \; ? \; \binom{n}{k+1}. \tag{C.13}$$

Rearranging the above formula, we get the following one which is equivalent

$$k \; ? \; \frac{n-1}{2}. \tag{C.14}$$

So if $k < \frac{n-1}{2}$, then $\binom{n}{k} < \binom{n}{k+1}$; if $k = \frac{n-1}{2}$, then $\binom{n}{k} = \binom{n}{k+1}$ (the latter is the case for the two entries in the middle if n is odd); and if $k > \frac{n-1}{2}$, then $\binom{n}{k} > \binom{n}{k+1}$.
A (finite) sequence $a_1, a_2, \ldots, a_n$ of numbers is called unimodal if there is an index k such that

$$a_1 \leq a_2 \leq \ldots \leq a_{k-1} \leq a_k \geq a_{k+1} \geq \ldots \geq a_{n-1} \geq a_n. \tag{C.15}$$

And so we proved

Theorem C.2.3 *The sequence $\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n}$ of binomial coefficients is unimodal, whereby the middle element is the largest. If n is even, there is a unique middle; if n is odd, then there are two equal middle elements.*

In general it is difficult to compute the binomial coefficients for large values of n and k, but in many cases only upper or lower bounds are of interest. Several simple calculations gave the following result.

Theorem C.2.4

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \frac{1}{e}\left(\frac{en}{k}\right)^k. \tag{C.16}$$

Consider the identity

$$(1+x)^m(1+x)^n = (1+x)^{m+n}. \tag{C.17}$$

Using C.2.2 for both sides and comparing the coefficients gives us

$$\binom{m+n}{j} = \sum_{k=0}^{j} \binom{m}{k}\binom{n}{j-k}. \tag{C.18}$$

Specification for $j = n$ gives us

Theorem C.2.5 *(Vandermonde's convolution) Let m and n be positive integers. Then*

$$\sum_k \binom{n}{k}\binom{m}{k} = \binom{m+n}{n}. \tag{C.19}$$

Differentiation of both sides of C.2.2 gives us

$$n(1+y)^{n-1} = \sum_{k=1}^{n} k\binom{n}{k} y^{k-1}, \tag{C.20}$$

and by setting $y = 1$ we obtain

Theorem C.2.6

$$\sum_{k=0}^{n} k\binom{n}{k} = n2^{n-1}. \tag{C.21}$$

C.3 Multinomial coefficients

The number of arrangements of the four letters in BALL is not $24 = 4!$, since we do not have four distinct letters to arrange. The letter L occurs twice, and we have to count $4!/2 = 12$. Generalizing this idea we solved a new type of problem by relating it to the previous enumeration principles:

Lemma C.3.1 *If there are n objects of k types with n_i of the ith type, $i = 1, \ldots, k$, where $n_1 + \ldots + n_k = n$, then the number of arrangements equals*

$$\binom{n}{n_1 n_2 \ldots n_k} = \frac{n!}{n_1! \cdot n_2! \cdots n_k!}. \tag{C.22}$$

Proof.

$$\begin{aligned}
&\binom{n}{n_1} \cdot \binom{n-n_1}{n_2} \cdots \binom{n-\sum_{i=1}^{k-1} n_i}{n_k} \\
&= \frac{n!}{n_1!(n-n_1)!} \cdot \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \cdots \frac{(n-\sum_{i=1}^{k-1} n_i)!}{n_k!} \\
&= \frac{n!}{n_1! \cdot n_2! \cdots n_k!}.
\end{aligned}$$

□

For instance an RNA is a chain consisting of bases, each link of which is one of four possible chemical components: a,c,g,u. The number of such chains, where n_i, $i \in \{a, c, g, u\}$ denotes the number of the components, is given by

$$\frac{(n_a + n_c + n_g + n_u)!}{n_a! \cdot n_c! \cdot n_g! \cdot n_u!}. \tag{C.23}$$

The quantity

$$\binom{n}{n_1 n_2 \ldots n_k} = \frac{n!}{n_1! \cdot n_2! \cdots n_k!} \tag{C.24}$$

is called the multinomial coefficient, since it is a generalization of the binomial coefficients:

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} = \binom{n}{k \quad n-k}. \tag{C.25}$$

In expanding $(x_1 + x_2 + x_3)^7$, we think of writing down seven $(x_1 + x_2 + x_3)$ terms in a row, and then adding up $x_1^i x_2^j x_3^k$ for all ways of selecting x_1 from i of the terms, x_2 from j of the terms and x_3 from k of the terms. Note that $i + j + k = 7$ in each case. In view of C.3.1 we get:

Theorem C.3.2 *(The multinomial theorem) Let $x_1, \ldots, x_m$ be any real numbers and let n be a nonnegative integer, then*

$$(x_1 + \ldots + x_m)^n = \sum_{k_1 + \ldots + k_m = n} \frac{n!}{k_1! \cdots k_m!} x_1^{k_1} \cdots x_m^{k_m}, \tag{C.26}$$

where we read the sum sign that appears in the formula as "the sum over all lists $k_1, \ldots, k_m$ such that $k_1 + \ldots + k_m = n$".

With $x_1 = \ldots = x_m = 1$ we get

$$\sum_{k_1 + \ldots + k_m = n} \frac{n!}{k_1! \cdots k_m!} = m^n. \tag{C.27}$$

For further reading see [43], [87], [105], [108] or [170].

Appendix D

Geometric Series

D.1 The Towers of Hanoi

We consider the well-known Towers of Hanoi: Given n discs, all of different sizes. A collection of discs forms a tower if they are ordered according to their size, the largest at the bottom. Now

(i) All the n discs form a tower on position 1.

(ii) There are two other positions: 2 and 3.

(iii) Move the original tower from position 1 to position 3 by moving exactly one disc in one step, and by producing only towers at the diferent positions.

What is minimum number of steps required?[1]
Let a_n be the smallest number of steps required to move the n discs. It is easy to see that $a_1 = 1$, $a_2 = 3$ and $a_3 = 7$ (Exercise).
What about a_n? Forget the bottom disc and move the remaining $n-1$ discs to position 2. To get this stage, a_{n-1} steps are needed. Then move the disc from position 1 to position 3: one step. Now move the tower from position 2 to position 3. Altogether we need $a_{n-1} + 1 + a_{n-1}$ steps. Hence we have to solve

$$a_n = 2 \cdot a_{n-1} + 1, \tag{D.1}$$

with $a_1 = 1$ (and $a_0 = 0$).[2] Then $a_n = 2^n - 1$.

[1] Lucas, compare [105], furnished this toy with a romantic legend. The tower of Brahama, which supposedly has 64 disks of pure gold. At the beginning of the time, God placed these disks on the first needle and ordained that a group of priests should transfer them to the third, according to the rules above. The priests work day and night at their task. When they finish, the tower will crumble and the world will end.

[2] In reality, we only proved $a_n \leq 2 \cdot a_{n-1} + 1$, since these moves suffice. As exercise show that so many moves are also necessary.

D.2 The finite case

Suppose that the population of a colony of ants doubles in each successive year. A colony is established with an initial population of $a_0 = a$ ants. How many ants will this colony have after n years?
Let a_n denote this number. Then

$$a_n = 2 \cdot a_{n-1} = 2^2 \cdot a_{n-2} \ldots = 2^{n-1} \cdot a_1 = 2^n \cdot a.$$

Theorem D.2.1 *Let*

$$a_n = ca_{n-1} + g, \tag{D.2}$$

$n \geq 1$, *with given constants* c *and* g, *and an initial condition* a_0, *be a (nonhomogeneous) recurrence relation. Then*

$$a_n = \begin{cases} c^n a_0 + \frac{c^n - 1}{c-1} g & : \quad c \neq 1 \\ a_0 + ng & : \quad c = 1 \end{cases}$$

For $0 < c < 1$ *we get*

$$a_n \to \frac{1}{1-c} g.$$

Proof.

$$\begin{aligned} a_n &= c \cdot a_{n-1} + g \\ &= c \cdot (c \cdot a_{n-2} + g) + g = c^2 \cdot a_{n-2} + cg + g \\ &= c^2 \cdot (c \cdot a_{n-3} + g) + cg + g = c^3 \cdot a_{n-3} + c^2 g + cg + g \\ &\vdots \\ &= c^{n-1} \cdot a_1 + c^{n-2} g + \ldots + c^2 g + cg + g \\ &= c^n a_0 + c^{n-1} g + \ldots + c^2 g + cg + g \\ &= c^n a_0 + (c^{n-1} + \ldots + c^2 + c + 1) \cdot g. \end{aligned}$$

Now, we distinguish between $c = 1$ and $c \neq 1$. □

More about population biology we find in [128] and [233].

As an reformulation of D.2.1 we have for a real number x

$$R_n = \sum_{k=1}^{n} x^k = x \frac{1 - x^n}{1 - x}. \tag{D.3}$$

Now consider

$$S_n = \sum_{k=1}^{n} k x^k. \tag{D.4}$$

Then

$$\begin{aligned}(1-x)S_n &= \sum_{k=1}^{n} x^k - nx^{n+1} \\ &= R_n - nx^{n+1} \\ &= x\frac{1-x^n}{1-x} - nx^{n+1} \\ &= \frac{x-(n+1)x^{n+1}+nx^{n+2}}{1-x}.\end{aligned}$$

Corollary D.2.2

$$\sum_{k=1}^{n} kx^k = \frac{x-(n+1)x^{n+1}+nx^{n+2}}{1-x}. \tag{D.5}$$

D.3 Infinite series

For $0 < c < 1$ we have a convergent geometric series:

Theorem D.3.1 *Let*

$$a_n = ca_{n-1} + g, \tag{D.6}$$

$n \geq 1$, with given constants $0 < c < 1$ and g, and an initial condition a_0, be a (nonhomogeneous) recurrence relation. Then

$$a_n \to \frac{1}{1-c}g.$$

And,

Corollary D.3.2

$$\sum_{k=1}^{\infty} kx^k = \frac{x}{(1-x)^2}. \tag{D.7}$$

Appendix E

Polynomials and its Zeros

We consider polynomials of degree n, which are functions of the kind

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0, \tag{E.1}$$

where a_i are real numbers with $a_n \neq 0$. We will call this representation the monomial form of the polynomial.
One of the most important questions to manipulate polynomials is to determine the functional value(s) at one or more points.

Algorithm E.0.3 *If a polynomial p is given in its monomial form, then we can determine the values of p by Horner's method:*

1. $p := a_n$;
2. *For* $i = 1, 2, \ldots, n$ *do* $p := x \cdot p + a_{n-i}$.

E.1 Roots of polynomials

Usually the zero of a polynomial is called its root. The following fact is well-known from algebra, compare [161].

Lemma E.1.1 *The number α is a root of the polynomial $p(x)$ if and only if $p(x)$ is divisible by $x - \alpha$.*

In other terms, finding the roots of a polynomial is equivalent for finding its linear factors.
In examining the roots of polynomials we did not pose the question of whether every polynomial possesses roots. This is indeed true, when we not only restrict ourself to real numbers. The fundamental theorem of algebra is the most important discovery dealing with the roots of a polynomial.

Theorem E.1.2 *A polynomial of degree n has exactly n roots, that means the equation*

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0 = 0 \tag{E.2}$$

has exactly n solutions, provided each root is counted as many times as its multiplicity.[1]

Corollary E.1.3 *If the polynomials $p(x)$ and $q(x)$ whose degrees do not exceed n have equal values for more than n distinct values of the unknown, then $p(x) = q(x)$.*

Proof. The polynomial $p(x) - q(x)$ has degree at most n and more roots than n. By E.1.2 the equation $p(x) - q(x) = 0$ must be true. □

In view of E.1.2, there are unique real or complex numbers $\alpha_1, \ldots, \alpha_k$ and positive integers $m_1, \ldots, m_k$, with $m_1 + \ldots + m_k = n$, which make it possible to factorize p:

$$p(x) = a_n (x - \alpha_1)^{m_1} \cdots (x - \alpha_k)^{m_k}. \tag{E.3}$$

Let there be given the polynomial with leading coefficient $a_n = 1$ and let its roots by counting multiplicities, that means we have the following expansion

$$p(x) = (x - \alpha_1) \cdots (x - \alpha_n). \tag{E.4}$$

Multiplying out the parantheses on the right, and then collecting like terms and comparing the resulting coefficients with the coefficients of the polynomials its given form, we get the following formulas.

Theorem E.1.4 *Let*

$$p(x) = x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0 \tag{E.5}$$

a polynomial with the roots $\alpha_1, \ldots, \alpha_n$. Then Vieta's formulas hold true:

$$\begin{aligned}
a_0 &= (-1)^n \alpha_1 \cdots \alpha_n, \\
a_1 &= (-1)^{n-1} (\alpha_1 \alpha_2 \cdots \alpha_{n-1} + \alpha_1 \alpha_2 \cdots \alpha_{n-2} \alpha_n + \ldots + \alpha_2 \alpha_3 \cdots \alpha_n), \\
&\vdots \\
a_{n-2} &= \alpha_1 \alpha_2 + \alpha_1 \alpha_3 + \ldots + \alpha_1 \alpha_n + \alpha_2 \alpha_3 + \ldots + \alpha_{n-1} \alpha_n, \\
a_{n-1} &= -(\alpha_1 + \ldots + \alpha_n).
\end{aligned}$$

If the leading coefficient a_n of the polynomial $p(x)$ is different from unity, then in order to make use Vieta's formulas, it is first necessary to divide all the coefficients by a_n; this has no effect on the roots of the polynomial.

[1] Note that the theorem holds true for $n = 0$ as well, since a polynomial of zero degree has no roots. The theorem is not applicable to the polynomial 0, which has no degree, and is equal to zero for any value.

For polynomials of degree $n \leq 4$, there exist formulas for determination of the roots. In the 1820's Abel demonstrated that no such formulas can be found for for the nth-degree equations where $n \geq 5$. For instance the equation

$$x^5 - 4x - 2 = 0$$

is not solvable by radicals.
In the 1830's Galois made a complete investigation of the conditions under which a given equation is solvable by radicals. For the so-called Galois theory, see Artin [11] or Stewart [228].
In general the roots of a polynomial can be found only approximately. In practice, however, iterative methods are employed even for solving polynomial equations of third and fourth degree.

E.2 Estimating of roots

We are interested in bounds within which lie the roots of a polynomial, and to determine the number of the roots.

Theorem E.2.1 *Let*

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0 \tag{E.6}$$

be a polynomial ($a_n \neq 0$). Then

$$|\alpha| \leq 1 + \frac{A}{|a_n|}, \tag{E.7}$$

holds true for any root α of $p(x)$, where $A = \max\{|a_0|, \ldots, |a_{n-1}|\}$.

Proof.

$$a_n x^n = p(x) - (a_{n-1} x^{n-1} + \ldots + a_1 x + a_0)$$

implies

$$|a_n x^n| \leq |p(x)| + |a_{n-1} x^{n-1} + \ldots + a_1 x + a_0|.$$

In view of this inequality we find

$$\begin{aligned}
|p(x)| &\geq |a_n x^n| - |a_{n-1} x^{n-1} + \ldots + a_1 x + a_0| \\
&\geq |a_n| \cdot |x|^n - A(|x|^{n-1} + \ldots + |x| + 1) \\
&= |a_n| \cdot |x|^n - A \frac{|x|^n - 1}{|x| - 1} \\
&> |a_n| \cdot |x|^n - \frac{A}{|x| - 1} |x|^n \quad \text{assuming } |x| > 1 \\
&= (|a_n| - \frac{A}{|x| - 1}) |x|^n.
\end{aligned}$$

The factor $|a_n| - A/(|x|-1)$ is strictly positiv if and only if

$$|x| > \frac{A}{|a_n|} + 1,$$

which is also satisfied for $|x| \leq 1$. □

Corollary E.2.2 *There is a number c, such that for $|x| > c$ the value of the polynomial $p(x)$ has the same sign as the leading term $a_n x^n$.*

This implies the following, a little bit unexpected, fact.

Corollary E.2.3 *A polynomial of odd degree has at least one root.*

As an exercise prove the following, more exact, version of these corollaries.

Theorem E.2.4 *Let*

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0 \tag{E.8}$$

be a polynomial of odd degree. Assume that $a_n > 0$, then

$$\left\{ \begin{array}{l} p(x) > 0 \\ p(x) < 0 \end{array} \right\} \text{ for } \left\{ \begin{array}{l} x > 1 + \frac{B}{a_n} \\ x < -(1 + \frac{B}{a_n}), \end{array} \right\}$$

where $B = |a_{n-1}| + \ldots + |a_1| + |a_0|$. Similarly for $a_n < 0$.

E.3 A randomized algorithms

Computers appear to behave far too unpredictably as it is. Adding randomness would seemingly be a disadvantage, adding further complications to the already challenging task of efficiently utilizing computers.
Randomized algorithmss are algorithms that make random choice during their execution. In practice, a randomized program would use values generated by a random number generator to decide the next step at several branches of its execution.

Mitzenmacher and Upfal [175] consider the following problem of polynomial identity:
Given: Two polynomials $f(x)$ and $g(x)$.
Verify: The identity, that means $f(x) \equiv g(x)$. It is well-known, that the following algorithm verifies the identity $f(x) \equiv g(x)$:

1. Convert the two polynomials to their canonical form[2];

[2] that means of the form $\sum_{k=0}^{n} c_k x^k$

2. $f(x)$ and $g(x)$ are equivalent if and only if all the coefficients in their canonical form are equal.

In particular, if $f(x)$ is given as a product

$$f(x) = \Pi_{k=1}^{n}(x - a_k)$$

and $g(x)$ is given in its canonical form. Transforming $f(x)$ to its canonical form by consecutively multiplying of the monomials requires $\Theta(n^2)$ multiplications of coefficients. Assuming that multiplying and adding numbers can be performed in constant time we have an algorithm which requires quadratic time. Let us instead utilize randomness to obtain a faster method to verify the identity. Our new algorithm is the following:

Algorithm E.3.1 *1. Determine the largest exponent n of x in $f(x)$ and $g(x)$;*

2. Choose a positive integer m;

3. Chooses an integer r uniformly at random in the range $\{1, \ldots, mn\}$;

4. Compute $f(r)$ and $g(r)$;

5. If $f(r) \neq g(r)$ the two functions are not equivalent; otherwise they are equivalent.

The algorithm runs in linear time (Exercise). But E.3.1 may give a wrong answer. More exactly:

a) If $f(x) \equiv g(x)$, then the algorithm gives the correct answer, since it will find $f(x) = g(x)$ for any value of x;

b) If $f(x) \not\equiv g(x)$ and $f(r) \neq g(r)$, then the algorithm gives the correct answer, since it has found a case where the functions disagree;

c) If $f(x) \not\equiv g(x)$ and $f(r) = g(r)$, then the algorithm gives the wrong answer.

We take a closer look at the third case. For this error to occur r must be a root of the equation

$$f(x) - g(x) = 0. \tag{E.9}$$

The degree of the polynomial on the left side is at most n and, by the fundamental theorem of algebra, $f(x) - g(x)$ has no more than n roots. And therefore no more than n values in $\{1, \ldots, mn\}$ for which the polynomials are equal.

Theorem E.3.2 *The algorithm E.3.1 running in linear time gives a wrong answer with probability $1/m$, where the integer m can be arbitrarily chosen.*

For all questions arising randomized algorithms see Motwani and Raghavan [176].

Appendix F

Recurrence Relations

> Recursion: see recursion.
>
> David Darling in The Universal Book of Mathematics [62]

F.1 Fibonacci's rabbits

In his famous book *Liber Abaci*, Fibonacci raised the following question

> A certain man put a pair of rabbits in a place surround on all sides by a wall. How many pairs of rabbits can be produced from that pair in a year if it is supposed that every month each pair begets a new pair which from the second month on becomes productive.

For convenience, we will count the rabbits in male-female pair. F_0 represents the initial population, and F_i represents the population in the ith month. Let $f_i = |F_i|$ denote the total number of pairs in the ith generation.

generation	F_0	F_1	F_2	F_3	F_4	F_5	F_6
number of mature pairs	0	1	1	2	3	5	8
number of baby pairs	1	0	1	1	2	3	5
f_i	1	1	2	3	5	8	13

And further

f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
21	34	55	89	144	233	377	610	987	1597

We can see from this table that

$$f_n = f_{n-1} + f_{n-2}, \tag{F.1}$$

for $n \geq 2$ with $f_0 = 1$ and $f_1 = 1$.

F.2 Handle recurrences with care

Consider the following sequence

$$a_n = \begin{cases} 1 + a_{\frac{n}{2}} & : \text{ n even} \\ 1 + a_{3n-1} & : \text{ otherwise} \end{cases}$$

with $a_1 = 1$. We get

$$a_3 = 1 + a_8 = 1 + (1 + a_4) = 2 + a_4 = 2 + (1 + a_2) = 3 + a_2 = 3 + (1 + a_1) = 5,$$

but

$$a_5 = 1 + a_{14} = 2 + a_7 = 3 + a_{20} = 4 + a_{10} = 5 + a_5,$$

which means that a_5 is not defined.
Another example, first posed by Collatz in 1937, starts with a positive integer z such that $a_1 = z$ and

$$a_n = \begin{cases} a_{n-1}/2 & : \ a_{n-1} \text{ even} \\ 3a_{n-1} + 1 & : \text{ otherwise} \end{cases}$$

For $z = 7$ we get $7, 22, 11, 34, 17, 52, 26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1$ and the last three digits $4, 2, 1$ being continually repeated; exact the same situation we find for $z = 6$: $6, 3, 10, \ldots, 4, 2, 1$.
It is an open problem whether or not for each choice of z the sequence has a period $4, 2, 1$. As an exercise start the sequence with $z = 27$. For a nice description of this question compare [151] or [152].

F.3 Recurrence relations of second order

Now, we concentrate on the recurrence relation

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} \tag{F.2}$$

with given a_0, a_1, and constants c_1, c_2, where $c_2 \neq 0$. There is a very neat method for solving such relations.
Substituting α^n for a_n with $\alpha \neq 0$ in (F.2) gives $\alpha^n = c_1\alpha^{n-1} + c_2\alpha^{n-2}$, that is $\alpha^2 = c_1\alpha + c_2$. Consequently, α^n is a solution of (F.2) if and only if α is a solution of the so-called characteristic equation

$$x^2 = c_1 x + c_2. \tag{F.3}$$

We have to distinguish between two cases:

1. We assume that α and β are distinct solutions of (F.3). Then α^n and β^n and also their linear combination satisfy (F.2):

   $$a_n = d_1\alpha^n + d_2\beta^n. \tag{F.4}$$

   Choose d_1, d_2 so that

   $$\begin{aligned} a_0 &= d_1 + d_2 \\ a_1 &= d_1\alpha + d_2\beta. \end{aligned}$$

2. On the other hand, when the characteristic equation has a repeated root α, that means of multiplicity two, then

$$x^2 - c_1 x - c_2 = (x - \alpha)^2 = x^2 - 2\alpha x + \alpha^2 \tag{F.5}$$

so that $c_1 = 2\alpha$ and $c_2 = -\alpha^2$. $n\alpha^n$ also satisfies (F.2), since

$$\begin{aligned} c_1 a_{n-1} + c_2 a_{n-2} &= c_1(n-1)\alpha^{n-1} + c_2(n-2)\alpha^{n-2} \\ &= 2(n-1)\alpha^n - (n-2)\alpha^n \\ &= n\alpha^n \\ &= a_n. \end{aligned}$$

Then we consider the linear combination

$$a_n = d_1\alpha^n + d_2 n\alpha^n. \tag{F.6}$$

Choose d_1, d_2 so that

$$\begin{aligned} a_0 &= d_1 \\ a_1 &= d_1\alpha + d_2\alpha. \end{aligned}$$

Solving these equations we get the following theorem.

Theorem F.3.1 *Suppose that $\{a_n\}$ satisfies the recurrence relation*

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} \tag{F.7}$$

with given a_0, a_1.
Let α and β be the roots of the characteristic equation

$$x^2 - c_1 x - c_2 = 0. \tag{F.8}$$

a) *If $\alpha \neq \beta$ then*

$$a_n = \frac{a_1 - a_0\beta}{\alpha - \beta} \cdot \alpha^n + \frac{a_0\alpha - a_1}{\alpha - \beta} \cdot \beta^n. \tag{F.9}$$

b) *If $\alpha = \beta$ then*

$$a_n = ((1-n) \cdot a_0 + n \cdot \frac{a_1}{\alpha}) \cdot \alpha^n. \tag{F.10}$$

F.4 Phylotaxis

Recall Fibonacci's rabbits to find the characteristic equation in $x^2 - x - 1 = 0$, with the roots

$$\alpha, \beta = \frac{1 \pm \sqrt{5}}{2},$$

Example F.4.1 *The Fibonacci sequence*

$$f_n = f_{n-1} + f_{n-2}, \tag{F.11}$$

for $n \geq 2$ *with* $f_0 = 1$ *and* $f_1 = 1$*, has the solution*

$$f_n = \frac{1}{\sqrt{5}} \left(\left(\frac{1+\sqrt{5}}{2} \right)^{n+1} - \left(\frac{1-\sqrt{5}}{2} \right)^{n+1} \right). \tag{F.12}$$

This result is strange, since f_n is in any case an integer.
Note that for specific values of n it is easier to determine f_n in the preceding example by using the recurrence relation starting with the initial conditions than to solve the equations.

The number

$$\frac{1+\sqrt{5}}{2} = 1.61803\ldots \tag{F.13}$$

is important in many parts of mathematics as well as in the art world since ancient times.[1]
In the nineteenth century Fibonnacci numbers were discovered in many natural forms. For example, many types of flower have a Fibonacci number of petals: certain types of daisies tend to have 34 or 55 petals, while sunflowers have 89 or 144.
The understanding of these relations is called phylotaxis, see [58] or [233].

F.5 A general solution method

Generalizing the method used in F.3.1 we outline the theory of solving recurrence relations of the form

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \ldots + c_k a_{n-k} \tag{F.14}$$

with given

$$a_0, \ldots, a_{k-1}, \tag{F.15}$$

where the c_i's are given constants, $c_k \neq 0$.

The recurrence allows us to compute a_n for any n we like, but it only gives indirect information. A solution to the recurrence in a "closed form" helps us to understand what a_n really stands for, compare our investigations about Fibonacci's rabbits.
There is a general solution technique involving a sum of individual solutions of the form $n^r \cdot \alpha^n$.

Algorithm F.5.1 *Let a recurrence relation (F.14), (F.15) be given. Then*

[1] Therefore it has a special name, the Golden Ratio, and in general denoted by Φ, in honor of Phidas, compare [105].

1. *Solve the characteristic equation*

$$x^k - c_1 x^{k-1} - c_2 x^{k-2} - \ldots - c_k = 0. \tag{F.16}$$

It has k roots, some of which may be multiple.[2]

2. *If $\alpha_1, \ldots, \alpha_r$ are the roots of (F.16), then $a_n = \alpha_i^n$ is a solution of the recurrence equation (F.14).*
Compose a linear combination for the roots in the following sense: If root α has multiplicity m then use

$$\alpha^n, n\alpha^n, n^2\alpha^n, \ldots, n^{m-1}\alpha^n. \tag{F.17}$$

3. *We need to be given the initial conditions of the first k values (F.15).*
The k equations can be solved if we insert these conditions. This forms a system of k linear equations with k unknowns, which is simple to solve.

F.6 Triangles of numbers

Recursion is a process that wraps back on itself and feeds the output of a process back in as the input.
It often happens that in studying a sequence of numbers, a connection between the current value and several of the previous values is obtained. Recall the triangles of the binomial coefficients $b(n,k) = \binom{n}{k}$, the Stirling numbers $s(n,k)$ of the first and $S(n,k)$ of the second kind. We constructed these triangles by

$$b(n,k) = b(n-1,k) + b(n-1,k-1), \tag{F.18}$$

$$s(n,k) = (n-1) \cdot s(n-1,k) + s(n-1,k-1), \tag{F.19}$$

$$S(n,k) = k \cdot S(n-1,k) + S(n-1,k-1), \tag{F.20}$$

A nice description of the coefficients $b(n,k)$ is given by the so-called Pascal's triangle, which displays C.12:

row 0						1						
row 1					1		1					
row 2				1		2		1				
row 3			1		3		3		1			
row 4		1		4		6		4		1		
row 5	1		5		10		10		5		1	
row 6	1	6		15		20		15		6		1
$\vdots$												$\ddots$

For the Stirling number of the first and the second kind in the form of a right triangle will be give below.
The numbers in such triangle satisfy, practically speaking, infinitely many identities for the recurrence relation.

[2] A solution is explicitly possible if $k \leq 4$, but maybe not for the case $k > 4$. For considerations concerning these questions, the so-called Galois theory, see Artin [11].

Appendix G

Inequalities

G.1 Bernoulli's inequality

Theorem G.1.1 *Let a be a positive real number and let $n \geq 2$ be an integer. Then*

$$(1+a)^n > 1 + na. \tag{G.1}$$

Proof. In view of C.2.2 we have

$$(1+a)^n = \binom{n}{0}a^0 + \binom{n}{1}a^1 + \underbrace{\binom{n}{2}a^2 + \ldots}_{>0} > 1 + na.$$

□

G.2 The Cauchy-Schwarz inequality

Theorem G.2.1 *Let v and w be vectors in a space with an inner product $(.,.)$. Then it holds the so-called Cauchy-Schwarz inequality:*

$$|(v,w)|^2 \leq (v,v) \cdot (w,w). \tag{G.2}$$

Equality holds if and only if v and w are linearly dependent.

Proof. For all real numbers α, we have

$$\begin{aligned} 0 &\leq (v - \alpha w, v - \alpha w) \\ &= (v,v) - 2\alpha(v,w) + \alpha^2(w,w). \end{aligned}$$

Taking $\alpha = (v,w)/(w,w)$, we get

$$0 \leq (v,v) - 2\frac{(v,w)}{(w,w)}(v,w) + \frac{(v,w)^2}{(w,w)^2}(w,w)$$

$$
\begin{aligned}
&= (v,v) - 2\frac{(v,w)^2}{(w,w)} + \frac{(v,w)^2}{(w,w)} \\
&= (v,v) - \frac{(v,w)^2}{(w,w)}.
\end{aligned}
$$

Hence,

$$(v,w)^2 \leq (v,v)\cdot(w,w). \tag{G.3}$$

□

In an inner product space we define a map from the space into the real numbers by $||v|| = \sqrt{(v,v)}$. Consequently,

$$|(v,w)| \leq ||v||\cdot||w|| \tag{G.4}$$

G.3 Arithmetic and geometric means

Let $x_1,\ldots,x_n$ be nonnegative real numbers. The geometric mean for these numbers is defined by

$$G(x_1,\ldots,x_n) = \sqrt[n]{x_1\cdots x_n}, \tag{G.5}$$

and the arithmetic mean by

$$A(x_1,\ldots,x_n) = \frac{x_1+\ldots+x_n}{n}. \tag{G.6}$$

Observation G.3.1 *If $x_1 \geq x_2 \geq \ldots \geq x_n$, then*

$$G(x_1,\ldots,x_k) \geq G(x_1,\ldots,x_n) \tag{G.7}$$

and

$$A(x_1,\ldots,x_k) \geq A(x_1,\ldots,x_n) \tag{G.8}$$

for $1 \leq k \leq n$.

Proof. Since $x_1 \geq x_2 \geq \ldots \geq x_k \geq x_{k+1}$ we have

$$x_1\cdots x_k \geq x_{k+1}^k. \tag{G.9}$$

Furthermore, in view of (G.9),

$$(x_1\cdots x_k)^{k+1} = (x_1\cdots x_k)^k(x_1\cdots x_k) \geq (x_1\cdots x_k)^k x_{k+1}^k = (x_1\cdots x_{k+1})^k,$$

which gives the asserted result.
The second inequality follows with similar arguments. □

As an exercise determine the limit of the sequence a_n for which the general term is the arithmetic mean of its two preceding terms: $a_{n+2} = A(a_n, a_{n+1})$.
Now, we will prove the very important inequality $G \leq A$. Many different and ingenious proofs of this general result have been devised. The simplest way is the following.

Theorem G.3.2 *Let $x_1, \ldots, x_n$ be nonnegative real numbers. Then*

$$G(x_1, \ldots, x_n) \leq A(x_1, \ldots, x_n), \tag{G.10}$$

where equality holds if and only if $x_1 = \ldots = x_n$.

Proof. Let $s = \sum_{i=1}^{n} x_i$. Consider the function

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} x_i. \tag{G.11}$$

Since the set $\{(x_1, \ldots, x_n) : x_i \geq 0, \sum_{i=1}^{n} = s\}$ is compact, the quantity Max$f$ exists. We may assume that all x_i are positive.
Let $x_1 \neq x_2$, then for $y_1 = y_2 = \frac{x_1+x_2}{2}$ it holds

$$y_1 + y_2 + x_3 + \ldots + x_n = s, \tag{G.12}$$

and

$$y_1 y_2 - x_1 x_2 = \left(\frac{x_1 + x_2}{2}\right)^2 - x_1 x_2 = \left(\frac{x_1 - x_2}{2}\right)^2 > 0. \tag{G.13}$$

Consequently, $y_1 y_2 x_3 \cdots x_n > x_1 \cdots x_n$. In the same way we can prove that $x_1 = x_i$, where x_i is any one of the x's and we may assume that Max$f$ is achieved if $x_i = x_j = x$ for all i, j. Then $s = nx$.

$$\prod_{i=1}^{n} x_i = f(x_1, \ldots, x_n) \leq f\left(\frac{s}{n}, \ldots, \frac{s}{n}\right) = \left(\frac{s}{n}\right)^n = \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^n.$$

This gives the assertion. □

G.4 Means generated by integrals

Chen [46] and Eves [82] give a surprising generalization of many different means. Let a and b two distinct positive real numbers. We define a general mean by

$$F(a, b)(t) = \frac{\int_a^b x^{t+1} dx}{\int_a^b x^t dx}. \tag{G.14}$$

This definition encompass the following specific (and well-known) means. Verify as an exercise:

$$\text{Harmonic mean: } F(a, b)(-3) = H(a, b) = \frac{2ab}{a + b}; \tag{G.15}$$

$$\text{Geometric mean: } F(a, b)(-3/2) = G(a, b) = \sqrt{ab}; \tag{G.16}$$

$$\text{Logarithmic mean: } F(a, b)(-1) = L(a, b) = \frac{b - a}{\ln b - \ln a}; \tag{G.17}$$

$$\text{Herionian mean: } F(a,b)(-1/2) = N(a,b) = \frac{a+\sqrt{ab}+b}{3}; \tag{G.18}$$

$$\text{Arithmetic mean: } F(a,b)(0) = A(a,b) = \frac{a+b}{2}; \quad \text{and} \tag{G.19}$$

$$\text{Centroidal mean: } F(a,b)(1) = T(a,b) = \frac{2(a^2+ab+b^2}{3(a+b)}. \tag{G.20}$$

Theorem G.4.1 *The function $F(a,b)(t)$ is strictly increasing.*

Proof. To prove that $F(t) = F(a,b)(t)$ is strictly increasing for $0 < a < b$, we show that $F'(t) > 0$.
By the quotient rule,

$$F'(t) = \frac{\int_a^b x^{t+1}\ln x dx \cdot \int_a^b x^t dx - \int_a^b x^{t+1}dx \cdot \int_a^b x^t \ln x dx}{(\int_a^b x^t dx)^2}. \tag{G.21}$$

Since the bounds of the definite integrals are constant, the numerator of this quotient can be written as

$$\begin{aligned} &= \int_a^b x^{t+1}\ln x dx \cdot \int_a^b y^t dy - \int_a^b y^{t+1}dy \cdot \int_a^b x^t \ln x dx \\ &= \int_a^b \int_a^b x^t y^t \ln x(x-y) dxdy. \end{aligned}$$

Substituting in a different manner, we write the same numerator as

$$\begin{aligned} &= \int_a^b y^{t+1}\ln y dy \cdot \int_a^b x^t dx - \int_a^b x^{t+1}dx \cdot \int_a^b y^t \ln y dy \\ &= \int_a^b \int_a^b x^t y^t \ln y(y-x) dxdy. \end{aligned}$$

Averaging the two equivalent expressions shows that this numerator is

$$\frac{1}{2}\int_a^b \int_a^b x^t y^t (x-y)(\ln x - \ln y) dxdy > 0,$$

as long as $0 < a < b$. In view of (G.21), this implies that $F'(t) >= 0$ as desired. □

An immediate consequence is the following result.

Corollary G.4.2

$$H(a,b) \le G(a,b) \le L(a,b) \le N(a,b) \le A(a,b) \le T(a,b). \tag{G.22}$$

Equality holds if and only if $a = b$.

As an exercise compute $F(a,b)(-2)$ and find further inequalities for means.

Appendix H

The Harmonic Numbers

H.1 The sequence of harmonic numbers

We are interested in the sequence $\{H_n\}$ of the sum of the reciprocals:

$$H_n = \sum_{k=1}^{n} \frac{1}{k}, \tag{H.1}$$

which is called the nth harmonic number.

About the convergence behavior of $\{H_n\}_{n=1,2,\ldots}$ we know.

Theorem H.1.1 *The sequence $\{H_n\}$ of the harmonic numbers diverges.*

Proof. We group the terms of H_n according to the powers of 2.

$$\underbrace{\frac{1}{1}}_{\text{group 1}} + \underbrace{\frac{1}{2}+\frac{1}{3}}_{\text{group 2}} + \underbrace{\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}}_{\text{group 3}} + \underbrace{\frac{1}{8}+\frac{1}{9}+\ldots+\frac{1}{14}+\frac{1}{15}}_{\text{group 4}} + \ldots.$$

Each of the 2^{k-1} terms in group k is between 2^{1-k} and 2^{-k}; hence, the sum of each group is between $1/2$ and 1.
This procedure shows us that if n is in group k, we must have $H_n > k/2$; thus $H_n \to \infty$. □

On the other hand, consider the Riemann zeta function $\varsigma(s)$ defined for any real $s > 1$ by

$$\varsigma(s) = \sum_{k\geq 1} \frac{1}{k^s}. \tag{H.2}$$

For real $s > 1$ the series converges, since for

$$\varsigma_n(s) - 1 = \sum_{k=1}^{n} \frac{1}{k^s} - 1 = \sum_{k=2}^{n} \frac{1}{k^s} \leq \int_1^n \frac{dx}{x^s} = \frac{1}{1-s}(n^{1-s} - 1),$$

and n running to infinity, we get:

$$\varsigma(s) \leq \frac{s}{s-1}. \tag{H.3}$$

Several exact values of the function are known:

s	2	4	6	8
$\varsigma(s)$	$\frac{\pi^2}{6}$	$\frac{\pi^4}{90}$	$\frac{\pi^6}{945}$	$\frac{\pi^8}{9450}$

H.2 Approximations

H_n is unbounded, but how fast does it increase? The proof of H.1.1 gave a hint of an approximation; we will create a better one. Consider the function $f(x) = 1/x$. The area under the curve between 1 and n, which is $\int_1^n \frac{dx}{x}$, is greater than the area of the lower rectangles and less than the area of the upper trapezium:

$$\sum_{k=2}^{n} \frac{1}{k} \leq \int_1^n \frac{dx}{x} \leq \sum_{k=1}^{n-1} \frac{1}{2}\left(\frac{1}{k} + \frac{1}{k+1}\right)$$

$$\sum_{k=1}^{n} \frac{1}{k} - 1 \leq \ln n \leq \frac{1}{2}\left(\sum_{k=1}^{n-1} \frac{1}{k} + \sum_{k=1}^{n-1} \frac{1}{k+1}\right)$$

$$H_n - 1 \leq \ln n \leq \frac{1}{2}\left(H_n - \frac{1}{n} + H_n - 1\right)$$

and consequently

Theorem H.2.1 *The harmonic number H_n is bounded by*

$$\ln n + \frac{1}{2} + \frac{1}{2n} \leq H_n \leq \ln n + 1. \tag{H.4}$$

In other terms

$$H_n = \ln n + \Theta(1). \tag{H.5}$$

This theorem allows us to conclude that the millionth harmonic number is

$$H_{1,000,000} = 14.39272\ldots.$$

Furthermore, it holds

$$H_n \approx \ln n + \frac{1}{2n} + \gamma, \tag{H.6}$$

where the quantity

$$\gamma = \lim_{n\to\infty}(H_n - \ln n) \tag{H.7}$$

is called Euler's constant or Mascheroni's constant and is calculated approximately by $\gamma = 0.57721\ldots$. This number is one of the most mysterious of all arithmetic constants. In particular, it is not even known whether γ is irrational.

For more and deeper facts about harmonic numbers and Euler's constant compare [105] and [129].[1]

[1] An amusing problem which shows the strangeness of the world is the so-called "worm on the rubber band", [95], [96], [129]. A slow, but persistent worm W starts at one end of a one meter-long rubber band and crawls one centimeter per minute toward the other end. At the end of each minute, an (equally persistent) keeper K stretches the band on one meter. During the stretching W maintains his relative position. After W crawls for another minute, K stretches the band again for one meter. And so on. Does the worm ever reach the end? At first glance it seems as if not; he keeps moving, but the goal seems to move away even faster.
When K stretches the rubber band, the fraction of it that W has crawled stays the same. Thus he crawls along 1/100th of the rubber band in the first minute, along 1/200th in the second; and so on along $1/n \cdot 100$th in the nth. After n minutes the fraction of the band that W has crawled is

$$\frac{1}{100} + \frac{1}{200} + \ldots + \frac{1}{n \cdot 100} = \frac{H_n}{100}.$$

Consequently, W reaches the finish if ever H_n surpass 100. In view of H.1.1 this must happen at some time.
On the other hand, our facts about the growing of the harmonic numbers tell us that this event will happen when

$$\ln n + \gamma \geq 100,$$

that means when n is approximately $e^{100-\gamma}$. We can imagine W's triumph when he crosses the finish line at last: 287 decillion centuries after starting his long crawl.

Look for an old problem which has a deepth consequence, compare [129]: You have to cross the desert by jeep. There are the following limiting facts for your trip:

(i) There are no sources of fuel in the desert.

(ii) You cannot carry enough fuel in a jeep in order to make the crossing in one go.

(iii) You do not have time to establish fuel dumps.

On the other hand, you have a large supply of jeeps (and drivers), but none of which you want to lose.
Can you cross the desert, and if the answer is "yes" what is the minimum amount of fuel? In a first view we assume that the answer will be "no"; but our "worm" let us be carefully.

Appendix I

The Order of Magnitude of the Factorials

I.1 Stirling's inequalities

The quantity $n!$, spoken "n factorial", increases very quickly. We describe the order of growing by the following considerations: In calculus, an integral can be regarded as the area under a curve, and we can approximate this area by adding up long, "skinny" rectangles that touch the curve. Consider the function $\ln x$. Then

$$\sum_{k=1}^{n-1} \ln k \leq \int_1^n \ln x \, dx \leq \sum_{k=2}^{n} \ln k$$

$$\ln \prod_{k=1}^{n-1} k \leq n \ln n - n + 1 \leq \ln \prod_{k=2}^{n} k$$

$$\ln \frac{n!}{n} \leq \ln n^n - n + 1 \leq \ln n!$$

exponentiating

$$\frac{n!}{n} \leq e\frac{n^n}{e^n} \leq n!$$

rearranging gives the

Observation I.1.1 *(Stirling's inequalities)*

$$e\frac{n^n}{e^n} \leq n! \leq en\frac{n^n}{e^n}. \tag{I.1}$$

I.2 Approximations

When we use the inequality

$$\sum_{k=1}^{n-1} \frac{1}{2}(\ln(k+1) + \ln k) \leq \int_1^n \ln x \, dx, \tag{I.2}$$

coming from an approximation of $\int \ln x \, dx$ by trapezium, we get the better bound

$$n! \leq e\sqrt{n} \cdot \frac{n^n}{e^n}. \tag{I.3}$$

An approximation from below with the same order uses the approximation of $\int \ln x \, dx$ by trapezium under the lines $y = \frac{1}{k}x + \ln k - 1$. We get

$$\begin{aligned}
\int_1^n \ln x \, dx &\leq \sum_{k=2}^{n} \frac{1}{2}\left(\left(\frac{k-1}{k} + \ln k - 1\right) + (1 + \ln k - 1)\right) \\
&= \sum_{k=2}^{n} \left(\ln k - \frac{1}{2k}\right) \\
&= \sum_{k=2}^{n} \ln k - \frac{1}{2}\sum_{k=2}^{n} \frac{1}{k} \\
&= \sum_{k=1}^{n} \ln k - \frac{1}{2}(H_n - 1) \\
&= \ln n! - \frac{1}{2}H_n + \frac{1}{2}.
\end{aligned}$$

Hence,

$$n \ln n - n + 1 \leq \ln n! - \frac{1}{2}H_n + \frac{1}{2}.$$

Further going

$$\begin{aligned}
\ln n! &\geq n \ln n - n + \frac{1}{2}H_n + \frac{1}{2} \\
&\geq n \ln n - n + \frac{1}{2}\left(\ln n + \frac{1}{2}\right) + \frac{1}{2} \quad \text{in view of H.2.1} \\
&= n \ln n - n + \frac{1}{2}\ln n + \frac{3}{4}
\end{aligned}$$

which immediately gives

$$n! \geq e^{3/4}\sqrt{n} \cdot \frac{n^n}{e^n}. \tag{I.4}$$

Altogether,

Theorem I.2.1 *For the order of growing of the factorials the inequalities (I.3) and (I.4) hold. Consequently,*

$$n! = \Theta\left(\sqrt{n} \cdot \frac{n^n}{e^n}\right). \tag{I.5}$$

And

$$n! = O(2^{(n+0.5)\log n}). \tag{I.6}$$

In view of

$$e^{3/4} = 2.11701\ldots \tag{I.7}$$

$$\sqrt{2\pi} = 2.5066\ldots \tag{I.8}$$

$$e = 2.7183\ldots, \tag{I.9}$$

this is not far from

Remark I.2.2 *(Stirling's equality)*

$$n! \approx \sqrt{2\pi n} \cdot \frac{n^n}{e^n}. \tag{I.10}$$

For further reading see [166].

Appendix J

Decomposition of permutations

J.1 The Stirling Number of the first kind

Each permutation can be written as a product of cycles. For instance:

$$\begin{array}{ccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 4 & 3 & 7 & 6 & 5 & 1 & 9 & 8 \end{array}$$

is (1247)(3)(56)(89).
Conversely, every cycle arrangement defines a permutation if we reverse the construction. In other words, permutations and cycle arrangements are essentially the same thing.
The Stirling number of the first kind is defined as follows: $s(n,k)$ is the number of permutations of $1,\ldots,n$ consisting of exactly k cycles.
As an example consider the permutations of the set $\{1,2,3,4\}$. Clearly, these 24 permutations can be classified according to the cycles they have as follows.

1. There are 6 permutations with exactly one cycle: (1234), (1324), (2134), (2314), (3124) and (3214).
2. There are 11 permutations with exactly two cycles: (123)(4), (124)(3), (134)(2), (234)(1), (132)(4), (142)(3), (143)(2), (243)(1), (12)(34), (13)(24) and (14)(23).
3. There are 6 permutations with exactly three cycles: (12)(3)(4), (13)(2)(4), (14)(2)(3), (23)(1)(4), (24)(1)(3) and (34)(1)(2).
4. There is only one permutation with four cycles: (1)(2)(3)(4).

Here are several elementary facts about the Stirling number of the first kind: $s(n,n) = 1$ and

$$\sum_{k=1}^{n} s(n,k) = n!. \tag{J.1}$$

Theorem J.1.1

$$s(n,1) = (n-1)!. \tag{J.2}$$

Proof. If we select the first $n-1$ elements, then the nth is determined. Selecting $n-1$ elements can be done in $(n-1)!$ ways. □

Theorem J.1.2

$$s(n,n-1) = \binom{n}{2}. \tag{J.3}$$

The proof remains as an exercise for the reader.

Theorem J.1.3 *For $n,k \geq 2$ it holds*

$$s(n,k) = (n-1)\cdot s(n-1,k) + s(n-1,k-1). \tag{J.4}$$

Proof. Consider n elements and the nth element explicitly. We distinguish two cases.
Case 1: n forms a 1-cycle on its own.
There are $s(n-1,k-1)$ ways to do this.
Case 2: n can be slotted into a cycle.
There are $s(n-1,k)$ of such permutations of $1,\ldots,n-1$ with k cycles. Then n can be inserted after one of the numbers $1,\ldots,n-1$. □

It is customary to write the Stirling numbers as Stirling's triangle (of the first kind) in the form of a right triangle.

$n \setminus k$	1	2	3	4	5	6	7	8
1	1							
2	1	1						
3	2	3	1					
4	6	11	6	1				
5	24	50	35	10	1			
6	120	274	225	85	15	1		
7	720	1764	1624	735	175	21	1	
8	5040	13068	13132	6769	1960	322	28	1

J.2 $s(n,2)$ and the harmonic numbers

In view of J.1.3 and J.1.1 we have

$$\begin{aligned} s(n,2) &= (n-1)s(n-1,2) + (n-2)! && \text{(J.5)} \\ \text{with } s(2,2) &= 1. && \text{(J.6)} \end{aligned}$$

Repeatedly applications give

$$\begin{aligned}
s(n,2) &= (n-1)s(n-1,2) + (n-2)! \\
&= (n-1)(n-2)s(n-2,2) + (n-1)(n-3)! + (n-2)! \\
&= (n-1)(n-2)(n-3)s(n-3,2) + (n-1)(n-2)(n-4)! \\
&\quad +(n-1)(n-3)! + (n-2)! \\
&= (n-1)(n-2)(n-3)s(n-3,2) + \frac{(n-1)!}{(n-3)} + \frac{(n-1)!}{(n-2)} + \frac{(n-1)!}{(n-1)} \\
&\vdots \\
&= (n-1)(n-2)\cdots 2 \cdot s(2,2) + (n-1)!\left(\frac{1}{n-1} + \frac{1}{n-2} + \ldots + \frac{1}{2}\right) \\
&= (n-1)! + (n-1)!\sum_{k=2}^{n-1}\frac{1}{k} \\
&= (n-1)! \cdot \sum_{k=1}^{n-1}\frac{1}{k}.
\end{aligned}$$

When H_n denotes the nth Harmonic number, we obtain

Theorem J.2.1

$$s(n,2) = (n-1)! \cdot H_{n-1}. \tag{J.7}$$

J.3 Benford's paradox

If a number is chosen at random from a large table of data or statistics we assert that distribution of first significant digits is

$$\log_{10}\left(1 + \frac{1}{d}\right) \tag{J.8}$$

has become known as Benford's law. In view of its conterintuitive nature of the law it is also called Benford's paradox.

d	intuitive probability	suggested probability
1	$0.111\ldots$	$0.30103\ldots$
2	$0.111\ldots$	$0.17609\ldots$
3	$0.111\ldots$	$0.12494\ldots$
4	$0.111\ldots$	$0.09691\ldots$
5	$0.111\ldots$	$0.07918\ldots$
6	$0.111\ldots$	$0.06695\ldots$
7	$0.111\ldots$	$0.05799\ldots$
8	$0.111\ldots$	$0.05115\ldots$
9	$0.111\ldots$	$0.04578\ldots$

Appendix K

The Partition of Sets

K.1 Partitions and equivalence relations

A partition of a set S is a collection of subsets $S_1, \ldots, S_k$ of S such that

- $S_i \neq \emptyset$ for all $i = 1, \ldots, k$;
- $S_i \cap S_j = \emptyset$ for $i \neq j$; and
- $\bigcup_{i=1}^{k} S_i = S$.

The subsets S_i are called the parts of the partition.

Observation K.1.1 *There is a one-to-one correspondence between the set of equivalence relations and the collection of partitions.*

A direct consequence of the multiplication principle is

Observation K.1.2 *Suppose that $\sim$ is an equivalence relation on a set S with n elements and each equivalence class has the same number m of elements. Then $\sim$ has n/m equivalence classes.*

K.2 Partitions of a given size

Now we are interested in the number of partitions with specified, but not necessarily equal, part sizes. Consider a set S of n elements and a partition of S into α_1 parts of size 1, α_2 parts of size 2, up to α_n parts of size n, where, of course, $1 \leq i \leq n$ and

$$\sum_{i=1}^{n} i\alpha_i = n. \tag{K.1}$$

Such a partition is called of type $[1]\alpha_1[2]\alpha_2 \ldots [n]\alpha_n$. Recall C.3.1. The n elements can be placed in $n!$ ways. To count distinct partitions we have to take into account

the ways of ordering the elements within the parts and the ways of ordering the parts of the same size i. Hence,

Theorem K.2.1 *The number of partitions of type* $[1]\alpha_1[2]\alpha_2 \dots [n]\alpha_n$ *is*

$$\frac{n!}{\prod_{i=1}^{n}(i!)^{\alpha_i} \cdot \alpha_i!}. \tag{K.2}$$

In particular,

Corollary K.2.2 *The number of partitions of a set into pairs, with* $m = \frac{n}{2}$*, is*

$$\frac{n!}{2^m \cdot m!}. \tag{K.3}$$

K.3 The Stirling number of the second kind

The Stirling number $S(n, k)$ of the second kind denotes the number of ways of partitioning of a set of n elements into exactly k parts.

As an example consider the partitions of the set $\{1, 2, 3, 4\}$. These partitions can be classified according to the number of parts they have, as follows.

1. There is only one partition with exactly one part: $\{\{1, 2, 3, 4\}\}$.
2. There are 7 partitions with exactly two parts:

$$\begin{array}{llll} \{\{1,2,3\},\{4\}\} & \{\{1,2,4\},\{3\}\} & \{\{1,3,4\},\{2\}\} & \{\{2,3,4\},\{1\}\} \\ \{\{1,2\},\{3,4\}\} & \{\{1,3\},\{2,4\}\} & \{\{1,4\},\{2,3\}\} & \end{array}$$

3. There are 6 partitions with exactly three parts:

$$\begin{array}{lll} \{\{1,2\},\{3\},\{4\}\} & \{\{1,3\},\{2\},\{4\}\} & \{\{1,4\},\{2\},\{3\}\} \\ \{\{2,3\},\{1\},\{4\}\} & \{\{2,4\},\{1\},\{3\}\} & \{\{3,4\},\{1\},\{2\}\} \end{array}$$

4. There is only one partition with four parts: $\{\{1\},\{2\},\{3\},\{4\}\}$.

For all $n \geq 2$,

$S(n, 1) = S(n, n) = 1$.

$S(n, 2) = 2^{n-1} - 1$.

$S(n, n-1) = \binom{n}{2}$.

Theorem K.3.1 *Whenever* $1 < k < n$*,*

$$S(n, k) = S(n-1, k-1) + k \cdot S(n-1, k). \tag{K.4}$$

Proof. Consider a partition of $\{1, \ldots, n\}$ into k parts; consider the element n.
Case 1: n appears by itself as a 1-element part.
Then the remaining $n-1$ elements have to form a partition of $\{1, \ldots, n-1\}$ into $k-1$ subsets. There are $S(n-1, k-1)$ ways in which this can be done.
Case 2: n is in a part of size at least two.
Then we can think of partitioning $\{1, \ldots, n-1\}$ into k sets (which can be done in $S(n-1, k)$ ways) and then of adding n in one of the k sets (there are k ways of doing this). □

There is an explicit formula for the Stirling number, namely

Theorem K.3.2

$$S(n,k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^n. \tag{K.5}$$

The *proof* uses 3.4.4(b) and □

It is customary to write the Stirling numbers as Stirling's triangle in the form of a right triangle, where the second column is the number of splits and the diagonal under the main diagonal contains the binomial coefficients.

$n \setminus k$	1	2	3	4	5	6	7	8
1	1							
2	1	1						
3	1	3	1					
4	1	7	6	1				
5	1	15	25	10	1			
6	1	31	90	65	15	1		
7	1	63	301	350	140	21	1	
8	1	127	966	1701	1050	266	28	1

K.4 Bell numbers

$B(n)$ is the total number of partitions of a set of n elements, and is called a Bell number:

$$B(n) = \sum_{k=1}^{n} S(n,k) \tag{K.6}$$

$$= \sum_{k=1}^{n} \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^n. \tag{K.7}$$

where we define $B(0) = 1 = S(0,0)$. In other terms, the Stirling numbers $S(n,k)$, $k = 1, \ldots, n$ form a "partition" of the nth Bell number.

Theorem K.4.1 $B(n) \leq n!$.

Outline of the *proof.* We introduced the Stirling number $s(n,k)$ of the first kind, which is the number of permutations of $1,\ldots,n$ consisting of exactly k cycles. Obviously,

$$S(n,k) \leq s(n,k). \tag{K.8}$$

Then we have

$$B(n) = \sum_{k=1}^{n} S(n,k) \leq \sum_{k=1}^{n} s(n,k) = n!. \tag{K.9}$$

□

We can obtain the following recursive relation for the Bell numbers.

Theorem K.4.2 *For all* $n \geq 1$,

$$B(n) = \sum_{k=0}^{n-1} \binom{n-1}{k} B(k). \tag{K.10}$$

Proof. Consider the nth element of a set which is partitioned. It is in one of the parts of the partition with $j \geq 0$ other elements. There are $\binom{n-1}{j}$ ways of choosing these j elements. The remaining $n-1-j$ elements can be partitioned in $B(n-1-j)$ ways. Hence,

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

putting $n-1-j=k$. □

For further reading see [118].

Appendix L

The Partition of Integers

L.1 Composition of integers

Remember 3.2.2 which said that the number of solutions for the equation $x_1 + \ldots + x_n = k$ in nonnegative integers x_i equals

$$\binom{n+k-1}{k}. \tag{L.1}$$

As an example consider

$$\begin{aligned} x+y+z &= 8 \\ \text{subject to} \quad x &\geq 2 \\ y &\geq 4 \end{aligned}$$

We substitute $x = 2 + u$, $y = 4 + v$ and solve $u + v + z = 2$ with help of 3.2.2 to find $\binom{3+2-1}{2} = 6$ solutions.

At this point it is crucial that we recognize the equivalence of the following statements.

- The number of integer solutions of the equation $x_1 + \ldots + x_n = k$, $x_i \geq 0$.
- The number of selections, with repetition, of size k from a collection of size n.
- The number of ways k identical objects can be distributed among n distinct containers.

It is an important exercise for the reader to restate a problem given in one of the above formulations in the other two.

Let us determine all the different ways in which we can write the number 4 as a sum of positive integers, where the order of the summands is considered relevant:

$$4 = 3+1 = 1+3 = 2+2 = 2+1+1 = 1+2+1 = 1+1+2 = 1+1+1+1.$$

We find eight compositions in total. Now suppose that we wish to count the number of compositions for the positive integer n.

1. For one summand there is only one composition, namely n itself.

2. If there are two (positive) summands, we want to count the number of integer solutions for $z_1 + z_2 = n$, which is the number

$$\binom{2 + (n-2) - 1}{n-2} = \binom{n-1}{n-2}. \tag{L.2}$$

3. Continuing with our next case, we examine the composition with three (positive) summands and find

$$\binom{3 + (n-3) - 1}{n-3} = \binom{n-1}{n-3} \tag{L.3}$$

possibilities.

When we summarize the cases we get

$$\sum_{j=0}^{n-1} \binom{n-1}{j} = 2^{n-1}. \tag{L.4}$$

Theorem L.1.1 *The number of compositions for the positive integer n, where the order of the summands is considered relevant, equals 2^{n-1}.*

L.2 The partition numbers

If n objects are indistinguishable, then the number of ways of grouping them is called the partition number $p(n)$. To determine the bipartition number $p_2(n)$, which counts the number of grouping in exactly two integers is simple: If counting $k = 1, 2, 3, \ldots, n-1$ then in the same time we select $n - k$ and have all pairs, each twice.

Theorem L.2.1 *For the bipartition numbers it holds*

$$p_2(n) = \begin{cases} \frac{n}{2} & : \textit{ if } n \textit{ even} \\ \frac{n-1}{2} & : \textit{ otherwise} \end{cases}$$

In general the question is much harder. A partition of a nonnegative integer n is a finite list of nonnegative integers with sum n, the order of the summands is not important.

n	partitions	$p(n)$
1	1	1
2	2, $1+1$	2
3	3, $2+1$, $1+1+1$	3
4	4, $3+1$, $2+2$, $2+1+1$, $1+1+1+1$	5
5	5, $4+1$, $3+2$, $3+1+1$, $2+2+1$, $2+1+1+1$, $1+1+1+1+1$	7
6	6, $5+1$, $4+2$, $4+1+1$, $3+3$, $3+2+1$, $3+1+1+1$, $2+2+2$, $2+2+1+1$, $2+1+1+1+1$, $1+1+1+1+1+1$	11
7	Exercise	15
8		22
9		30
10		42
11		56

In view of L.1.1 we have $p(n) \leq 2^{n-1}$; but, paying attention [170], a better bound is given by the following considerations:

$$p(n) \leq \frac{1}{x^n} \cdot \Pi_{i=1}^{n} \frac{1}{1-x^i} \tag{L.5}$$

holds true for any real number x between 0 and 1. Thus,

$$\ln p(n) \leq -n \ln x - \sum_{i=1}^{n} \ln(1-x^i). \tag{L.6}$$

In view of the series

$$-\ln(1-y) = y + \frac{y^2}{2} + \frac{y^3}{3} + \frac{y^4}{4} + \dots$$

we can transform the second term of the former equation (L.6):

$$\begin{aligned}
-\sum_{i=1}^{n} \ln(1-x^i) &= \sum_{i=1}^{n} \sum_{j=1}^{\infty} \frac{x^{ij}}{j} = \sum_{j=1}^{\infty} \frac{1}{j} \sum_{i=1}^{n} x^{ij} \\
&\leq \sum_{j=1}^{\infty} \frac{1}{j} \sum_{i=1}^{\infty} x^{ij} = \sum_{j=1}^{\infty} \frac{1}{j} \frac{x^j}{1-x^j} \\
&\leq \sum_{j=1}^{\infty} \frac{1}{j} \frac{x^j}{(1-x) j x^{j-1}} \quad \text{(Exercise)} \\
&= \frac{x}{1-x} \sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{x}{1-x} \varsigma(2) \\
&= \frac{x}{1-x} \frac{\pi^2}{6} \quad \text{in view of the } \zeta \text{ -function.}
\end{aligned}$$

Altogether we find

$$\ln p(n) \leq -n \ln x + \frac{x}{1-x} \frac{\pi^2}{6}, \tag{L.7}$$

which is a function of the real number x. With the help of a little bit of calculus, this implies

$$\ln p(n) \leq \pi \sqrt{2n/3}. \tag{L.8}$$

Consequently,

Theorem L.2.2 *For the partition numbers there is the inequality*

$$p(n) \leq e^{\pi \sqrt{2n/3}} = e^{2.5650\ldots \cdot \sqrt{n}}. \tag{L.9}$$

There is no simple exact formula for p known, but there is a remarkable approximation given by Ramanujan, compare [58]:

$$p(n) \approx \frac{1}{4\sqrt{3}n} \cdot e^{\pi \sqrt{2n/3}}. \tag{L.10}$$

For further reading see Turan [238].

To find the values of $p(n)$ algorithmically, we can use the following observations. Let $q(k, m)$ be the number of ways of grouping k such that no summand is greater than m.

Theorem L.2.3 *(Barron [19]) It holds*

$$p(n) = q(n, n), \tag{L.11}$$

where we computed the values of $q(.,.)$ by the following recursion

$$q(k,m) = \begin{cases} 1 & : \quad k = 1 \\ 1 & : \quad m = 1 \\ q(k,k) & : \quad k < m \\ 1 + q(k, k-1) & : \quad k = m \\ q(k, m-1) + q(k-m, m) & : \quad k > m \end{cases}$$

For further reading see [43].

Appendix M

The Catalan Numbers

We introduce a sequence of numbers known as the Catalan numbers, which arise as the counting numbers of a remarkable number of different types of structure. They are named after the Belgian mathematician Catalan, but they had been studied earlier by several others.

M.1 Routes in grids

The set

$$\{0, \ldots, m\} \times \{0, \ldots, n\} \tag{M.1}$$

is called an $m \times n$-grid. An $n \times n$-grid is called a square of order n.

By an "up-right" route we mean a path from $(0, 0)$ to (m, n) following the edges of the grid always moving upwards or to the right, i.e. the only possible pairs following the pair (x, y) are $(x, y + 1)$ or $(x + 1, y)$.

Theorem M.1.1 *The number of up-right routes from the bottom left node to the top right node of an $m \times n$-grid is*

$$\binom{m+n}{n} = \frac{(m+n)!}{m! \cdot n!} = \binom{m+n}{m}. \tag{M.2}$$

Proof. Any route must consist of $m + n$ steps, m of which must be to the right and n upwards. □

Corollary M.1.2 *The number of up-right routes from the node (r, s) to the node (m, n), with $r \leq m$ and $s \leq n$ of a grid is given by*

$$\binom{m+n-r-s}{n-s}. \tag{M.3}$$

Proof. The calculation may be reduced by a parallel transfer of the system of axes: $x' = x - r$ and $y' = y - s$ to the calculation of the routes from $(0,0)$ to $(m-r, n-s)$. For this we use M.1.1. □

We consider a sequence w in $\{0,1\}^{2n}$, called a tree code (with respect to n), with the following properties:

1. In each prefix of w the number of 1 is at least the number of 0;
   In particular, the first letter in w must be 1;
2. The number of 1 in w equals the number of 0;
   In particular, the last letter in w must be 0.

In other words we are interested in the number C_n of binary sequences of length $2n$ containing exactly n 0's and n 1's, such that at each position in the sequence the number of 0's up to that position never exceeds the number of 1's.

For such a sequence w consider a sequence of pairs

$$(x, y) \in \{0, \ldots, n\}^2, \tag{M.4}$$

where

$$x = \text{number of 1 in } w; \tag{M.5}$$
$$y = \text{number of 0 in } w. \tag{M.6}$$

Then w describes an up-right route from the bottom left node to the top right node of an $n \times n$-grid which never goes above the diagonal $\{(x,x) : 0 \le x \le n\}$, and vice versa.

Theorem M.1.3 *The number of tree codes C_n with respect to n is exactly the number of up-right routes from the bottom left node to the top right node of an $n \times n$-grid which never go above the diagonal.*

C_n is called the nth Catalan number.

M.2 A recurrence relation for the Catalan numbers

To determine the Catalan numbers we consider up-right routes from the bottom left node to the top right node of an $n \times n$-grid which never go above the diagonal. These routes are called correct routes.
Each correct route from $(0,0)$ to (n,n) must contain a pair (m,m) on the diagonal before (n,n), even if it is only $(0,0)$. Let (m,m) be the last point on the diagonal prior to the route reaching (n,n), $0 \le m < n$.
There are C_m possibilities for the part of the route from $(0,0)$ to (m,m).
After (m,m) the route must then proceed to $(m+1, m)$ and continue to $(n, n-1)$

with never going above the "line" from $(m+1, m)$ to $(n, n-1)$, since otherwise (m, m) would not have been the last meeting before (n, n). The points $(m+1, m)$ and $(n, n-1)$ are opposite nodes in a square of order $n-m-1$. There are C_{n-m-1} correct routes in this square.
Using the multiplication principle, we find out that the number of correct routes from $(0, 0)$ to (n, n) with (m, m) as the last point of contact with the diagonal before (n, n), is

$$C_m \cdot C_{n-m-1}. \tag{M.7}$$

Since m can have any value from 0 to $n-1$, it now follows from the addition principle

Theorem M.2.1 *The Catalan number C_n satisfies the recurrence relation*

$$C_n = \sum_{m=0}^{n-1} C_m \cdot C_{n-m-1}, \tag{M.8}$$

with $C_0 = 1$.

M.3 An explicit formula for the Catalan numbers

Following the considerations in the section before, Andre found an explicit description for the Catalan numbers.
Consider any incorrect route. There will be a first point on that route above the diagonal. Suppose that this is $(m, m+1)$. We reflect the part of the route from $(0, 0)$ to $(m, m+1)$ on the "line" $L = \{(x-1, x) : 1 \leq x \leq n\}$, this means

$$(x, y) \mapsto (y-1, x+1).$$

Together with the other part of the route, we get an up-right route from $(-1, 1)$ to (n, n) in an $(n+1) \times (n-1)$-grid. In view of M.1.2 there are

$$\binom{(n+1)+(n-1)}{n+1} = \binom{2n}{n+1} \tag{M.9}$$

such routes.
Conversely, any up-right route from $(-1, 1)$ to (n, n) must cross L somewhere, and must arise from precisely one incorrect route from $(0, 0)$ to (n, n). Hence, the number of incorrect routes from $(0, 0)$ to (n, n) is just the number of up-right routes from $(-1, 1)$ to (n, n), which was given by (M.9).
Altogether,[1]

$$\begin{aligned} \#\text{ correct routes} &= \#\text{ all routes } - \#\text{ incorrect routes} \\ &= \binom{2n}{n} - \binom{2n}{n+1} \end{aligned}$$

[1] where we use the abbreviation # for "the number of"

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

We proved

Theorem M.3.1 *The Catalan number C_n is*

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \frac{(2n)!}{(n+1)! \cdot n!}. \tag{M.10}$$

The sequence of Catalan numbers begins

n	0	1	2	3	4	5	6	7	8	9	10	11
C_n	1	1	2	5	14	42	132	429	1,430	4,862	16,796	58,786

From M.3.1 we can easily derive another recurrence relation for computing successive Catalan numbers, namely

$$C_{n+1} = \frac{4n+2}{n+2} C_n \tag{M.11}$$

for $n > 1$ with $C_0 = 1$. Consequently, we may assume that the Catalan numbers grow exponentially, where the base of the expression equals 4. And indeed, in view of I.1.1, (I.3) and I.2.2 we get the order and the asymptotic behavior of the Catalan numbers.

Corollary M.3.2

$$\frac{1}{e^{3/4}} \cdot \frac{\sqrt{2}}{\sqrt{e}} \cdot \frac{4^n}{(n+1)\sqrt{n}} \leq C_n \leq \frac{\sqrt{2}}{\sqrt{e}} \cdot \frac{4^n}{(n+1)\sqrt{n}}. \tag{M.12}$$

$$C_n = \Theta\left(\frac{4^n}{n^{3/2}}\right). \tag{M.13}$$

$$C_n \approx \frac{1}{\sqrt{\pi}} \cdot \frac{4^n}{n^{3/2}}. \tag{M.14}$$

For further reading see [9] and [43].

M.4 Applications

The Catalan numbers play an important role in many applications. In each case we find the number by using M.2.1.

- Suppose we have $n+1$ variables $x_0, x_1, \ldots, x_n$ whose product is to be computed by multiplying them in pairs. How many ways are there to insert parentheses into the product $\Pi_{i=0}^{n} x_i$ so that the order of multiplication is completely specified? The Catalan number C_n is the answer.

- The Catalan number count the number of ways of grouping any objects.
- How many noncrossing handshakes are possible with n pairs of people?

But the following problem was first proposed by Euler. It examines a given convex polygon of $n \geq 3$ sides. Euler wanted to count the number of ways the interior of the polygon can be subdivided into triangles by drawing diagonals that do not intersect. Let t_n be this number. As an exercise verify that

$$t_{n+1} = t_2 t_n + t_3 t_{n-1} + \ldots + t_{n-1} t_3 + t_n t_2.$$

Theorem M.4.1 *(Euler) C_{n-2} is the number of ways of dividing a convex polygon with $n \geq 4$ nodes into triangles by drawing $n-3$ non-intersection diogonals.*

Appendix N

Fixed Points in Permutations

N.1 Derangements

A derangement of a string of distinct elements is a permutation of the elements such that no element appears in its original position. Equivalently, a derangement is an injective function of a finite set onto itself without a fixed point, where a fixed point x of a function $f : X \to X$ being $f(x) = x$.

The number of derangements of $\{1, \ldots, n\}$ is denoted by D_n. For an explicit description of D_n we use the principle of inclusion and exclusion. Let P_i be the property: The element i is in the ith position. Then

$$N(i) = (n-1)!, \tag{N.1}$$

since i is fixed and the remaining $n-1$ elements can be permutated in any way. Because this is independent of the concrete choice of the position i we get $N_1 = (n-1)!$. Similarly, $N_k = (n-k)!$ is the number of permutation with at least k chosen fixed points. Then we can apply the principle of inclusion/exclusion:

$$\begin{aligned} D_n &= n! - \binom{n}{1}(n-1)! + \binom{n}{2}(n-2)! \pm \ldots + (-1)^n \binom{n}{n} 0! \\ &= n! - \frac{n!}{1!} + \frac{n!}{2!} \pm \ldots + (-1)^n \frac{n!}{n!} \\ &= n! \left(1 - \frac{1}{1!} + \frac{1}{2!} \pm \ldots + (-1)^n \frac{1}{n!}\right). \end{aligned}$$

This gives us the following values for the derangement numbers:

n	1	2	3	4	5	6	7	8	9	10
D_n	0	1	2	9	44	265	1,854	14,831	133,496	1,334,961

In view of $e^x = \sum_{k=0}^{\infty} \frac{1}{k!} x^k$ for any real number x, we find

$$\frac{1}{e} = e^{-1} = \sum_{k=0}^{\infty} (-1)^k \frac{1}{k!}. \tag{N.2}$$

Consequently $D_n/n!$ is very close to the number $1/e$.

Theorem N.1.1 *For the number D_n of derangements the following holds*

$$\begin{aligned} D_n &= n! \cdot \sum_{k=0}^{n} (-1)^k \frac{1}{k!} && \text{(N.3)} \\ &\approx \frac{n!}{e}. && \text{(N.4)} \end{aligned}$$

For further reading see [87].

N.2 A given number of fixed points

Now we consider $D_{n,k}$ the number of permutations of n elements with exactly k fixed points.[1] The following equations are not hard to see: $D_{n,0} = D_n$, $D_{n,n-1} = 0$ and $D_{n,n} = 1$. In general,

Theorem N.2.1 *Let $D_{n,k}$ be the number of permutations of $1, \ldots, n$ elements with exactly k fixed points. Then*

$$D_{n,k} = \binom{n}{k} D_{n-k} \approx \frac{n!}{ek!}. \tag{N.5}$$

Proof. There are $\binom{n}{k}$ ways of choosing the k numbers which are fixed. The remaining $n-k$ have to be deranged, and this can be done in D_{n-k} ways.
In view of N.1.1

$$D_{n,k} = \binom{n}{k} D_{n-k} \approx \frac{n!}{k!(n-k)!} \cdot \frac{(n-k)!}{e} = \frac{n!}{ek!}.$$

□

$n \setminus k$	0	1	2	3	4	5	6	7
1	0	1						
2	1	0	1					
3	2	3	0	1				
4	9	8	6	0	1			
5	44	45	20	10	0	1		
6	265	264	135	40	15	0	1	
7	1854	1855	924	315	70	21	0	1

[1] The problem to calculate $D_{n,k}$ is often called the "Rencontre" problem.

Corollary N.2.2 *The probability that a permutation has exactly k fixed points is approximately $1/ek!$.*

Study the following table with care and note that the sum is not dependent on the value of n.

n	0	1	2	3	4
$\sum_{k=0}^{n} \frac{1}{ek!}$	$0.3678\ldots$	$0.7357\ldots$	$0.9196\ldots$	$0.9809\ldots$	$0.9962\ldots$

Knowing all these facts, we obtain that it is a rare event that a permutation has many fixed positions, and we can determine the average number of fixed points in a permutation.

Theorem N.2.3 *On average there is one fixed position in a randomly chosen permutation (independently of n).*

Appendix O

Elements of Group Theory

O.1 Groups

Let Γ be a nonempty set and a (binary) map $\cdot$ on Γ given; then $(\Gamma, \cdot)$ is called a group if the following conditions are satisfied:

Closure: For all $a, b \in \Gamma$, $a \cdot b \in \Gamma$;

Associativity: For all $a, b, c \in \Gamma$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$;

Identity: There exists $e \in \Gamma$ such that for any $a \in \Gamma$ it is $a \cdot e = e \cdot a = a$;

Inverse: For each $a \in \Gamma$ there is an element $b \in \Gamma$ such that $a \cdot b = b \cdot a = e$.

If, in addition,

Commutativity: For all $a, b \in \Gamma$, $a \cdot b = b \cdot a$, then Γ is called a commutative, or Abelian, group.

As example consider the following groups:

G_1	a	b	c	d
a	a	b	c	d
b	b	a	d	c
c	c	d	a	b
d	d	c	b	a

G_2	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

G_3	$\emptyset$	R	S	N
$\emptyset$	$\emptyset$	R	S	N
R	R	$\emptyset$	N	S
S	S	N	$\emptyset$	R
N	N	S	R	$\emptyset$

G_1 and G_3 are isomorphic, and is called the Klein group, but this group and G_2 cannot be isomorphic. Do you find a third non-isomorphic group with four elements? If not, prove that such a group does not exist.

A group Γ is called cyclic if there is an element $a \in \Gamma$ such that

$$\Gamma = \{a^n : n \text{ an integer}\}. \tag{O.1}$$

Let $(\Gamma, \cdot)$ be group. If H is a subset of Γ such that H is a group with the operation $\cdot$, then H is called a subgraph of Γ. Every group has two trivial subgroups, the group consisting of the identity alone and the whole group.

Theorem O.1.1 *(Lagrange's theorem) The order of a subgroup of a finite group* Γ *is a factor of the order of* Γ.

The converse statement is not true: The alternating group A_4, which is of order 12, has no subgroup of order 6.

Corollary O.1.2 *Every group of prime order is cyclic.*

A complete introduction into the theory of groups is given by Gallian [94].

O.2 The number of finite groups

Obviously, a finite cyclic group Γ with n elements is isomorphic to $\mathbf{\Gamma}_n$. Consequently, in view of O.1.2, we have only list the number of groups for non-primes.[1]

n	4	6	8	9	10	12	14	15	16	18	20
number of groups	2	2	5	2	2	5	2	1	14	5	5

n	21	22	24	25	26	27	28	30	32	33	34
number of groups	2	2	15	2	2	5	4	4	51	1	2

n	35	36	38	39	40	42	44	45	46	48	49
number of groups	1	14	2	2	14	6	4	2	2	52	2

Further reading: Speiser [224].

[1] But note, that there are non-primes with only one group of this order.

O.3 Permutation groups

Recall that a permutation of a set S is a function from S to S that is both one-to-one and onto. A permutation group of a set S is a set of permutations of S that forms a group under function composition.
We will often focus on the case where S is finite; the symmetric group S_n of all permutations of n elements.

Theorem O.3.1 *(Cayley) Every group is isomorphic to a group of permutations.*

Let Γ be a group of permutations of a set S. We define the following concepts:

- For each $i \in S$, let
$$\text{stab}(i) = \{\pi \in \Gamma : \pi(i) = i\}, \tag{O.2}$$
called the stabilizer of i in Γ.
- For each $i \in S$, let
$$\text{orb}(i) = \{\pi(i) : \pi \in \Gamma\}, \tag{O.3}$$
called the orbit of i under Γ.

We leave the reader to show that $\text{stab}(i)$ is a subgroup of Γ of S for any $i \in S$.

Theorem O.3.2 *Let Γ be a (finite) permutation group of a set S. Then for any $i \in S$,*
$$|\Gamma| = |orb(i)| \cdot |stab(i)|. \tag{O.4}$$

Hence, we are interested in counting the number of orbits.

Lemma O.3.3 *Let Γ be a finite permutation group of a set S with n orbits, then*
$$n = \sum_{i \in S} \frac{1}{|orb(i)|}. \tag{O.5}$$

Proof. Suppose that $X_1, X_2, \ldots, X_n$ are the orbits. Then
$$S = X_1 \cup X_2 \cup \ldots \cup X_n.$$
It follows that
$$\sum_{i \in S} \frac{1}{|\text{orb}(i)|} = \sum_{j=1}^{n} \sum_{i \in X_j} \frac{1}{|\text{orb}(i)|} = \sum_{j=1}^{n} \sum_{i \in X_j} \frac{1}{|X_j|} = \sum_{j=1}^{n} \frac{1}{|X_j|} \sum_{i \in X_j} 1 = \sum_{j=1}^{n} \frac{1}{|X_j|} |X_j| = n.$$
□

For any group Γ of permutations on a set S and any $\pi \in \Gamma$, we let
$$\text{fix}(\pi) = \{i \in S : \pi(i) = i\}, \tag{O.6}$$
called the set of fixed points of π.[2] The double counting principle gives

[2] Recall that we counted the number of permutations with a given number of fixed points in N.2.1.

Lemma O.3.4 *Let Γ be a finite permutation group of a set S, then*

$$\sum_{i \in S} |\mathit{stab}(i)| = \sum_{\pi \in \Gamma} |\mathit{fix}(\pi)|. \tag{O.7}$$

Then

Theorem O.3.5 *(Burnside's lemma) If Γ is a finite permutation group of a set S with n orbits. Then*

$$n = \frac{1}{|\Gamma|} \sum_{\pi \in \Gamma} |\mathit{fix}(\pi)|. \tag{O.8}$$

Proof. Our considerations above establish the following chain of equalities:

$$\frac{1}{|\Gamma|} \sum_{\pi \in \Gamma} |\mathrm{fix}(\pi)| = \frac{1}{|\Gamma|} \sum_{i \in S} |\mathrm{stab}(i)| = \frac{1}{|\Gamma|} \sum_{i \in S} \frac{|\Gamma|}{|\mathrm{orb}(i)|} = \sum_{i \in S} \frac{1}{|\mathrm{orb}(i)|} = n.$$

□

Appendix P

Latin squares

Latin squares are of interest arose through their use in statistical experimental design, but they are also in pure theoretical sense, a trivial example is the fact that the composition table of a finite group is a Latin square.[1]

P.1 Finite fields

Theorem P.1.1 *Each finite field is of order of a power of a prime.*
For each prime p and each positive integer n, there is, up to isomorphism, a unique field of order p^n.

The unique field is called a Galois field $\mathrm{GF}(p^n)$. If $n = 1$ and only in this case $\mathrm{GF}(p) = \Gamma_p$. Furthermore, the structure of Galois fields is completely described.

Remark P.1.2 *Consider the Galois field GF(p^n).*

a) *As a group under addition, GF(p^n) is isomorphic to*

$$\underbrace{\mathbf{\Gamma}_p \oplus \mathbf{\Gamma}_p \oplus \ldots \oplus \mathbf{\Gamma}_p}_{n-times}. \tag{P.1}$$

b) *As a group under multiplication, the set of nonzero elements of GF(p^n) is isomorphic to*

$$\mathbf{\Gamma}_{p^n-1}. \tag{P.2}$$

P.2 The Existence of Latin squares

An $n \times n$ Latin square of order n is a $n \times n$ matrix of symbols, usually $1, 2, \ldots, n$, where each symbol appears exactly once in each row and each column of the matrix.[2]

[1] As a popular example Latin squares of order 3 play an important role in the play called "Sodoku".

[2] Latin squares may also be regarded as bicolored graphs $K_{n,n}$ in which the edges are also colored. the vertices of the first color correspond to the rows of the Latin square while the vertices of the

The following is an example for a 4×4 Latin square.

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{array}$$

May be the reader find this square well-known; and indeed we considered a matrix as the operation table for (finite) groups which is essentially a Latin square.[3] This observation gives the following

Lemma P.2.1 *For all $n \geq 2$, we can obtain an $n \times n$ Latin square from the table of the group $(\mathbf{\Gamma}_n, +)$ if we replace the occurences of 0 by the value of n.*

In other terms, for each $n \geq 2$ an $n \times n$ Latin square exists. But how much? The following theorem seems intuitively true, but is not simple to prove, compare [142].

Theorem P.2.2 *(M.Hall) There are at least $l(n) = \prod_{k=1}^{n} k!$ Latin squares of order n.*

n	1	2	3	4	5	6	7
$l(n)$	1	2	12	288	34,560	24,883,200	124,913,664,000

We estimate the order of $l(n)$. For an even number $n \geq 4$

$$\prod_{k=1}^{n} k! \geq 2^{n/2} \left(\frac{n}{2}\right)!^{n/2} \geq 2^{n/2} \left(e \left(\frac{n/2}{e} \right)^{n/2} \right)^{n/2} = (2e)^{n/2} \left(\frac{n}{2e} \right)^{n^2/4} = \frac{n^{n^2/4}}{(2e)^{n/2}}.$$

The bound $l(n)$ is extremely bad. In [118] we find a better estkimated values for the number $l^{\text{real}}(n)$ of Latin squares. $l(n)$ grows exponentially in n; is $l^{\text{real}}(n)$ a superexponential function?

An $r \times n$ ($r \leq n$) Latin rectangle based on $1, \ldots, n$ is an $r \times n$ matrix that each entry is one of the numbers $1, \ldots, n$ and each number occurs in each row and column at most once.

Theorem P.2.3 *([142])*

a) *Each $r \times n$ Latin rectangle can extend to a Latin square of order n.*

b) *There at least $\prod_{k=1}^{n-r+1} k!$ Latin rectangles.*

---

second color stand for the columns. Every edge is colored with one of the n colors so that each vertex is incident with one edge of each color.

[3] The converse cannot be true (Why?).

P.3 Orthogonal Latin squares

There are several other 4×4 Latin squares, than are given above:

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{array}$$

and,

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \\ 2 & 1 & 4 & 3 \end{array}$$

From these two Latin squares we are able to produce all of the ordered pairs in $\{1, 2, 3, 4\}^2$.

$$\begin{array}{cccc} (1,1) & (2,2) & (3,3) & (4,4) \\ (2,3) & (1,4) & (4,1) & (3,2) \\ (3,4) & (4,3) & (1,2) & (2,1) \\ (4,2) & (3,1) & (2,4) & (1,3) \end{array}$$

showing that all the 16 pairs are distinct. We now question whether or not we can do this for $n \times n$ Latin squares in general.
Let $L_k = (a_{ij}^{(k)}, k = 1, 2$ be two $n \times n$ Latin squares. If the n^2 ordered pairs $(a_{ij}^{(1)}, a_{ij}^{(2)})$ are distinct, then L_1 and L_2 are called a pair of orthogonal Latin squares. As exercise show

1. There is no pair of 2×2 orthogonal Latin squares.
2. There is a pair of 3×3 orthogonal Latin squares.
3. (Already shown above:) There are three pairs of 4×4 orthogonal Latin squares.

Can we continue this sequence? First we create an upper bound for the number of orthogonal Latin squares.

Theorem P.3.1 *For $n \geq 2$ the largest possible number of $n \times n$ Latin squares that are orthogonal in pairs is $n - 1$.*

Proof. Let

$$L_m = (a_{ij}^{(m)}), \tag{P.3}$$

$m = 1, \ldots, k$ be distinct $n \times n$ Latin squares. We may assume that the squares are in standard form, which means that its first row is $12 \ldots n$. In other words,

$$a_{1j}^{(m)} = j \tag{P.4}$$

for all $1 \leq m \leq k$.
Now consider $a_{21}^{(m)}$ for all $1 \leq m \leq k$. These entries in the second row and first column must be different from 1. Furthermore, if there exists $1 \leq l < m \leq k$ with $a_{21}^{(l)} = a_{21}^{(m)}$, then the pair L_l and L_m cannot be an orthogonal pair (Why?). Consequently, there are at most $n-1$ choices for the entries $a_{21}^{(m)}$ in any of our Latin squares. □

Secondly, we give a construction $n-1$ of $n \times n$ Latin squares for specific values of n: Let $n > 2$ be a power of a prime, that is $n = p^t$ where p is a prime and t a positive integer. Let

$$F = \text{ GF}(p^t) = \{x_1, x_2, \ldots, x_n\} \tag{P.5}$$

be the Galois field of order n, where x_1 is the unity and x_n the zero.
For each $1 \leq m \leq n-1$ let

$$L_m = (a_{ij}^{(m)}), \tag{P.6}$$

be the $n \times n$ matrix with

$$a_{ij}^{(m)} = x_m \cdot x_i + x_j. \tag{P.7}$$

First we show that each L_m is a Latin square. If not there are two identical elements in the same row or column. Suppose it occurs in a column. Then

$$x_m \cdot x_r + x_j = a_{rj}^{(m)} = a_{sj}^{(m)} = x_m \cdot x_s + x_j.$$

this implies that $x_m \cdot x_r = x_m \cdot x_s$, by the cancellation for addition in F. Since $m \neq n$ the element x_m is not the zero in F. Hence, $x_r = x_s$ and $r = s$. Similar for the rows in L_m.
Now we have $n-1$ Latin squares. We have to prove that they orthogonal in pairs. Assuming not. Let $1 \leq m < k \leq n-1$ with

$$\begin{aligned} a_{ij}^{(m)} &= a_{rs}^{(m)} \\ a_{ij}^{(k)} &= a_{rs}^{(k)} \end{aligned}$$

and $(i,j) \neq (r,s)$. Equivalently

$$\begin{aligned} x_m \cdot x_i + x_j &= x_m \cdot x_r + x_s \\ x_k \cdot x_i + x_j &= x_k \cdot x_r + x_s. \end{aligned}$$

Subtracting these equations, we find

$$(x_m - x_k) \cdot x_i = (x_m - x_k) \cdot x_r.$$

With $m \neq k$ it follows that $x_m - x_k$ cannot be the zero. Thus $x_i = x_r$. Putting this back into either of the above equations, we find $x_j = x_s$. Consequently, $i = r$ and $j = s$.
We proved (in a constructive way):

Theorem P.3.2 *Let $n > 2$ be a power of a prime. Then there are $n-1$ of $n \times n$ Latin squares that are orthogonal in pairs.*

The first natural number which is not a power of a prime is 6. Euler in 1779 observed the following problem:

> Thirty-six officers of six ranks and from six different regiments are to march in a square formation 6×6. Each row and each column of the formation is to contain one and only one officer of each rank and one and only one officer from each regiment. Is such formation possible?

Euler's conjecture, that there was no solution, around 1900, Tarry showed correct.

n	2	3	4	5	6	7
Does GF(n) exist?	yes	yes	yes	yes	no	yes
number of orthogonal Latin squares	1	2	3		1	

n	8	9	10	11	12	13
Does GF(n) exist?	yes	yes	no	yes	no	yes
number of orthogonal Latin squares	7	8				

Appendix Q

Hadamard matrices

In 1893 Hadamard showed that any $n \times n$ matrix H, whose entries h_{ij} all satisfies $|h_{ij}| \leq 1$, has determinant at most $n^{n/2}$, where equality occurring only if $HH^T = nE$.

We introduce a family of matrices all whose entries are ± 1. A $n \times n$ matrix with entries ± 1 is called a Hadamard matrix of order n if

$$H \cdot H^T = H^T \cdot H = nE. \tag{Q.1}$$

From the definition the following facts follow immediately (Exercise).

Observation Q.0.3 *Let H be a Hadamard matrix of order n. Then*

a) *H^T is also a Hadamard matrix.*

b) *Any two rows of H are orthogonal; and any two columns of H are orthogonal.*

c)

$$detH = \pm n^{n/2}, \tag{Q.2}$$

d) *H is invertible with*

$$H^{-1} = \frac{1}{n} H^T. \tag{Q.3}$$

Theorem Q.0.4 *Let H be a Hadamard matrix of order n. Then $n = 1, 2$ or $n = 2m$ for some positive integer m.*

Proof. □

Examples for the specific orders n:

$$H_0 = (1).$$

$$H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$H_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

And, in general, there is a straightforward way of constructing Hadamard matrices.

Observation Q.0.5 *Starting with the Hadamard matrix $H_0 = (1)$, we define for each $m \geq 1$ recursively the Hadamard matrices*

$$H_m = \begin{pmatrix} H_{m-1} & H_{m-1} \\ H_{m-1} & -H_{m-1} \end{pmatrix} \tag{Q.4}$$

Now we have Hadamard matrices of order $0, 1, 2, 4, \ldots, 2^m, \ldots$. Does a Hadamard matrix for other values of n exist?

Appendix R

Metric Spaces

Distance is the mathematical description of the idea of proximity, and consequently, will play an important role in mathematics. A metric space is a kind of of space in which the concept of distance has meaning.

R.1 Distances

The following term was introduced by Fréchet in 1906: A pair (X, ρ) is called a metric space if X is a nonempty set of elements called the points, and $\rho : X \times X \to \mathbb{R}$ is a real-valued function satisfying:

(i) $\rho(x, y) \geq 0$ for all x, y in X;

(ii) $\rho(x, y) = 0$ if and only if $x = y$;

(iii) $\rho(x, y) = \rho(y, x)$ for all x, y in X; and

(iv) $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$ for all x, y, z in X (triangle inequality).

Usually, such a function ρ is called a metric.

Note that the axioms are not independent: (i) is a consequence of (iv). On the other hand,

Observation R.1.1 *A metric ρ can be defined equivalently by*

(ii) $\rho(x, y) = 0$ if and only if $x = y$; and

(iv') $\rho(x, y) \leq \rho(x, z) + \rho(y, z)$ for all x, y, z in X.

We will say that the quantity $\rho(x, y)$ is the distance between the points x and y.

In the biological context the equality of words makes no sense, since mutations do not allow identical sequences in reality. On the other hand, in biomolecular sequences,

high sequence similarity usually implies significant functional and structural similarity. Consequently, the following variants of "metric approaches" will be also of interest:

- If ρ satisfies (ii) only in the weaker form

  (ii') $\rho(x,x) = 0$ for all x in X;

  we say that ρ is a pseudometric.

- If the function ρ satisfies the conditions (i),(ii') and (iii) it is called a dissimilarity. It will be the dual of the approach of "similarity".

- A metric ρ is called an ultrametric if

$$\rho(v,w) \leq \max\{\rho(v,u), \rho(w,u)\} \tag{R.1}$$

  for any points u, v, w.
  Ultrametric distances very useful in phylogenetics when implying of a constant rate of evolution. Furthermore, if distances between sequences are ultrametric then the most similar sequences are also the closely related.
  It is easy to see that

  Observation R.1.2 *The following is true for all ultrametric spaces (X,ρ):*

$$\textit{If } \rho(v,u) \neq \rho(w,u), \textit{ then } \rho(v,w) = \max\{\rho(v,u), \rho(w,u)\}.$$

  That means that all triangles in (X,ρ) are isosceles triangles where the base is the shorter side.

A metric, pseudometric, ultrametric or dissimilarity ρ on a finite set X of n points can be specified by an $n \times n$ matrix of (nonnegative) real numbers. (Actually $\binom{n}{2}$ numbers suffice because of (ii') and (iii).)

R.2 Examples

We will find distances for many sets which are of great importance.

- The function

$$\rho(x,y) = \begin{cases} 1 & : \quad x \neq y \\ 0 & : \quad x = y \end{cases}$$

  defines a metric for any set X.[1]

- The field of real numbers is considered as a metric space with the metric $\rho(x,y) = |x-y|$.

[1] This generality makes the space not very informative.

- The Euclidean plane is defined in the affine plane with the Euclidean metric $\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$ between the points (x_1,y_1) and (x_2,y_2) derived from a norm $||.||$:

$$||(x,y)|| = \sqrt{x^2+y^2}. \tag{R.2}$$

- Let $\mathcal{F}([a,b],\mathbb{R})$ be the set of all continuous real functions over the real numbers between a and b.

$$\rho(f,g) = \sup\{|f(x)-g(x)| : a \leq x \leq b\} \tag{R.3}$$

defines a metric for $f,g \in \mathcal{F}([a,b],\mathbb{R})$.

- For $x=\{x_i\}, y=\{y_i\} \in \mathbb{N}^{\mathbb{N}}$ the function

$$\rho(x,y) = \begin{cases} 0 & : \quad x=y \\ \frac{1}{n} & : \quad \text{otherwise, where } n=\min\{n : x_i \neq y_i\} \end{cases}$$

creates a (complete) ultrametric.

- Using the binary operation $\triangle$ -the symmetric difference between sets- we find a metric for sets

Observation R.2.1 *$|S_1 \triangle S_2|$ is a metric.*

It is sufficient to show the triangle inequality.

$$S_1 \triangle S_2 \subseteq S_1 \triangle S_3 \cup S_3 \triangle S_2. \tag{R.4}$$

Moreover,

$$S_1 \triangle S_3 \cap S_3 \triangle S_2 = ((S_1 \cap S_2) \setminus S_3) \cup (S_3 \setminus (S_1 \cup S_2)). \tag{R.5}$$

That means, if an element is in $S_1 \triangle S_3 \cap S_3 \triangle S_2$, then it cannot be in $S_1 \triangle S_2$. □

- The Hamming distance between v and w in A^n, for an alphabet A is the number of positions in which v and w disagree:

$$\rho_H((a_1,\ldots,a_n),(b_1,\ldots,b_n)) = |\{i : a_i \neq b_i \text{ for } i=1,\ldots,n\}|, \tag{R.6}$$

for $a_i, b_i \in A$.

- Consider the set $A^\star$ of all words over the alphabet A. The edit distance ρ_L, between two words of not necessarily equal length is the minimal number of "edit operations" required to change one word into the other, where an edit operation is a deletion, insertion, or substitution of a single letter in either word.[2]

[2]At first glance, it seems that the spaces with Hamming distance are subspaces of the space with Levenshtein distance, but this is not true: Consider the two words $v=(ab)^n$ and $w=(ba)^n$; then $\rho_L(v,w)=2$ but $\rho_H(v,w)=2n$.

- To extend the Hamming distance to a metric for all words we may proceed in the following way: Add a "dummy" letter "-" to A. We define a map

$$cl : (A \cup \{-\})^\star \to A^\star \tag{R.7}$$

deleting all dummies in a word from $(A \cup \{-\})^\star$. Then for two words w and w' in $A^\star$ we define the extended Hamming-distance as

$$\begin{aligned} \rho(w, w') \quad = \min\{\rho_H(\underline{w}, \underline{w'}) : \quad & \underline{w}, \underline{w'} \in (A \cup \{-\})^\star, |\underline{w}| = |\underline{w'}|, \\ & cl(\underline{w}) = w, cl(\underline{w'}) = w'\}. \end{aligned} \tag{R.8}$$

Observation R.2.2 *The extended Hamming-distance coincides with the Levenshtein metric.*

- We create a metric ρ for the set S_n of all permutations of n elements by

$$\rho(\pi, \kappa) = i(\pi \circ \kappa^{-1}). \tag{R.9}$$

- Each graph is a metric space, and conversely

Observation R.2.3 *(Hakimi, Yau [116], [143]) Each finite metric space is equivalent to some network.*

When we restrict to metric spaces with integer-valued distances we have a deeper result.

Observation R.2.4 *(Kay, Chartrand [148]) Let (X, ρ) be a finite metric spaces where all values of ρ are integers. Then (X, ρ) is a distance space of some graph if and only if for any two points u and v with $\rho(u, v) \geq 2$ there is a third point w such that*

$$\rho(u, w) + \rho(w, v) = \rho(u, v).$$

- One of the major challenges of shape analysis is object recognition, but there is to distinguish between thousands of object categories, each characterized by tremendous variability. At the heart of this work is Gromov's approach using the Hausdorff distance: For the sets A and B in a metric space (X, ρ) we define:

$$\rho(A, B) = \max\{\sup_{a \in A} \inf_{b \in B} \rho(a, b), \sup_{b \in B} \inf_{a \in A} \rho(a, b)\}. \tag{R.10}$$

Equivalently

$$\rho(A, B) = \inf\{r > 0 : A \subseteq U_r(B), B \subseteq U_r(A)\}, \tag{R.11}$$

where $U_r(\cdot)$ stands for the r-dilation of the set. More by Sapiro [212].

R.3 Topological spaces

One calls topological space a set X equipped with a family $\mathcal{O}$ of subsets of X, called the open sets of X satisfying the following conditions:

(i) $\emptyset \in \mathcal{O}$ and $X \in \mathcal{O}$;

(ii) Every union of open sets is open; and

(iii) Every finite intersection of open sets is open.

We will find topologies for many sets and of great importance. Firstly, we consider the following examples, we will meet many more later.

- Let X be a nonempty set, then we may define the open sets by $\mathcal{O} = \mathcal{P}(X)$, which is called the discrete topology.
- Let X be a nonempty set, then we may define the open sets by $\mathcal{O} = \{\emptyset, X\}$, called the coarse topology.
- Every metric space is automatically a topological space considering its open sets.

Let $(X, \mathcal{O})$ be a topological space. A subset S of X is called closed if $X \setminus S$ is open.

Theorem R.3.1 *Let $(X, \mathcal{O})$ be a topological space. Then*

(i) *$\emptyset$ and X are closed;*

(ii) *Every intersection of closed sets is closed; and*

(iii) *Every finite union of closed sets is closed.*

The *proof* follows immediately by passage the definition of open sets to complements. □

R.4 Radon's lemma

Theorem R.4.1 *Let $\mathcal{C} = \{C_\alpha\}_{\alpha \in A}$ be a collection of closed sets of $\mathbb{R}^d$ where*

- *For at least one index α_0 the set C_{α_0} is bounded, consequently compact.*
- *Each finite subcollection of $\mathcal{C}$ has a nonempty intersection.*

Then

$$\bigcap_{\alpha \in A} C_\alpha \neq \emptyset. \tag{R.12}$$

Proof. Assume that

$$\emptyset = \bigcap_{\alpha \in A} C_\alpha = \bigcap_{\alpha \in A \setminus \{\alpha_0\}} C_\alpha \cap C_{\alpha_0}. \tag{R.13}$$

Then

$$\bigcup_{\alpha \in A \setminus \{\alpha_0\}} (\mathbb{R}^d \setminus C_\alpha) \supseteq C_{\alpha_0} \tag{R.14}$$

says that

$$\{\mathbb{R}^d \setminus C_\alpha\}_{\alpha \in A \setminus \{\alpha_0\}} \tag{R.15}$$

is an open covering of C_{α_0}. Since this set is compact, a finite subcollection

$$\{\mathbb{R}^d \setminus C_i\}_{i=1}^k \tag{R.16}$$

is a covering too. Equivalently,

$$\bigcap_{i=1}^k C_i \cap C_{\alpha_0} = \emptyset, \tag{R.17}$$

which contradicts the conditions on $\mathcal{C}$. □

Appendix S

Minimum Spanning Trees

The minimum spanning tree problem is one of the most typical problems of combinatorial optimization; methods for its solution have generated important ideas of modern combinatorics and have played a central role in the design of computer algorithms. The problem is usually stated as follows:

> Given a weighted (connected) graph one would then wish to select for construction a set of communication links that would connect all the vertices and have minimal total cost.

S.1 A greedy strategy

Starting with Boruvka in 1926, Kruskal in 1956 and Prim in 1957, Minimum Spanning Trees have a well-documented history [104] and effective constructions [47]. In view of the many contributions to the problem of constructing minimum spanning trees, its popularity through the ages, and its natural applications to various practical questions, it is hopeless to expect a complete list of the many facets of the problem. In other terms, the problem has an interest in its own.[1]

A minimum spanning tree in a graph can be found with the help of Kruskal's method[2]. For an introduction to this algorithm the reader should prove the following facts:

Observation S.1.1 *Let $G = (V, E, f)$ be a network, where all edge lengths are distinct. Then it holds*

[1] It seems to be the first network optimization problem ever studied. Its history dates back to at least 1926. Boruvka [32] produced the first fully realized minimum spanning tree algorithm by a parallel technique, and it has been rediscovered several times, Sollin in [24]

[2] This cheapest-link algorithm is the mother of all greedy algorithms, that is to takes the best choice and run, [158].
Another method, created by Prim [193] and Dijkstra [67], is a typical example for dynamic programming.

a) *Let $V' \subseteq V$ be a nonempty subset of vertices, and let e be an edge of minimal length and with one endvertex in V' and the other not. Then every minimum spanning tree contains e.*

b) *Let C be a cycle in G, and let e be the longest edge belonging to C. Then e does not belong to any minimum spanning tree of G.*

Now let us have a look at the algorithm.

Algorithm S.1.2 *(Kruskal [159]) A minimum spanning tree in a graph $G = (N, E)$ with a positive length-function $f : E \to \mathbb{R}$ can be found*

1. *Start with the forest $T = (N, \emptyset)$;*
2. *Sequentially choose the shortest edge that does not form a circle with already chosen edges;*
3. *Stop when all vertices are connected, that is when $|N|-1$ edges have been chosen.*

Or in a dualistic version:

Algorithm S.1.3 *Given a network $G = (V, E, f)$, a minimum spanning tree T for G can be found by the following procedure:*

1. *Start with the graph $G = (V, E)$;*
2. *Sequentially delete the longest edge that does not disconnect the remaining graph;*
3. *Stop when the graph does not contain a cycle, that is when $|V|-1$ edges remain.*

A nice description of the difference between the two techniques is given by Lovász et.al. [164]:

> There is this story about the pessimist and the optimist: They each get a box of assorted candies. The optimist always picks the best; the pessimist eats the worst (to save the better candies for late). So the optimist always eats the best available candy, and the pessimist always eats the worst available candy; and yet, they end up with eating same candies.

A complete discussion of minimum spanning tree strategies in networks are given by [234], [235], [253].

S.2 Shortest Connectivity

The problem of "Shortest Connectivity" has a long and convoluted history.[3] Usually, the problem is known as Steiner's Problem and it can be described more precisely in

[3]The history of Steiner's Problem started with P.Fermat early in the 17th century and C.F.Gauß in 1836. At first perhaps with the famous book *What is Mathematics* by R.Courant and H.Robbins in 1941, this problem became popularized under the name of Steiner.

the following way: Given a finite set of points in a metric space, search for a network that connects these points with the shortest possible length.

Steiner's Problem seems disarmingly simple, but it is rich with possibilities and difficulties. This is one of the reasons that an enormous volume of literature has been published, starting in the seventeenth century and continuing today. More and more real-life problems are given which use Steiner's Problem or one of its relatives as an application, as a subproblem or as a model, compare [52].

The most surprising application of Steiner's Problem is in the area of phylogenetics. Bern and Graham [26]:

> David Sankoff of the University of Montreal and other investigators defined a version of the Steiner problem in order to compute plausible phylogenetic trees. The workers first isolate a particular protein that is common to the organism they want to classify. For each organism they then determine the sequence of the amino acids that make up the protein and define a point at a position determined by the number of differences between the corresponding organism's protein and the protein of other organisms. Organisms with similar sequences are thus defined as being close together and organisms with dissimilar sequences are defined as being far apart. In a shortest network for this abstract arrangement of given points, the Steiner points correspond to the most plausible ancestors, and edges correspond to relations between organisms and ancestor that assume the fewest mutations.

The latter remark explains the importance of trees having the least possible length in phylogenetic spaces for evolutionary relation investigation. This approach to Evolution Theory was suggested first by Fitch [85] in 1971, and also explicitly written by Foulds et al. [91], [222] in 1979. Unfortunately, this idea does not give a simple method.[4]

The central question of "Shortest Connectivity" in networks was originally formulated by Hakimi [117] in 1971:

Steiner's Problem in Graphs

Given: A connected graph $G = (V, E)$ with a length-function $f : E \to \mathbb{R}$, and a nonempty subset N of V.

Find: A connected subgraph $G' = (V', E')$ of G such that

$$L(G') = \sum_{e \in E'} f(e) \tag{S.1}$$

is minimal.

[4] And seems to have been rather forgotten in the field of biology after tree-building program packages became widely available.

Two specific cases are well-known:

$|N| = 2$**:** We search a shortest path interconnecting the two points in N. Here there does not exist a Steiner point, so any internal vertex on the path has degree 2. To find such paths we use the dynamic programming strategy.

$N = V$**:** Here Steiner points are not necessary; we look for a minimum spanning tree. This is easy to do using the greedy strategy.

Two algorithms, which generalize our specific cases, create an SMT in graphs.
The Dreyfus and Wagner solution method [71] breaks the problem down into subproblems, and each of these subproblems themselves into subproblems etc., until the subproblems can be solved with help of a shortest path technique. The time complexity is $O(3^n k + 2^n k^2 + k^3)$, where $n = |N|$ and $k = |V|$. Hence, the algorithm is exponential in the number of given points and polynomial in the number of other vertices.
On the other hand, Hakimi [117]) proposed that a minimum spanning tree be calculated for each of the possible subsets of vertices, from just the set of given points through to the complete set of vertices. The time complexity of the algorithm is $O(n^2 \cdot 2^{k-n} + k^3)$, where $n = |N|$ and $k = |V|$. Hence, the algorithm is polynomial in the number of given points and exponential in the number of the other vertices.
All known exact algorithms for Steiner's Problem in graphs are in some way enumerative algorithms. However, they differ in how the enumeration is done and how clever their strategies for avoiding total enumeration are.[5] Consequently, all of these algorithms need exponential time. But this is not a surprise, since Steiner's Problem in graphs is $\mathcal{NP}$-hard, [147].

Surveys on Steiner's Problem in graphs can be found in [137], [194].

[5]For the problem of enumerating all solutions see [51].

Appendix T

Matroids

In 1935 Whitney introduced the concept of a matroid. His intention was to elaborate fundamental properties of dependence which are common to graphs and matrices.

Many combinatorial optimization problems can be formulated as follows: Given a set system $(E, \mathcal{F})$ for a finite set E, and a cost function $c : \mathcal{F} \to \mathbb{R}$, find a set X in $\mathcal{F}$ whose total cost

$$c(X) = \sum_{e \in X} c(e) \tag{T.1}$$

is minimumal.

If we restrict ourselves to those combinatorial optimization problems, where $(E, \mathcal{F})$ describes an independence system we can generate a general and useful theory for such problems.

T.1 Independence systems

A set system $(E, \mathcal{F})$ for a finite set E is called a matroid if

1. $\emptyset \in \mathcal{F}$;
2. If $Y \in \mathcal{F}$ and $X \subseteq Y$ then $X \in \mathcal{F}$;
3. If $X, Y \in \mathcal{F}$ and $|X| > |Y|$, then there is an element $x \in X \setminus Y$ with $Y \cup \{x\} \in \mathcal{F}$.

If only (i) and (ii) are satisfied we will speak about an independence system.[1]
The members of $\mathcal{F}$ are called independent. For $X \subseteq E$, the maximal independent subsets of X are called bases of X.

[1] The name matroid points out that these structures are generalizations of matrices. Namely, the set of columns of a matrix over some field which are linearly independent form a matroid. (Why?)

Theorem T.1.1 *Let $(E, \mathcal{F})$ be an independence system. Then the following statements are (pairwise) equivalent:*

- *If $X, Y \in \mathcal{F}$ and $|X| > |Y|$, then there is an element $x \in X \setminus Y$ with $Y \cup \{x\} \in \mathcal{F}$.*
- *If $X, Y \in \mathcal{F}$ and $|X| = |Y| + 1$, then there is an element $x \in X \setminus Y$ with $Y \cup \{x\} \in \mathcal{F}$.*
- *For each $X \subseteq E$, all bases of X have the same size.*

Proof. The implications (a)⇒(b)⇒(c) are obvious.
To prove (c)⇒(a), let $X, Y \in \mathcal{F}$ and $|X| > |Y|$. By (c), Y cannot be a basis of $X \cup Y$. So there must be an $x \in (X \cup Y) \setminus Y = X \setminus Y$ such that $Y \cup \{x\} \in \mathcal{F}$. □

T.2 The greedy algorithm

A greedy algorithm is one in which we make the best choice possible at each step, regardless of the subsequent effect of that choice.
An immediate generalization of S.1.2 is the following strategy.

Algorithm T.2.1 *(Greedy algorithm) Given an independence system $(E, \mathcal{F})$ for a finite set E, and a cost function $c : \mathcal{F} \to \mathbb{R}$. Consider the following procedure:*

1. *Sort $E = \{e_1, e_2, \ldots, e_n\}$ such that $c(e_1) \leq c(e_2) \leq \ldots \leq c(e_n)$;*
2. *Set $F := \emptyset$;*
3. **for** $i := 1$ **to** n **do**:
   if $F \cup \{e_i\} \in \mathcal{F}$ **then** *set* $F := F \cup \{e_i\}$.

In step 1, the complexity of sorting the n values depends, of course, on the choice of the sorting algorithm, but it never takes more than quadratic time. The other steps need linear time (Exercise). Altogether, we get that the greedy strategy T.2.1 is an efficient algorithm.

Now, the most interesting questions is: Where does T.2.1 work exactly?

Theorem T.2.2 *An independence system $(E, \mathcal{F})$ is a matroid if and only if T.2.1 finds a minimum solution for all cost functions c.*

Proof. First assume that $(E, \mathcal{F})$ is a matroid, but there exists a cost function c for which the greedy algorithm is not optimal, which means that the greedy algorithm stops with a set $F = \{e_1, \ldots, e_k\}$, but there exists a maximal independent set $F' = \{f_1, \ldots, f_{k'}\}$ such that $c(F') < c(F)$. Since all maximal independent sets have the same cardinality, we have $k' = k$.

By construction of F it holds $c(e_1) \leq \ldots \leq c(e_k)$. Then the assumption $c(F') < c(F)$ implies that there exists an index $1 \leq i \leq k$ such that $c(e_i) > c(f_i)$. Let

$$\begin{aligned} G &= \{e_1, \ldots, e_{i-1}\} \\ G' &= \{f_1, \ldots, f_{i-1}, f_i\}. \end{aligned}$$

As G and G' are both independent and $|G'| > |G|$ there exists an $x \in G' \setminus G$ such that $G \cup \{x\} \in \mathcal{F}$. But because of our choice of i and the definition of G', we have

$$c(x) \leq c(f_i) \leq c(e_i).$$

This cannot be, the case though the greedy algorithm would have then chosen x.
To show the other direction assume now that the greedy algorithm is optimal for all cost functions. We have to valid the third axiom.
Let $X, Y \in \mathcal{F}$ with $|X| = |Y| + 1$, but for any element $x \in X \setminus Y$ we have $Y \cup \{x\} \notin \mathcal{F}$. We define a cost function in the following way

$$c(x) = \begin{cases} -|Y| - 2 & : \quad x \in Y \\ -|Y| - 1 & : \quad x \in X \setminus Y \\ 0 & : \quad \text{otherwise} \end{cases}$$

Let F the output of the algorithm. Y is independent; and since Y contains all smallest elements we have $Y \subseteq F$. On the other hand, it holds $Y \cup \{x\} \notin \mathcal{F}$ for all $x \in X \setminus Y$, which implies $F \cap (X \setminus Y) = \emptyset$. Hence,

$$\begin{aligned} c(F) &= c(Y) \\ &= |Y| \cdot (-|Y| - 2) \\ &= -(|Y|^2 + 2|Y|) \\ &> -(|Y|^2 + 2|Y| + 1) \\ &= -(|Y| + 1)^2 \end{aligned}$$

Furthermore,

$$\begin{aligned} c(F) &\leq |X| \cdot (-|Y| - 1) \\ &= (|Y| + 1) \cdot (-|Y| - 1) \\ &= -(|Y| + 1)^2 \end{aligned}$$

Altogether we find $c(X) < c(F)$, which cannot be. $\square$

Appendix U

Computational Complexity

U.1 Sources for algorithms in graph theory

As discrete objects we are interested to handle these algorithmically.

1. Böckenhauer, Bongartz: Algoritmische Grundlagen der Bioinformatik; [31].
2. Chartrand, Oellermann: Applied and Algorithmic Graph Theory; [44].
3. Christofides: Graph Theory - An Algorithmic Approach; [48].
4. Jungnickel: Graphen, Netzwerke und Algorithmen; [143].
5. Valiente: Algorithms on Trees and Graphs; [240].

U.2 $\mathcal{P}$ versus $\mathcal{NP}$

The class of problems which is solvable by an algorithm running in polynomially bounded time is usually defined as $\mathcal{P}$.
In theoretical computer science a problem is said to be efficiently solvable if it is in $\mathcal{P}$. This observation has led to the widely accepted consensus that feasible problems should have polynomial time complexity. This is reasonable, as polynomial time complexity does not depend on the machine model provided realistic machines are considered, [2].[1] A problem for which it is conjectured that no polynomial algorithm exists is said to be intractable. For instance, we saw that the problem of a shortest path in a network is in $\mathcal{P}$; but the problem of a longest path is intractable, see Garey

[1] The natural answer that a linear time algorithm is efficient, and an exponential time one not is to be read with care: Consider two algorithms whose running times are $t_1(n) = c \cdot n$ and $t_2(n) = 2^{n/c}$, where c is a "very large" number. Then the second algorithm is faster for all practical purposes.
As another example consider chess. There is only a finite number of possible games, seeing as follows: First we have at most 32 figures on the 64 arrays on the board. Hence, in view of D.2.2 we have at most $\sum -k = 2^{32}64^k \leq 64^{33} \leq 10^{60}$ possible positions. second, a game is finite sequence of positions. Since cycles are not are not of interest, we can estiamate the number of games by

and Johnson [97].

The class $\mathcal{NP}$ is the class of decision problems that can be solved in polynomially bounded time in a nondeterministic way. A nondeterministic algorithm

- Has the property that a state may determine many successor states, and each of these followed up on simultaneously; or equivalently,
- Has two stages: First it guesses a structure of a potential solution; Secondly it checks whether it is really a solution.

In other words, $\mathcal{NP}$ is the class of problems for which it is "easy", i.e. achievable in polynomially bounded time, to check the correctness of a claimed solution; while $\mathcal{P}$ is the class of problems that are "easy" to solve.

$$\mathcal{P} \subseteq \mathcal{NP}. \tag{U.1}$$

A problem is $\mathcal{NP}$-hard if it is as "hard" as any problem in $\mathcal{NP}$; it is $\mathcal{NP}$-complete if it is both $\mathcal{NP}$-hard and in $\mathcal{NP}$. More exactly, a problem in $\mathcal{NP}$ is defined to be $\mathcal{NP}$-complete if all other problems in $\mathcal{NP}$ can be reduced to it with the help of a transformation which takes polynomial time.
There is a straightforward strategy for proving new $\mathcal{NP}$-complete problems, once we have at least one (suitably chosen) known $\mathcal{NP}$-complete problem available. To prove that the problem Π_1 is $\mathcal{NP}$-complete, we merely show that

1. $\Pi_1 \in \mathcal{NP}$; and
2. Some known $\mathcal{NP}$-complete problem Π_2 can be transformed to Π_1, using at most polynomial time.

$\mathcal{NPC}$ denotes the class of all $\mathcal{NP}$-complete problems. All the problems in this class are believed to be intractable.
An important open question in the theory of computation is whether the containment of these classes is proper; meaning, is $\mathcal{P} \subset \mathcal{NP}$? Usually, this statement is held to be true, and is called Cook's hypothesis, first stated in 1971 [59]. Note that the statements

- $\mathcal{P} \neq \mathcal{NP}$, i.e. $\mathcal{P} \subset \mathcal{NP}$;
- $\mathcal{NPC} \cap \mathcal{P} = \emptyset$; and
- $\mathcal{NPC} \cup \mathcal{P} \subset \mathcal{NP}$, i.e. $\mathcal{NPC} \cup \mathcal{P} \neq \mathcal{NP}$;

are pairwise equivalent, compare Garey and Johnson [97]. Roughly speaking, the class of $\mathcal{NPC}$ problems has the following properties:

$\kappa = 1060!$. Then

$$\lg \kappa \leq \lg e + 30 + 10^{60} \lg \frac{10^{60}}{e} < 31 + 10^{60} 60 \leq 10^{62}.$$

Hence, there are at most $10^{\kappa} = 10^{10^{62}}$ possible games, and the question "Has white a winning strategy?" can be solved in constant time!

1. If an efficient solution is found for one, then it will work for all;

2. No such general solution has been found for any; but

3. There is no proof that an efficient solution cannot exist.

We assume that Cook's hypothesis is true. By now there are several thousands of problems known to be $\mathcal{NP}$-complete. For none of these a polynomial algorithm has been found. Furthermore, Strassen [230]:

> "The evidence in favor of Cook's hypothesis is so overwhelming, and the consequences of their failure are so grotesque, that their status may perhaps be compared to that of physical laws rather than that of ordinary mathematical conjectures."

Remember Church thesis; now we give the **Polynomial-Time Church-Turing thesis:** The class $\mathcal{P}$ captures the true notion of those problems that are computable in polynomial time by sequential machines, and is the same for any physically relevant model and minimally reasonable time measure of sequential computation that will ever be devised. In other terms, in "our world" $\mathcal{P} \neq \mathcal{NP}$ holds.[2]

The set

$$\mathcal{NPI} := \mathcal{NP} \setminus (\mathcal{P} \cup \mathcal{NPC}) \tag{U.2}$$

consists of the problems having "intermediate" difficulty between $\mathcal{P}$ and $\mathcal{NPC}$. It is reasonable to ask if there is any "usual" problem that is a candidate for membership in $\mathcal{NPI}$. A potential member is the problem of graph isomorphism.

U.3 The asymmetry of $\mathcal{NP}$

Note that the definition of efficient computation, and hence of $\mathcal{NP}$, is essentially asymmetric. That means: When we have a "yes" solution, we can provide a relatively short proof of this fact. But when we have "no" solution, no such short proof is guaranteed.
For each problem Π, there is a natural complementary problem Π^c: For all inputs x, we say $x \in \Pi^c$ if and only if $x \notin \Pi$. Of course, If $\Pi \in \mathcal{P}$ then $\Pi^c \in \mathcal{P}$.
Such a result for $\mathcal{NP}$ is far from to be clear. There is a class related to $\mathcal{NP}$ that is designed to model this issue, called co-$\mathcal{NP}$, defined by $\Pi \in$ co-$\mathcal{NP}$ if and only if $\Pi^c \in \mathcal{NP}$. It is unknown whether $\mathcal{NP}$ and co-$\mathcal{NP}$ are different.

[2] There are "worlds" in which $\mathcal{P} = \mathcal{NP}$ and others in which $\mathcal{P} \neq \mathcal{NP}$. Furthermore, if a "world" is chosen at random, the probability is 1 that it will be a world in which $\mathcal{P} \neq \mathcal{NP}$. For a proof and a broader discussion see Schöning and Pruim [216].

U.4 The complexity of enumeration problems

Sometimes we need the number of solutions to a problem. Enumeration problems provide natural candidates for the type of problems that might be intractable even if $\mathcal{P} = \mathcal{NP}$. Many such problems appear to be quite difficult. Clearly,

Observation U.4.1 *An enumeration problem associated with an $\mathcal{NP}$-complete problem is $\mathcal{NP}$-hard.*

The class $\#\mathcal{P}$ (read: number-$\mathcal{P}$) captures the problems of counting the number of solutions of $\mathcal{NP}$-problems. Moreover, some enumeration problems seem to be even harder than the corresponding decision problems. On the basis of such observations we have the class of the $\#\mathcal{P}$-complete problems (read: number-$\mathcal{P}$-complete) which is designed to reflect the difficulty of enumeration; see Garey and Johnson [97]. For instance,

Remark U.4.2 *(Jerrum [140]) Counting the number of trees with a given number of vertices is $\#\mathcal{P}$-complete.*

On the other hand, some nontrivial enumeration problems can be solved in polynomial time. Consider the following problem:

The number of spanning trees
Given: A graph G.
Question: How many distinct spanning trees are there for G?

This question can be solved in polynomial time using Kirchhoff's theorem 13.6.1. Counting spanning trees is one of the few enumeration problems which has a polynomially bounded time algorithm.

There are problems which decision version is in $\mathcal{P}$, but the counting version is $\#\mathcal{P}$-complete. For instance Vazirani [241] named:

- Perfect matching in general graphs.
- Number of trees in an undirected graph.[3]
- Counting graphs with a given degree sequence.

U.5 The spectrum of computational complexity

Tarjan [235] and [234] illustrates what one calls the "Spectrum of computational complexity", a plot of problems, versus the complexities of their fastest known algorithms. We generalize this concept to three regions, dividing all problems in the tractable, the intractable ones and in a class for which the complexity is still unknown.[4]

[3] Not necessarily spanning trees and connected graphs.
[4] Provided $\mathcal{P} \neq \mathcal{NPC}$.

Tractable	$\log n$	Searching in an ordered universe
	n	Selection
		Searching
		Planarity
	$n \log n$	Sorting
	n^2	Optimal pairwise alignment
		Eulerian cycle
		Shortest path
		Tree isomorphism
		Dynamic programing
	$n^2 \log n$	Minimum spanning tree
	$n^{\log 7}$	Matrix multiplication
	n^3	Metric closure
	polynomially bounded	Linear programming
Intermediate	?	Graph isomorphism
Intractable		$\mathcal{NP}$-complete problems
	exponential	Traveling salesman problem
		Hamiltonian cycle
		Longest path
		Shortest common superstring
		Steiner's problem
		Chromatic number
	superexponential	Presburger arithmetic
	undecidable	The halting problem
		Hilbert's tenth Problem

At the bottom of the plot are the undecidable problems, those with no algorithms at all. Above these are the problems that do have algorithms but only inefficient ones. These problems form the subject matter of high-level complexity. The emphasis in such a class is on proving non-polynomial lower bounds on the time requirements.[5] At the top is the class of low-level complexity. For such problems lower bounds are almost nonexistent; the emphasis is on obtaining even faster algorithms.
For a broader discussion of almost all these problems see Korte, Vygen [158] or Papadimitiou, Steiglitz [186].

U.6 Bioinformatics

Computers are useless. They can only give you answers.

[5] In literature there is a little bit confusion about the term "exponential". We use it in the sense, that there is a polynomial p such that the function grows how $2^{O(p(n))}$, since for any polynomial q we have

$$q(n)^{p(n)} = 2^{p(n)\cdot \log q(n)} = 2^{O(p(n))}.$$

But, on the other hand, we are already interested in the degree of the polynomial $p(.)$.

Pablo Picasso

It is extremely remarkable that the molecules which are the carriers of information and the operational units which make life work are all linear polymers. Such polymers can be written as sequences or words; and exactly these entities are the subjects which can be handled by computers.

Bioinformatic stands for discussing biological questions with a computer, for example about

- Searching in biological databases, in particular using public databases;
- Comparing sequences, in particular alignment sequences;
- Looking at protein structures;
- Phylogenetic analysis.

It may be of importance here to note that the culture of computational biology differs from the culture of bioinformatics, Konopka, Crabbe [157]:

> Sequence analysis plays an important role in both fields, but its methods and goals are understood differently by computational biologists and by bioinformaticians. Computational biology originally attracted a considerable number of practically minded theoretical biologists in the 1970s and 1980s who were both curious about the phenomenon of life and mathematical literate. They wanted to study nucleic acid and protein sequences in order to better understand life itself. In contrast, bioinformatics has attracted a large number of skilled computer enthusiasts with knowledge of computer programs that could serve as tools for laboratory biologists. ... Today's split between computational biology and bioinformatics appears to be a reflection of a profound cultural clash between curiosity-driven attitude of computational scientists and adversarial competitivness of molecular biology software providers.

Appendix V

The genetic code

A block code is a code having all its words of the same length; this number of letters. Of course, a block code is a prefix code.
The famous genetic code hardwired into every cell in your body is a good example for another type of a block code. Because there are four possible nucleic acids: adenine (a), cytosine (c), guanine (g), and uracil (u), that can appear at each location in a code word. 20 amino acids: alanine, arginine, ..., valine, are coded. Hence, each code word must be of length 3.

	u	c	a	g	
u	phenylalanine	serine	tyrosine	cysteine	u
	phenylalanine	serine	tyrosine	cysteine	c
	leucine	serine	*punctuation*	*punctuation*	a
	leucine	serine	*punctuation*	tryptophan	g
c	leucine	proline	histidine	arginine	u
	leucine	proline	histidine	arginine	c
	leucine	proline	glutamine	arginine	a
	leucine	proline	glutamine	arginine	g
a	isoleucine	threonine	asparagine	serine	u
	isoleucine	threonine	asparagine	serine	c
	isoleucine	threonine	lysine	arginine	a
	methionine	threonine	lysine	arginine	g
g	valine	alanine	aspartic acid	glycine	u
	valine	alanine	aspartic acid	glycine	c
	valine	alanine	glutamic acid	glycine	a
	valine	alanine	glutamic acid	glycine	g

Nature uses a similar approach, namely using "supersymbols", for the genetic code. In the genetic code each codeword has a length of 6 bits, but this is not necessary,

when we additionally use the binary alphabet $A' = \{r, y\}$ in which r codes for a purine (a or g), y codes for a pyrimidine (c or u), each of 1 bit, and $-$ codes for any one of 0 bit.[1]

	u	c	a	g
u	y: phenylalanine r: leucine	-: serine	y: tyrosine r: *punctuation*	y: cysteine a: *punctuation* g: tryptophan
c	-: leucine	-: proline	y: histidine r: glutamine	-: arginine
a	y: isoleucine a: isoleucine g: methionine	-: threonine	y: asparagine r: lysine	y: serine r: arginine
g	-: valine	-: alanine	y: aspartic acid r: glutamic acid	-: glycine

Glancing at this structure, it is clear that the genetic code is fault-tolerant, in the sense that transcription errors in the third codon position frequently do not influence the amino acid expressed. This is called the wobble-hypothesis-hypothesis.[2]

[1] $-$ is a "dummy" letter.

[2] Consider the amino acid composition given by the Swiss-Prot protein sequence data bank *www.expasy.ch*

Appendix W

The Linnaeus' System

Classifications are of great relevance in biology. Here a class is defined as a group of entities which are similar and related. In the book *The System of Nature* Linnaeus introduced a system still in use today. We divide life into

- Domain: There are three domains. The first two, Bacteria and Archea, are made up of many microscopic single-celled organisms. The third domain, Eukarya, is diverse.
- Kingdom: Part of the Eukarya, namely protists, fungi, animals and plants;
- Phylum: Organisms built to the same underlying plan.
- Class: (not our mathematical sense of a class.) Part of a phylum. Contains organisms that share important features.
- Order: Part of a class. Organisms in an order are usually similar in shape.
- Family: Part of an order. Organisms in a family have similar ways of life.
- Genus: Part of a family. A number of different species that are very closely related.
- Species: A maximal group of individual organisms that are able to interbreed and produce fertile offspring.

More or less all these groups are artificial, insofar as their members are categorized according to agreed-upon levels of similarity rather than precise definitions. The exceptions are species.[1]
This task is more complicated than it seems at first glance; Gould [102] wrote:

> When systematists, also known as taxonomists, set out to reconstruct the phylogeny (evolutionary history) of a group of species that they think

[1] A nice illustration of this point of view is given by Gould and Keeton [102]:

are related, they have before them the species living today and the fossil record. To reconstruct a phylogenetic history as closely as possible, they must make inferences based on observational and experimental data. The difficulty is that what can be measured is *similarity*, whereas the goal is to determine *relatedness*.

Note that the definition of similarity cannot be the problem of the mathematical analysis. This is, in any case, the task of the biological sciences. But mathematics can help to check if the choice was not false.

Biological	Postal
Domain	Old/New World
Kingdom	Country
Phylum	State/Province
Class	City
Order	Street
Family	Number
Genus	Last name
Species	First name

Bibliography

[1] R. Ahlswede, N. Cai, and Z. Zhang. A Recursive Bound for the Number of Complete K-Subgraphs of a Graph. In R. Bodendieck and R. Henn, editors, *Topics in Combinatorics and Graph Theory*, Physica-Verlag, Heidelberg, 1990, 37–39.

[2] A.V. Aho, J.E. Hopcroft, and J.D. Ullmann. *The Design and Analysis of Computer Algorithms*. Addison-Weseley, 1974.

[3] M. Aigner. *Graphentheorie*. Teubner, 1984.

[4] M. Aigner. *Diskrete Mathematik*. Vieweg, 1993.

[5] M. Aigner and G.M. Ziegler. *Proofs from The Book*. Springer, 1998.

[6] P.S. Alexandroff. *Einführung in die Mengenlehre und die Theorie der reellen Funktionen*. Deutscher Verlag der Wissenschaften, Berlin, 1973.

[7] K. Al-Knaifes and H. Sachs. Graphs, Linear Equations, Determinants, and the Number of Perfect Matchings. *Contemporary Methods in Graph Theory*, Bibliographisches Institut (BI), Mannheim, 1990, 47–71.

[8] S.F. Altschul. A Protein Alignment Scoring System Sensitive at All Evolutionary Distances. *J. Molecular Evolution*, 36:290–300, 1993.

[9] I. Anderson. *A First Course in Discrete Mathematics*. Springer, 2001.

[10] K. Appel and W.Haken. The solution of the four-color map problem. *Sci. Am.*, 237:108–121, 1977.

[11] E. Artin. *Galoissche Theorie*. Teubner Verlagsgesellschaft, Leipzig, 1965.

[12] W.G. Aschkinuse. Vielecke und Vielflache. In *Enzyklopädie der Elementarmathematik*. Band IV, Deutscher Verlag der Wissenschaften, Berlin, 1980.

[13] F.J. Ayala and A.A. Escalaute. The evolution of human populations: A molecular perspective. *Mol. Phyl. Evol.*, 5:188–201, 1996.

[14] M. Balinski. On the graph structure of of convex polyhedra in n-space. *Pacific Journal of Mathematics*, 11:431–434, 1961.

[15] H.-J. Bandelt and A. Dress. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics*, 7:309–343, 1986.

[16] H.-J. Bandelt, P. Forster, B.C. Sykes, and M.B. Richards. Mitochondrial Portraits of Human Populations Using Median Networks. *Genetics*, 141:743–753, 1995.

[17] J. Bang-Jensen and G. Gutin. *Digraphs*. Springer, 2002.

[18] D.W. Barnette. Trees in polyhedral graphs. *Canad. J. Math.*, 18:731–736, 1966.

[19] D.W. Barron. *Rekursive Techniken in der Programmierung*. Leipzig, 1971.

[20] I.W. Beineke. Decomposition of complete graphs in forests. *Magyar Tud. Adad. Mat. Kutato Int. Közl*, 9:589–594, 1964.

[21] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[22] E.A. Bender. The number of Three-dimensional Convex Polyhedra. *Am. Math. Monthly*, 94(1987), 7–21.

[23] E.A. Bender, Z. Gao, and N.C.Warmold. The number of 2-connected labelled planar graphs. *Electron. J. Combin.*, 9(2002), 104–117.

[24] C. Berge and A. Ghouila-Houri. *Programmes, Jeux et Reseaux de Transport.* Paris, 1962.

[25] C. Berge. *Graphs*. Elsevier Science Publishers, 1985.

[26] M.W. Bern and R.L. Graham. The Shortest Network Problem. *Scientific American*, 260:84–89, 1989.

[27] L.J. Billera, S.P.Holmes, and K.Vogtmann. Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics*, 27:733–767, 2001.

[28] K.P. Bogart. *Introductory Combinatorics*. HBJ, 1990.

[29] B. Bollobas. *Graph Theory*. Springer, 1979.

[30] B. Bollobas. *Random Graphs.* Cambridge University Press, 2001.

[31] H.-J. Böckenhauer and D.Bongartz. *Algoritmische Grundlagen der Bioinformatik*. Teubner, 2003.

[32] O. Boruvka. O jistem problemu minimalnim. *Acta Societ. Scient. Natur. Moravice*, 3:37–58, 1926.

[33] A. Brondsted. *An Introduction to Convex Polytopes*. Springer, 1983.

[34] P. Buneman. The recovery of trees from measures of dissimilarity. In F.R. Hodson, D.G. Kendall, and P. Tautu, editors, *Mathematics in the Archaeological and Historical Sciences*, pages 387–395. Edinburgh University Press, 1971.

[35] P. Buneman. A note of metric properties of trees. *J. Comb. Theory B.*, 1:48–50, 1974.

[36] R.M. Cann, M. Stoneking, and A. Wilson. Mitochondrial DNA and Human Evolution. *Nature*, 325:31–36, 1987.

[37] M. Carter, M. Hendy, D. Penny, L.A. Székely, and N.C. Wormald. On the distribution of lengths of evolutionary trees. *SIAM J. Disc. Math.*, 3:38–47, 1990.

[38] L.L. Cavalli-Sforza and A.W.F. Edwards. Phylogenetic analysis: models and estimation procedures. *Evolution*, 21:550–570, 1967.

[39] L.L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The History and Geography of Human Genes*. Princeton University Press, 1994.

[40] L.L. Cavalli-Sforza. Stammbäume von Völkern und Sprachen. In B. Streit, editor, *Evolution des Menschen*, pages 118–125, Spektrum Akademischer Verlag, 1995.

[41] A. Cayley. A theorem on trees. *Quart. Math.*, 23:376–378, 1889.

[42] S. Chan, A. Wong, and D. Chiu. A survey of multiple sequence comparison methods. *Bull. Math. Biol.*, 54:563–598, 1992.

[43] C.A. Charalambides. *Enumerative Combinatorics*. Chapman Hall/CRC, 2002.

[44] G. Chartrand and O.R. Oellermann. *Applied and Algorithmic Graph Theory*. McGraw-Hill, 1993.

[45] G. Chartrand and L. Lesniak. *Graphs and Digraphs*. Wadsworth and Brooks/Coole, 1996.

[46] H. Chen. Means Generated by an Integral. *Mathematics Magazin*, 78:397–399, 2005.

[47] D. Cheriton and R.E. Tarjan. Finding Minimum Spanning Trees. *SIAM J. Comp.*, 5:724-742, 1976.

[48] N. Christofides. *Graph Theory - An Algorithmic Approach.* New York, 1975.

[49] F. Chung. Graph Theory in the Information Age. In *Notice of the AMS*, 57:726–732, 2010.

[50] D. Cieslik, A. Ivanov, and A. Tuzhilin. Melzak's Algorithm for Phylogenetic Spaces (Russian). *Vestnik Moskov. Univ. Ser. I. Mat. Mech.*, 3:22–28, 2002. English translation in *Moscow Univ. Bull.*, 57:22-28, 2002.

[51] D. Cieslik, A. Dress, K.T. Huber, and V. Moulton. Connectivity Calculus. *Appl. Math. Letters*, 16:395–399, 2003.

[52] D. Cieslik. *Shortest Connectivity - An Introduction with Applications in Phylogeny*. Springer, 2005.

[53] D. Cieslik. What does Ockham's Razor in Network Design really mean? *Proceeedings of the IPSI-2005 Slovenia*, (electronic version), 2005.

[54] D. Cieslik. *Discrete structures in biomathematics*. Shaker Verlag, 2006.

[55] D. Cieslik. *Counting Graphs - An Introduction with Specific Interest in Phylogeny*. Shaker Verlag, 2012.

[56] P. Clote and R. Backofen. *Computational Molecular Biology*. John Wiley & Sons, 2000.

[57] B. Comrie, S. Matthews, and M. Polinsky. *The Atlas of Languages*. Quarto Publishing Plc., 1996.

[58] J.H. Conway and R.K. Guy. *The Book of Numbers*. Springer, 1996.

[59] S.A. Cook. The complexity of theorem-proving procedures. In *3rd Annual Symp. on Foundations of Computer Sciences*, pages 431–439, 1971.

[60] H.T. Croft, K.F. Falconer, and R.K. Guy. *Unsolved Problems in Geometry*. Springer, 1991.

[61] D.M. Cvetkovic, M.Doob, and H.Sachs. *Spectra of Graphs*. Heidelberg, 1985.

[62] D. Darling. *The Universal Book of Mathematics*. John Wiley & Sons, Inc., 2004.

[63] C. Darwin. *The Origin of Species*. London, 1859.

[64] R. Dawkins. *Geschichten vom Ursprung des Lebens*. Ullstein, 2008.

[65] M.O. Dayhoff. Atlas of Protein Sequence and Structure. Technical Report 5, National Biomedical Research Foundation, Washington, D.C., 1978.

[66] R. Diestel. *Graph Theory*. Springer, 1997.

[67] E.W. Dijkstra. A note on two problems in connection with graphs. *Numer. Math.*, 1:269–271, 1959.

[68] W.F. Doolittle. Stammbaum des Lebens. *Spektrum der Wissenschaft*, pages 52–57, April 2000.

[69] D. Dossing, V. Liebscher, H. Wagner, and S. Walcher. Evolution, Bäume und Algorithmen. *MNU*, 60/2:68–75, 2007.

[70] A. Dress, K.E.Biebler, D.Cieslik, G.Füllen, M.Haase, and B.Jäger (Eds.). *Phylogenetic combinatorics*. Shaker Verlag, 2008.

[71] S.E. Dreyfus and R.A. Wagner. The Steiner Problem in Graphs. *Networks*, 1:195–207, 1972.

[72] V. Eberhard. *Zur Morphologie der Polyeder*. Leipzig, 1891.

[73] M. Eigen. Das Urgen. Nova Acta Leopoldina 243/52, Deutsche Akademie der Naturforscher Leopoldina, 1980.

[74] M. Eigen. *Stufen zum Leben*. Serie Piper, 1992.

[75] K. Engel and H.D.O.F. Gronau. *Sperner Theory in Partially Ordered Sets*. Leipzig, 1985.

[76] E.M. Engels. *Charles Darwin*. beck'sche reihe, 2007.

[77] P. Erdös and P.Szekeres. A combinatorial problem in geometry. *Compositio Math.*, 2:463–470, 1935.

[78] P. Erdös and T.Gallai. Graphs with prescribed degrees of vertices (Hungarian). *Mat. Lapok*, 11:264–274, 1960.

[79] P. Erdös, C.Ko, and R.Rado. Intersection theorems for systems of of finite sets. *Quart. J. Math.*, 12:313–320, 1961.

[80] P. Erdös, P.Frankl and V.Rödl. The Asymptotic Number of Graphs not Containing a Fixed Subgraph and a Problem for Hypergraphs Having no Exponent. *Graphs and Combinatorics*, 2:113–121, 1968.

[81] B.S. Everitt. *Cluster Analysis*. Arnold, 1993.

[82] H. Eves. Means appearing in geometric figures. *Mathematics Magazin*, 76:292–294, 2003.

[83] J. Felsenstein. The Number of Evolutionary Trees. *Systematic Zoology*, 27:27–33, 1978.

[84] W. Fitch and E. Margoliash. Construction of Phylogenetic Trees. *Science*, 155:279–284, 1967.

[85] W.M. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20:406–416, 1971.

[86] W.M. Fitch. An Introduction to Molecular Biology for Mathematicians and Computer Programmers. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 47:1–31, 1999.

[87] J. Flachsmeyer. *Kombinatorik*. Deutscher Verlag der Wissenschaften, Berlin, 1972.

[88] H. Fleischner. The square of every two-connected graph is Hamiltonian. *J. Combin. Theory (B)*, 16(1974), 29–34.

[89] R.W. Floyd. Algorithm 97, shortest path. *Comm. ACM*, 5:345, 1962.

[90] S. Földes and P.L.Hammer. Split Graphs. *8th SE Conf., Louisiana State University, Baton Rouge.*, 311–315, 1977.

[91] L.R. Foulds, M.D. Hendy, and D. Penny. A graph theoretic approach to the development of minimal phylogenetic trees. *J. Mol. Evol.*, 13:127–149, 1979.

[92] R. Frucht. Herstellung von Graphen mit vorgegebener abstrakten Gruppe. *Compositio Math.*, 6:239–250, 1938.

[93] H.N. Gabow and R.E. Tarjan. Efficient algorithms for a family of matroid intersection problems. *J. of Algorithms*, 5:80–131, 1984.

[94] J.A. Gallian. *Contemporary Abstract Algebra.* Houghton Mifflin Company, 2002.

[95] M. Gardner. *Gotcha.* Hugendubel, 1985.

[96] M. Gardner. *Gotcha.* W.H.Freeman and Company, 2000.

[97] M.R. Garey and D.S. Johnson. *Computers and Intractibility.* San Francisco, 1979.

[98] O. Gimenez and M.Noy. Asymptotic enumeration and limit laws of planar graphs. *J. Am. Math. Soc.*, (to appear).

[99] M. Glaubrecht. *Die ganze Welt ist eine Insel* Hirzel Verlag, 2002.

[100] M.J. Golin, Y.C. Leung, and Y. Wang. Counting Spanning Trees and Other Structures in Non-constant jump Circulant Graphs. *LNCS*, 3341:508–521(2004).

[101] M.C. Golumbic. *Algorithmic Graph Theory and Perfect Graphs.* Academic Press, 1980.

[102] J.L. Gould and W.T. Keeton. *Biological Sciences.* W.W.Norton and Company, 1996.

[103] R.L. Graham and L.R. Foulds. Unlikelihood That Minimal Phylogenies for a Realistic Biological Study Can be Constructed in Reasonable Computational Time. *Mathematical Biosciences*, 60:133–142, 1982.

[104] R.L. Graham and P. Hell. On the History of the Minimum Spanning Tree Problem. *Ann. Hist. Comp.*, 7:43–57, 1985.

[105] R.L. Graham, D.E. Knuth, and O. Patashnik. *Concrete Mathematics.* Addison-Wesley, Boston, 1989.

[106] R.L. Graham, B.L. Rothschild, and J.H. Spencer. *Ramsey Theory.* John Wiley and Sons, 1990.

[107] D. Graur and W.H. Li. *Fundamentals of Molecular Evolution.* Sinauer Associates, Inc., 1999.

[108] R.P. Grimaldi. *Discrete and Combinatorial Mathematics.* Addison-Weseley, 1999.

[109] J. Gross and J. Yellen. *Graph Theory and its Applications.* CRC Press, 1999.

[110] B. Grünbaum. *Convex Polytopes.* 1966.

[111] M. Guan. Graphic programming using odd and even points. *Chinese Math.*, 1:273–277, 1962.

[112] D. Gusfield. *Algorithms on Strings, Trees, and Sequences.* Cambridge University Press, 1997.

[113] J. Haß, S. Matuszewski, D. Cieslik, and M. Haase. The Role of Swine as "Mixing Vessel" for Interspecies Transmission of the influenca A subtype H1N1: A Simultaneous Bayesian Inference of Phylogeny and Ancestral Hosts. *Infection, Genetics and Evolution*, (11) 2011.
Also in: A. Dress, K.E. Biebler, D. Cieslik, and A.Spillner (Eds.). *The Math of Flu*, Shaker Verlag, 2010, 77–92.

[114] O. Häggström. *Streifzüge durch die Wahrscheinlichkeitstheorie.* Springer, 2006.

[115] A. von Haeseler and D.Liebers. *Molekulare Evolution.* Fischer Verlag, 2003.

[116] S.L. Hakimi and S.S. Yau. Distance matrix of a graph and its realizability. *Quart. Appl. Math.*, 22:305–317, 1964.

[117] S.B. Hakimi. Steiner's Problem in Graphs and its Implications. *Networks*, 1:113–133, 1971.

[118] H.-R. Halder and W.Heise. *Einführung in die Kombinatorik.* Akademie Verlag, Berlin, 1977.

[119] B.G. Hall. *Phylogenetic Trees Made Easy.* Sinauer Ass., Inc., 2001.

[120] P.L. Hammer and B.Simeone. The splittance of a graph. *Combinatorics.*, 1:275–284, 1981.

[121] F. Harary. *Graph Theory.* Perseus Book Publishing, 1969.

[122] F. Harary and E.M. Palmer. *Graphical Enumeration.* Academic Press, 1973.

[123] F. Harary and G.Prins. The number of homeomorphically irreducible trees and other species. *Acta Math. Uppsala*, 101:141–162, 1959.

[124] F. Harary and A.J.Schwenk. The number of caterpillars. *Discrete Mathematics*, 6:359–365, 1973.

[125] C.W. Harper. Phylogenetic inference in paleontology. *J. Paleont.*, 50:180–193, 1976.

[126] J.A. Hartigan. Minimum mutation fits to a given tree. *Biometrics*, 29:53–65, 1973.

[127] N. Hartsfield and G.Ringel. *Pearls in Graph Theory* . Dover Publications, Inc., 1990.

[128] A. Hastings. *Population Biology*. Springer, 1997.

[129] J. Havil. *Gamma*. Princeton University Press, 2003.

[130] M. Hendy and D. Penny. Branch and bound algorithms to determine minimal evolutionary trees. Math. Biosci., 59:277–290, 1982.

[131] M. Hendy, C.H.C. Little, and D. Penny. Comparing trees with pendant vertices labelled. *SIAM J. Appl. Math.*, 44:1054–1065, 1984.

[132] C. Hierholzer. Über die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren. *Math. Ann.*, 6:30–32, 1873.

[133] S. Hildebrandt and A. Tromba. *The Parsimonious Universe*. Springer, 1996.

[134] C. Hoffmann. *Graph-theoretic algorithms and graph isomorphism*. Number 136, Lecture Notes in Computer Science. Springer-Verlag, 1982.

[135] I. Holyer. The NP-completeness of edge-coloring. *SIAM J. Computing*, 10:718–720, 1971.

[136] D. Huson, R.Rupp, and C.Scornavacca. *Phylogenetic Networks*. Cambridge University Press, 2010.

[137] F.K. Hwang, D.S. Richards, and P. Winter. *The Steiner Tree Problem*. North-Holland, 1992.

[138] T. Ihringer. *Diskrete Mathematik*. Teubner Stuttgart, 1994.

[139] A. Isaev. *Introduction to Mathematical Methods in Bioinformatics*. Springer, 2004.

[140] M.R. Jerrum. Counting trees in a graph is $\#p$-complete. *Inf. Proc. Letters*, 51:111–116, 1994.

[141] K. Jacobs. *Einführung in die Kombinatorik*. de Gruyter, 1983.

[142] D. Jungnickel. *Transversaltheorie*. Akademischer Verlag Geest & Portig, Leipzig, 1982.

[143] D. Jungnickel. *Graphen, Netzwerke und Algorithmen*. BI Wissenschaftsverlag, Mannheim, 1994.

[144] M. Kanehisa. *Post-genome Informatics*. Oxford University Press, 2000.

[145] S. Kapoor and H. Ramesh. Algorithms for Enumerating All Spanning Trees of undirected and weighted Graphs. *SIAM J. Comp.*, 24:247–265, 1995.

[146] J.J. Karaganis. On the cube of a graph. *Canad. Math. Bull.*, 11:295–296, 1968.

[147] R.M. Karp. Reducibility among combinatorial problems. In R.E. Miller and J.W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103, New York, 1972.

[148] D.C. Kay and G.Chartrand. A characterization of certain ptolemaic graphs. *Canad. J. Math.*, 17:342–346, 1965.

[149] A.K. Kelmans. On properties of the characteristic polynomial of a graph. In *Kibernetiku - Na Sluzbu Kommunizmu*. Gosenergoizdat, Moscow, 1967.

[150] B.W. Kernighan and S. Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49:291–307, 1970.

[151] V. Klee and S. Wagon. *Old and New Unsolved Problems in Plane Geometry and Number Theory*. Mathematical Association of America, Washington, 1991.

[152] V. Klee and S. Wagon. *Alte und neue ungelöste Probleme in der Zahlentheorie und Geometrie der Ebene*. Birkhäuser, Basel, 1997.

[153] J. Kleinberg. The Small-World Phenomenon and Decentralized Searched. *SIAM News*, April 2004:14, 2004.

[154] J. Kleinberg and E. Tardos. *Algorithm Design*. Pearson, Addison Wesley, 2006.

[155] D.J. Kleitman and D.B. West. Spanning trees with many leaves. *SIAM J. Disc. Math.*, 4:99–106, 1991.

[156] V. Knoop and K. Müller. *Gene und Stammbäume*. Spektrum, 2006.

[157] A.K. Konopka and M.J.C. Crabbe. *Compact Handbook of Computational Biology*. Marcel Dekker, 2004.

[158] B. Korte and J. Vygen. *Combinatorial Optimization*. Springer, 2000.

[159] J.B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. of the Am. Math. Soc.*, 7:48–50, 1956.

[160] A. Kurek. Arboricity and Star Arboricity of Graphs. *Ann. Discrete Mathematics*, 51:171–173, 1992.

[161] A. Kurosch. *Higher Algebra*. Mir Publishers, 1980.

[162] A. Lifschitz and A.Redstone. A determination of the number of caterpillars. *South East Asian Bull. Math.*, (9)1985, 50–52.

[163] B. Lomborg. *The sceptical environmentalist*. Cambridge University Press, 2002.

[164] L. Lovász, and J.Pelikán, and K.Vestergombi. *Discrete Mathematics*. Springer, 2003.

[165] D. Lubell. A short proof of Sperner's theorem. *J. Comb. Theory*, 1:299, 1966.

[166] H. Lüneburg. *Tools and fundamental constructions of combinatorial mathematics*. BI Wissenschaftsverlag, 1989.

[167] M. Marcus. *A survey of finite mathematics*. Dover Publications, 1969.

[168] L. Margulis. *Symbiotic Planet*. Weidenfeld and Nicolson/Orion Publishing, 1998.

[169] T. Margush and F.R.Morris. Consensus n-Trees. *Bull. Math. Biology*, 43:239–244, 1981.

[170] J. Matoušek and J. Nešetřil. *Diskrete Mathematik*. Springer, 2002.

[171] C. McDiarmid, A.Steger, and A.Welsh. Random Planar Graphs. *J.Combin.Theory*. Ser. B, 2005.

[172] W. Mantel. Problem 28. *Wiskunde Opgaven*, 10:60–61, 1907.

[173] G.E. Martin. *Counting: The Art of Enumerative Combinatorics*. Springer, 2001.

[174] G. Mink. Editing and Genealogical Studies: the New Testament. *Literary and Linguistic Computing*, 15:51–56, 2000.

[175] M. Mitzenmacher and E. Upfal. *Probability and Computing*. Cambridge University Press, 2005.

[176] R Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[177] G. Nägler and F.Stopp. *Graphen und Anwendungen*. Teubner, 1996.

[178] C.St.J.A. Nash-Williams. Edge-disjoint spanning trees of finite graphs. *J. London Math. Soc.*. 50:445–450, 1961.

[179] M.E.J. Newman. *Networks*. Oxford University Press, 2010.

[180] S.D. Nikolopoulos and P.Rondogiannis. On the the Number of Spanning Trees of Multi-Star Related Graphs. *Inf. Proc. Letters*, 65:183–188, 1998.

[181] M.A. Nowak *Evolutionary Dynamics*. The Belknap Press of Harvard University Press, 2006.

[182] W. Oberschelp. Kombinatorische Anzahlbestimmungen in Relationen. *Math. Ann.*, 174:53–78, 1967.

[183] R. Onadera. The number of trees in a complete n-partite graph. *Matrix Tensor Quart.*, 23:142–146, 1972/73.

[184] R. Otter. The Number of Trees. *Ann. Math.*, 49:583–599, 1948.

[185] R.D.M. Page and E.C. Holmes. *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, 1998.

[186] C.H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization*. Prentice-Hall, 1982.

[187] E. Pennisi. Modernizing the Tree of Life. *Science*, 300:1692–1697, 2003.

[188] D. Penny and M.Hendy. Testing methods of evolutionary tree construction. *Cladistics*, 1:266–272, 1985.

[189] D. Penny, 2001. private communication.

[190] D. Penny, M.D. Hendy, and A. Poole. Testing fundamental evolutionary hypotheses. *J. Theor. Biology*, 223:377–385, 2003.

[191] J. Petersen. Die Theorie der regulären Graphen. *Acta Math.*, 15:193–220, 1891.

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

[193] R.C. Prim. Shortest communication networks and some generalizations. *Bell Syst. Techn. J.*, 31:1398–1401, 1957.

[194] H.J. Prömel and A.Steger. *The Steiner Tree Problem*. Vieweg, 2002.

[195] H.J. Prömel. Graphentheorie. In G. Walz, editor, *Faszination Mathematik*, pages 184–194. spectrum, 2003.

[196] H. Prüfer. Ein neuer Beweis eines Satzes über Permutationen. *Arch. Math. Phys.*, 27:742–744, 1918.

[197] F.P. Ramsey. On a problem of formal logic. *Proc. London Math. Soc.*, 30:246–286, 1930.

[198] R.C. Read. The number of k-colored graphs on labelled nodes. *Canad. J. Math.*, 12:409–413, 1960.

[199] J.C. Redfield. The theory of group-reduced distributions. *Amer. J. Math.*, 49:433–455, 1927.

[200] A. Renyi. Some remarks on the theory of trees. *Publ. Math. Inst. Hungar. Acad. Sc.*, 4:73–85, 1959.

[201] A. Renyi and G. Szekeres. On the height of trees. *J. Austral. Math. Soc.*, 7:497–507, 1967.

[202] L.B. Richmond, R.W.Robinson, and N.C.Warmold. On Hamilton cycles in 3-connected cubic maps. *Ann. Discrete Math.*, 27:141–149, 1985.

[203] J. Riordan. *An Introduction to Combinatorial Analysis*. New York - London, 1958.

[204] F. Roberts. *Discrete Mathematical Models, with Applications to Social, Biological, and Environmental Problems*. Prentice Hall, 1976.

[205] F. Roberts. *Graph Theory and its Applications to Problem of Society*. SIAM, Philadelphia, 1978.

[206] R.W. Robinson, and N.C.Warmold. Almost all regular graphs are Hamiltonian. Preprint, 1991.

[207] R.W. Robinson, and N.C.Warmold. Almost all cubic graphs are Hamiltonian. *Random Structures and Algorithms*, 3:117–125, 1992.

[208] B. Russell. *The Principles of Mathematics*. George Allen and Unwin, 1903.

[209] H. Sachs. *Einführung in die Theorie der endlichen Graphen*, volume 1. Teubner Verlagsgesellschaft, Leipzig, 1970.

[210] H. Sachs. Einige Gedanken zur Geschichte und zur Entwicklung der Graphentheorie. *Mitteilungen der Mathematischen Gesellschaft in Hamburg*, 11:623–641, 1989.

[211] D. Sankoff. Minimal Mutation Trees of Sequences. *SIAM J. Appl. Math.*, 28:35–42, 1975.

[212] G. Sapiro. The Gromov-Hausdorff Distance and Shape Analysis. *SIAM News*, Juli/August 2009, 2009.

[213] A.A. Saposhenko and K.F.E. Weber. On the Diameter of Random Subgraphs of the n-Cube. *Contemporary Methods in Graph Theory*, Bibliographisches Institut (BI), Mannheim, 1990, 507–524.

[214] V.M. Sarich and A.C. Wilson. Immunological time scale for hominoid evolution. *Science*, 158:1200–1203, 1967.

[215] R. Schimming. A Partition of the Catalan Numbers and Enumeration of Genealogical Trees. *Discussiones Mathematicae*, 16:181–195, 1996.

[216] U. Schöning and R. Pruim. *Gems of Theoretical Computer Science*. Springer, 1998.

[217] R.-H. Schulz. *Codierungstheorie*. Vieweg, 1991.

[218] J. Sedlacek. Ungerichtete Graphen und ihre Gerüste. *Beiträge zur Graphentheorie*, Teubner, Leipzig, 1968.

[219] J. Sedlacek. *Einführung in die Graphentheorie*. Leipzig, 1972.

[220] C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, 2003.

[221] J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. PWS Publishing Company, 1997.

[222] M.L. Shore, L.R. Foulds, and P.B. Gibbons. An algorithm for the Steiner problem in graphs. *Networks*, 12:323–333, 1982.

[223] T.F. Smith, M.S. Waterman, and W.M. Fitch. Comparative Biosequence Metrics. *J. Molecular Evolution*, 18:38–46, 1981.

[224] A. Speiser. *Die Theorie der Gruppen endlicher Ordnung*. Berlin, 1937.

[225] E. Sperner. Ein Satz über Untermengen einer endlichen Menge. *Math. Zeitschrift*, 27:588–548, 1928.

[226] R.P. Stanley. *Enumerative Combinatorics*. Vol. 1, Cambridge University Press, 1997.

[227] R.P. Stanley. *Enumerative Combinatorics*. Vol. 2, Cambridge University Press, 1997.

[228] I. Stewart. *Galois Theory*. Chapman & Mathematics, 1973.

[229] L.J. Stockmeyer. Planar 3-colorability is np-complete. *SIGACT News*, 5:19–25, 1973.

[230] V. Strassen. The Work of Leslie G. Valiant. In *Proc. of the International Congress of Mathematicians, Berkeley*, 1986.

[231] K. Strimmer and A. von Haeseler. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, 13:964–969, 1996.

[232] D.L. Swofford. *PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods*, (software). Sinauer Associates, Sunderland, MA, 2000.

[233] P. Tannenbaum and R. Arnold. *Excursions in Modern Mathematics*. Prentice Hall, 2001.

[234] R.E. Tarjan. *Data Structures and Network Algorithms*. SIAM, Philadelphia, 1983.

[235] R.E. Tarjan. Efficient Algorithms for Network Optimization. In *Proceedings of the International Congress of Mathematicians*, pages 1619–1635, Warszawa, 1983.

[236] C. Thomassen. Planarity and duality of finite and infinite graphs. *J. Comb. Theory B.*, 29:244–271, 1980.

[237] P. Turan. Eine Extremalaufgabe aus der Graphentheorie. *Mat. Fiz. Lapok* (Hungarian), 48:436–452, 1941.

[238] P. Turan. Ein sonderbarer Lebensweg, Ramanujan. In Robert Freud, editor, *Grosse Augenblicke aus der Geschichte der Mathematik*. BI Wissenschaftsverlag, Mannheim, 1990.

[239] W.T. Tutte. A theorem on planar graphs. *Trans. Amer. Math. Soc.*, 82:99–116, 1956.

[240] G. Valiente. *Algorithms on Trees and Graphs*. Springer, 2002.

[241] V.V. Vazirani. *Approximation Algorithms*. Springer, 2001.

[242] M. Vingron, H.-P. Lenhof, and P. Mutzel. Computational Molecular Biology. In M. Dell'Amico, F. Maffioli, and S. Martello, editors, *Annotated Bibliographies in Combinatorial Optimization*, pages 445–471. John Wiley and Sons, 1997.

[243] M. Vingron. Sequence Alignment and Phylogeny Construction. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 47:53–63, 1999.

[244] H. Walther and H.J.Voß. *Über Kreise in Graphen*. Berlin, 1974.

[245] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *J. of Computational Biology*, 1:337–348, 1994.

[246] M.S. Waterman. Sequence Alignments. In M.S. Waterman, editor, *Mathematical Methods for DNA-Sequencing*, pages 53–92. CRC Press, 1989.

[247] M.S. Waterman. Applications of Combinatorics to Molecular Biology. In R.L. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics*, pages 1983–2001. Elsevier Science B.V., 1995.

[248] M.S. Waterman. *Introduction to Computational Biology*. Chapman & Heil, 1995.

[249] F.J. Wetuchnowski. Graphen und Netze. In S.W. Jablonski and O.B. Lupanow, editors, *Diskrete Mathematik und mathematische Fragen der Kybernetik*, pages 145–197. Akademie-Verlag Berlin, 1980.

[250] J. Whitfield. Born in a watery commune. *Nature*, 427:674–676, 2004.

[251] H. Whitney. Congruent graphs and the connectivity of graphs. *American J. Math.*, 54:150–168, 1932.

[252] J.A. Winn. *Asymptotic Bounds for Classical Ramsey Numbers*. Polygonal Publishing House, 1988.

[253] B.Y. Wu and K.-M. Chao. *Spanning Trees and Optimization Problems*. Chapman and Hall, 2004.

[254] H. Wußing. *6000 Jahre Mathematik - 1. Von den Anfängen bis Leibniz und Newton.* Springer, 2008.

[255] H. Wußing. *6000 Jahre Mathematik - 2. Von Euler bis zur Gegenwart.* Springer, 2009.

[256] H. Yockey. *Information Theory and Molecular Biology.* Cambridge University Press, 1992.

[257] G.M. Ziegler. *Lectures on Polytopes.* Springer, 1995.

[258] A.A. Zykov. *Theory of Finite Graphs.* (Russian) Novosibirsk, 1969.

Index