Short Cycles

Minimum Cycle Bases of Graphs from Chemistry and Biochemistry

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium

an der Fakultät für Naturwissenschaften und Mathematik
der Universität Wien

Vorgelegt von

Petra Manuela Gleiss

im September 2001

An dieser Stelle möchte ich mich herzlich bei all jenen bedanken, die zum Entstehen der vorliegenden Arbeit beigetragen haben.

Allen voran Peter Stadler, der mich durch seine wissenschaftliche Leitung, sein überwältigendes Wissen und seine Geduld unterstützte, sowie Josef Leydold, ohne den ich so manch tieferen Einblick in die Mathematik nicht gewonnen hätte. Ivo Hofacker, dermich oftmals aus den unendlichen Weiten des "Computer Universums" rettete.

Meinem Bruder Jürgen Gleiss, für die Einführung und Hilfstellungen bei meinen Kampf mit C++.

Daniela Dorigoni, die die Daten der atmosphärischen Netzwerke in den Computer eingeben hat.

Allen Kolleginnen und Kollegen vom Institut, für die Hilfsbereitschaft.

Meine Eltern Erika und Franz Gleiss, die mir durch ihre Unterstützung ein Studium ermöglichten. Meiner Oma Maria Fischer, für den immerwährenden Glauben an mich. Meinen Schwiegereltern Irmtraud und Günther Scharner, für die oftmalige Betreuung meiner Kinder.

Zum Schluss Roland Scharner, Florian und Sarah Gleiss, meinen drei Liebsten, die mich immer wieder aufbauten und in die reale Welt zurückführten.

Ich wurde teilweise vom österreichischem *Fonds zur Förderung der Wissenschaftlichen Forschung*, Proj.No. P14094-MAT finanziell unterstüzt.

Zusammenfassung

In der Biochemie werden Kreis-Basen nicht nur bei der Betrachtung kleiner einfacher organischer Moleküle, sondern auch bei Struktur Untersuchungen hoch komplexer Biomoleküle, sowie zur Veranschaulichung chemische Reaktionsnetzwerke herangezogen.
Die kleinste kanonische Menge von Kreisen zur Beschreibung der zyklischen Struktur eines ungerichteten Graphen ist die Menge der *relevanten Kreis* (Vereingungsmenge aller minimaler Kreis-Basen). Die relevanten Kreise sind diejenigen, die nicht als Summe kürzerer Kreise dargestellt werden können. Auf Grund fehlender Algorithmen zur Berechnung der relevanten Kreise wurden in der chemischen Literatur lange Zeit erweiterte minimale Kreis-Basen verwendet. Diese Erweiterungen sind normalerweise nicht eindeutig. Wir verdeutlichen die Zusammenhänge zwischen den am häufigsten verwendeten "Ring-Sets".
Wir führen einen neuen, von Pfaden im Graphen aufgespannten Vektorraum ein, der in der graphentheoretischen Untersuchung von chemischen Reaktionsnetzwerken ein Anwendungsgebiet hat. Die Endpunkte dieser sgn. U-Pfade bilden eine Teilmenge der Knotenmenge des Graphen. Diese Pfade spannen gemeinsam mit den Kreisen wieder einen Vektorraum, den U-Raum, auf. Wir verallgemeinern den Begriff der relevante Kreise auf diesen U-Raum.
Weiters stellen wir eine Partition der Menge aller relevanten Kreise vor, der Art, dass Kreise aus einer Klasse nur durch Kreise aus der selben Klasse und echt kürzeren dargestellte werden können. Jede minimale Kreis-Base enthält immer die gleiche Zahl an Repräsentanten einer Klasse. Wir können eine Erweiterung dieser Partition auf den U-Raum vornehmen. Diese Äquivalenzklassen sowie die Menge aller relevanten Kreise und U-Pfade lassen sich mit unserem Algorithmen in polynomialer Zeit ausrechnen.
Da beim Betrachten von chemischen Reaktionsnetzwerken, die Richtung des Flusses eine wichtige Rolle spielt, übertragen wir das Konzept der relevanten Kreise auf die relevanten Zyklen (gerichtete Kreise). Wie allgemein bekannt, hat jeder stark zusammenhängende gerichtete Graph eine (gerichtete) Zyklen-Basis. Wir beweisen, dass eine minimale Zyklen-Basis in polynomialer Zeit ausgerechnet werden kann.
Als Anwendung der mathematischen Konzepte, zeigen wir, dass jede Äquivalenzklasse der relevanten Kreise jeweils ein einzelnes Strukturelement der mit Pseudoknoten erweiterten RNA Sekundärstruktur darstellt.

Abstract

In chemistry cycle bases are not only suitable for the analysis of small simple organic molecules, but also for structural studies of highly complex biomolecules and the visualization of chemical reaction networks.
The smallest canonical set of cycles that describes the cyclic structure of a graph is the union of all minimum cycle bases, the so-called set of relevant cycles. These relevant cycles can not be represented as the sum of shorter cycles. Since no efficient algorithm was known to calculate the set of relevant cycles, many investigators dealt with the definition of extended minimum cycles bases. These sets are in general not canonical, however. We clarify the mutual relationships of some of the more frequently used ring sets.
We introduce a new vector space, spanned by paths of the graph. The endpoints of these so called U-paths form a subset of the vertex set. This construction is of interest in the context of chemical reaction networks. The U-paths and the cycles of the graphs form an extended vector space, the U-space. This extended vector space is the union of the well known cycle space and our new vector space. Thus we generalize the notion of relevant cycles to the U-space and give a polynomial time algorithm to calculate the relevant cycles and U-paths.
Furthermore, we introduce a partition of the set of relevant cycles, called interchangeability, such that each class contains cycles of the same length, which can be represented through cycles of the same class and shorter ones. We show that each minimum cycle basis contains the same number of representatives from each class. In addition, we give a polynomial time algorithm to compute this partition. Moreover, this partition is extended to the U-space.
When analyzing chemical reaction networks, the direction of the flux plays an important role. Therefore we extend the notion of the relevant cycles to the circuits. It is well known that every strongly connected digraph has a circuit basis. We show that a minimum circuit basis of a strongly connected digraph can be computed in polynomial time.
Finally, we give a biochemical application of the mathematical concept of interchangeability. We can show that each interchangeability class corresponds to a single structural element of the RNA secondary structure containing pseudo-knots.

Contents

1 Introduction **1**

1.1 General Context . . . 1

1.2 Chemical Applications of Graph Theory . . . 2

1.2.1 Ring Perception . . . 3

1.2.2 Biopolymer Graphs . . . 5

1.2.3 Reaction Graphs . . . 7

1.3 Organization of the Thesis . . . 10

2 Basic Definitions **12**

2.1 Graphs . . . 12

2.1.1 Basic Definitions . . . 12

2.1.2 Paths, Circuits, Trees and Cuts . . . 13

2.1.3 Distance . . . 14

2.1.4 Trees and Cuts . . . 14

2.2 Vector Spaces . . . 15

2.2.1 Groups and Fields . . . 15

2.2.2 Vector Space . . . 16

2.2.3 Bases of Vector Spaces . . . 16

2.2.4 Subspaces . . . 17

2.2.5 Vector Space over GF(2) . . . 17

2.2.6 Vector Spaces on a Graph . . . 18

2.3 Matroids . . . 19

2.3.1 Independent Sets, Bases and Circuits . . . 20

2.3.2 The Cycle Matroid of a Graph . . . 21

2.3.3 Binary Matroids . . . 22

2.3.4 The Greedy Algorithm . . . 23

3 Cycle Bases of Undirected Graphs 25

3.1 Fundamental Cycle Bases . . . 25
3.2 Isometric, Short, Shortest, Edge-Short Cycles . . . 26
3.3 Minimum Cycle Bases . . . 29
3.4 Relevant Cycles . . . 30
3.5 Vismara's Prototypes of Relevant Cycles . . . 32
3.6 Essential Cycles . . . 34

4 Ring Sets for Chemical Applications 38

4.1 All Cycles and Simple-Cycles . . . 38
4.2 Smallest Set of Smallest Rings (SSSR) . . . 39
4.3 K-rings . . . 41
4.4 β-ring . . . 41
4.5 Essential Set of Essential Rings (ESER) . . . 41
4.6 Minimal Planar Cycle Bases . . . 43
4.7 Extended Set of Smallest Rings (ESSR) . . . 43
4.8 Set of Elementary Rings (SER) . . . 45

5 U-Bases 46

5.1 Dimension of the U-space $\mathfrak{U}(\mathcal{G})$ . . . 46
5.2 Minimal U-Bases . . . 47
5.3 Relevant U-Paths . . . 48
5.4 U-Path Prototypes . . . 49

6 Interchangeability of Relevant Cycles 52

6.1 A Partition of $\mathcal{R}$ . . . 52
6.2 Some Examples . . . 57
6.3 The Number of Minimal Cycle Bases . . . 59
6.4 Computing Interchangeability Classes . . . 61
6.5 Stronger Interchangeability . . . 62
6.6 Interchangeability of U-Path . . . 64

7 Circuit Bases of Digraphs 65

7.1 Circuit Space . . . 65
7.2 Elementary Circuits . . . 66
7.3 Circuit Bases . . . 68
7.4 Minimum Circuit Bases and Relevant Circuits . . . 70
7.5 Short and Isometric Circuits . . . 73

8 Computations **78**
8.1 Programs 78
8.2 RNA Structures 79
8.2.1 Background 79
8.2.2 *Escherichia coli* α-operon mRNA 84
8.2.3 tmRNA 85
8.2.4 Ribonuclease P 89
8.3 Chemical Networks 92
8.3.1 Reaction Networks 92
8.3.2 A Metabolic Network 96
8.3.3 Planetary Atmospheres 97
8.3.4 Comparison of the Chemical Networks 98

9 Conclusion and Outlook **101**

References **104**

Chapter 1

Introduction

1.1 General Context

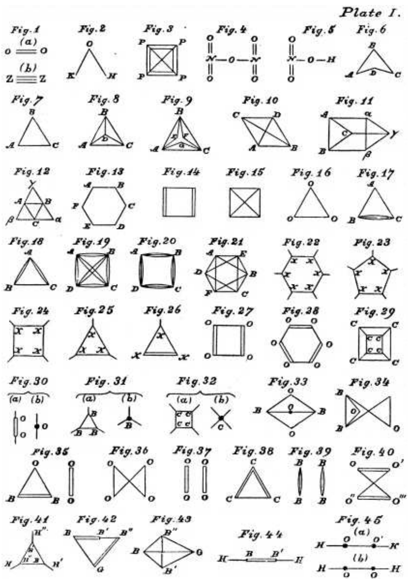


Figure 1.1. Reproductions of chemical graphs studied by Sylvester (1878).

The mathematization of chemistry has a long and colorful history extending back well over two centuries. At any period in the development of chemistry the extent of the mathematization process roughly parallels the progress of chemistry as a whole. Thus, in 1786 the German philosopher Immanuel Kant [85] observed that the chemistry of his day could not qualify as one of the natural sciences because of its insufficient degree of mathematization.

It has frequently been remarked that mathematics is a more effective tool in the natural sciences than might be reasonably expected [164]. One of the functions of a model in science is to replace the actual elements in a given set by an idealized set of mathematical abstraction that approximate these elements. The closer the approximation, the better the model. In the case of chemical compounds, for example, the use of graph theoretical notions (node = atom, edge = chemical bond) is a convenient first step in building the model.

Graph theory itself is one of the few branches of mathematics that may be said to have a precise starting date. In 1736, Euler [44] solved a celebrated problem, known as the *Königsberg Bridges Problem*. Since its beginnings, graph theory has been exploited for the solution of numerous practical problems, and today still retains an applied character.

In the early days, important progress was made in the development of graph theory by the investigation of some very concrete problems, e.g. Kirchhoff's study of electrical circuits [87], and Cayley's attempts to enumerate chemical isomers [19]. Further details on the history of graph theory may be obtained form the monograph by Biggs *et al.* [12].

The term *graph* was introduced by Sylvester 1878 [131], referring to diagrams showing analogies between the chemical bonds in molecules and graphical representations of mathematical invariants (see Figure 1.1). For a mathematician, a graph is the application of a set on itself (i.e. a collection of elements of the set, and of binary relations between these elements). For a chemist, however, the geometrical realization of a graph is more appealing, namely a collection of *points* (i.e. elements of the set) and of *lines* joining some of these points either to other points or to themselves.

Since usually no specification is made as to the shape or length of lines, or to angles between lines, graphs are topological rather than geometrical objects, having as the most important feature the adjacency relationships between points.

Having chemistry as one of the breeding grounds, graph theory is well adapted for solving chemical problems, both by the high degree of abstraction evidenced by the generality of such concepts as points, lines and neighbors, as well as by the combinatorial derivation of many graph-theoretical concepts which correspond to the essence of chemistry viewed as the study of combinations between atoms.

1.2 Chemical Applications of Graph Theory

Mainly, two kinds of correspondence between graphs and chemical categories have found numerous applications: (i) a graph corresponds to a molecule or group of molecules, i.e., points symbolize atoms and lines symbolize chemical covalent bonds (*structural* or *constitutional graphs*, Fig. 1.2), and (ii) a graph corresponds to a reaction mixture, i.e., points symbolize chemical species and lines symbolize conversions between these species (*reaction graphs*, Fig. 1.5). The former type of graph gave Cayley the incentive to develop a procedure for counting the constitutional isomers of alkanes [19]. Reaction graphs play an ever increasing role in explaining and rationalizing rearrangements. The methods of graph theory were used for the first time in kinetic studies in the works of Vol'kenshtein and Gol'dshtein (see [150, 151, 152, 153]).

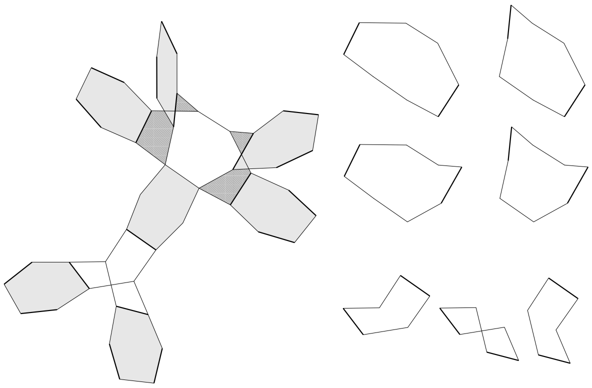


Figure 1.2. Organic carbon compounds may exhibit elaborate polycyclic structures. The example shown here is Compound 8 from [84] (aromatic "double bonds" are indicated by thick lines).

1.2.1 Ring Perception

The identification of cyclic substructures from connectivity information is a critical aspect in the solution of problems as diverse as the analysis of electrical circuits, analysis of communication networks, and analysis of chemical structures. Chemical applications of ring identification algorithms are also quite diverse. Software for the prediction of physical and chemical properties, suggestion of synthetic strategies, chemical database management, substructural searching, structure elucidation and three-dimensional coordinate generation all require a fast and accurate method for identification of the "chemically meaningful" rings among the potentially large number of cyclic subgraphs embedded in the molecular structure.

Ring identification methods can be classified into three categories of increasing complexity:

(i) *Ring detection*: identification of a basis set of the "ring space" of the structure

(ii) *Ring perception*: identification of a minimal basis set of the ring space of the structure (i. e., a **S**mallest **S**et of **S**mallest **R**ings, an SSSR)

(iii) *Canonical ring perception*: identification of the "preferred" SSSR based upon a collection of application-dependent ring properties

A simple ring detection method may be sufficient in some cases but the vast majority of applications requires a ring perception or a canonical ring perception method.

In 1957, Gould [58] noticed that the cycles of a graph generate a finite-dimensional linear space closed with respect to the Boolean sum (exclusive-OR) of their edges. Welch [159] recognized the need for a ring detection method that would identify a basis set of the ring space, thus allowing the exhaustive generation of all possible cycles in a structure [53]. Although other investigators have presented methods for the direct identification of all the rings without an intermediate ring detection step [4, 142], all of these exhaustive algorithms suffer from the very serious intrinsic inefficiency of generating and storing a very large number of rings. To overcome this problem, several investigators explored more compact and efficient models for representing the ring space: The "maximum proper covering set of rings" [28], the "synthetically significant rings" [27], the "chemically interesting rings" [166], and the "extended set of smallest rings" [37, 38] are just a few of the empirical models proposed in the chemical literature to minimize the use of computer resources while providing the ring information needed in various types of chemical applications.

Quite interestingly, all of these models essentially describe small supersets of the "smallest set of smallest rings" originally used in *The Ring Index* [110] and by the Wiswesser line notation [126]. The commonly accepted, although imprecise, definition of SSSR as "a basis set of rings which consists of the smallest rings that can form a basis set" [166] is equivalent to Balducci's abstract definition of a "minimal basis set of the ring space" [5] and, in most cases, is also equivalent to the graph-theoretical definition of the "fundamental cycles of a minimal spanning tree" [30] - although, in general, not every SSSR corresponds to a minimal spanning tree. The SSSR concept, by any definition, minimizes the usage of computer resources while providing an accurate description of the cyclic nature of the structure and, consequently, is the most widely accepted ring model for general purpose applications.

More extensive reviews on the subject of ring perception have been published by Gray [60] and by Downs *et al.* [37], revealing that virtually all the ring perception algorithms make use of some combination of the following two basic approaches:

(i) *graph-theoretical* techniques, using a graph representation of the structure, for sequential exploration and manipulation of nodes and edges walking along connected paths (breadth-first [30, 79] or depth-first [121] searches, graph reversals [52, 79] etc.), and

(ii) *linear-algebraic* techniques, using a matrix representation of the structure, for non-sequential exploration and manipulation of structural features (column or row exchanges to reorder the structural elements [73, 79], linear independence tests [77], etc.).

For complex ring systems, such as multiply-bridged systems and cage structures, a wide range of cycles can be perceived, starting from the minimum number necessary to include all vertices and edges to the maximum number of cycles in the ring system. This range varies in terms of the number, size and atom/bond composition of the cycles. Different applications have different requirements, and so a variety of ring sets (the particular sets chosen from within the wide range available) have been defined (see Chapter 4).

The problem then becomes one of choosing a ring set that is in some way “optimum” for the particular application. The main factor is usually that the ring set should be unique for a given structure and invariant, i. e., processing or ordering the graph in a different way should not produce a different set, or a choice between several sets. Given the large number of rings that could be included in a set, a general aim is to include the minimum number of rings necessary to describe the ring system and also to include sufficient rings to describe the ring system adequately for a given application.

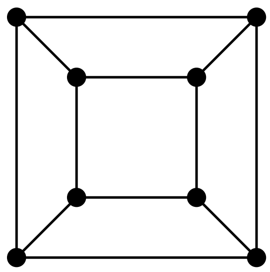


Figure 1.3. Cubane

For instance, in cubane, Fig. 1.3, there are 28 cycles, 14 chordless cycles, 6 chordless faces, 6 relevant cycles (see section 3.4), the dimension of the cycle space is 5 and all edges and vertices can be included by using just four of the relevant cycles.

1.2.2 Biopolymer Graphs

Biopolymers, such as RNA, DNA, or proteins form well-defined three dimensional structures. These are of utmost importance for their biological function. The most salient features of these structures are captured by their *contact graphs* which have the atoms of small molecules or the monomers of a biopolymer as their vertices, and edges that connect spatially adjacent objects. While this simplification of the 3D shape obviously neglects a wealth of structural details, it encapsulates the type of structural information that can be obtained by a variety of experimental and computational methods.

Biopolymers share a number of common features distinguishing them from other classes of the molecular contact graphs. In particular, they have a spanning path $\mathcal{T}$ corresponding to the covalent backbone. The remaining non-covalent bonds $B = E \setminus \mathcal{T}$ then determine the “fold” or three-dimensional structure of the molecule. Nucleic acids, both RNA and DNA, form a special type of contact structures known as *secondary structures*.

A particular type of cycles, which is commonly termed *loops* in the RNA literature, plays an important role for RNA (and DNA) secondary structures: the energy of a secondary structure can be computed as the sum of energy contributions of the *loops*. This secondary structures are outerplanar graphs $\mathcal{G}$; hence these *loops* form the unique minimum cycle basis $\mathcal{B}(\mathcal{G})$ of the contact graph [93]. Experimental energy parameters are available for the contribution of an individual *loop* as a function of its size, of the type of bonds that are contained in it, and on the monomers (nucleotides) that it is composed of [49, 99]. Based on this energy model it is possible to compute the secondary structure with minimal energy given the sequence of nucleotides using a dynamic programming technique [156, 171]. Two public domain program packages are available [75, 170] on the internet.

The RNA secondary structure prediction problem can be rephrased as minimizing the energy function $E(\mathcal{G}) = \sum_{C\in\mathcal{B}(\mathcal{G})} E(C)$ over the class of secondary structure graphs (i.e., the sub-cubic outerplanar graphs satisfying a few further restrictions, see e.g. [128, 156] for details).

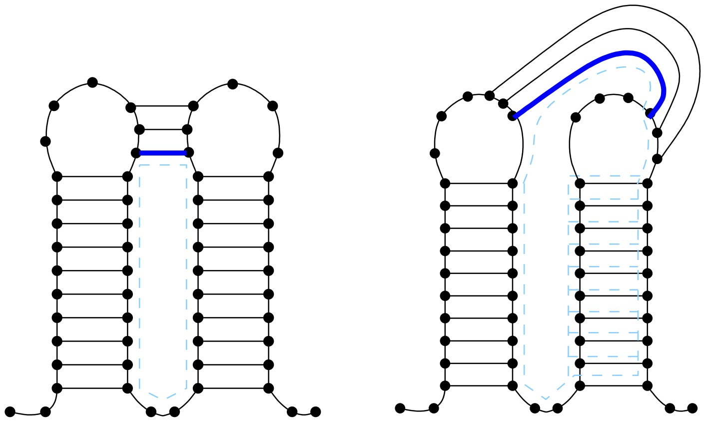


Figure 1.4. Two RNA secondary structures with a pseudo-knot. The only difference between the two structures is the exact location of the “middle stem” consisting of the three base pairs that connect the two hairpin loops. The energy contribution for the “pseudo-knot formation” should be attributed to the relevant cycle(s) associated with the “closing pair” of the “middle stem”, indicated by a thick line.
On the l.h.s. there is a unique relevant cycle (indicated by the dashed line) associated with the “closing pair” of the middle “stem”. In the example on the r.h.s. we find ten relevant cycles that differ by the ring sum of one or more of the 4-cycles of the rightmost “stem”. It seems natural therefore to associate an energy contribution not with an individual relevant cycle, but rather with an equivalence class of cycles, in this case with the class of equal length cycles indicated by the dashed lines on the r.h.s.

In recent years, however, there has been increasing evidence that so-called *pseudo-knots* play an important role, see e.g. [61]. These structural elements violate outerplanarity and — in the simplest case — lead to the *bisecondary structures* introduced in [128]. The minimum cycle basis is not unique for most graphs, including most non-trivial bisecondary structures, Figure 1.4. The set $\mathcal{R}$ of relevant cycles, i.e., the union of all minimum cycle bases [149] seems to be a good candidate for extending the energy model. However, as the example in Figure 1.4 shows, sometimes there is a large class of relevant cycles associated with what biophysically is a single structural element. It seems natural therefore to average over contributions of an equivalence class of cycles of the same length or to define the energy parameters in such a way that all cycles of this class contribute the same energy.

1.2.3 Reaction Graphs

Metabolic networks form a particular class of chemical reaction networks, i.e. the graph in Fig. 1.5 represents the reactions in the planetary atmosphere of the Jovian satellite Io (datas from the book [167]).

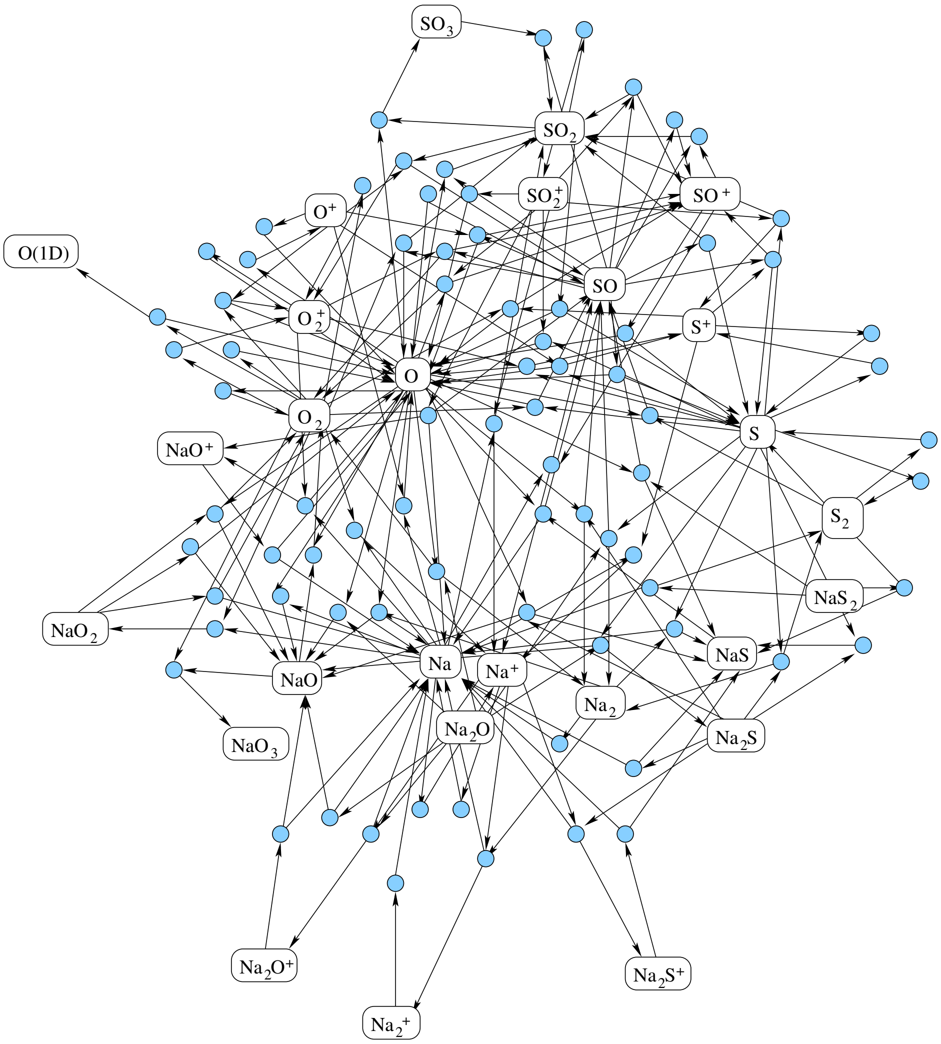


Figure 1.5. Chemical network of the planetary atmosphere of the Jovian satellite Io.

Because it is germane to the functional analysis of the metabolic networks, we first point out a nexus between graph representations of metabolic network, and metabolic flux analysis (MFA), the most generic framework to analyze the biological function of metabolic networks.

The key ingredient of MFA is the *stoichiometric matrix* $\mathbf{S}$. Its entries are the stoichiometric coefficients s_{kr}, i.e., the number of molecules of species k produced ($s_{kr} > 0$) or consumed ($s_{kr} < 0$) in each reaction r. Reversible reactions are entered as two separate reactions in most references. In general, additional "pseudo-reactions" are added to describe the interface of the metabolic reaction network with its environment.

The dynamics of the concentration of metabolite k may be generally described by

$$\frac{\mathrm{d}c_k}{\mathrm{d}t} = \sum_r s_{kr} J_r - \upsilon(t) c_k \tag{1.1}$$

where the flux J_r through reaction r depends on the kinetic properties of the participating enzymes, on the concentrations of metabolites and on environmental parameters such as temperature and pH. The enzymes are generally subject to complex regulations by inhibition and activation. The assumption of a steady state and neglecting the dilution fluxes $\upsilon(t)c_k$ as a consequence of low concentrations of intermediates yields the homogeneous, time-independent system of linear equations

$$\mathbf{S}J = \vec{o} \tag{1.2}$$

for the flux vector J. Consequently, the steady state flux vectors are elements of the null-space $\mathsf{Null}(\mathbf{S})$. Using the constraint that we have must have $J_r \geq 0$ for each reaction r, we see that J is a steady state flux vector if and only if

$$J \in \mathsf{Null}(\mathbf{S}) \cap \mathbb{R}_+^{|V|}. \tag{1.3}$$

The extremal rays of this cone are usually called the *elementary flux modes* and are closely associated with the relevant metabolic pathways, see e.g. [26, 41, 45, 72, 120, 122] for further details on MFA.

It is not hard to see that, if all reactions are mono-molecular, then $\mathbf{S}$ is the incidence matrix of a directed graph: $s_{kr} = 1$ for the single product k formed in reaction r and $s_{kr} = -1$ for the single metabolite used in reaction r, i.e., $\mathbf{S}$ is the incidence matrix of the digraph $\mathcal{G}$ whose vertices are the chemical species and whose edges denote the reactions. Such networks were studied already in the 1960s [3]. It is well known that x is an element of the cycle space of $\mathcal{G}$ if and only if $\mathbf{S}x = \vec{o}$, i.e., the circuit space of $\mathcal{G}$ is $\mathsf{Null}(\mathbf{S})$ [14]. The stationary flux vectors are therefore cycles of $\mathcal{G}$.

In general, $\mathbf{S}$ represents a *directed hypergraph* [169]. Equivalently, one may use a bipartite graph in which one class of vertices represents the substrates and the other class of vertices denotes the reactions. Arcs point from the educts to the reaction node and from the reaction node to the products, Fig. 1.6. A very simple graph representation of chemical networks, which is sufficient for our purposes, is the *substrate graph* Σ introduced in [154]. Its vertices are the molecular compounds (substrates); two substrates k and l are adjacent in Σ if they participate in the same reaction r. The substrate graph is a straight-forward approximation of the directed hypergraph representing $\mathbf{S}$: a directed hyper-edge is replaced by a clique on the same set of vertices. As a consequence, the stationary flux vectors are closely related to the cycles of the substrate graph.

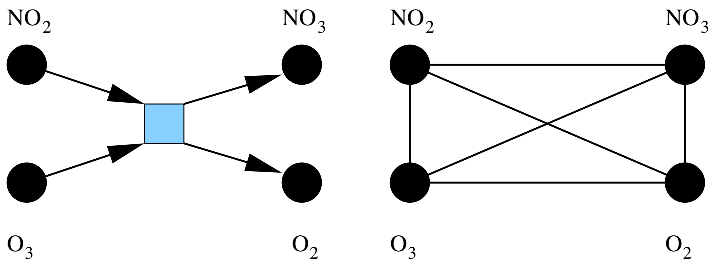


Figure 1.6. Representations of the reaction $NO_2 + O_3 \rightarrow NO_3 + O_2$ in hypergraph form drawn as the equivalent directed bipartite graph (l.h.s) and as part of a substrate graph (r.h.s).

The undirected substrate graphs are considered in some applications, because directed graphs would not properly represent the propagation of perturbations: even for irreversible reactions the product concentration may affect the reaction rate, for instance by product occupancy of the enzyme's active site; this in turn affects the substrate concentration. Thus, perturbations may travel backwards even from irreversible reactions. A similar argument for considering undirected graphs can be derived from metabolic control theory [123]. A number of more complicated graph representations for chemical reaction networks are discussed e.g. in the book [138].

The matrix $\mathbf{S}$ does not identify the input and output metabolites. This information is added in the form of additional "I/O-vertices" and "pseudo-reactions" representing flux in and out of the reaction network in MFA applications, see e.g. [45]. The corresponding extension of the cycle space of the network graph is the vector space spanned by all cycles of the reaction network and all paths connecting pairs of "I/O-vertices". The generalization of the notion of relevant cycles to this extended vector space was explored in a different context by Hartvigsen [66] and will be explored further in chapter 5.

1.3 Organization of the Thesis

In chapter 2, we give a brief introduction of the most important concepts of graph theory that will be used throughout this thesis. Furthermore, the relation between graphs, vector spaces and the matroid theory are explained, because some of our main results make use of these connections.

Our main interest concerns the cyclic structure of graphs, representing biopolymers on the one hand and chemical networks on the other hand, so chapter 3 deals with this topic. Naturally, short cycles are particularly useful for the purpose of the description of cycles structures. We will discuss the different types of cycles and their relation to each other.

Minimum cycle bases are of particular practical interest because they encapsulate the entire cycle space in concise manner. The energy model of RNA secondary structure is established on the cycles of the minimum cycle bases. For the most structures with pseudo-knots the minimum cycle basis is not unique. The union of all minimum cycle bases, on the other hand, does not really fit with biophysically structural elements. these shortcoming can be overcome by introducing a partition of this cycle set (chapter 6).

Till Vismara's thesis [148], no efficient algorithm for calculating the relevant cycles was known. Hence, a many investigators dealt with extended minimum cycles bases. We give a brief summary of what chemists call "ring sets" in chapter 4 and reveal some faults in the chemical literature.

When analyzing chemical networks [56] the cycles of the network graphs do not hold any informations about the input and output of the network. This information is added in the form of "I/O-vertices" and "pseudo-reactions". In chapter 5 we introduce the corresponding extension of the cycle space of the network graph and the generalization of the notion of relevant cycles to this extended vector space.

In metabolic flux analysis one is mostly concerned with the propagation of mass through the network. In this case directionality is crucial and directed graph models are required. In particular, the directed circuit bases produces our interest. Therefore, in chapter 7 we extended the cycle space of undirected graphs to a circuit space of directed graphs.

All algorithms presented here are implemented and part of a ANSI C++ program, which will be briefly discussed in chapter 8. Furthermore, this chapter gives some biochemical examples as applications of the mathematical concepts introduced before. Since, our starting motivation for these graph theoretical contribution arises from the search for a suitable energy model for RNA secondary structure computations in the presence of pseudo-knots, the results of chapter 6 - the partitioning of the set of relevant cycles - will be discussed on two well studied examples of RNAs with pseudo-knots: tmRNA and RNaseP RNA.

On the other hand chemical reaction networks found our interest. In section 1.2.3 we have briefly outlined the relationship between the cycle structure of a reaction network and the *Chemical Flux Analysis*. In section 8.3 the distribution of triangles and longer relevant cycles is discussed for uncorrelated random graphs as well as for small world models. In the following section, we compare two classes of chemical reaction networks here: (1) Metabolic networks in which all reactions are mediated by specific enzymes, and (2) the reaction networks of planetary atmospheres which lack specific catalysis. Surprisingly, their global structure is quite similar.

A discussion of our results and open problems concludes this work.

Chapter 2

Basic Definitions

2.1 Graphs

Intuitively speaking, a *graph* is a set of points, and a set of arrows, with each arrow joining one point to another. The points are called the *vertices* of the graph, and the arrows are called the *edges* of the graph.

2.1.1 Basic Definitions

A graph $\mathcal{G}$ is a pair (V, E) of a set of vertices V and a set of edges E together with two maps $i : E \mapsto V$ and $t : E \mapsto V$ assigning to each edge $e \in E$ its initial vertex $i(e) \in V$ and its terminal vertex $t(e) \in V$. The number n of vertices is the *order* of $\mathcal{G}$, and the number of edges is denoted by m.

A *loop* is an edge of the type (x, x). Two edges are called *parallel* or *multiple* edges, if they have common endpoints and are not loops. If this graph $\mathcal{G}$ does neither contain multiple edges nor loops, it is called a *simple graph.* For this work, a graph $\mathcal{G}$ is always a simple graph.

In graphs without multiple edges we can regard E simply as a subset $V \times V$ with edges being pairs of vertices $e = (i(e), t(e))$. We distinguish *undirected* graphs, where edges are non-ordered pairs of vertices and *directed* graphs where edges are considered as ordered pairs. More formally, a *digraph* $\mathcal{G}(V, A)$ consists of a finite non-empty set V of elements called vertices and a finite set A of distinct *ordered* pairs of distinct elements called *arcs.* A *simple digraph* is a digraph with no loops or multiple arcs.

The *degree* $\deg(v)$ of a vertex v is the number of edges with v as an endpoint, the *out-degree* $\mathrm{outdeg}(v)$ is the number of arcs of the form (v, w), and the *in-degree* $\mathrm{indeg}(v)$ is the number of arcs of the form (w, v).

A *subgraph* of $\mathcal{G} = (V, E)$ is a graph $\mathcal{H} = (V', E')$ such that $V' \subseteq V$ and $E' \subseteq E$. If $V' = V$, then $\mathcal{H}$ is a *spanning* subgraph of $\mathcal{G}$. If W is any set of vertices in $\mathcal{G}$, then the *subgraph induced by* W is the subgraph of $\mathcal{G}$ obtained by joining those pairs of vertices in W that are joined in $\mathcal{G}$. An *induced subgraph* $\mathcal{G}[W]$ of $\mathcal{G}$ is a subgraph that is induced by some subset W of V.

If $\mathcal{G}$ is a digraph, then the *underlying graph* $\mathcal{G}^\circ$ of $\mathcal{G}$ is the graph obtained from $\mathcal{G}$ by replacing each arc by an undirected edge joining the same pair of vertices and removing the multiple edges.

2.1.2 Paths, Circuits, Trees and Cuts

The following definitions can be used for both the directed and undirected graph, for the undirected graph we use edge instead of arc.

A *chain* in $\mathcal{G}$ is a sequence

$$\mathbf{c} = (x_0, e_1, x_1, e_2, x_2 \ldots, e_{q-1}, x_{q-1}, e_q, x_q) \tag{2.1}$$

of vertices and arcs such that x_k is an end-vertex of both the preceeding arc e_k and the succeeding arc e_{k+1}. The vertices x_0 and x_q are the initial and terminal vertex of the chain, respectively. A chain that does not encounter the same vertex twice is called *elementary*. If it does not contain the same arc twice it is called *simple*. An elementary chain is of course simple. The length $|\mathbf{c}|$ of a chain is the number q of its arcs. The concatenation $\mathbf{c}' * \mathbf{c}''$ of a chain $\mathbf{c}'$ with initial vertex x and terminal vertex x' and a chain $\mathbf{c}''$ with initial vertex x' and terminal vertex x'' is defined in the obvious way. Any chain can therefore be regarded as the concatenation of its individual steps (x_i, e_{i+1}, x_{i+1}).

A *walk* is a chain in which $e_k = (x_{k-1}, x_k) \in E$ for all k, i.e., in which each arc is traversed in forward direction. A *path* is a simple walk, which in the undirected graph is the same as a simple chain.

A simple chain is *closed*, if the endpoints are the same vertex. A close simple chain is called a *cycle*, a closed path is called a *circuit*, which in an undirected graph is the same as a cycle. A cycle or circuit C is *proper* if $(x, y) \in C$ implies $(y, x) \notin C$. Proper cycles therefore have length $|C| \geq 3$. A circuit of length 2 is also called a *double edge*.

A cycle C is *elementary* if each vertex has degree 2. A *(generalized) cycle* is an arc-disjoint union of elementary cycles. A *chord* of C is an edge $e = \{x, y\} \in E$ such that $e \notin C$, but both x and y are vertices of C. A cycle C is *simple* or *chordless* if it is elementary and has no chord. A cycle C is *tied* if it is elementary and has exactly one chord.

2.1.3 Distance

The *distance* $d(x, y)$ is the minimum length of a path connecting x and y. It satisfies

(D0) $d(x, y) = 0$ implies $x = y$.
(D1) $d(x, x) = 0$ for all $x \in V$.
(D2) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality) for all $x, y, z \in V$
(D3) $d(x, y) = d(y, x)$ for all $x, y \in V$.

(D3) holds in general only for undirected graphs.

A subgraph $\mathcal{H}$ of $\mathcal{G}$ is *isometric* if $d_H(x, y) = d_G(x, y)$ for all vertices $x, y \in V_H$.

A graph $\mathcal{G}$ is *connected* if there is a chain joining each pair of vertices of $\mathcal{G}$. Every disconnected graph can be split into maximal connected subgraphs called *components*. Similar definition can be given for digraphs.

A digraph $\mathcal{G}(V, A)$ is *weakly connected*, if the underlying undirected graph $\mathcal{G}^\circ(V, A^\circ)$ is connected.

A digraph $\mathcal{G}(V, A)$ is *strongly connected* if for all $x, y \in V$ there is a path from x to y and a path from y to x. It is well known that $\mathcal{G}(V, A)$ is strongly connected if and only each arc is contained in a circuit [10].

2.1.4 Trees and Cuts

A graph is said to be *acyclic* if it has no circuits. A *tree* is a connected acyclic graph. A *forest* is defined to be a graph whose connected components are trees.

A *spanning* tree of a graph $\mathcal{G}$ is a tree of $\mathcal{G}$ having all the vertices of $\mathcal{G}$. A *cospanning tree* $\mathcal{T}^*$ of a spanning tree $\mathcal{T}$ of a graph $\mathcal{G}$ is the subgraph of $\mathcal{G}$ having all the vertices of $\mathcal{G}$ and exactly those edges of $\mathcal{G}$ that are not in $\mathcal{T}$. The edges of a spanning tree $\mathcal{T}$ are called the *branches* of $\mathcal{T}$, and those of the corresponding cospanning tree $\mathcal{T}^*$ are called *links*.

An edge is a *cut edge*, if its removal disconnects the graph and increases the number of components; hence it cannot be part of a cycle. Similarly, a vertex is a *cut vertex*, if its removal increases the number of components. Obviously each vertex incident to a cut edge is a cut vertex, but a cut vertex can also be part of a cycle.

The *union* of two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, denoted as $\mathcal{G}_1 \cup \mathcal{G}_2$, is the graph $\mathcal{G}_3 = (V_1 \cup V_2, E_1 \cup E_2)$; that is, the vertex set of $\mathcal{G}_3$ is the union of V_1 and V_2, and the edge set of $\mathcal{G}_3$ is the union of E_1 and E_2.

The *intersection* of two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, denoted as $\mathcal{G}_1 \cap \mathcal{G}_2$, is the graph $\mathcal{G}_3 = (V_1 \cap V_2, E_1 \cap E_2)$; that is, the vertex set of $\mathcal{G}_3$ consists of only those vertices present in both $\mathcal{G}_1$ and $\mathcal{G}_2$, and the edge set of $\mathcal{G}_3$ consists of only those edges present in both $\mathcal{G}_1$ and $\mathcal{G}_2$.

The *ring sum* or *symmetric difference* of two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$, denoted as $\mathcal{G}_1 \oplus \mathcal{G}_2$, is the induced graph $\mathcal{G}_3$ on the edge set $E_1 \oplus E_2 = (E_1 - E_2) \cup (E_2 - E_1)$.

It can be verified easily that the three operations $(\cup, \cap, \oplus)$ defined above are associative and commutative.

2.2 Vector Spaces

Identifying the algebraic structure associated with a set of objects has been found to be very useful since the powerful and elegant results relating to the algebraic structure can then be brought to bear upon the study of such a set. We gave here a very brief introduction to some elementary algebraic concepts and results that will be used later are given. For more detailed discussions of these concepts and related results in linear algebra see [63, 76, 95].

2.2.1 Groups and Fields

Let $\mathcal{S}$ be a nonempty set with a binary operation denoted by $+$. Then $\mathcal{S}$ is called a *group* if the following axioms hold:

1. *Closure:* $(a+b) \in \mathcal{S}$ for all $a, b \in \mathcal{S}$.
2. *Associative law:* $a + (b + c) = (a + b) + c$ for all $a, b, c \in \mathcal{S}$.
3. *Identity element:* There exists a unique element e called *identity* in $\mathcal{S}$ such that $a + e = e + a = a$ for all a in $\mathcal{S}$.
4. *Inverses element:* For each element a in $\mathcal{S}$ there exists a unique element i called *inverse* such that $a + i = i + a = e$. Clearly the identity element e is its own inverse.

A group is said to be *abelian*, if the *commutative law* holds, i.e. if $a + b = b + a$ for all $a, b \in \mathcal{S}$.

A common example of a group is the set $\mathbb{Z}$ of all integers, with $+$ defined as the usual addition operation. Due to the missing inverse element, the set $\mathbb{Z}$ with the multiplication operation is not a group.

A set $\mathcal{F}$ with two operation $+$ and $\cdot$, called addition and multiplication, is a *field* if the following postulates are satisfied:

1. $\mathcal{F}$ is an abelian group under $+$, with the identity element e.
2. The set $\mathcal{F} - \{e\}$ is an abelian group under $\cdot$.

3. Distributive law of $\cdot$ over $+$: $a \cdot (b + c) = (a \cdot b) + (a \cdot c) = (b + c) \cdot a$ for all $a, b, c \in \mathcal{S}$.

As an example, consider the set $\mathbb{Z}_p = 0, 1, 2, ...p - 1$ of integers with modulo p addition and modulo p multiplication as the two operations. $\mathbb{Z}_p$ is an abelian group under modulo p addition, with 0 as the identity element. It can be shown that the set $\mathbb{Z}_p - 0$ is a group under modulo p multiplication if and only if p is prime. Also the fact that modulo p multiplication is distributive with respect to modulo p addition may be easily verified. Thus the set $\mathbb{Z}_p$ is a field if and only if p is prime. The field $\mathbb{Z}_p$ is usually denoted as GF(p) and is called a *Galois field*. A field that is of special interest is GF(2), the set of integers modulo 2.

2.2.2 Vector Space

A *vector space* over the field $\mathcal{F}$ with elements called *scalars* is a set $\mathcal{V}$ of elements called *vectors* together with an operation $\boxplus : \mathcal{V} \times \mathcal{V} \mapsto \mathcal{V}$ called addition and an operation $* : \mathcal{F} \times \mathcal{V} \mapsto \mathcal{V}$ called *scalar multiplication* such that

1. $\mathcal{V}$ is an abelian group under $\boxplus$.

2. $(a \boxplus b) * v = (a * v) \boxplus (b * v)$ and $a * (v \boxplus w) = (a * v) \boxplus (a * w)$ for all $a, b \in \mathcal{F}$ and for all $v, w \in \mathcal{V}$.

3. $(a \cdot b) * v = a * (b * v)$ for all $a, b \in \mathcal{F}$ and for all $v \in \mathcal{V}$.

4. $1 * s = s$ for all $s \in \mathcal{S}$ and 1 as the multiplicative identity in $\mathcal{F}$.

If an element v in $\mathcal{V}$ is expressible as $v = (a_1 * v_1) \boxplus (a_2 * v_2) \boxplus \cdots \boxplus (a_j * v_j)$, where v_i's are vectors and a_i's are scalars, the v is said to be a *linear combination* of $v_1, v_2, \ldots, v_j$. The elements $v_1, v_2, \ldots, v_j$ in a vector space are *linearly independent* if no vector in this set is expressible as a linear combination of the remaining vectors in the set, otherwise the vectors are *linearly dependent*.

2.2.3 Bases of Vector Spaces

Vectors $v_1, \ldots, v_k$ from a *basis* in the vector space $\mathfrak{V}$, if they are linearly independent and every vector in $\mathfrak{V}$ is expressible as a linear combination of these vectors, which are called *basis vectors*.

It can be shown that the representation of a vector as a linear combination of basis vectors is unique for a given basis. The following basic property of vector spaces will be used later.

Proposition 1. [24] *Let $\mathcal{B}$ be a basis of a vector space $\mathfrak{V}$. If any vector v in $\mathcal{B}$ is replaced by the sum of v and a linear combination of the vectors in $\mathcal{B} \setminus \{v\}$, then the resulting set of vectors is again a basis of $\mathfrak{V}$.*

A vector space may have more than one basis. However, it can be proved that all the bases have the same number of vectors, called the *dimension* of the vector space $\mathfrak{V}$, denoted as $\dim(\mathfrak{V})$.

2.2.4 Subspaces

If $\mathfrak{V}'$ is a subset of the vector space $\mathfrak{V}$ over $\mathcal{F}$, then $\mathfrak{V}'$ is a *subspace* of $\mathfrak{V}$ if $\mathfrak{V}'$ is also a vector space over $\mathcal{F}$.

The *direct sum* $\mathfrak{V}_1 \boxplus \mathfrak{V}_2$, of two subspaces $\mathfrak{V}_1$ and $\mathfrak{V}_2$ of $\mathfrak{V}$ is the set of all vectors of the form $v_1 \boxplus v_2$, where $v_1 \in \mathfrak{V}_1$ and $v_2 \in \mathfrak{V}_2$ and is again a subspace of $\mathfrak{V}$. The dimension is given by $\dim(\mathfrak{V}_1 \boxplus \mathfrak{V}_2) = \dim(\mathfrak{V}_1) + \dim(\mathfrak{V}_2) - \dim(\mathfrak{V}_1 \cap \mathfrak{V}_2)$.

Let $\mathfrak{V}$ and $\mathfrak{V}'$ be two n-dimensional vector spaces over a field $\mathcal{F}$. Then $\mathfrak{V}$ and $\mathfrak{V}'$ are said to be *isomorphic* if there exists a one-to-one correspondence between $\mathfrak{V}$ and $\mathfrak{V}'$ such that the following holds true.

1. If the vectors v_1 and v_2 of $\mathfrak{V}$ correspond to the vectors v_1' and v_2' of $\mathfrak{V}'$, then the vector $v_1 \boxplus v_2$ corresponds to the vector $v_1' \oplus v_2'$, where $\boxplus$ and $\oplus$ are corresponding operations in $\mathfrak{V}$ and $\mathfrak{V}'$.

2. For any α in $\mathcal{F}$, the vector $\alpha * s$ corresponds to the vector $\alpha \otimes s'$ if s corresponds to s', where $*$ and $\otimes$ are corresponding operations in $\mathfrak{V}$ and $\mathfrak{V}'$.

All vector spaces (over the same field $\mathcal{F}$) of the same dimension are isomorphic.

2.2.5 Vector Space over GF(2)

Proposition 2. [24] *Let $\{v_1, \ldots, v_k\}$ be a basis of a vector space $\mathfrak{V}$ over* GF(2) *and let $\{u_1, \ldots, u_k\}$ be another basis of $\mathfrak{V}$. Then, there exists a permutation Θ of $\{1, \ldots, k\}$ such that for $i = 1, \ldots, k$ each $u_{\Theta(i)}$ can be written as the sum of v_i and a linear combination of $\{v_1, \ldots, v_k\} \setminus \{v_i\}$.*

Define the *length* of a vector v over GF(2), denoted by $|v|$, to be the number of 1's that it contains. The *shortest basis* of a vector space $\mathfrak{V}$ is a basis $\mathcal{B}$ in which the sum of the lengths of all vectors in $\mathcal{B}$ is minimized. Chickering [24] showed that the length-distribution of the vectors equals for all shortest basis:

Proposition 3. [24] *Let $\{u_1, \ldots, u_k\}$ and $\{v_1, \ldots, u_k\}$ each be a shortest basis of a vector space over* GF(2) *having length $|u_1| \leq |u_2| \leq \cdots \leq |u_k|$, and $|v_1| \leq |v_2| \leq \cdots \leq |v_k|$, respectively. Then, for $i = 1, \ldots, k$, $|u_i| = |v_i|$.*

Let $L(\mathcal{B})$ denote the length of the longest vector in a basis $\mathcal{B}$. A basis of $\mathfrak{V}$ with minimum longest vector is a basis $\mathcal{B}'$ such that $L(\mathcal{B}')$ is minimized over all bases of $\mathfrak{V}$. Chickering further proved that every algorithm that finds a shortest basis also finds a basis with the minimum longest vector [24].

Proposition 4. [24] *Let $\mathcal{B}$ be a shortest basis of a vector space $\mathfrak{V}$ over* GF(2)*, $L(\mathcal{B})$ denote the length of the longest vector in $\mathcal{B}$ and let $\mathcal{B}'$ be a basis of $\mathfrak{V}$ with minimum longest vector with length $L(\mathcal{B}')$. Then, $L(\mathcal{B}) = L(\mathcal{B}')$.*

2.2.6 Vector Spaces on a Graph

A vector space can be associated with a graph $\mathcal{G}(V, E)$ in the following way [59, 124]:

Let $\mathcal{P}(E)$ denote the collection of all subsets of E, including the empty set $\emptyset$. It is easy to see that $\mathcal{P}(E)$ is an abelian group under $\oplus$, the ring sum operation ("exclusive or") between sets. Furthermore, for any $D \in \mathcal{P}(E)$, $D \oplus \emptyset = D$ and $D \oplus D = \emptyset$. The multiplication $*$ of an element of the field GF(2) and an element D of $\mathcal{P}(E)$ is defined as follows: For any $D \in \mathcal{P}(E)$, $1 * D = D$ and $0 * D = \emptyset$. One easily verifies that $(\mathcal{P}(E), \oplus, *)$ is a vector space, [59, 124].

If $E = e_1, e_2, \ldots, e_m$, then the subsets $\{e_1\}, \{e_2\}, \ldots, \{e_m\}$ will constitute a basis for $\mathcal{P}(E)$. Hence the dimension of $\mathcal{P}(E)$ is equal to m, the number of edges in $\mathcal{G}$.

Since each edge-induced subgraph of $\mathcal{G}$ corresponds to a unique subset of E, and by definition the ring sum of any two edge-induced subgraphs corresponds to the ring sum of their corresponding edge sets, it is clear that the set of all edge-induced subgraphs of $\mathcal{G}$ is also a vector space over GF(2), if the multiplication operation $*$ is defined as follows: For any edge-induced subgraph $\mathcal{G}_i$ of $\mathcal{G}$, $1 * \mathcal{G}_i = \mathcal{G}_i$ and $0 * \mathcal{G}_i = \emptyset$, the null graph having no vertices and no edges. This vector space will also be referred to by the symbol $\mathcal{P}(E)$.

Proposition 5. [141] *For a graph $\mathcal{G}$ with m edges $\mathcal{P}(E)$ is an $|E|$-dimensional vector space over* GF(2)*.*

The following subset of $\mathcal{P}(E)$ is also a subspace of $\mathcal{P}(E)$ and there for also a vector space over GF(2):

Proposition 6. [141] *$\mathcal{C}$, the set of all cycles (including the null graph $\emptyset$) and unions of edge-disjoint cycles of $\mathcal{G}$, is a subspace of the vector space $\mathcal{P}(E)$ of $\mathcal{G}$.*

$\mathcal{C}$ will be referred to as the *cycle space* of the graph $\mathcal{G}$.

Proposition 7. [141] *The dimension of the cycle space of $\mathcal{G}$ is equal to $|E|-|V|+k = \nu(\mathcal{G})$, the nullity or the cyclomatic number or first Betti number of $\mathcal{G}$, where k is the number of components.*

It is obvious that the cycle space of a graph is the direct sum of the cycle spaces of its 2-connected components. It will be sufficient therefore to consider only 2-connected graphs throughout of this work.

In chapter 7 we discuss the circuit space of a digraph as a vector space over $\mathbb{R}$, the propositions 5 and 7 remain valid.

2.3 Matroids

Matroids were introduced by Whitney [163] in 1935, with the aim of capturing the fundamental properties of dependence that are common to graphs and matrices. A matroid consists of a collection of subsets of a finite set which, loosely speaking, behave like a finite collection of vectors. Matroids also arise naturally from matrices and projective geometries. A matroid may be defined in many different but equivalent ways, several of which were described in Whitney's original paper.

Definition 8 (Independence Axioms). [163] *A matroid $\mathcal{M}$ is a finite set $\mathcal{S}$ and a collection $\mathscr{I}$ of subsets of $\mathcal{S}$ (called* independent *sets), such that (I1)-(I3) are satisfied.*

(I1) $\emptyset \in \mathscr{I}$.

(I2) *If $X \in \mathscr{I}$ and $Y \subseteq X$ then $Y \in \mathscr{I}$.*

(I3) *If $X, Y \in \mathscr{I}$ with $|X| = |Y| + 1$ there exists $x \in X \setminus Y$ such that $Y \cup x \in \mathscr{I}$.*

A *basis* of $\mathcal{M}$ is a maximal independent subset of $\mathcal{S}$, the collection of the bases is denoted by $\mathscr{B}(\mathcal{M})$ or simply $\mathscr{B}$.

A subset of $\mathcal{S}$ not belonging to $\mathscr{I}$ is called *dependent*.

The *rank function* ρ of $\mathcal{M}$ is a function: $\rho : 2^{\mathcal{S}} \mapsto \mathbb{Z}$, defined by

$$\rho(A) = \max\left(|X| : X \subseteq A, X \in \mathscr{I}\right).$$

The *rank* of the matroid is the rank of the set $\mathcal{S}$.

A subset $A \subseteq \mathcal{S}$ is *closed* or a *flat* or a *subspace* of the matroid $\mathcal{M}$, if for all $x \in \mathcal{S} \setminus A$ hold $\rho(A \cup x) = \rho(A) + 1$. In other words no element can be added to A without increasing its rank.

A *circuit* of $\mathcal{M}$ is defined as a minimal dependent subset of $\mathcal{S}$. The collection of circuits is denoted by $\mathscr{C}(\mathcal{M})$ or $\mathscr{C}$.

The *dual* $\mathcal{M}^*$ of a matroid $\mathcal{M}$ is defined as followed:

Proposition 9. [163] *Let $\mathscr{B}$ be the set of bases of a matroid $\mathcal{M}$ on a set $\mathcal{S}$. Then $\mathscr{B}^* = \{\mathcal{S} \setminus \mathcal{B} \mid \mathcal{B} \in \mathscr{B}\}$ is the set of bases of the dual matroid $\mathcal{M}^*$.*

The *cocircuit* of $\mathcal{M}$ is a circuit of $\mathcal{M}^*$.

The knowledge of the bases or circuits or rank function is sufficient to uniquely determine the matroid. Hence it is not surprising that there exists axiom systems for a matroid in terms of each of these concepts:

Proposition 10 (Basis axioms). [163] *A non-empty collection $\mathscr{B}$ of subsets of $\mathcal{S}$ is the set of bases of a matroid on $\mathcal{S}$ if and only if it satisfies the following condition:*

(B) *If $\mathcal{B}_1, \mathcal{B}_2 \in \mathscr{B}$ and $x \in \mathcal{B}_1 \setminus \mathcal{B}_2$, $\exists\, y \in \mathcal{B}_2 \setminus \mathcal{B}_1$ such that $(\mathcal{B}_1 \cup y) \setminus x \in \mathscr{B}$.*

Proposition 11 (Circuit axioms). [145] *A collection $\mathscr{C}$ of subsets of $\mathcal{S}$ is the set of circuits of a matroid on $\mathcal{S}$ if and only if condition (C1) and (C2) are satisfied.*

(C1) *If $X \neq Y \in \mathscr{C}$, then $X \nsubseteq Y$.*

(C2) *If C_1, C_2 are distinct members of $\mathscr{C}$ and $z \in C_1 \cap C_2$, there exists $C_3 \in \mathscr{C}$ such that $C_3 \subseteq (C_1 \cup C_2) \setminus z$.*

Two matroids $\mathcal{M}_1$ and $\mathcal{M}_2$ on $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively are *isomorphic* - denoted by $\mathcal{M}_1 \backsimeq \mathcal{M}_2$ if there is a bijection $\phi : \mathcal{S}_1 \rightarrow \mathcal{S}_2$ which preserves independence. It is clear that equivalently ϕ is an isomorphism if and only if it preserves the rank function, circuits and so on.

2.3.1 Independent Sets, Bases and Circuits

It is clear that if a subset A is independent there exists a basis $\mathcal{B}$ such that $A \subseteq \mathcal{B}$. The following stronger result is used extensively.

Proposition 12 (Augmentation Theorem). *Suppose that X, Y are independent in $\mathcal{M}$ and that $|X| < |Y|$. Then there exists $Z \subseteq Y \setminus X$ such that $|X \cup Z| = |Y|$ and $X \cup Z$ is independent in $\mathcal{M}$.*

An immediate consequence of this is the following result, which extends the well known property of bases of a vector space.

Corollary 13. *All bases of a matroid on $\mathcal{S}$ have the same cardinality, which is the rank of $\mathcal{S}$.*

For graph theorist the most natural way to define a matroid is by its circuit axioms (theorem 11). This is the approach used by Tutte [145].

Proposition 14. [145] *If A is independent in $\mathcal{M}$, then for $x \in \mathcal{S}, A \cup x$ contains at most one circuit.*

Corollary 15. [145] *If $\mathcal{B}$ is a basis of $\mathcal{M}$ and $x \in \mathcal{S}$ then there exists a unique circuit $C = C(x, \mathcal{B})$ such that $X \in C \subseteq \mathcal{B} \cup x$.*

This circuit $C(x, \mathcal{B})$ is called the fundamental circuit of x in the basis $\mathcal{B}$. In fact there is a stronger result.

Proposition 16. [145] *If $\mathcal{M}$ is a matroid on $\mathcal{S}$, and $\mathcal{B}$ is a basis of $\mathcal{M}$, then for any $x \in \mathcal{S} \setminus \mathcal{B}$, $(\mathcal{B} \setminus y) \cup x$ is a basis of $\mathcal{M}$ if and only if $y \in C(x, \mathcal{B})$ or $y = x$.*

A much stronger statement than theorem 11 (C2) can be made about the circuits of a matroid:

Proposition 17. [145] *If C_1, C_2 are distinct circuits of a matroid $\mathcal{M}$ and $x \in C_1 \cap C_2$, then for any element y of $C_1 \setminus C_2$ there exists a circuit C such that $y \in C \subseteq (C_1 \cup C_2) \setminus x$.*

Whitney [163] used the following condition (C3) and theorem 11 (C1) as his circuit axioms, where (C3) is what is sometimes known as the *strong circuit axiom*. The equivalence of these with the apparently weaker (C1) and (C2) was proved by Lehman [92].

(C3) *If C_1, C_2 are distinct members of $\mathscr{C}$ and $y \in C_1 \setminus C_2$ then for each $x \in C_1 \cap C_2$, there exists $C_3 \in \mathscr{C}$ such that $y \in C_3 \subseteq (C_1 \cup C_2) \setminus x$.*

The following two results of Tutte [145] are very useful.

Proposition 18. [145] *Let $\mathscr{D}$ be a collection of non-null subsets of $\mathcal{S}$ such that for any two distinct members X, Y of $\mathscr{D}$ such that $x \in X \cup Y, y \in X \setminus Y$, there exists $Z \in \mathscr{D}$ such that $y \in Z \subseteq (X \cup Y) \setminus x$.*

Proposition 19. [145] *Let $\mathscr{D}$ be a family of subsets satisfying the hypotheses of theorem 18. Then if $a \in W \in \mathscr{D}$ there exists $V \in \mathscr{D}$ such that $a \in V \subseteq W$.*

2.3.2 The Cycle Matroid of a Graph

Let $\mathcal{G}$ be an undirected graph with vertex set V and edge set E.

Proposition 20. *If $\mathcal{G}$ is a graph, the cycles of $\mathcal{G}$ are the circuits of a matroid $\mathcal{M}(\mathcal{G})$ on the edge set E.*

This matroid $\mathcal{M}(\mathcal{G})$ is call the *cycle matroid* of $\mathcal{G}$ (or in Tutte's work [145] the *polygon matroid* of $\mathcal{G}$).

By definition a maximal subgraph of $\mathcal{G}$ which contains no cycle is a spanning forest. Hence following basic properties of $\mathcal{M}(\mathcal{G})$ can be listed:

1. If $\mathcal{G}$ is disconnected, the bases of $\mathcal{M}(\mathcal{G})$ are the spanning forests of $\mathcal{G}$.

2. If $\mathcal{G}$ is connected, the bases of $\mathcal{M}(\mathcal{G})$ are the spanning trees of $\mathcal{G}$.

3. A set X of edges of $\mathcal{G}$ is independent in $\mathcal{M}(\mathcal{G})$, if and only if X contains no cycle, that is, iff X is a forest.

4. The rank of the matroid $\mathcal{M}(\mathcal{G})$ is $|V| - k$, where k is the number of connected components of $\mathcal{G}$.

5. For any subset $A \subseteq E$ the rank of A in $\mathcal{M}(\mathcal{G})$ is given by $\rho(A) = |V_A| - k_A$, where k_A is the number of components in the subgraph generated by A.

However despite the power of matroid theory as a tool in the clarification of certain graphical ideas, many problems of graph theory cannot even be posed in matroid language. Crudely this is because there is no simple exact counterpart of a vertex in a matroid. Also, non-isomorphic graphs may have isomorphic cycle matroids (see Fig. 2.1).

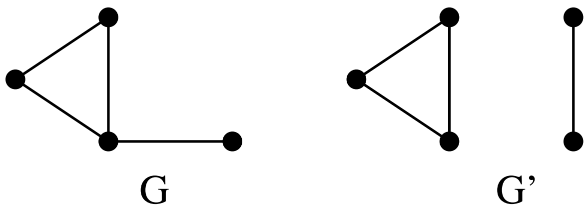


Figure 2.1. The two graphs $\mathcal{G}$ and $\mathcal{G}'$ are clearly non-isomorphic, however it is easy to check that the map $x \rightarrow x'$ is an isomorphism between $\mathcal{M}$ and $\mathcal{M}'$.

A matroid $\mathcal{M}$ on $\mathcal{S}$ is called *graphic*, if it is isomorphic to the cycle matroid of a graph $\mathcal{G}$.

2.3.3 Binary Matroids

Whitney was the first who considered the problem of finding necessary and sufficient conditions for a matroid to be graphic. An obvious starting point is to find properties of graphic matroids which do not hold for matroids in general. A definitive solution of the problem of "when is a matroid graphic" was given by Tutte [144], who introduced the concept of a *binary* matroid as a matroid determined by a chain group over the field GF(2) of integers modulo 2.

It turns out that there are many equivalent ways of defining a binary matroid. In view of the adjective "binary" probably the most appropriate is to define $\mathcal{M}$ to be binary if it is representable over the field GF(2).

Proposition 21. [145] *A matroid $\mathcal{M}$ is binary if and only if its dual $\mathcal{M}^*$ is binary.*

Proposition 22. [92, 102, 145] *The following statements about a matroid $\mathcal{M}$ are equivalent.*

(i) *For any circuit C and cocircuit C^*, $|C \cap C^*|$ is even.*

(ii) *The symmetric difference of any collection of distinct circuits of $\mathcal{M}$ is the union of disjoint circuits of $\mathcal{M}$.*

(iii) *If C_1, C_2 are distinct circuits of $\mathcal{M}$, the symmetric difference $C_1 \oplus C_2$ contains a circuit C.*

(iv) *For any basis $\mathcal{B}$ and circuit C of $\mathcal{M}$, if $C \setminus \mathcal{B} = \{e_1, \ldots, e_i\}$ and if $C_i = C(e_i, \mathcal{B})$ is the fundamental circuit of e_i in the basis $\mathcal{B}$, then $C = C_1 \oplus \cdots \oplus C_i$.*

(v) *$\mathcal{M}$ is binary.*

Let $\mathcal{M}$ be a binary matroid on $\mathcal{S}$, $|\mathcal{S}| = n$ and let $\mathfrak{V}$ be the vector space of rank n over GF(2). The circuit space of $\mathcal{M}$ is the subspace of $\mathfrak{V}$ generated by the incidence vectors of the circuits of $\mathcal{M}$. Similarly the cocircuit space of $\mathcal{M}$ is the subspace of $\mathfrak{V}$ generated by the cocircuits of $\mathcal{M}$.

Proposition 23. [102] *The rank of the circuit space $\mathscr{C}(\mathcal{M})$ of the binary matroid $\mathcal{M}$ is $\rho(\mathcal{M}^*)$ and the incidence vectors of the set of fundamental circuits of any basis of $\mathcal{M}$ form a basis of the circuit space.*

Definition 24. *An element C of the circuit space of $\mathcal{M}$ is* elementary, *if it coincides with the incidence vector of a fundamental circuit of some basis of $\mathcal{M}$.*

For graphs, these elementary elements of the circuit vector space $\mathscr{C}(\mathcal{M}(\mathcal{G}))$ of the cycle matroid $\mathcal{M}(\mathcal{G})$ of $\mathcal{G}$ are exactly the elementary cycles.

2.3.4 The Greedy Algorithm

The best-known algorithmic property of matroids is their intimate relationship with what has been termed the “greedy algorithm”. Loosely speaking the greedy algorithm makes maximum improvements in an objective function at each stage and never back tracks.

The basic idea for graphs is a will known result of Kruskal [89]. The extension to matroids was first carried out by Rado [116], but Edmonds [40] and Welsh [160] independently also developed an algorithm for the formation of a minimal basis of a matroid, known as the above mentioned greedy algorithm.

Consider a set $\mathcal{S}$ whose elements s_i have been assigned nonnegative weights $w(e_i)$. The weight of a subset of $\mathcal{S}$ is defined as equal to the sum of the weights of all the elements in the subset. Let $\mathscr{I}$ be a collection of subsets of $\mathcal{S}$.

The problem is to find a subset X_{opt} of $\mathcal{S}$ such that $X_{opt} \in \mathscr{I}$ is minimum (or maximum) over all elements of $\mathcal{S}$. Just call this *problem* $(\mathscr{I}, w)$. The *greedy algorithm* for the problem $(\mathscr{I}, w)$ is an automatic routine for selecting a minimal neighbour of $\mathscr{I}$. It proceeds as follows:

Algorithm 1 The Greedy Algorithm

Sort $\mathcal{S}$ by weight such that $w(x_i) \leq w(x_j)$ for $i < j$.
$\mathcal{B} \leftarrow \emptyset$
for all $k = 1, \ldots, |\mathcal{S}|$ **do**
  if $\mathcal{B} \cup \{x_k\} \in \mathcal{J}$ **then**
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{x_k\}$

Proposition 25. *Let $\mathscr{I}$ be a collection of subsets of $\mathcal{S}$ with the property that $A \in \mathscr{I}, B \subseteq A \Rightarrow B \in \mathscr{I}$. Then the greedy algorithm works for $(\mathscr{I}, w)$ for all non-negative weight functions w only if $\mathscr{I}$ is the collection of independent sets of a matroid on $\mathcal{S}$.*

Notice that proposition 25 gives the following useful characterization of matroids.

Proposition 26. *A non-empty collection $\mathscr{I}$ of subsets $\mathcal{S}$ is the set of independent sets of a matroid on $\mathcal{S}$ if and only if*

(i) $X \in \mathscr{I}, Y \subseteq X \Rightarrow Y \in \mathscr{I}$,

(ii) for all non-negative weight functions $w : \mathcal{S} \to \mathbb{R}_0^+$, the greedy algorithm selects a member A of $\mathscr{I}$ with

$$\sum_{e \in A} w(e) \geq \sum_{e \in B} w(e)$$

for all members B of $\mathscr{I}$.

Chapter 3

Cycle Bases of Undirected Graphs

To each cycle C in a graph $\mathcal{G}$ we associate an incidence vector x, where $x_e = 1$ if e is an edge of C and $x_e = 0$ otherwise. The vector space over GF(2) generated by these incidence vectors of cycles is called the *cycle space* of $\mathcal{G}$ (see chapter 2.2.6) with the dimension ν. A collection of cycles whose incidence vectors forms a basis for the cycle space of a graph is called a *cycle basis* [29].

A cycle basis is used to examine the cyclic structure of a graph. For example, many algorithms use cycle bases to list all simple cycles in a graph [98, 134] or look for the longest cycle [33]. Cycle bases have been also used to solve electrical networks since the time of Kirchhoff [25]. Brief surveys and extensive references can be found in [70, 77].

Proposition 27. *If $\mathcal{B}$ is a cycle basis for a graph, C is a cycle in $\mathcal{B}$, and $C = C_1 \oplus C_2$, then either $\mathcal{B} \setminus \{C\} \cup \{C_1\}$ is a cycle basis or $\mathcal{B} \setminus \{C\} \cup \{C_2\}$ is a cycle basis.*

3.1 Fundamental Cycle Bases

A fundamental cycle set is used by the organic chemists interested in the coding of ring compounds (see Chapter 1.2.1).

Whitney [163] introduced the following definition (in the more general setting of matroids).

Definition 28. [163] *Let $\mathcal{G}$ be a graph and let ν be the dimension of its cycle space. Then a collection of cycles $\mathscr{P}$ in $\mathcal{G}$, where $|\mathscr{P}| = \nu$, is called* fundamental*, if there exists an ordering of the cycles in $\mathscr{P}$ such that*

$$C_j \setminus (C_1 \cup \cdots \cup C_{j-1}) \neq \emptyset \text{ for } 2 \leq j \leq \nu. \tag{3.1}$$

If every ordering is fundamental, $\mathscr{P}$ is called strictly fundamental.

A well-known special case of a fundamental collection $\mathscr{P}$ of cycles for a graph $\mathcal{G} = (V, E)$ is the following definition, introduced by Kirchhoff in 1847:

Definition 29. [87] *Let $\mathcal{T} = (V, E')$ be a maximal spanning forest for $\mathcal{G}$ and let the cycles in $\mathscr{P}$ consist of the unique cycles in $e \cup E'$ for each $e \in E \setminus E'$. $\mathscr{P}$ is called a* Kirchhoff *cycle basis.*

These Kirchhoff cycle bases are exactly the cycle bases corresponding to the fundamental circuits of the matroid $\mathcal{M}(\mathcal{G})$.

Sysło [133] proved that a cycle basis is Kirchhoff iff each cycle in the basis contains an edges that is in no other cycle of this basis. Hartvigsen [67] showed:

Proposition 30. [67] *A cycle basis $\mathscr{P}$ is strictly fundamental iff it is Kirchhoff.*

It is a simple observation that a fundamental collection of cycles for a graph $\mathcal{G}$ is a cycle basis for $\mathcal{G}$, but not every cycle basis is strictly fundamental (see Fig. 3.1). Furthermore the graph in Fig. 3.1 shows not every graph has a strictly fundamental cycle basis. Hartvigsen *et al.* [71] gave a characterization of those graphs for which every cycle basis is fundamental. An example of a non-fundamental minimum cycle basis is discussed in section 4.2 (Fig. 4.1).

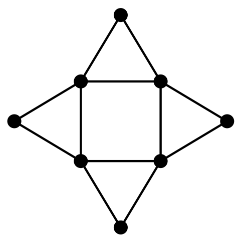


Figure 3.1. The minimum cycle basis of this planar graph consists of the three triangles C_1 through C_4 and the central square C_5. The ordering $(C_1, C_2, C_3, C_4, C_5)$ does not satisfy the equ.(3.1): in fact, $C_5 \subseteq C_1 \cup C_2 \cup C_3 \cup C_4 = E$. Alternatively one easily checks directly that no spanning tree generates the minimum cycle basis (from [78]).

3.2 Isometric, Short, Shortest, Edge-Short Cycles

Definition 31. *A cycle C is called* short*, if for every pair of nodes u and v in C a shortest path from u to v or a shortest path from v to u in $\mathcal{G}$ is contained in C.*

A cycle C is called isometric*, if for any two of its vertices u and v, it contains a shortest path from u to v* and *a shortest path from v to u.*

A cycle C in a graph $\mathcal{G}$ is called edge-short*, if $\mathcal{G}$ contains a node w, an edge $e = (u, v)$, a shortest path from u to w and a shortest path from v to w such that C is the edge disjoint union of e and the two paths.*

A cycle C is strictly edge-short*, if for each vertex x of C there is an edge $e = (u, v) \in C$, such that C consists of e, a shortest path from x to u and a shortest path from v to x.*

These definitions for undirected graphs can be used in digraphs as well. One simple has to replace the word “edge” by “arc”, see section 7.5.

Theorem 32. *In an undirected graph a cycle is isometric if and only if it is short.*

Proof. From the definition follows directly that an isometric cycle is short. Suppose C is short. We know that for every pair of vertex $x \neq y, x, y \in C$ there exists a shortest path from x to y or in the other direction. In undirected graphs a shortest path P from x to y is also a shortest path from y to x. Thus C is isometric. □

Theorem 33. *Every strictly edge-short cycle is edge-short. A cycle is short if and only if it is strictly edge-short.*

Proof. From the definition follows directly that a strictly edge-short cycle is edge-short. Horton [77] showed that every short cycle is also strictly edge-short. Thus it remains to show that strictly edge-short and short is equivalent.

For two distinct vertices $x \neq y$ in C, we denote the path from x to y in C by C' and C''. Furthermore we write S_{xy} for a path from x to y in $\mathcal{G}$ that is shorter than C' or C''. We call S_{xy} a shortcut from x to y.

Suppose C is not short. We show that C is not strictly edge-short. If C is not short then there are two vertices $x \neq y$ such that there exists a shortcut S_{xy} from x to y. Then it is impossible to find an edge $e = (u, v) \in C$ such that there is neither a shortcut S_{xu} nor a shortcut S_{xv}. Suppose $u \in C'$, such that the path P from x to u is contained in C'. Then the shortest path from x to v must contain S_{xy}. Hence C is not strictly edge-short. □

The converse that every edge-short cycle is also strictly edge-short is not true, Fig. 3.2 gives a counterexample.

Definition 34. *A cycle C is* shortest *in $\mathcal{G}$ if there is an edge $e = (x, y) \in C$ such that C is a shortest cycle containing e, i.e., $C = e \cup P^e_{xy}$ where P^e_{xy} is a shortest path in $\mathcal{G} \setminus e$, the graph obtained from $\mathcal{G}$ by deleting the edge e. The* set of all shortest cycles *is denoted by $\mathcal{S}(\mathcal{G})$ or just $\mathcal{S}$.*

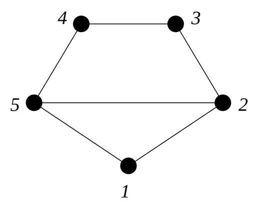


Figure 3.2. Consider the edge-short pentagon C. For vertex 2, it is impossible to find an edge e, such that the shortest paths from both endpoints of e are contained in C. Besides C is not chordless.

However, as the graph in Fig. 3.4 shows, the shortest cycles do not convey the complete information about the graph. The hexagon (bold edges in Fig. 3.4) cannot be reconstructed from the collection of triangles, but determines the diameter (maximal distance between two vertices) of the graph. It is crucial for the network structure since the local information conveyed by the twelve triangles does not allow the reconstruction of the hexagonal overall-structure.

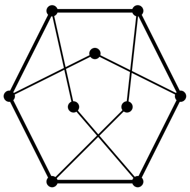


Figure 3.3. The hexagon (bold) is a chordless cycle, but it is impossible to find a vertex x and an edge e such that the shortest oaths from both endpoints of e are contained in the hexagon.

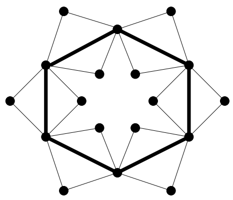


Figure 3.4. The hexagon (bold lines) is a short cycle, but is is not a shortest cycle through any of its edges (from [56]).

The diagram in Fig. 3.5 shows the relationships between the different types of cycles, in undirected graphs. Below the one-sided arrows, standing for one-sided implications, the number of the figure of counterexample for the inversion is given. For the definition of relevant and essential cycles see section 3.4 and 3.6.

$$
\begin{tikzcd}
\textit{unique shortest} \arrow[r, "\mathit{Fig.}\ 3.1"'] \arrow[d, "\mathit{Fig.}\ 3.10"'] & \textit{essential} \arrow[d, "\mathit{Fig.}\ 3.10"'] & \\
\textit{shortest} \arrow[r, "\mathit{Fig.}\ 3.4"'] & \textit{relevant} \arrow[d, "\mathit{Fig.}\ 3.6"'] & \\
\textit{strictly edge}-\textit{short} \arrow[r, leftrightarrow] & \textit{short} \arrow[r, leftrightarrow] \arrow[d, "\mathit{Fig.}\ 3.2"'] & \textit{isometric} \\
& \textit{edge}-\textit{short} \arrow[d, dotted, no head, "\mathit{Fig.}\ 3.2"', "\mathit{Fig.}\ 3.3"] & \\
& \textit{chordless} &
\end{tikzcd}
$$

Figure 3.5. The relations between the different type of cycles. Two-sided arrows implies an iff relation. Below the one-sided arrow the number of the figure of the counterexample for the reverse implication is given. The dotted line stands for a "no implication" with the number of the counterexamples.

3.3 Minimum Cycle Bases

The length of a cycle is the number of its edges. In terms of the vector representation we have $|C| = \sum_a |C_a|$. For each collection $\mathcal{B}$ of cycles we define the length

$$\ell(\mathcal{B}) = \sum_{C \in \mathcal{B}} |C|. \tag{3.2}$$

A *minimum cycle basis* (MCB) is a cycle basis with minimum length.

The problem of finding an MCB was first considered by Stepanec [129] and Zykov [172]. They proposed an algorithm that generates an independent set of cycles which are minimum length cycles through various edges in the graph, i.e., a set of shortest cycles. Hubicka and Sysło [78] showed that this algorithm does not always produce an MCB. The shortest cycle through an edge is always in a minimum cycle basis [78]. Deo [30] showed that the problem of finding a strictly fundamental cycle basis of minimal length is $\mathcal{NP}$ complete, but the graph in Fig. 3.1 shows that not every fundamental cycle basis is minimal.

Finally in 1987, Horton [77] presented a polynomial time algorithm ($\mathcal{O}(|E|^3|V| = \mathcal{O}(d^3|V|^4)$ with d as maximum vertex degree) for finding an MCB in a non-negative edge weighted graph. In 1991, Hartvigsen and Mardon [67] gave an algorithm for finding an MCB in a simple planar graph in $\mathcal{O}(|V|^2 \log |V|)$ times. Balducci and Pearlman [6] presented an algorithm for finding an MCB with at most order $\mathcal{O}(d^3|V|^2 \log |V|)$, where d is the maximum vertex degree, but Vismara [148] showed that this performance estimate is wrong. In fact, the worst case performance is $\mathcal{O}(\mu(\mathcal{G})|E|^2|V|d)$.

The problem of finding an MCB for a graph is a matroid optimization problem over the set of cycles of the graph. If the edges weights were restricted to be non-negative, then we could state this equivalently as a matroid optimization problem over the entire cycle space since, in this case, an optimal solution would still consist of (elementary) cycles. In either case, an MCB can be found using the greedy algorithm. We restate this fact in the following form:

Proposition 35 (Matroid Property). *Let $\mathcal{Q}$ be a set of cycles containing a minimum cycle basis. Then a minimum cycle basis $\mathcal{M}$ can be extracted from $\mathcal{Q}$ by a* greedy procedure *in the following way: (i) Sort $\mathcal{Q}$ by cycle length and set $\mathcal{M} = \emptyset$. (ii) Traversing $\mathcal{Q}$ in the established order, set $\mathcal{M} \leftarrow \mathcal{M} \cup \{C\}$ whenever $\mathcal{M} \cup \{C\}$ is linearly independent.*

MCBs have the property that their longest cycle is at most as long as the longest cycle of any basis of the cycle space [24]. An MCB therefore contains the salient information about the cyclic structure of a graph in its most compressed form. It appears natural to consider the cyclic structure of a graph in terms of its MCBs.

In general, graphs do not have unique MCBs. In fact, the known classes of graphs with unique MCB have a very simple structure: they are outerplanar [93], Halin graphs [127], or certain series-parallel graphs [100]. However, the distribution of cycle sizes is the same in all MCBs of a graph $\mathcal{G}$. In fact, we can restate proposition 3 in the form:

Corollary 36. *Suppose $\mathcal{M}$ is an MCB of $\mathcal{G}$ containing n_k cycles of length k. Then every MCB of $\mathcal{G}$ has exactly n_k cycles of length k.*

David Hartvigsen and Russel Mardon considered the minimum cycle basis problem for graph with *perturbed edge weights* $w(e)$, $e \in E$ which are chosen such that any two distinct edge-(multi)sets have different total weights [68, 69]. In this setting the MCB becomes unique for all graphs. Translated to unweighted graphs, this means that no two minimum cycle bases contain all edges in the same number of cycles. Hence, given a minimum cycle basis $\mathcal{M}$ of $\mathcal{G}$, there is a perturbed edge weighting such that $\mathcal{M}$ is the unique MCB of the edge-weighted version.

3.4 Relevant Cycles

The main shortcoming of minimum cycle bases for the characterization of graphs is the fact that the minimum cycle basis is not unique in general. A natural way to avoid ambiguities is to consider the union of all minimum cycles bases.

Definition 37. [115] *A cycle C is* relevant *if it cannot be represented as an $\oplus$-sum of shorter cycles.*

We write $\mathcal{R}$ or $\mathcal{R}(\mathcal{G})$ for the set of relevant cycles.

Proposition 38. [149] *A cycle C is relevant if and only if it is contained in a minimum cycle basis.*

Proposition 39. [129, 172] *If C is shortest, then it is relevant. Moreover, any minimum cycle basis must contain some shortest cycle through e for each $e \in E$.*

As an immediate consequence of Proposition 39 the shortest cycles through an edge are relevant, i.e., $\mathcal{S} \subseteq \mathcal{R}$.

Horton [77] showed that the cycles in a minimum cycle basis satisfy a very simple necessary condition. This allows one to polynomially compile a list of cycles that must contain a minimum cycle basis. Then a solution can be extracted from this list with the greedy algorithm.

Proposition 40. [77] *If C is contained in an MCB then it is short.*

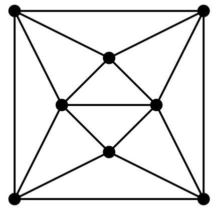


Figure 3.6. A counterexample to the converse of Proposition 40 (from [77]).

Note that not all short cycles belong to a minimum cycle basis. A counterexample from [77] is given in Fig. 3.6. The minimum cycle basis consists of all ten triangles. However, the quadrangle satisfy the condition of Theorem 40, nevertheless it is not contained in any minimum cycle basis.

Vismara [148, 149] proposed an algorithm for computing $\mathcal{R}$ that works by first extracting so-called prototypes (see below) from a set of short cycles similar to Horton's algorithm for finding *one* MCB [77]. The computation of the prototypes requires $\mathcal{O}(d^4|V|^4)$ operations, where d is the maximum vertex degree. The set $\mathcal{R}$ is then obtained by a backtracking procedure from the prototypes with $\mathcal{O}(|V|\,|\mathcal{R}|)$ operations.

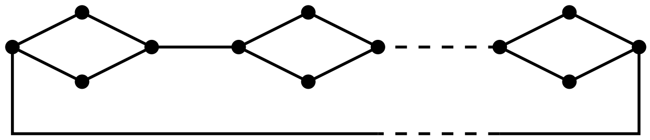


Figure 3.7. The set of relevant cycles of this graph contains $2^{\frac{|V|}{4}}$ cycles with $\frac{3|V|}{4}$ vertices and all $\frac{|V|}{4}$ quadrangles (from [149]).

For *some* classes of graphs $|\mathcal{R}|$ grows exponentially with $|V|$, i.e., Fig 3.7. However, in section 8.3.1 we report computational evidence that, typically, $|\mathcal{R}|$ is not too much larger than the minimum possible value $\nu(\mathcal{G})$ (Fig. 8.12).

Lemma 41. *If $\mathcal{G}$ contains K_4 as a subgraphs then $\mathcal{R}$ is dependent, i.e., $|\mathcal{R}| > \nu(\mathcal{G})$.*

Proof. K_4 contains four triangles, each of which is a relevant cycle of any graph containing the K_4. From $\nu(K_4) = 3$ we conclude immediately that one of them is the $\oplus$-sum of the other three. □

Further, the question arises, how well is $\mathcal{R}$ approximated by the shortest cycles $\mathcal{S}$. We briefly mention two infinite classes of graphs for which $\mathcal{R} = \mathcal{S}$.

Theorem 1.2 of [69] characterizes the perturbed graphs for which the MCB consists only of shortest cycles as the planar graphs without a dual containing a double claw. A *double claw* with ends x and y is a subgraph that consisting of three internally node disjoint paths from x to y (see Fig. 3.8). In the unweighted case this implies:

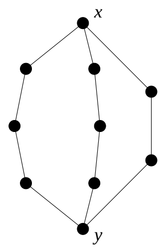


Figure 3.8.
A double claw

Corollary 42. *If $\mathcal{G}$ planar and none of its duals contains a double claw, then $\mathcal{S} = \mathcal{R}$.*

The converse is not true. For instance, all triangles in a complete graph (which for $|V| > 4$ is not planar) are relevant, and of course they are shortest cycles.

A graph is *null-homotopic* [15, 39, 81] if it has a cycle basis consisting only of triangles. This is the case for instance for chordal graphs (in which every cycle C of length $|C| \geq 4$ contains a chord.), and in particular for complete graphs K_m, $m \geq 3$. Trivially, if $\mathcal{G}$ is null-homotopic, then $\mathcal{R} = \mathcal{S}$.

3.5 Vismara's Prototypes of Relevant Cycles

Vismara [149] describes an algorithm for constructing the set of relevant cycles $\mathcal{R}$ that makes use of a partitioning of $\mathcal{R}$ into *cycle families*. Let $\preceq$ be an arbitrary ordering of the vertex set V of $\mathcal{G}$ and $\check{P}_{rx}$ denotes for a shortest path between r and x, that only passed through vertices preceding r in the ordering $\preceq$. Set $V_r = \{x \in V | x \preceq r \text{ and } \exists\, \check{P}_{rx}\}$ and $U_r = \{\text{ arc } (y,z) | (y,z) \in \check{P}_{ry}\}$.

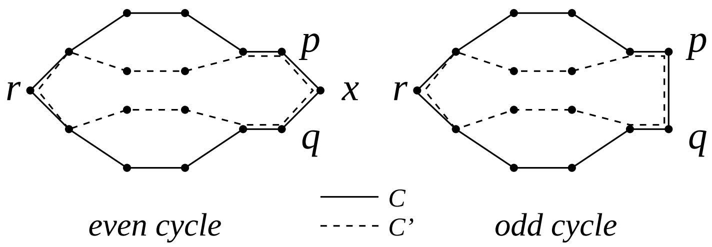
Proposition 43. [149] *Let C be a relevant cycle, and let r be the vertex of C that is maximal w.r.t. the order $\preceq$. Then there are vertices $p, q \in V_r$ such that C consists of two disjoint shortest paths $(r \dots p)$ and $(r \dots q)$ of the same lengths linked by the edge $\{p,q\}$ if $|C|$ is odd or a path (p,x,q), $x \in V_r$, if $|C|$ is even.*

Definition 44. [149] *Let C^r_{pqx} be a cycle as described in proposition 43. The cycle family $\mathcal{F}^r_{pqx}$ consists of all cycles C satisfying the following conditions:*

(i) $|C| = |C^r_{pqx}|$*;*

(ii) *C contains the vertex r as well as the edge $\{p,q\}$ or the path (p,x,q);*

(iii) *There are two shortest paths $(p \dots r)$ and $(q \dots r)$ in C that pass only through vertices $\preceq$-smaller than r, i.e., that are contained in V_r.*



Note that the cycle families $\mathcal{F}^r_{pqx}$ explicitly depend of the order $\preceq$ on V. Vismara shows that $\{\mathcal{F}^r_{pqx} | C^r_{pqx} \text{ is relevant}\}$ forms a partition of $\mathcal{R}$ for any order $\preceq$ on V.

The algorithm Vismara proposes to compute cycle prototypes C^r_{pqx} is based on the converse of proposition 43. This converse is not necessarily true but it gives a

Algorithm 2 Calculation of initial set $\mathcal{C}'_I$

for all $r \in V$ **do**
  Compute V_r and $\forall t \in V_r$ find a shortest path $\breve{P}_{rt}$ from r to t
  $U_r \leftarrow \emptyset$
  for all $y \in V_r$ **do**
    $S \leftarrow \emptyset$
    for all $z \in V_r$ such that z is adjacent to y **do**
      if $d(r,z) + 1 = d(r,y)$ **then**
        $S \leftarrow S \cup \{z\}$
        $U_r \leftarrow U_r \cup \{(y,z)\}$
      else if $d(r,z) \neq d(r,y) + 1$ and $z \preceq y$ and $\breve{P}_{ry} \cap \breve{P}_{rz} = \{r\}$ **then**
        Add to $\mathcal{C}'_I$ the odd cycle $C = \breve{P}_{ry} + \breve{P}_{rz} + (z,y)$
    for any pair of vertices $p, q \in S$ such that $\breve{P}_{rp} \cap \breve{P}_{rq} = \{r\}$ **do**
      Add to $\mathcal{C}'_I$ the even cycle $C = \breve{P}_{rp} + \breve{P}_{rq} + (p,y,q)$

Algorithm 3 Extraction of prototypes $\mathcal{P}$ from $\mathcal{C}'_I$

sort the cycles of $\mathcal{C}'_I$ by length
$k = 3$; $\mathcal{M}_< \leftarrow \emptyset$; $\mathcal{M}_= \leftarrow \emptyset$; $\mathcal{P} \leftarrow \emptyset$;
for all $C \in \mathcal{C}'_I$ **do**
  if $|C| \neq k$ **then**
    $k \leftarrow |C|$ $\mathcal{M}_< \leftarrow \mathcal{M}_< \cup \mathcal{M}_=$; $\mathcal{M}_= \leftarrow \emptyset$;
    if $|\mathcal{M}_<| = \nu$ **then**
      stop, when complete MCB is found
  if $\mathcal{M}_< \cup \{C\}$ is independent **then**
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{C\}$
    if $\mathcal{M}_< \cup \mathcal{M}_= \cup \{C\}$ is independent **then**
      $\mathcal{M}_= \leftarrow \mathcal{M}_= \cup \{C\}$

Algorithm 4 ListPaths(r, x, P)

$P \leftarrow P \cup \{x\}$.
if $x = r$ **then**
  Return(P).
else
  $Result \leftarrow \emptyset$
  for all $z \in V_r$ such that $(x,z) \in U_r$ **do**
    $Result \leftarrow Result \cup$ ListPaths(r, z, P)
  Return$(Result)$

strong condition on cycle relevance. In the first step the set $\mathcal{C}'_I$ including one cycle for each triple r, p, q (or quadruple r, p, q, x) satisfying the condition of proposition 43 is calculated (see Algorithm 2), then in the second step the relevant cycles are extracted from $\mathcal{C}'_I$ by using a greedy algorithm (Algorithm 3).

For the greedy algorithm during processing of a cycle C, $\mathcal{M}_<$ and $\mathcal{M}_=$ denotes the subsets of cycles in the current sub-basis whose length are less than C and equal to C, respectively.

From this set of prototypes $\mathcal{P}$ the set of relevant cycles $\mathcal{R}$ can easily be listed. To generate the cycle family $\mathcal{F}^r_{pqx}$, the digraph $\mathcal{D}_r = (V_r, U_r)$ associated with the vertex r has already been calculated in Algorithm 2.

To list all the paths from x to r in $\mathcal{D}_r$, a backtracking function which is based on a deep first search from x is used (see Algorithm 4). To compute $\mathcal{F}^r_{pqx}$, first the path $(p \dots r)$ in C^r_{pqx} is replaced by each one of the paths returned by the call of $\texttt{ListPaths}(r, p, \emptyset)$. Then, in each cycle generated this way, the path $(r \dots q)$ is again replaced by each one of the paths resulting from $\texttt{ListPaths}(r, q, \emptyset)$. So each cycle in $\mathcal{F}^r_{pqx}$ corresponds to a pair of paths $(p \dots r), (q \dots r)$.

3.6 Essential Cycles

Definition 45. [55] *A cycle C in $\mathcal{G}$ is* essential *if it is contained in every minimum cycle basis of $\mathcal{G}$.*

Therefore, the set $\mathcal{J}(\mathcal{G})$ or short $\mathcal{J}$ of essential cycles is the intersection of all minimum cycles bases. Note that $\mathcal{J}$ can be empty. As an example consider the complete graph K_4, see Lemma 41. Similarly, $\mathcal{J} = \emptyset$ for larger complete graphs. Not surprisingly, $\mathcal{J}$ can be computed rather easily from $\mathcal{R}$.

Lemma 46. *Let $\mathcal{M}$ be a minimum cycle basis of $\mathcal{G}$,*

$$\mathcal{M}_k = \{|C| \in \mathcal{M} : |C| < k\},\ \mathcal{R}_k = \{|C| \in \mathcal{R} : |C| = k\} \text{ and } C \in \mathcal{R}_k.$$

Then $C \in \mathcal{J}$ if and only if $\operatorname{rank}[\mathcal{M}_k \cup \mathcal{R}_k \setminus \{C\}] < |\mathcal{M}_{k+1}|$.

Proof. By definition, $C \in \mathcal{J}$ if and only if $\mathcal{R} \setminus \{C\}$ does not contain a cycle basis. If $|C| = k$, it follows from the matroid properties of the cycle space that we have to consider only cycles up to length k. With $\mathcal{R}_{\leq k} = \bigcup_{j \leq k} \mathcal{R}_j$ we have $C \in \mathcal{J}$ if and only if $\operatorname{rank}[\mathcal{R}_{\leq k} \setminus \{C\}] < \operatorname{rank}[\mathcal{R}_{\leq k}]$. The lemma now follows from $\operatorname{rank}[\mathcal{R}_{\leq k}] = \operatorname{rank}[\mathcal{M}_k \cup \mathcal{R}_k]$. □

The algorithm for computing $\mathcal{J}$ from $\mathcal{R}$ is summarized in Algorithm 5. Its worst case complexity is determined by the $|\nu(\mathcal{G})|$ rank computations in step 7, which in

Algorithm 5 Extract essential from relevant cycles

Input: $\mathcal{R}$

$k = 3$; $\mathcal{M} \leftarrow \emptyset$; $\mathcal{R}_= \leftarrow \emptyset$; $\mathcal{J} \leftarrow \emptyset$;
repeat
  $C \leftarrow$ cycle with minimal $|C|$ in $\mathcal{R}$; $\mathcal{R} \leftarrow \mathcal{R} \setminus \{C\}$.
  if $|C| > k$ or $\mathcal{R} = \emptyset$ **then**
    $r = |\mathcal{M}|$ /* Rank of $\{C \in \mathrm{MCB} : |C| \leq k\}$ */
    for all $C' \in \mathcal{M}_=$ **do**
      if $\mathrm{rank}[(\mathcal{M} \cup \mathcal{R}_=) \setminus \{C'\}] < r$ **then**
        $\mathcal{J} \leftarrow \mathcal{J} \cup \{C'\}$
    $k \leftarrow |C|$; $\mathcal{R}_= \leftarrow \emptyset$; $\mathcal{M}_= \leftarrow \emptyset$;
  if $\mathcal{R} = \emptyset$ **then**
    return $\mathcal{J}$
  $\mathcal{R}_= \leftarrow \mathcal{R}_= \cup \{C\}$;
  /* Extract an MCB */
  if $\mathcal{M} \cup \{C\}$ independent **then**
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{C\}$; $\mathcal{M}_= \leftarrow \mathcal{M}_= \cup \{C\}$;
until

practice can be divided into two parts. Let $\mathcal{M}_{=k} = \mathcal{M}_{k+1} \setminus \mathcal{M}_k$ denote the set of cycles with length k in the minimum cycle basis. For each length k it suffices to perform a Gaussian elimination on $\mathcal{M}_k \cup (\mathcal{R}_k \setminus \mathcal{M}_{=k})$ once. This step requires at most $\mathcal{O}(|\mathcal{R}|\,|\mathcal{M}|\,|E|)$ operations. The ranks can now be computed by performing Gaussian elimination on the union of the result of the first step (which has only $\mathcal{O}(|\mathcal{M}|)$ rows) and $\mathcal{M}_{=k} \setminus \{C\}$ for each $C \in \mathcal{M}_{=k}$. For each C, this can be done with no more than $\mathcal{O}(|\mathcal{M}|^2|E|)$ steps. In the worst case, hence, $\mathcal{J}$ can be obtained in $\mathcal{O}(|\mathcal{R}|\,\nu(\mathcal{G})^2\,|E|)$ operations.

Since the cycles in $\mathcal{S}$ are not independent in general (Fig. 3.10), it seems natural to consider the set

$$\mathcal{S}^*(\mathcal{G}) = \{C \mid \exists\, e \in E : C \text{ is the } \textit{unique} \text{ shortest cycle containing } e\}$$

instead. For short we write $\mathcal{S}^*$ instead of $\mathcal{S}^*(\mathcal{G})$. As each cycle C in $\mathcal{S}^*$ is the unique shortest cycle for any perturbed edge weighting the discussion in [69] implies that C is contained in all minimum cycle bases, i.e., $\mathcal{S}^* \subseteq \mathcal{J}$. Trivially, if $\mathcal{S}^*$ is a cycle basis, then the MCB is unique and $\mathcal{S}^* = \mathcal{J} = \mathcal{S} = \mathcal{R}$. This provides a sometimes convenient way to establish the uniqueness of the MCB, see Figure 3.9.

However, uniqueness of the minimum cycle basis, i.e., $\mathcal{J} = \mathcal{R}$, in general does not imply that $\mathcal{S}^* = \mathcal{J}$. The example in Figure 3.4 is outerplanar and hence has a unique

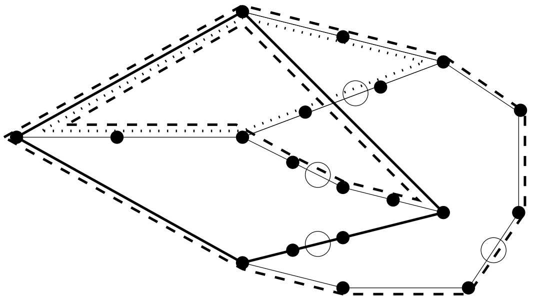


Figure 3.9. $\mathcal{G}$ is a subdivision of $K_{3,3}$, hence $\nu(\mathcal{G}) = 4$. Since $\mathcal{S}^*$ contains the four marked cycles, $\mathcal{G}$ has a unique minimum cycle basis. For each of these cycles, an edge for which the cycle is the unique shortest one, is indicated by a circle.

MCB, but $\mathcal{S}^*$ contains only the four triangles. In a more restricted setting, however, which includes secondary structure graphs, we have

Lemma 47. *Let $\mathcal{G}$ be sub-cubic 2-connected outerplanar graph. Then $\mathcal{S}^*$ is the minimum cycle basis.*

Proof. $\mathcal{G}$ has a unique Hamiltonian H cycle which forms the boundary of the planar embedding [132]. The minimum cycle basis is given by the cells of planar embedding [93]. For each edge $e \in H$ there is unique shortest cycle, namely the cell in which it is contained. Since the vertex degree is at most 3, each cycle $|C|$ must contain at least one boundary edge e, i.e., $\mathcal{S}^*$ is the collection of all cells, and hence the MCB. □

Biopolymer graphs of nucleic acid, represented by secondary structures, bisecondary structures, or even more elaborate models, do not contain triangles. Furthermore, the only class of quadrangles is formed by so-called base-pairing stacks, along which edges from T and B alternate. It is easy to verify that each quadrangle is the unique shortest cycle for each of the two backbone edges $e, e' \in T$. Thus $\mathcal{S}^*$ contains all base-pairing stacks which correspond to the stabilizing structural elements.

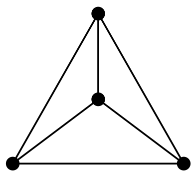


Figure 3.10. All four triangles of the tetraeder are shortest cycles, but none of them is a unique shortest one.

Lemma 48. *Let $\mathcal{G}$ be a planar graph with a unique minimum cycle basis. Then $\mathcal{S}^* \neq \emptyset$.*

Proof. By [93, Cor.13], any MCB of a planar graph is fundamental and hence there is an edge e that is contained in exactly one cycle C. Since all shortest cycles containing e are contained in $\mathcal{R}$ by Prop. 39, we conclude from the uniqueness of the MCB that C is the unique shortest cycle containing e, i.e., $C \in \mathcal{S}^*$. □

The lemma immediately generalizes to graphs with a unique MCB that is fundamental. Not all graphs has this purpose as the example Fig. 4.1 shows. We *conjecture* that uniqueness of the MCB implies $\mathcal{S}^* \neq \emptyset$.

Chapter 4

Ring Sets for Chemical Applications

Until Vismaras' paper [149] no efficient algorithm was known to find all relevant cycles. Not surprisingly therefore many investigations deal with the definition of extended minimum cycle bases which are generally not canonical. This section covers the main ring sets used in applications processing chemical structure graphs, for further details see [37] and references there.

A plane graph $\hat{\mathcal{G}}$ consists of a set V of points in $\mathbb{R}^2$ and a sets E of line segments in $\mathbb{R}^2$ connecting exactly two points in V such that two lines intersect only in the points that they connect. A graph $\mathcal{G}$ is *planar* if there is a plane graph $\hat{\mathcal{G}}$ that is isomorphic with $\mathcal{G}$. We say that $\hat{\mathcal{G}}$ is an embedding of $\mathcal{G}$ in the plane. The connected components of $\mathbb{R}^2 \setminus \hat{\mathcal{G}}$ are the regions of $\hat{G}$. There is exactly one infinite region, all other regions are finite. Planar embeddings are equivalent to embeddings on the surface of a sphere $\mathbb{S}$. Here all regions are of course finite. Each region of the spherical embedding can be made to the infinite region of the plane embedding. If $\hat{\mathcal{G}}$ is 2-connected then the regions are delimited by a unique cycle in $\mathcal{G}$ which we call a *face*. Note, that the planar embedding is usually not unique.

Since each edge of $\hat{\mathcal{G}}$ appears in exactly two faces the $\oplus$-sum of all faces is 0. The set of all faces except one, on the other hand, are a basis of the cycle space. In other words, the faces belonging to the finite regions of every planar embedding of $\mathcal{G}$ are a cycle basis for $\mathcal{G}$.

4.1 All Cycles and Simple-Cycles

The set of all cycles, and its related subset of all chordless cycles, is unique for a given structure, but generally contains far more cycles than are necessary to describe the ring system. For complex ring systems, processing to find the number of cycles can grow exponentially with the number of vertices. The most recent algorithm is by Hanser *et al.* [64], which uses graph reduction to make the processing fast and easy to implement.

Owing to the large number of cycles to be found in complex ring systems, in the worst case processing will be slow. However, for certain applications, such as structure display optimization and the automatic generation of chemical names, the set of all cycles or all chordless cycles is required to ensure complete description of a ring system.

4.2 Smallest Set of Smallest Rings (SSSR)

Originally, the **S**mallest **S**et of **S**mallest **R**ings was defined as a minimum length Kirchhoff basis [124, 159]. In 1982 Deo *et al.* [30] showed that the problem of finding a strictly fundamental cycle basis with minimum length is $\mathcal{NP}$ complete. In the more recent literature [6, 47, 111] the term SSSR is used mostly as another word for *minimum cycle basis*. This discrepancy seems to stem from the wide-spread misconception that every cycle basis or at least every minimum cycle basis is strictly fundamental.

Every *elementary* cycle appears in some Kirchhoff-fundamental cycle basis of $\mathcal{G}$. The term "fundamental cycle" which frequently appears in the ring-perception literature therefore simply means "elementary cycle", when used without reference to a particular spanning tree $\mathcal{G}$.

Not all cycle bases of a graph are fundamental [71]. It is shown in [93, Cor.13], however, that every minimum cycle basis of a *planar* graph is fundamental. Cubane (Fig. 1.3) is a chemical example. Nevertheless, there are planar graphs, such as the one in Fig. 3.1, that do not have a strictly fundamental minimum cycle basis, i.e., for which no minimum cycle basis can be derived from a spanning tree.

For non-planar graphs an even stronger negative result holds: There are graphs with non-fundamental minimum cycle bases. A non-fundamental MCB of the complete graph K_9 is described in [93]. The example in Fig. 4.1, which is due to Champetier [21], has a unique minimum cycle basis consisting entirely of triangles.

A number of polynomial-time algorithms for computing minimum cycle bases have been published in recent years. Their worst-case complexity is rather high. Nevertheless, the average performance on problems of practical interest seem to be much more favorable. In Table 4.1 we summarize the performance bounds of some of these approaches for general graphs and for planar graphs, where $|E| \leq 3|V| - 6$.

For chemical applications the main problem with the SSSR is, that it is not unique [119, 168]. Clearly this fact introduces a potential ambiguity in the definition of which SSSR is selected to model the ring of a particular structure. Many applications may just ignore this problem and use whatever SSSR is first perceived, implicitly depending upon the ordering of the connectivity data used by the ring perception algorithm. In some

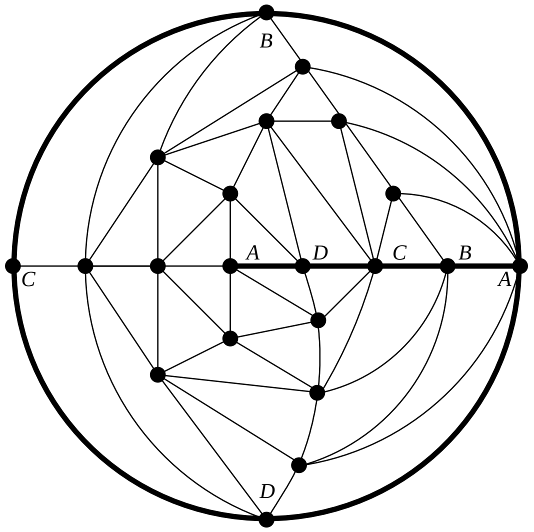


Figure 4.1. Champetier's graph [21]:
Note, that in the graph $\mathcal{G}$ the vertices with the same label are identified, thus the bold 4-cycle and the bold line are the same cycle in the real graph.
$\mathcal{G}$ is null-homotopic (i.e., it has a cycle basis consisting exclusively of triangles) and all triangles are part of the unique minimum cycle basis. Each edge is contained in two triangles with the exception of the edges of the 4-cycle ABCD which are contained in 3 triangles. Thus there is no ordering of the triangles satisfying equ.(3.1).

Table 4.1. Worst Case Behavior of some Minimum Cycle Basis Algorithms.
The maximum vertex degree is denoted by d, the cyclomatic number is $\nu = |E| - |V| + 1$.

Algorithm		general	planar
Horton	[77]	$\mathcal{O}(\nu^2 \vert E\vert\vert V\vert)$	$\mathcal{O}(\vert V\vert^4)$
Balducci & Pearlman#	[6]	$\mathcal{O}(d\nu \vert E\vert^2 \vert V\vert)$	$\leq \mathcal{O}(\vert V\vert^5)$
Hartvigsen (planar)	[70]	—	$\mathcal{O}(\vert V\vert^2 \log \vert V\vert)$
Vismara*	[149]	$\mathcal{O}(\mu \vert E\vert^3)$	$\mathcal{O}(\vert V\vert^4)$

* Vismara's algorithm computes "prototypes" for the set of all relevant cycles and produces a minimum cycle basis as by-product.
# The estimate of the worst case complexity in [6] is incorrect. We give here the bound derived in [148]. The planar estimate is obtained by setting $|E| = \mathcal{O}(|V|)$ and $\nu = \mathcal{O}(|V|)$ and $d = \mathcal{O}(|V|)$.

cases, however, it is necessary to make an "intelligent" decision regarding which SSSR to use in further processing based upon specific (application dependent) properties of the rings involved. This ambiguity presents the implementor with three options when there is more than one SSSR:

1. Compute an arbitrary minimum cycle basis.

2. Select the minimum cycle basis with a preferred ordering of the cycles, e.g. by using an ordering of the vertices that is inspired by the chemical intuition for algorithm 2.

3. Include more rings to find a superset of the SSSR.

4.3 K-rings

The set of K-rings was defined by Plotkin [115] and is the set containing all possible SSSR rings. The set of K-rings avoids arbitrary exclusions of rings when there is more than one SSSR for a structure. With the current usage of SSSR the K-rings are the set $\mathcal{R}$ of relevant cycles.

4.4 β-ring

The set of β-rings [107] is one of the earliest attempts to extend the SSSR to include the hexagon of norbornane (Fig. 4.2) but without including much larger "envelope" rings.

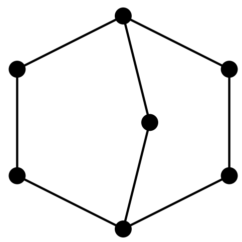


Figure 4.2. Norbornane

A β-set is obtained from the length-sorted list of all chordless faces in the same way as the relevant cycles, except that (i) all three and four-cycles are in β irrespective of linear dependence, and (ii) the test for the linear independence of the set $\mathcal{M}_< \cup \{C\}$ is replaced by checking whether $\{C, C', C'', C'''\}$ is independent for some $\{C', C'', C'''\} \subseteq \beta_<$ in algorithm 2. The problem here is that not all planar graphs have a minimum cycle basis consisting of faces. Hence β-rings as defined here does **not** always contain a minimum cycle basis (Fig. 4.4).

The variation β^* of the definition of β-rings that uses all chordless cycles instead of only the faces in a particular embedding in the plane, clearly is a super-set of the relevant cycles because linear independence of $\mathcal{M}_< \cup \{C\}$ implies the three-cycle condition. Thus $\mathcal{R} \subseteq \beta^*$.

4.5 Essential Set of Essential Rings (ESER)

The **E**ssential **S**et of **E**ssential **R**ings was defined by Fujita [50, 51] specifically for use with the imaginary transitions structure construct for representing reaction-site changes during organic reactions. Depending on atom types cycles are classified by the atoms that they contain as *carbon*, *heteroatom* (N, O, S, P), and *abnormal* (all other atoms). The definitions below are a rephrasing of those given in the review [37] and in Vismara's dissertation [148], respectively.

For each simple cycle C define $\mathscr{T}[C]$ as the set of all tied cycles C' belonging to the same atom-type class as C with at most the same number of heteroatoms and abnormal atoms, respectively, as C, that satisfy (i) $|C'| \leq |C|$ and (ii) $C' \cap C \neq \emptyset$. If

(ii) is replaced by the stronger condition (iii) $2|C' \cap C| \geq |C'|$, i.e., at least half of the edges of C' are in C, we write $\mathscr{T}^*[C]$.

A simple cycle C is *ESER*-dependent if there is a subset $\mathscr{T}' \subseteq \mathscr{T}^*[C]$ such that $C \subseteq E[\mathscr{T}']$. The reviews [35, 37] give a slightly different definition: A simple cycle C is *DESER*-dependent if there is a subset $\mathscr{T}' \subseteq \mathscr{T}[C]$ such that (i) $C \subseteq E[\mathscr{T}']$, i.e., $\mathscr{T}'$ covers C, (ii) $2|C| < |E[\mathscr{T}']|$. Finally, ESER($\mathcal{G}$) (DESER($\mathcal{G}$)) is the set of all simple cycles in $\mathcal{G}$ that are not ESER-dependent (DESER-dependent).

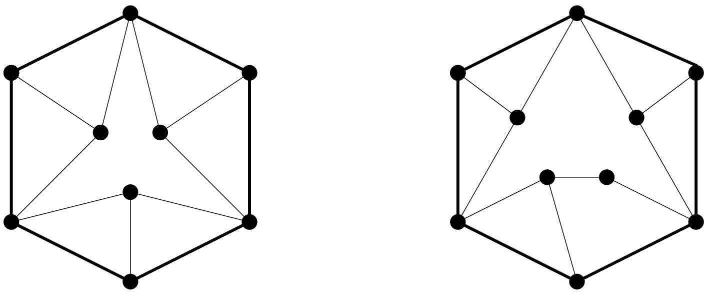


Figure 4.3. DESER and ESER are unrelated. The outer hexagonal H (bold) of the graph $\mathcal{G}'$ is ESER-dependent and DESER-independent. The situation is reversed in $\mathcal{G}''$. The hexagon H (bold) is relevant for $\mathcal{G}'$ and $\mathcal{G}''$, hence neither ESER nor DESER is a superset of an SSSR in general. For details see text.

Consider the outer hexagon, H, of the graph $\mathcal{G}'$ in Figure 4.3. This example is taken from [148, p.78]. The set $\mathscr{T}[H]$ consists of the 3 squares. The only subset of $\mathscr{T}[H]$ that covers H is $\mathscr{T}[H]$ itself. Furthermore $\mathscr{T}^*[H] = \mathscr{T}[H]$. Each square has two of its edges in common with H, hence H is ESER-dependent. On the other hand, $|E[\mathscr{T}[H]]| = 3 \times 4 = 12 \not> 2|H| = 12$, hence $H \in$ DESER($\mathcal{G}$).

Now consider the outer hexagon, H, of the graph $\mathcal{G}''$ in Figure 4.3. The set $\mathscr{T}[H]$ now consists of the two tied squares Q_1 and Q_2 and the tied pentagon P. We have $|E[\mathscr{T}[H]]| = 2 \times 4 + 5 = 13 > 2|H| = 12$, thus H is DESER-dependent. However, $\mathscr{T}^*[H] = \{Q_1, Q_2\}$ does not cover H, hence $H \in$ ESER($\mathcal{G}$).

Thus the definitions of ESER and DESER give unrelated cycle sets. Downs [35] mentions that "the ESER is in general always a superset of an SSSR". Similarly, Fujita [51, Fig.2] claims that ESER($\mathcal{G}$) contains the SSSR. This is incorrect as the $\mathcal{G}'$ shows for ESER and $\mathcal{G}''$ shows for DESER. The relevant cycles of $\mathcal{G}'$ are the 6 triangles (all of which are essential) and all $2^3 = 8$ interchangeable hexagons (of which one is contained in every minimum cycle basis). The relevant cycles of $\mathcal{G}''$ are the 5 triangles and the square contained in P (each of these cycles is essential), and the $2^2 = 4$ hexagons (one of which is contained in every minimum cycle basis). Thus, neither ESER($\mathcal{G}'$) nor DESER($\mathcal{G}''$) contain a minimum cycle basis.

4.6 Minimal Planar Cycle Bases

Some authors focus entirely on planar graphs in the context of chemical ring perception, see e.g., [42, 38]. Let us call a cycle basis of $\mathcal{G}$ *planar* if it consists of the faces belonging to the finite regions of a planar embedding of $\mathcal{G}$. A planar cycle basis has minimal length if and only if the length of face F_∞ belonging to the unbounded region is maximal. This is true because $\ell(\mathcal{B}) = 2|E| - |F_\infty|$ for any planar cycle basis.

Planar embeddings can be computed in $\mathcal{O}(|V|)$ time, see e.g., [23, 125]. Unfortunately, the embedding of a planar graph on the sphere $\mathbb{S}$ is in general not unique, unless $\mathcal{G}$ is tri-connected [162]. Algorithms are available that can produce all embeddings of $\mathcal{G}$ on the sphere [18]. The computation of one or all minimal planar cycle bases can be achieved e.g. by the integer linear programming approach outlined in [104].

We denote by Faces($\mathcal{G}$) the set of all possible faces in $\mathcal{G}$, i.e., the set of all cycles that are faces in some embedding on $\mathbb{S}$.

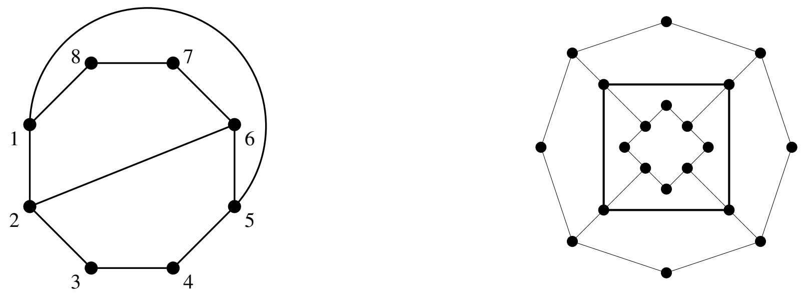


Figure 4.4. Two examples of planar graphs for which the minimal cycle basis is non-planar. The l.h.s. example is taken from [93]: No planar embedding has the face $Q = (1, 2, 6, 5)$. A minimal cycle basis contains Q and two of the cycles $(2, 3, 4, 5, 6)$, $(1, 2, 6, 7, 8)$, $(1, 2, 3, 4, 5)$, and $(1, 5, 6, 7, 8)$ and hence has length $\ell = 14$. The planar bases have length 15.
The bold square Q in the shortest cycle in the r.h.s. graph, which is taken from [38, Fig.12], and hence contained in every minimal cycle basis. It cannot appear as a face, however.

It is well known that Faces($\mathcal{G}$) in general does not contain a minimal cycle basis. In Fig. 4.4 we give two examples.

4.7 Extended Set of Smallest Rings (ESSR)

The **E**xtended **S**et of **S**mallest **R**ings was introduced by Downs *et al.* [38] as an approach to design an optimal ring set for retrieval purposes. ESSR by definition is limited to planar graphs. Paraphrasing the original definition, a cycle C is in ESSR($\mathcal{G}$) provided it satisfies at least one of the following conditions:

(i) There is a planar embedding of $\mathcal{G}$ such that C is a face.

(ii) C is a shortest cycle through at least one of its edges.

(iii) There is a planar embedding of $\mathcal{G}$ such that $|C| \geq |C'|$ for all faces C' adjacent to C, and there is at least one adjacent face C'' for which $|C| = |C''|$.

An algorithm for computing the ESSR is described in [36]. Downs *at al.* [38, p.192] claims that $\mathrm{ESSR}(\mathcal{G}) = \mathrm{Faces}(\mathcal{G}) \cup \mathcal{R}(\mathcal{G})$. The graph $\mathcal{G}$ in Figure 4.5, which is used to give an example of a "tertiary cut-face" in [38, Fig.16] and [35, Fig.7], however, provides an example of a graph for which $\mathcal{R}(\mathcal{G}) \not\subseteq \mathrm{ESSR}(\mathcal{G})$.

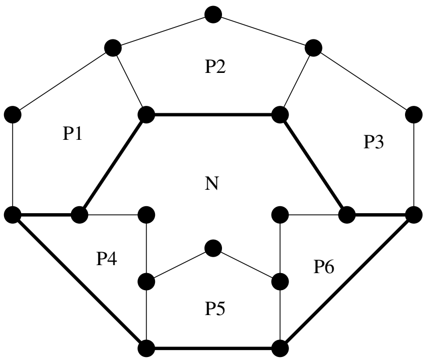


Figure 4.5. Not all relevant cycles are in ESSR. The graph $\mathcal{G}$ has $|E| = 24$, $|V| = 18$, i.e., $\mu(\mathcal{G}) = 7$. The relevant cycles are the six pentagons P_i, $i = 1, \ldots, 6$, which form $\mathcal{S}(\mathcal{G})$ and the octagon O shown in bold. O is obviously linearly independent from the pentagons and it is the next-shortest cycle in $\mathcal{G}$. On the other hand, there is no planar embedding in which O is face and O does not have an adjacent octagon. Thus $O \notin \mathrm{ESSR}(\mathcal{G})$.

This restriction to planar graphs is much to restrictive, since there exits also non-planar graphs in the organic chemistry (Fig. 4.6 and Fig. 4.7).

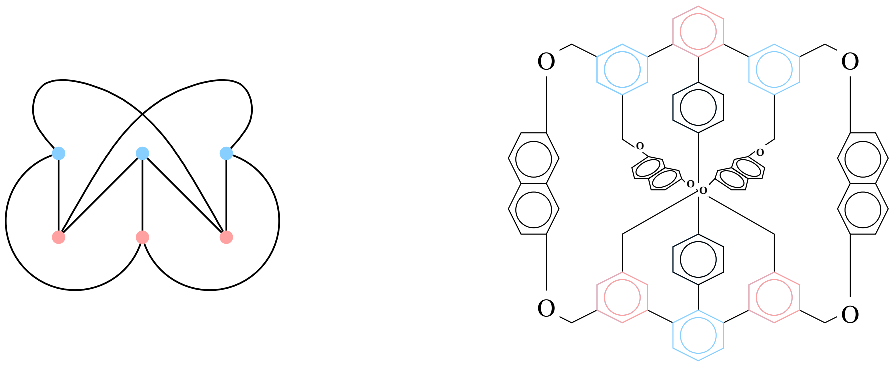


Figure 4.6. The Kuratowski-Cyclophan[††] [22] (r.h.s.) contains the $K_{3,3}$ (l.h.s.) as structural element.

[††]This Cyclophan was named after Casimir Kuratowski, who firstly proved, that K_5 and $K_{3,3}$ are fundamental non-planar graphs and that each non-planar graph has to contain at least one of them. [91]

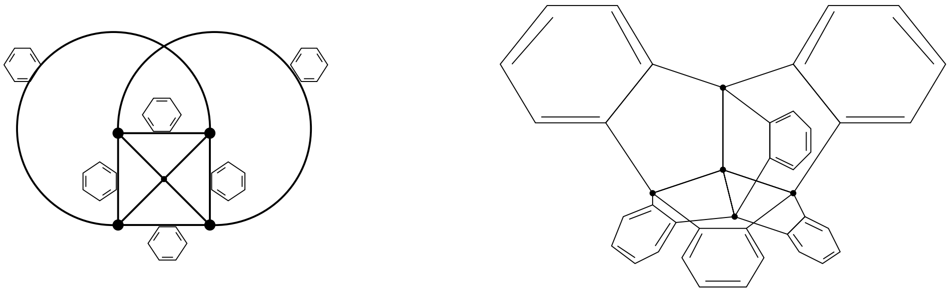


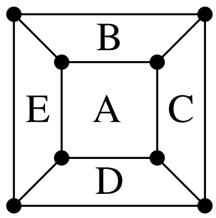
Figure 4.7. Centrohexaindan [90] was the first synthesized hydrocarbon with a non-planar molecule structure.

4.8 Set of Elementary Rings (SER)

The **S**et of **E**lementary **R**ings was defined by Takahashi [135] and based on the notion of θ-graphs. A θ-graph consists of two vertices x and y and three disjoint paths connecting x and y, i.e., a double claw. The SER is calculated as followed: Start with SER=MCB (or SSSR). For C_1, C_2 construct the subgraph induced by $C_1 \cup C_2$. If this is a θ-graph then add $C_1 \oplus C_2$ to SER. Iterate until no such cycles can be found any more.

In the table in Fig. 4.8 the different cycles of cubane and the membership in the ring sets are given.

cycle type	length	chordless	relevant	β-rings	SSSR
A, B, C, D, E	4	x	x	x	x
$A \oplus B \oplus C \oplus D \oplus E$	4	x	x	x	-
$A \oplus B$ etc.	6	-	-	-	-
$A \oplus B \oplus C$ etc.	6	x	-	-	-
$A \oplus B \oplus C \oplus D$ etc	6	-	-	-	-
$E \oplus A \oplus C$	8	-	-	-	-



cycle type	length	K-rings	ESSR	ESSR	SER
A, B, C, D, E	4	x	x	x	x
$A \oplus B \oplus C \oplus D \oplus E$	4	x	x	x	x
$A \oplus B$ etc.	6	-	x	-	x
$A \oplus B \oplus C$ etc.	6	-	-	x	x
$A \oplus B \oplus C \oplus D$ etc	6	-	-	x	x
$E \oplus A \oplus C$	8	-	-	-	-

Figure 4.8. The ring sets of cubane, partly taken from [38].

Chapter 5

U-Bases

A *uv-path* P in $\mathcal{G}$ is a connected subgraph that has exactly two degree-one vertices, u and v, called its *end-nodes*, while all other vertices, called the *interior vertices* of P have even degree. A uv-path is elementary if all its interior vertices have degree 2. One can easily check that a uv-path is an edge-disjoint union of an elementary uv-path and a collection of elementary cycles. Obviously, a uu-path is a cycle through u.

Let $U \subseteq V$ and consider the vector space $\mathfrak{U}^*$ generated by the incidence vectors of the uv-paths with $u, v \in U$. This construction is of interest for example in the context of chemical reaction networks, where a subset U of all chemical species V is fed into the system from the outside or it harvested from the system. The uv-paths hence correspond to productive pathways [45, 56, 138]. Hartvigsen [66] introduced the U-space $\mathfrak{U}(\mathcal{G})$ as the union of $\mathfrak{U}^*$ and the cycle space $\mathcal{C}(\mathcal{G})$. He gives an algorithm for computing the a minimum length basis of $\mathfrak{U}(\mathcal{G})$, a *minimal U-basis* for short, in polynomial time that extends a previous algorithm by Horton [77] for minimum length bases of the $\mathcal{C}(\mathcal{G})$.

5.1 Dimension of the U-space $\mathfrak{U}(\mathcal{G})$

The dimension of the cycle space $\mathcal{C}(\mathcal{G})$ is the cyclomatic number

Theorem 49. *If $\mathcal{G}$ is connected then* $\dim(\mathfrak{U}(\mathcal{G})) = \nu(\mathcal{G}) + |U| - 1$.

Proof. Let $C = C_1 \oplus C_2 \oplus \cdots \oplus C_k$. Then for any vertex $x \in V$ the degree of x in C is even if and only if $\sum_{i=1}^{k} \deg_{C_i}(x)$ is even. In particular, the $\oplus$-sum of two paths between two vertices x and y is a cycle.

We proceed by induction on the number of vertices in U. Trivially, if $|U| = 1$ then $\mathfrak{U}(\mathcal{G}) = \mathcal{C}(\mathcal{G})$. Hence assume $U = \{x, y\}$. To construct a basis for $\mathfrak{U}$, we need a path

$P(x, y)$ in addition to the cycle basis, since all paths from between x and y are obtained as $\oplus$-sums of the path $P(x, y)$ and some cycles. Hence $\dim(\mathfrak{U}) = \nu + 1$.

Now assume the proposition holds for $U \subset V$ and consider $U' = U \cup \{v\}$ for some $v \in V \setminus U$. Since there is no path with endpoint v in the basis of $\mathfrak{U}$ the degree of v is even for every $\oplus$-sum of elements in $\mathfrak{U}$. Thus $\dim(\mathfrak{U}') > \dim(\mathfrak{U})$. To obtain a basis for $\mathfrak{U}'$ we have to add a path $P(v, x)$ for some $x \in U$ to the basis of $\mathfrak{U}$. Clearly, $P(v, x) \oplus P(x, y)$ is the edge-disjoint union of a path $P(v, y)$ and a (possibly empty) collection of elementary cycles. All other paths from v to $y \in U$ can now be obtained as the $\oplus$-sum of the path $P(v, y)$ and an appropriate set of cycles. Hence $\dim(\mathfrak{U}') = \dim(\mathfrak{U}) + 1$ and the theorem follows. □

We immediately find the following

Corollary 50. *If $\mathcal{G}$ is a simple connected graph $\mathcal{G}$ and $U \subset V$ is non-empty, then* $\dim(\mathfrak{U}) = |E| - |V| + |U|$.

Notice that this result also holds for graphs $\mathcal{G}$ that are not connected provided that each component of $\mathcal{G}$ contains at least one vertex of U.

5.2 Minimal U-Bases

A *minimum U-basis* is a U-basis of minimum length. Horton's [77] algorithm, the first polynomial algorithm to find a minimum cycle basis, is based on a very simple necessary condition, that cycles must fulfill to belong to minimum cycle basis (see 3.3 and proposition 40). From this set of cycles a minimum cycle basis can be extracted with the greedy algorithm. Hartvigsen [66] also found a very simple necessary condition that U-paths and cycles in a minimum U-basis must satisfy, by generalizing proposition 40 from Horton. This allows to polynomially compile a list of U-paths and cycles from which a minimum U-basis can be extracted with the greedy algorithm.

For short P is called a *shortest uv-path* if it is a shortest path from u to v. In the weighted situation, P is a *shortest uv-path* if it is a path of minimum weight connecting u and v. We reserve the symbol P_{xy} for **a shortest** path between x and y, while $P(x, y)$ may be any path between x and y.

If p and q are vertices in the path P, we write $P[p, q]$ for the subpath of P connecting p and q. Note that P_{uv} is a shortest uv-path if and only if $P[x, y]$ is a shortest xy-path for all vertices x and y in P.

Definition 51. *A uv-Path P is* short *if for every vertex w in P, $P[u, w]$ is a shortest uw-path or $P[v, w]$ is a shortest vw-path.*

Definition 52. *A uv-path P is* edge-short *if there is an edge $e = \{x, y\}$ such that both $P[u, x]$ is a shortest ux-path and $P[y, v]$ is shortest yv-path.*

Lemma 53. *A uv-path P is short if and only if it is edge-short.*

Proof. This result is given without proof in [66]. We give the simple proof for completeness.

Suppose P is edge-short. Since each w in P is contained either in $P[u, x]$ or in $P[v, y]$, there is a wu- or a wv-shortest subpath of P for each w in P, and hence P is short.

Suppose P is short. Suppose $P[u, x]$ is a shortest ux-path, and y is a vertex in $P[u, v]$; then $P[u, y]$ is a shortest uy-path. Thus there is a maximal vertex x such that $P[u, x]$ is a shortest ux-path and a maximal vertex y such that $P[v, y]$ is a shortest vy-path. Since P is short, every vertex w in P is contained in at least one of $P[u, x]$ and $P[v, y]$. Thus x and y are either (i) separated by the single edge $e = \{x, y\}$, (ii) $x = y$, or (iii) $P[u, x]$ and $P[v, y]$ contain at least one common edge. Clearly, in each case P is edge short. ☐

5.3 Relevant U-Paths

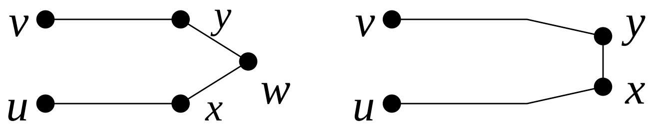
Proposition 54. [66] *A U-path P can be relevant only if it is short.*

We denote the union of all relevant cycles and U-paths $\mathcal{U}_{\mathcal{R}}$ and follow Vismara's approach [149] for the initial set of $\mathcal{C}_{\mathcal{R}}$ further.

Definition 55. *Let $u, v \in U$ and let P_{uv} be a short uv-path. P_{uv} is called* balanced *if either*

(even) *there is a vertex w in P_{uv} such that $|P_{u,w}| = |P_{w,v}|$ and $P_{u,w}$ and $P_{v,w}$ are shortest uw- and wv-paths, respectively, or*

(odd) *there is an edge $e = \{x, y\} \in P$ such that $|P_{u,x}| < \frac{1}{2}|P|$ and $|P_{v,y}| < \frac{1}{2}|P|$, and $P_{u,x}$ and $P_{v,y}$ are shortest ux- and vy-paths respectively.*



even balanced *odd balanced*

Theorem 56. *Any relevant U-path P consists of two disjoint shortest paths $(u \dots x)$ and $(v \dots y)$ of the same length, linked by the edge (x, y) if P is odd balanced or by the path (x, w, y) if P is even balanced.*

Proof. We know that P must be short if it is relevant (see proposition 54). Suppose P is short but not balanced.

In the even case, either $P_{u,w}$ or $P_{v,w}$ is therefore not a shortest path. W.l.o.g. we assume that $P_{u,w}$ is not uw-shortest and write therefore $P[u, w]$. Let Q be a uw-shortest path. Then $|Q| < |P[u, w]| = \frac{1}{2}|P|$.
Set $C = Q \oplus P[u, w]$. Clearly C is a cycle or an edge disjoint union of cycles and $|C| \leq |Q| + |P[u, w]| < |P|$. Now consider the path $P' = Q \oplus P_{w,v}$; it is a short U-path satisfying $|P'| = |Q| + |P_{w,v}| < |P[u, w]| + |P_{w,v}| = |P|$. We have

$$P = P[u, w] \oplus P_{w,v} = P[u, w] \oplus Q \oplus Q \oplus P_{w,v} = C \oplus P' \tag{5.1}$$

Hence P can be written as $\oplus$-sum of strictly shorter elements of the U-space, hence it cannot be relevant.

In the odd case either $P_{u,x}$ or $P_{v,y}$ is not shortest. W.l.o.g. we assume that $P_{u,x}$ is not ux-shortest, write then $P[u, x]$ and consider a shortest ux-path Q. In this case we have $|Q| < |P[u, x]| < \frac{1}{2}|P|$.
Set $C = Q \oplus P[u, x]$. Clearly C is a cycle or an edge disjoint union of cycles and $|C| \leq |Q| + |P[u, x]| < |P|$. Now consider the path $P' = Q \oplus e(x, y) \oplus P_{y,v}$; it is a short U-path satisfying $|P'| = |Q| + |e(x, y)| + |P_{y,v}| < |P[u, x]| + |e(x, y)| + |P_{y,v}| = |P|$. We have

$$P = P[u, x] \oplus e(x, y) \oplus P_{y,v} = P[u, x] \oplus Q \oplus Q \oplus e(x, y) \oplus P_{y,v} = C \oplus P' \tag{5.2}$$

Hence again P can be written as $\oplus$-sum of strictly shorter elements of the U-space and therefore cannot be relevant. □

In other words:

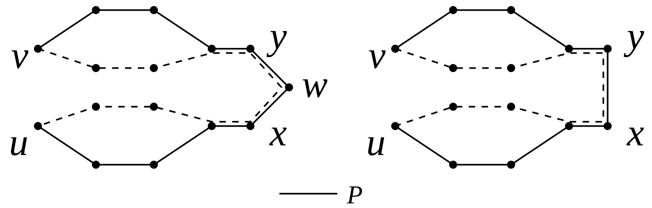
Corollary 57. *If P is a relevant U-path, then it is a balanced short U-path.*

5.4 U-Path Prototypes

We are now in the position to construct *prototypes* of relevant U-path P^{uv}_{xwy} in the same manner as Vismara's prototypes of relevant cycles. For any relevant U-path P^{uv}_{xwy} including the vertices x, y and eventually w, as defined in theorem 56, we define the *U-path family* associated with P^{uv}_{xwy} as follows:

Definition 58. *The U-path family $\mathcal{F}^{uv}_{xwy}$ of the balanced U-path prototype P^{uv}_{xwy} with endpoints u, v and middle piece (x, y) or (x, w, y), respectively, is:*

$$\mathcal{F}^{uv}_{xwy} = \left\{ P' \in \mathcal{U}_{\mathcal{R}} \;\middle|\; \begin{array}{ll} \multicolumn{2}{l}{|P'| = |P^{uv}_{xwy}| \text{ and } P' \text{ consists of:}} \\ \bullet & \text{the vertices } u \text{ and } v \\ & \text{and the edge } (x, y) \text{ or the path } (x, w, y) \\ \bullet & \text{two shortest paths } (u \dots x) \text{ and } (v \dots y) \end{array} \right\}$$



even U−path *odd U−path*

Hence, two U-path P^{uv}_{xwy} and P' belonging to $\mathcal{F}^{uv}_{xwy}$ only differ on the shortest paths from u to x and from v to y that they include. It is clear that replacing any one of the two shortest paths $(u \dots x)$ or $(v \dots y)$ in definition 58 by an alternative one leads to another balanced short U-path P' differing from the original ones by $C_{ab} = P_{ab} \oplus P'_{ab}$, which is a cycle or a disjoint union of cycles. By construction each of the paths P_{ab} is strictly shorter than $P\{\dots\}/2$, whence $|C_{ab}| < |P\{\dots\}|$. Thus lemma 27 implies $P\{\dots\}$ is a relevant U-path either for all choices or for no choice of the shortest paths $(u \dots x)$ or $(v \dots y)$ in definition 58.

Theorem 59. *The set of all the relevant cycle and U-path families defines a partition of $\mathcal{U}_{\mathcal{R}}$.*

Proof. The proof for the relevant cycle families is given in [149].

By definition 58, the U-path family associated with a relevant U-path prototype P^{uv}_{xwy} is determined by the two U-nodes u and v, the two shortest paths $(u \dots x)$ and $(v \dots y)$ and the edge (x, y) or the pair of edges (x, w) and (w, y), respectively. Replacing one of the two shortest paths leads to the other relevant U-paths contained in this family. Since by theorem 56 each relevant U-path is uniquely defined of his endpoints u and v and the pair (x, y) or triplet (x, w, y) respectively, it is contained in exactly one family. ☐

Vismara [149] showed that the number of relevant cycle families is always polynomial, this is also true for the U-path families.

The algorithm to compute the U-path prototypes is based on the converse of theorem 56. This converse is not necessarily true but it gives a strong condition on U-path relevance.

Algorithm 6 Calculation of the initial set $\mathcal{U}_{\mathcal{I}}$

for all $(u, v) \in U$ **do**
  $\forall x \in V$ find ux- and vx-path
  /* calculate even prototypes: */
  for all $w \in V$ **do**
    if $|P_{uw}| = |P_{vw}|$ **then**
      for all $x \in V$ adjacent to w **do**
        for all $y \in V$ adjacent to w **do**
          if $|P_{ux}| + |P_{xw}| = |P_{uw}|$ and $|P_{vy}| + |P_{yw}| = |P_{vw}|$ **then**
            $P\{u, v; w, x, y\} = P_{ux} \oplus \{x, w\} \oplus \{w, y\} \oplus P_{yv}$
  /* calculate odd prototypes: */
  for all $e = \{x, y\} \in E$ **do**
    if $|P_{ux}|, |P_{vy}| < (|P_{ux}| + |P_{xy}| + |P_{yv}|)$ **then**
      $P\{u, v, x, y\} = P_{ux} \oplus \{x, y\} \oplus P_{yv}$
    if $|P_{uy}|, |P_{vx}| < (|P_{uy}| + |P_{yx}| + |P_{xv}|)$ **then**
      $P\{u, v, y, x\} = P_{uy} \oplus \{y, x\} \oplus P_{xv}$

The set $\mathcal{U}_{\mathcal{R}}$ are calculated in the following way:

- compute the initial set $\mathcal{U}'_{\mathcal{I}}$ (see Algorithm 6),
- compute the set $\mathcal{C}'_{\mathcal{I}}$ (see Algorithm 2)
- extract the prototypes from $\mathcal{U}'_{\mathcal{I}} \cup \mathcal{C}'_{\mathcal{I}}$ using a greedy procedure (see Algorithm 3)
- generate the set $\mathcal{U}_{\mathcal{R}}$ with a backtracking procedure (see section 3.5 and Algorithm 4).

To generate the U-paths from the U-path prototypes P^{uv}_{xwy}, first the path $(u \ldots x)$ in P^{uv}_{xwy} is replaced by each one of the resulting paths from `ListPaths`$(u, x, \emptyset)$ (Algorithm 4). Then, in each of the resulting U-path, the path $(v \ldots y)$ is replaced by each one of the paths returned by the call of `ListPath`$(v, y, \emptyset)$ (Algorithm 4). So each U-path in $\mathcal{F}^{uv}_{xwy}$ corresponds to a pair of paths $(u \ldots x), (v \ldots y)$.

The prototypes for the relevant $\mathfrak{U}$-elements can be computed by augmenting Vismara's algorithm 1 by Algorithm 6. The following greedy step on the collection of all balanced cycles and U-paths remains unchanged. Vismara [149] showed that the relevant cycle families can be computed in $\mathcal{O}(|E|^2|V|)$ steps. There are at most $|U|^2|E|$ families of relevant U-path, hence the algorithm remains polynomial. The relevant cycles and U-paths can be generated from the prototypes as described in [149].

Chapter 6

Interchangeability of Relevant Cycles

6.1 A Partition of $\mathcal{R}$

In order to simplify the notation we shall write

$$\bigoplus_{\mathcal{X}} = \bigoplus_{C \in \mathcal{X}} C \tag{6.1}$$

for $\mathcal{X} \subseteq \mathcal{P}(E)$. For a given length l we define $\mathcal{R}_< = \{C \in \mathcal{R} \big| |C| < l\}$ and $\mathcal{R}_= = \{C \in \mathcal{R} \big| |C| = l\}$.

Lemma 60. *For each relevant cycle $C \in \mathcal{R}$, exactly one of the following statements holds:*

(i) *C is essential, or*

(ii) *There is a cycle $C' \in \mathcal{R}$, $C' \neq C$, and a set of relevant cycles $\mathcal{X} \subseteq \mathcal{R} \setminus \{C, C'\}$ such that $\mathcal{X} \cup \{C'\}$ is linearly independent, $|C| = |C'|$, $|C''| \leq |C|$ for all $C'' \in \mathcal{X}$, and $C = C' \oplus \bigoplus_{\mathcal{X}}$.*

Proof. Let $\mathcal{Y} = \{C'' \in \mathcal{R} \big| |C''| \leq |C|\}$. If $\operatorname{rank}(\mathcal{Y}) > \operatorname{rank}(\mathcal{Y} \setminus \{C\})$, then C is contained in every minimum cycle basis as an immediate consequence of the matroid property (Theorem 35). In other words, C is essential.

Now assume $\operatorname{rank}(\mathcal{Y}) = \operatorname{rank}(\mathcal{Y} \setminus \{C\})$. Hence $C = \bigoplus_{\mathcal{Z}}$ for some $\mathcal{Z} \subseteq \mathcal{Y} \setminus \{C\}$. Without loss of generality we may assume that $\mathcal{Z}$ is an independent set of cycles. By the relevance of C, $\mathcal{Z}$ cannot consist only of cycles that are all strictly shorter than C, thus there is $C' \in \mathcal{Z}$ such that $|C'| = |C|$, and we can write

$$C = C' \oplus \bigoplus_{\mathcal{Z} \setminus \{C'\}}. \tag{6.2}$$

It remains to show that C is not essential in this case: Adding $C \oplus C'$ to both sides of equ.(6.2) yields $C' = C \oplus \bigoplus_{\mathcal{Z}\setminus\{C'\}}$. Thus we may extract two different minimum cycle bases from $\mathcal{R}$ one of which contains C but not C', while the other contains C' but not C, simply by ranking C before or after C' when sorting $\mathcal{R}$. Thus neither C nor C' is essential. $\square$

Definition 61. *Two relevant cycles $C', C'' \in \mathcal{R}$ are* interchangeable*, $C' \leftrightarrow C''$, if (i) $|C'| = |C''|$ and (ii) there exists a minimal linearly dependent set of relevant cycles that contains C' and C'' and with each of its elements not longer than C'.*

Theorem 62. *Interchangeability is an equivalence relation on $\mathcal{R}$.*

Proof. Trivially, we have $C \leftrightarrow C$; symmetry follows immediately from the proof of lemma 60.

Let us fix a length l. Then two cycles C_{j_1} and C_{j_2} of length l are interchangeable if and only if the equation

$$x_1C_1 \oplus \cdots \oplus x_MC_M \oplus \cdots \oplus x_{j_1}C_{j_1} \oplus \cdots \oplus x_{j_2}C_{j_2} \oplus \cdots \oplus x_NC_N = 0 \tag{6.3}$$

has a solution with $x_{j_1} = x_{j_2} = 1$ and with the following properties:

(1) $\{C_1, \ldots C_M\}$ is the intersection of $\mathcal{R}_<$ with an arbitrary but fixed minimum cycle basis, and $\{C_{M+1}, \ldots, C_N\} = \mathcal{R}_=$. The fact that instead of $\mathcal{R}_<$ we can restrict ourselves to a subset of a minimum cycle basis follows from the matroid property (Theorem 35).

(2) The solution is minimal in the following sense: if we take any strict subset of the coefficients with $x_k = 1$ then there is no solution with exactly these coefficients being nonzero. This is equivalent to the fact that we have a minimally linearly dependent set of cycles.

Let $A = (C_1, \ldots, C_M, C_{M+1}, \ldots, C_N)$ be the $(|E| \times N)$-matrix with the cycles C_k represented as column vectors. A can be transformed into the reduced row echelon form $\tilde{A}$ by Gauß-Jordan elimination. Then exactly the first $R = \operatorname{rank}(A)$ rows of $\tilde{A}$ are nonzero. Notice that the upper-left $M \times M$-matrix of $\tilde{A}$ is the identity matrix since $\{C_1, \ldots, C_M\}$ is a subset of a cycle basis by construction, see Tab. 6.1.

We introduce a coloring of the columns $M+1, \ldots, N$ of $\tilde{A}$:

(1) Two columns j' and j'' $(> M)$ have the same color if there exists a row i such that $\tilde{A}_{ij'} = \tilde{A}_{ij''} = 1$.

(2) Use as many colors as possible.

Definition 63. *Two relevant cycles $C', C'' \in \mathcal{R}$ are* color-related*, if (i) $|C'| = |C''|$ and (ii) they have the same color (as described above).*

It is clear from the definition that color-related is an equivalence relation. The definition of color-relatedness, however, depends explicitly on a prescribed ordering of the cycles $C_{M+1}, \ldots, C_N$. We proceed by showing that color-relatedness is in fact independent of this ordering and that it is equivalent to interchangeability.

C_1					C_M					C_{j_1}						C_{j_2}			
1	0	0	0	0	0	0	0	0	1	1	0	0	0						
0	1	0	0	0	0	0	0	0	0	1	0	1	0			0			
0	0	1	0	0	0	0	0	0	1	1	0	1	0						
0	0	0	1	0	0									0	0	0	1	0	1
0	0	0	0	1	0				0					0	0	0	1	1	1
0	0	0	0	0	1									0	0	0	1	0	0
						1	0	0	0	1	0	1	0						
						0	1	0	1	1	0	0	0						
		0				0	0	1	1	0	0	1	0			0			
						0	0	0	0	0	1	1	0						
						0	0	0	0	0	0	0	1						
														1	0	0	1	1	0
		0							0					0	1	0	0	1	1
														0	0	1	1	1	1

Table 6.1. Example of a reduced echelon form $\tilde{A}$ for the special case where the cycles of each color-equivalence class are consecutive in the chosen ordering. For the general case the situation is analogous with columns and rows permutated.

Lemma 64. *If two cycles C_{j_1} and C_{j_2} are interchangeable w.r.t. any ordering of the cycles then C_{j_1} and C_{j_2} are color-related.*

Proof. Fix an arbitrary ordering of the cycles and assume that two interchangeable cycles C_{j_1} and C_{j_2} are not color-related. Let $\mathcal{J}_1$ and $\mathcal{J}_2$ such that $\{C_i : i \in \mathcal{J}_1\}$ and $\{C_i : i \in \mathcal{J}_2\}$ are the respective color-equivalence classes of C_{j_1} and C_{j_2}. Then there is no row r in $\tilde{A}$ with two coefficients $\tilde{A}_{rk_1} = \tilde{A}_{rk_2} = 1$ such that $k_1 \in \mathcal{J}_1$ and $k_2 \in \mathcal{J}_2$, see Tab. 6.1.

Now suppose C_{j_1} and C_{j_2} are interchangeable. Then there exists a minimal solution of equ. (6.3) with $x_{j_1} = x_{j_2} = 1$. Set $x_k = 0$ for all $k \in \mathcal{J}_2$ in this solution (this includes $x_{j_2} = 0$). If the resulting vector (x'_i) is a solution of equ. (6.3), the original solution was not minimal, contradicting the assumption that C_{j_1} and C_{j_2} were interchangeable.

Hence we assume that the resulting vector (x'_i) may not be a solution any more. This happens when there is a row r with an odd number of coefficients $\tilde{A}_{rn}$ for which $x'_n \tilde{A}_{rn} = 1$. In this case, however, we must have $r \leq M$ and $x'_r \tilde{A}_{rr} = 1$. Hence we can set $x'_r = 0$, since the upper-left $M \times M$-matrix is the identity matrix. Since this holds for every such row r we end up with a new solution (x''_i) of equ. (6.3) with $x''_{j_1} = 1$

and $x''_{j_2} = 0$. Again the original solution (x_i) was not minimal, a contradiction to our assumption. □

Lemma 65. *If two cycles C_{j_1} and C_{j_2} are color-related w.r.t. a given ordering of the cycles, then C_{j_1} and C_{j_2} are interchangeable.*

Proof. Assume C_{j_1} and C_{j_2} are color-related and let $\mathcal{J}$ denote the set of indices of the cycles C_i in the color-equivalence class of C_{j_1}. Then there exists a sequence $\sigma = \{j_1 = k_0, k_1, \ldots, k_m = j_2\} \subseteq \mathcal{J}$, such that for each $i = 0, \ldots, m-1$ there exists a row r with $\tilde{A}_{r,k_i} = \tilde{A}_{r,k_{i+1}} = 1$ (otherwise the cycles C_{k_i} would not be color-related). Assume that our sequence is minimal (in the sense that no other sequence connecting j_1 and j_2 consists of fewer elements).

Set all $x_{k_i} = 1$ for $k_i \in \sigma$ and $x_p = 0$ for all other $p > M$. Then for each row $r > M$ there are only two (or zero) columns with $x_k \tilde{A}_{rk} = 1$ (i.e., $\neq 0$). If there were more such columns, say at k_1, k_3, k_9, then σ would not be minimal, since we could then remove $k_2, \ldots, k_8$ from σ. By the same argument there are at most two columns with $x_k \tilde{A}_{rk} = 1$ for $r \leq M$. For the rows $r \leq M$ with only one such column we set $x_r = 1$ and $x_r = 0$ otherwise. Thus (x_i) is a solution of equ. (6.3). Moreover (x_i) has the property that for each row r there are either 2 or 0 columns with $x_k \tilde{A}_{rk} = 1$ and for each column k there are either 2 or 0 rows with $x_k \tilde{A}_{rk} = 1$.

Now we show that this solution is minimal. If we change one of these x_k from 1 to 0 then we obtain a row r with an odd number of coefficients with $x_k \tilde{A}_{rk} = 1$, i.e., we do not have a solution any more. Thus, if we want to construct a new solution (x'_i) of equ. (6.3) by changing x_j from 1 to 0 we have to change the other x_i in row r with $x_i \tilde{A}_{ri} = 1$ from 1 to 0 as well. If we still find a row r' with an odd number of coefficients with $x_n \tilde{A}_{r'n} = 1$ we have to repeat this procedure. As a consequence, if $\tilde{A}_{r,k_i} = \tilde{A}_{r,k_{i+1}} = 1$ and $x_{k_i} = x_{k_{i+1}} = 1$ then any modified solution (x'_i) must satisfy $x'_{k_i} = x'_{k_{i+1}}$ and therefore all coordinates x_k for $k \in \sigma$ must be equal, i.e., either $(x'_i) = (x_i)$ or (x'_i) is the trivial solution. Hence the original solution was minimal. □

It follows that color-relatedness is independent of the ordering the cycles and the particular reduced echelon form $\tilde{A}$ that we have obtained by Gauß-Jordan elimination. Furthermore, color-relatedness and interchangeability are equivalent. Hence the theorem follows. □

Remark. The proofs of lemmata 64 and 65 explicitly uses the properties of a vector space over GF(2).

Corollary 66. *A relevant cycle C is essential if and only if it is not $\leftrightarrow$-interchangeable with any other cycle.*

Remark. We cannot assume that for the set $\mathcal{X} \subset \mathcal{R}$ in lemma 60, $\mathcal{X} \cup \{C'\}$ is a subset of a minimum cycle basis. Figure 6.1 gives a counter-example. In what follows let C_F, C'_F, C_G and C'_G denote the relevant cycles of length 6 through F and G, respectively, and let C_O be the cycle $\{O_1, \ldots, O_6\}$. $\mathcal{Z}$ always denotes an independent subset of $\mathcal{R} \setminus \{C_F, C'_F, C_G, C'_F, C_O\}$. Then $C_F = C_G \oplus (C'_G \oplus C'_F \oplus \bigoplus_{\mathcal{Z}})$, where the right hand side is linearly independent, i.e., $C_F \leftrightarrow C_G$. However, the r.h.s. contains both C_G and C'_G and hence it is not a subset of a minimum cycle basis. Moreover, C_F cannot be expressed as an $\oplus$-sum of an independent subset of relevant cycles that contains C_G but not C'_G.

The graph in figure 6.1 also demonstrates, that we cannot define a "stronger" interchangeability relation, $\leftrightarrow_s$, by replacing the condition that $\mathcal{X} \cup \{C'\}$ is independent by "$\mathcal{X} \cup \{C'\}$ is a subset of a minimum cycle basis" in definition 61. The relation $\leftrightarrow_s$ is not symmetric: We find $C_F = C_O \oplus (C'_F \oplus \bigoplus_{\mathcal{Z}})$, where the r.h.s. is a subset of a minimum cycle basis, i.e., $C_F \leftrightarrow_s C_O$. However, we always have $C_O = C_F \oplus (C'_F \oplus \bigoplus_{\mathcal{Z}})$ where the r.h.s is not a subset of a minimum cycle basis.

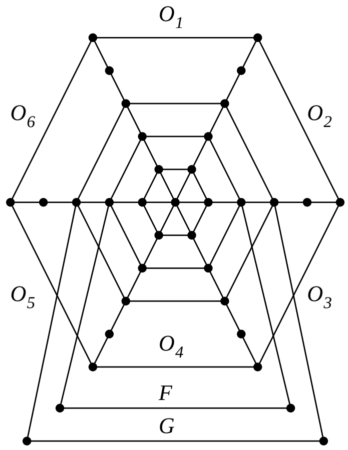


Figure 6.1. The set of relevant cycles of this graph consists of all triangles, all 4-cycles, two 6-cycles through F, two 6-cycles through G and the seven 6-cycles through at least one of the edges O_i. The three inner hexagons (thick lines) are not relevant, because they are the sum of triangles and 4-cycles. Notice that all 3- and 4-cycles are essential. Moreover, every minimum cycle basis contains exactly one 6-cycle through F and G, respectively, and six of the seven 6-cycles through at least one of the edges O_i. Moreover, no 6-cycle is essential.

Lemma 67. *Let C be a relevant cycle such that $C = \bigoplus_{\mathcal{X}}$ for a linearly independent set $\mathcal{X}$ of cycles with length less or equal $|C|$. Set $\mathcal{X}_= = \{C' \in \mathcal{X} \big| |C'| = |C|\}$. Then $C' \leftrightarrow C$ for each cycle $C' \in \mathcal{X}_=$.*

Proof. By corollary 66 $C \leftrightarrow C$. Assume there exists a $C' \in \mathcal{X}_= \setminus \{C\}$. Then $C' = C \oplus \bigoplus_{\mathcal{X}_= \setminus \{C\}}$, i.e., $C' \leftrightarrow C$ as proposed. ☐

Lemma 68. *Let $\mathcal{M}$ be a minimum cycle basis and let $\mathcal{W}$ be an $\leftrightarrow$-equivalence class of $\mathcal{R}$. Then $\mathcal{M} \cap \mathcal{W} \neq \emptyset$.*

Proof. Suppose there is a minimum cycle basis $\mathcal{M}$ and an $\leftrightarrow$-equivalence class $\mathcal{W}$ such that $\mathcal{W} \cap \mathcal{M} = \emptyset$. Choose $C \in \mathcal{W}$. By the matroid property there is an independent set of cycles $\mathcal{Q} = \mathcal{Q}_= \cup \mathcal{Q}_< \subseteq \mathcal{M}$ such that $C = \bigoplus_\mathcal{Q}$. By lemma 67 we have $\mathcal{Q}_= \subseteq \mathcal{W}$ which contradicts $\mathcal{M} \cap \mathcal{W} = \emptyset$ unless $\mathcal{Q}_= = \emptyset$. Thus $C = \bigoplus_{\mathcal{Q}_<}$ and hence $C \notin \mathcal{M}$ by proposition 38. ☐

Theorem 69. *Let $\mathcal{M}$ and $\mathcal{M}'$ be two minimum cycle bases and let $\mathcal{W}$ be an $\leftrightarrow$-equivalence class of $\mathcal{R}$. Then $|\mathcal{M} \cap \mathcal{W}| = |\mathcal{M}' \cap \mathcal{W}|$.*

Proof. Consider an $\leftrightarrow$-equivalence class $\mathcal{W}$ consisting of cycles of length l. Define $\mathcal{M}_= = \{C \in \mathcal{M} \big| |C| = l\}$, $\mathcal{M}_< = \{C \in \mathcal{M} \big| |C| < l\}$, and analogously for the second basis $\mathcal{M}'$. Assume $|\mathcal{M}' \cap \mathcal{W}| > |\mathcal{M} \cap \mathcal{W}|$ and set $\mathcal{W} \cap \mathcal{M} = \{C_1, \ldots, C_j\}$, $\mathcal{W} \cap \mathcal{M}' = \{D_1, \ldots, D_j, \ldots, D_k\}$. By lemma 68, $j > 0$. As a consequence of the matroid property we may assume $\mathcal{M}'_< = \mathcal{M}_<$ and we may write each D_i as a linear combination of cycles from $\mathcal{M}_< \cup \mathcal{M}_=$. Moreover by lemma 67 this linear combination cannot contain any cycles from $\mathcal{M}_= \setminus \mathcal{W}$. Since there are more than j cycles D_i there is a non-trivial linear combination

$$F = \bigoplus_{i \in I} D_i = \left[\bigoplus_{i \in J} C_i\right] \oplus \bigoplus\nolimits_{\mathcal{X} \subseteq \mathcal{M}'_<}$$

with $I \subseteq \{1, \ldots, k\}$ and $J \subseteq \{1, \ldots, j\}$ such that $\bigoplus_{i \in J} C_i = \mathsf{0}$. Thus

$$\left[\bigoplus_{i \in I} D_i\right] \oplus \bigoplus\nolimits_{\mathcal{X} \subseteq \mathcal{M}'_<} = \mathsf{0}$$

and hence $\{D_i | \in I\} \cup \mathcal{X} \subseteq \mathcal{M}'_= \cup \mathcal{M}'_<$ is linearly dependent, contradicting the assumption that $\mathcal{M}'$ is a basis. ☐

As an immediate consequence of theorem 69 we recover the well known fact [24, Thm. 3] (see Chapter 3.3), that any two minimum cycle bases contain the same number of cycles with given length.

Definition 70. *Let $\mathcal{M}$ be a minimum cycle basis and let $\mathcal{W}$ be an $\leftrightarrow$-equivalence class of $\mathcal{R}$. We call* $\operatorname{knar}(\mathcal{W}) = |\mathcal{M} \cap \mathcal{W}|$ *the* relative rank *of $\mathcal{W}$ in $\mathcal{R}$.*

Corollary 71. *Let $\mathcal{W}$ be an $\leftrightarrow$-equivalence class such that* $\operatorname{knar}(\mathcal{W}) = k$. *Then each $C \in \mathcal{W}$ can be written as $C = \bigoplus_\mathcal{Y} \oplus \bigoplus_\mathcal{Z}$ where $\mathcal{Z}$ consists only of cycles shorter than $|C|$ and $\mathcal{Y} \subseteq \mathcal{W} \setminus \{C\}$ has cardinality* $|\mathcal{Y}| \leq \operatorname{knar}(\mathcal{W})$.

6.2 Some Examples

Complete graphs. The relevant cycles of a K_n, $n \geq 3$, are its triangles. It follows immediately that all triangles are $\leftrightarrow$-equivalent.

Outerplanar graphs. Outerplanar graphs have a unique minimum cycle basis [93], i.e., each relevant cycle is essential. Thus there are $\nu(G)$ interchangeability classes consisting of a single cycle.

Triangulations. For each triangulation of the sphere all relevant cycles of the corresponding graph are triangles. Moreover, The $\oplus$-sum of all triangles equals $\mathsf{0}$, while any proper subset is independent. Thus there is a single $\leftrightarrow$-equivalence class with $\mathrm{knar}\,(\mathcal{W}) = |\mathcal{R}| - 1$.

If we change the situation a little bit, such that there is exactly one face cycle C of length $l > 3$, i.e., the graphs corresponds to a triangulation of the plane but not the sphere, then C is the $\oplus$-sum of all triangles and hence not relevant. Thus all triangles are essential, i.e., we have $|\mathcal{R}|$ $\leftrightarrow$-equivalence classes, all of $\mathrm{knar}\,(\mathcal{W}) = 1$. This example demonstrates that partitioning into $\leftrightarrow$-equivalence classes — similar to number and length of minimum cycle bases — can be very unstable against small changes in the geometry of graphs.

Chordal graphs. The next example shows that there are rather "irregular-looking" examples for which all relevant cycles are contained in the same $\leftrightarrow$-equivalence class. A graph is *chordal* (also called triangulated or rigid circuit) if all cycles of length $|C| \geq 4$ contain a chord, i.e., an edge connecting two of its non-adjacent vertices.

Let G be connected and let A be a minimal separating vertex set. Then there are two connected graphs $G_i = (V_i, E_i)$, $i = 1, 2$ such that $V = V_1 \cup V_2$, $E = E_1 \cup E_2$, and $A = V_1 \cap V_2$. If $\Sigma = (A, E_1 \cap E_2)$ is a complete graph, $G_1 \cup G_2$ is called a *simplicial decomposition* of G at A. This procedure can be repeated until no further separating complete graphs can be found. It can be shown that the resulting indecomposable subgraphs are independent of the order of the decomposition [140, Prop.4.1]. The resulting components are the *simplicial summands* of G. A graph is chordal if and only if all its simplicial summands are complete graphs [32].

Lemma 72. *If G is a 3-connected chordal graph then $\mathcal{R}$ consists of a single $\leftrightarrow$-equivalence class.*

Proof. Since $C \in \mathcal{R}$ only if it is chord-less, it follows that all relevant cycles of a chordal graph are triangles. If G is 3-connected, the minimum separating clique Σ contains a triangle. Let G_1 and G_2 be the two adjacent simplicial summands. Then all triangles in G_1 are contained in a single $\leftrightarrow$-equivalence class; the same is true for all triangles in G_2. Since the intersection of G_1 and G_2 contains at least one triangle by assumption, all triangles of their union are contained in the same $\leftrightarrow$-equivalence class, and the lemma follows by induction. $\square$

Lemma 73. *Let $\mathcal{F} \subseteq \mathcal{R}$ be a relevant cycle family, and let $\mathcal{M}$ be a minimum cycle basis. Then for all $C, C' \in \mathcal{F}$ there is an independent set $\mathcal{Y} \subseteq \mathcal{M}$ such that $C \oplus C' =$*

$\bigoplus_{\mathcal{Y}}$ *and* $|C''| < |C| = |C'|$ *for all* $C'' \in \mathcal{Y}$.

Proof. Let P, P' and Q, Q' be the paths connecting (r,p) and (r,q) in C and C', respectively. Then each of the combinations of paths $\{P,Q\}$, $\{P',Q\}$, $\{P,Q'\}$, $\{P',Q'\}$ belongs to a (possibly generalized) cycle in $\mathcal{F}$, which we denote by $C = C_{PQ}$, $C_{P'Q}$, $C_{PQ'}$, and $C' = C_{P'Q'}$ as outlined in [149]. Explicitly we have $C_{PQ} = P \oplus Q \oplus \{p,q\}$ if $|C|$ is odd and $C_{PQ} = P \oplus Q \oplus \{p,x\} \oplus \{x,q\}$ if $|C|$ is odd, etc. Note that the cycles $C_{P'Q}$ and $C_{PQ'}$ are not necessarily connected. Since P and P' have the same end points, their sum $P \oplus P'$ is an edge-disjoint union of cycles, which we denote by $\mathcal{A}$. Thus $C = C_{P'Q} \oplus \bigoplus_{\mathcal{A}}$ and analogously we obtain $C' = C_{P'Q'} = C_{P'Q} \oplus \bigoplus_{\mathcal{A}'}$, and thus $C' = C \oplus \bigoplus_{\mathcal{A}\triangle\mathcal{A}'}$. Since each cycle $C'' \in \mathcal{A}\triangle\mathcal{A}'$ satisfies $|C''| \leq 2d(r,p) = 2d(r,q) < |C|$, it follows from the matroid property that C'' can be written as an $\oplus$-sum of basis elements taken from $\mathcal{Y}$. ☐

Corollary 74. *For each relevant cycle family* $\mathcal{F}$ *there is an* $\leftrightarrow$*-equivalence class* $\mathcal{W}$ *such that* $\mathcal{F} \subseteq \mathcal{W}$.

Two cycles C, C' are *homotopic*, if there is a set $\mathscr{T}$ of triangles such that $C \oplus C' = \bigoplus_{T\in\mathcal{T}} T$ [39]. Obviously homotopic cycles belong to the same interchangeability class.

6.3 The Number of Minimal Cycle Bases

As an application of the $\leftrightarrow$-partition of $\mathcal{R}$ we derive bounds on the number of distinct minimum cycle bases of G.

Theorem 75. *Let* $\mathcal{R} = \bigcup_{i=1}^{m} \mathcal{W}_i$ *be the partition of the set of relevant cycles into* $\leftrightarrow$*-equivalence classes. Then the number* M *of distinct minimum cycle bases satisfies*

$$\prod_{i=1}^{m} |\mathcal{W}_i| \leq M \leq \prod_{i=1}^{m} \binom{|\mathcal{W}_i|}{\operatorname{knar}(\mathcal{W}_i)}. \tag{6.4}$$

Proof. The lower bound follows from the fact that, by lemma 68, each minimum cycle basis contains at least one element from each $\leftrightarrow$-equivalence class, and the fact that, by the matroid property, each element of $\mathcal{W}_i$ can be chosen. The upper bound follows directly from theorem 69 by assuming that the $\operatorname{knar}(\mathcal{W}_i)$ basis elements from $\mathcal{W}_i$ can be chosen freely. ☐

There even exists a universal bound that depends only on the number of relevant cycles and the cyclomatic number.

Corollary 76. *The number M of distinct minimum cycle bases satisfies*

$$M \leq \binom{|\mathcal{R}|}{\nu(G)}. \tag{6.5}$$

Proof. This upper bound follows immediately if we neglect any restrictions for the choice of $\nu(G)$ relevant cycles for a minimum cycle basis. $\square$

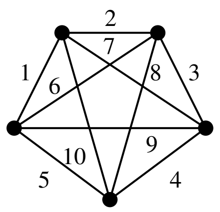
Corollary 77. *Upper and lower bound coincide in equ.(6.4) if all $\leftrightarrow$-equivalence classes satisfy* $\operatorname{knar}(\mathcal{W}) = 1$ *or* $\operatorname{knar}(\mathcal{W}) = |\mathcal{W}| - 1$.

It is tempting to speculate that the upper bound might be attained by all graphs. Equivalently, then we could choose $\operatorname{knar}(\mathcal{W})$ cycles from $\mathcal{W}$ without restrictions when extracting a minimum cycle basis from $\mathcal{R}$. Unfortunately, this is not the case as the following examples show.

The triangles of K_5. Figure 6.2 lists the 10 triangles of K_5. Each triangle is contained in two of the five induced K_4-subgraphs a to e. Thus there are 5 dependent four-sets of cycles:

$$A \oplus B \oplus G \oplus J = 0 \quad B \oplus C \oplus F \oplus H = 0 \quad A \oplus E \oplus F \oplus I = 0$$
$$C \oplus D \oplus G \oplus I = 0 \quad D \oplus E \oplus H \oplus J = 0$$

It is clear that all 10 cycles A through J are $\leftrightarrow$-equivalent forming a single equivalence class with $\operatorname{knar}(\text{triangles}) = \nu(K_5) = 6$. In general, it is clear that all triangles of a complete graph K_n, $n \geq 3$, belong to a single $\leftrightarrow$-equivalence class.



	1	2	3	4	5	6	7	8	9	10	a	b	c	d	e
A	*	*				*					*	*			
B		*	*				*				*				*
C			*	*				*						*	*
D				*	*				*				*	*	
E	*				*					*		*	*		
F		*						*		*		*			*
G			*			*			*		*			*	
H				*			*			*			*		*
I					*	*		*				*		*	
J	*						*		*		*		*		

Figure 6.2. The 10 triangles of K_5 cover the five sub-K_4s a through e twice.

More importantly, however, 5 of the $\binom{10}{4} = 210$ combinations of 4 cycles and hence at least $5\binom{6}{2} = 75$ of the $\binom{10}{6} = 210$ sets of six triangles are dependent. As a consequence, neither the upper nor the lower bound in equ.(6.4) is an equality for K_5.

Small relative ranks. The final example shows that Corollary 77 cannot be improved even if we restrict ourselves to graphs in which all $\leftrightarrow$-classes have small relative rank, or when only a single $\leftrightarrow$-class has $\operatorname{knar}(W) \geq 1$. The family of graphs in figure 6.3 shows that linearly dependent subsets $\mathcal{V} \subset \mathcal{W}$ with $|\mathcal{V}| \leq \operatorname{knar}(\mathcal{W})$ can be found even for $\operatorname{knar}(\mathcal{W}) = 2$.

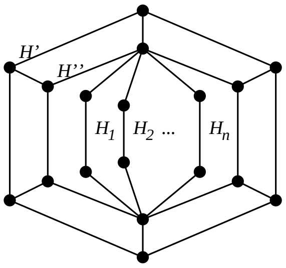


Figure 6.3. The six quadrangles are all essential. All hexagons are in one equivalence class $\mathcal{W}$ with $\operatorname{knar}(\mathcal{W}) = n + 1$ and $|\mathcal{W}| = n^2/2 + 3n/2 + 2$. The outer cycle H' (of length 6) can be expressed as $\oplus$-sum of all quadrangles and the inner hexagon H'' that does not contain any path H_i. Thus no minimum cycle basis can contain both H' and H''.

6.4 Computing Interchangeability Classes

To compute the interchangeability classes the definition 63 and the lemmatas 64 and 65 are used. Table 7 shows an algorithm to compile the classes.

Theorem 78. *The $\leftrightarrow$-equivalence classes of the set $\mathcal{R}$ can be computed by algorithm 7 in $\mathcal{O}(|\mathcal{M}|\,|\mathcal{R}|^2\,|E|)$ operations.*

Proof. Sorting $\mathcal{R}$ and $\mathcal{M} \subseteq \mathcal{R}$ by length requires $\mathcal{O}(|\mathcal{R}| \ln |\mathcal{R}|)$ operations. The Gauß-Jordan elimination requires at most $\mathcal{O}(|\mathcal{R}|^2\,|E|)$ operations. Coloring all the $\tilde{B}$'s needs at most $\mathcal{O}(|\mathcal{R}|\,|E|)$ comparisons. Note that this is only a rather crude upper bound for the worst case. The actual requirements are by far smaller for most graphs. ☐

Algorithm 7 Compute $\leftrightarrow$-partition $\mathfrak{P}$.

Input: $\mathcal{R}, \mathcal{M}$ /* relevant cycles and a minimum cycle basis */
Output: $\mathfrak{P}$ /* interchangeability partition */

Sort minimum cycle basis by length: $\{B_1, \ldots, B_\nu\}$.
Sort relevant cycles by length: $\{C_1, \ldots, C_n\}$.
$\mathfrak{P} \leftarrow \emptyset$.
for each cycle length l **do**
  $\mathcal{M}_< \leftarrow \{B \in \mathcal{M} \big| |B| < l\}$.
  $\mathcal{R}_= \leftarrow \{C \in \mathcal{R} \big| |C| = l\}$.
  $A \leftarrow (\mathcal{M}_<, \mathcal{R}_=)$. /* matrix of cycles */
  $\tilde{A} \leftarrow$ reduced row echelon form of A. /* Gauß-Jordan elimination */
  /* Color columns in submatrix $\tilde{B} = (\tilde{A}_{ij})$, $j > |\mathcal{M}_<|$ */
  Assign each column $j > |\mathcal{M}_<|$ a different color.
  for each row $i = 1, \ldots, \operatorname{rank}(A)$ **do**
    if $\tilde{A}_{ij'} = \tilde{A}_{ij''} = 1$ **then**
      Identify the colors of j' and j''.
  for each color c **do**
    $\mathfrak{P} \leftarrow \mathfrak{P} \cup \{\text{all cycles with color } c\}$.

We have assumed in algorithm 7 that a minimum cycle basis $\mathcal{M}$ is supplied as input since Vismara's algorithm for computing $\mathcal{R}$ also produces a minimum cycle basis. Of course it could be extracted from the Gauß-Jordan eliminations at virtually no extra cost. Algorithm 1 may require exponential time in terms of $|V|$ since the number of relevant cycles may grow exponentially (see Fig. 3.7) [149]. Typically, however, there are only $\mathcal{O}(|V|^3)$ relevant cycles [55].

It is not possible to determine $\leftrightarrow$-equivalence with the set $\widehat{R}$ of "cycle prototypes" that is computed in the first step of Vismara's algorithm. A counterexample is shown in Figure 6.4.

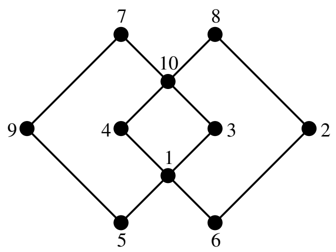


Figure 6.4. The set of relevant cycles consists of $C_0 = (1,3,10,4)$, $C_1 = (4,1,6,2,8,10)$, $C_1' = (3,1,6,2,8,10)$, $C_2 = (4,1,5,9,7,10)$, and $C_2' = (3,1,5,9,7,10)$. We observe $C_0 = C_1 \oplus C_1' = C_2 \oplus C_2'$. Thus $\{C_1, C_1', C_2, C_2'\}$ is an $\leftrightarrow$-equivalence class. Vismara's algorithm identifies C_0,C_1 and C_2 as cycle prototypes. Thus we cannot write $C_1 = C_2 \bigoplus_{\mathcal{Z}}$ such that $\mathcal{Z}$ contains only cycle prototypes.

6.5 Stronger Interchangeability

Definition 79. *Two relevant cycles C, C' are strongly exchangeable, $C \overset{s}{\leftrightarrow} C'$, if there is a set of cycles $\mathscr{C}$, such that $|C| > |C''|$ for all $C'' \in \mathscr{C}$ and $C' = C \oplus \bigoplus_{C'' \in \mathscr{C}} C''$.*

Lemma 80. *Two cycles $C, C' \in \mathcal{R}$ are strongly exchangeable, $C \overset{s}{\leftrightarrow} C'$, iff*

(i) $|C| = |C'|$

(ii) There is a set $\mathscr{C} \subseteq \mathcal{R}$ such that

(a) $|C''| < |C|$ *for all* $C'' \in \mathscr{C}$,

(b) $C \oplus C' = \bigoplus_{C'' \in \mathscr{C}} C''$,

(c) $\{C\} \cup \mathscr{C}$ *is linearly independent.*

Proof. From $C' \in \mathcal{R}$ and $C' = C \oplus \bigoplus_{\mathscr{C}}$, we know that $|C'| \leq |C|$. Adding $C \oplus C'$ on both sides, we get $C = C' \oplus \bigoplus_{\mathscr{C}}$, hence $|C| \leq |C'|$, i.e. $|C| = |C'|$.
Replacing each cycle in $\mathscr{C}$ by a sum of cycles from an MCB and then removing a maximal linearly dependent set, leads to $\mathscr{C} \subseteq \mathcal{R}$.
Suppose $\mathscr{C} \cup \{C\}$ is dependent. Then $C = C_1'' \oplus C_2'' \oplus \cdots \oplus C_n''$ with $C_j'' \in \mathscr{C}$, i.e., C is not relevant, contradicting the definition of $\overset{s}{\leftrightarrow}$. The converse implication is obvious. $\square$

Lemma 81. $\overset{s}{\leftrightarrow}$ *is an equivalence relation.*

Proof. $C \overset{s}{\leftrightarrow} C$ and $C \overset{s}{\leftrightarrow} C'$ if $C' \overset{s}{\leftrightarrow} C$ are trivial.
Suppose $C \overset{s}{\leftrightarrow} C'$ and $C' \overset{s}{\leftrightarrow} C''$. Then there are two sets $\mathscr{C}$ and $\mathscr{C}'$, such that $C = C' \oplus \bigoplus_{\mathscr{C}}$ and $C'' = C''' \oplus \bigoplus_{\mathscr{C}'}$. Hence $C = C''' \oplus \bigoplus_{\mathscr{C} \triangle \mathscr{C}'}$ is also trivial. □

Corollary 82. $C \overset{s}{\leftrightarrow} C'$ *implies* $C \leftrightarrow C'$.

Lemma 83. *An MCB contains at most one of two strong exchangeable cycles.*

Proof. This follows from the definition, that $C = C' \oplus \bigoplus_{\mathscr{C}}$. □

Restricting $\mathscr{C}$ to triangles gives an even finer partition, namely *homotopic* cycles, $C \overset{\equiv}{\leftrightarrow} C'$ [39].

The graph in Fig. 6.5 shows two homotopic cycles. Each of the two squares is the sum of the other square and all triangles.

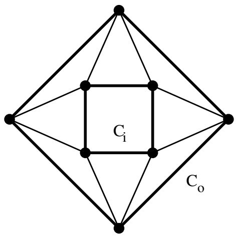


Figure 6.5. Consider the two squares C_i and C_o, then $C_i \overset{s}{\leftrightarrow} C_o$. But since $C_i = C_o \oplus \triangle$, the two squares also homotopic cycles: $C_i \overset{\equiv}{\leftrightarrow} C_o$.

Vismara's cycle families [149] can be viewed as a refinement of strong exchangeability. It is easy to see that all cycles contained in one of Vismara's cycle family are also in the same $\overset{s}{\leftrightarrow}$-class. The converse is not true, see Fig. 6.6.

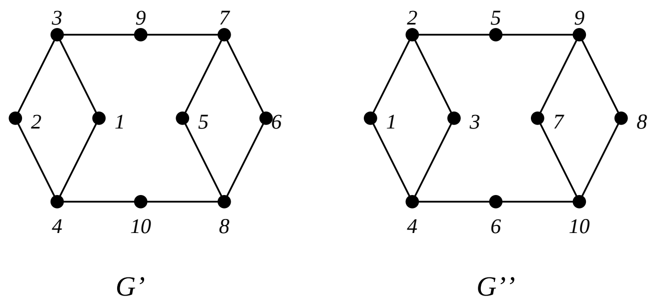


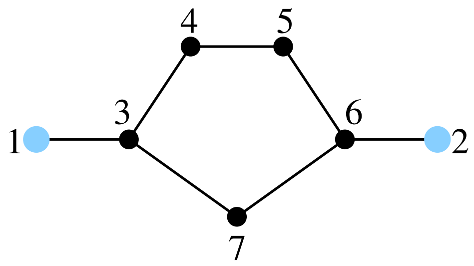
Figure 6.6. The only difference between $\mathcal{G}'$ and $\mathcal{G}''$ is the labeling of the nodes. For both $\mathcal{G}'$ and $\mathcal{G}''$, the 4 relevant 8-edged cycles belongs to one $\leftrightarrow$-class and to one $\overset{s}{\leftrightarrow}$-class. In $\mathcal{G}'$, there are two prototypes for the 8-edges cycles, each representing 2 relevant cycles: $\mathcal{F}^8_{2,9,3} = \{\{8,10,4,2,3,9,7,5\},\{8,10,4,2,3,9,7,6\}\}$ and $\mathcal{F}^8_{1,9,3} = \{\{8,10,4,1,3,9,7,5\},\{8,10,4,1,3,9,7,6\}\}$. But in $\mathcal{G}''$ all 4 relevant 8-edged cycles are contained in one cycle family: $\mathcal{F}^{10}_{1,5,2}$.

6.6 Interchangeability of U-Path

In chapter 5 the cycle space was extended to the U-space. The relevant U-paths can be calculated in a similarly manner as the relevant cycles. The theory about the interchangeability does not depend on the fact that one considers cycles; indeed it works for all finite vector spaces over $GF(2)$ and hence in particular for $\mathfrak{U}$-spaces. Hence we have

Proposition 84. *Let $\mathcal{M}$ be a minimum length $\mathfrak{U}$-basis and let $\mathcal{W}$ be a $\leftrightarrow$-equivalence class of relevant $\mathfrak{U}$-elements. Then $|\mathcal{W} \cap \mathcal{M}|$ is independent of the choice of the basis $\mathcal{M}$.*

It is tempting to speculate that the $\leftrightarrow$-partition might distinguish between cycles and paths. As the example below shows, however, this is not the case:



Here $U = \{1, 2\}$ and the relevant $\mathfrak{U}$-elements are the paths $P_1 = (1, 3, 7, 6, 3)$, $P_2 = (1, 3, 4, 5, 6, 2)$, and the cycle $C = (3, 4, 5, 6, 7, 3)$. with $|P_1| = 4$ and $|P_2| = |C| = 5$. Furthermore $C = P_2 \oplus P_1$, i.e., the path P_2 and the cycle C belong to the same $\leftrightarrow$-equivalence class.

If so, can a minimum length U-path basis of $\mathfrak{U}^*$ be computed efficiently? This might be of interest in the context of metabolic and other chemical reaction networks.

Chapter 7

Circuit Bases of Digraphs

7.1 Circuit Space

With each simple chain $\mathbf{c}$ in a digraph $\mathcal{G}(V,A)$ we associate a vector C indexed by the arcs in A such that

$$C(e) = \begin{cases} +1 & \exists k : e_k = e \quad \text{and} \quad e_k = (x_{k-1}, x_k) \\ -1 & \exists k : e_k = e \quad \text{and} \quad e_k = (x_k, x_{k-1}) \\ 0 & \text{otherwise} \end{cases} \tag{7.1}$$

In other words, $C(e) = +1$ if $e \in \mathbf{c}$ is traversed by $\mathbf{c}$ in forward direction, $C(e) = -1$ if $e \in \mathbf{c}$ is traversed in backwards direction and $C(e) = 0$ if $e \notin \mathbf{c}$. The vectors associated with simple paths and circuits therefore have no negative entries, see Fig. 7.1 for an example.

The concatenation of two simple chains $\mathbf{c} * \mathbf{c}' = \mathbf{c}''$ is again a simple chain, then we have $C + C' = C''$, since then the arc sets of $\mathbf{c}$ and $\mathbf{c}'$ must be disjoint. It is meaningful therefore to define the vector associated with an arbitrary chain $\mathbf{c}$ as the sum of the vectors associated with its individual steps. In other words $C(e)$ is the number of times in which $\mathbf{c}$ transverses $e \in A$ in forward direction minus the number of times in which $\mathbf{c}$ transverses e in backwards direction.

The *incidence matrix* $\mathbf{H}$ of the digraph $\mathcal{G}$ has the entries $\mathbf{H}_{ex} = +1$ if x is the terminal vertex of the arc e, $\mathbf{H}_{ex} = -1$ if x is the initial vertex of e, and 0 otherwise. The *circuit space* $\mathfrak{C}$ of $\mathcal{G}(V,A)$ is the subspace of $\mathbb{R}^{|A|}$ that is generated by the cycles of $\mathcal{G}(V,A)$. It is well known, see e.g. [14, II.3] that

$$U \in \mathfrak{C} \qquad \Longleftrightarrow \qquad \mathbf{H}U = 0 \tag{7.2}$$

A basis of the circuit space can be constructed just as in the case of undirected graphs (see section 3.1). Let T be a spanning forest of $\mathcal{G}$. For each $e \notin T$ there is a

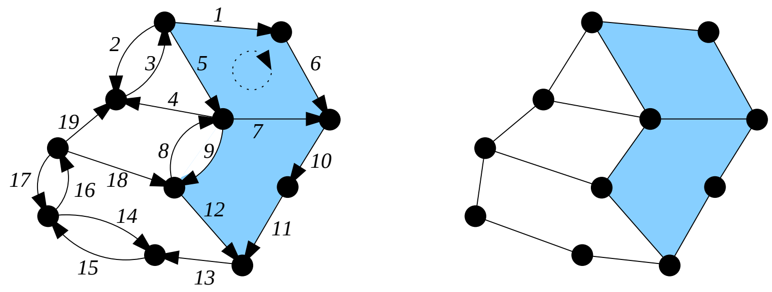


Figure 7.1. The cycle C delimiting the shaded region in the directed graph $\mathcal{G}$ on the l.h.s. and the orientation indicated by the arrow has the vector representation

$$(+1, 0, 0, 0, -1, +1, 0, 0, -1, +1, +1, -1, 0, 0, 0, 0, 0, 0, 0)\,.$$

The r.h.s. shows the graph $\mathcal{G}^\circ$, obtained from $\mathcal{G}$ by one edge of each 2-cycle and ignoring the direction of the arcs. We have for example $R = \{2, 8, 15, 16\}$ or $R' = \{3, 8, 14, 17\}$ as "reverse arcs" that are omitted in passing from $\mathcal{G}$ to $\mathcal{G}^\circ$ (from [54]).

unique cycle $\mathbf{c}^e$ in $T \cup \{e\}$. The cycles $\mathbf{c}^e$ are the *fundamental cycles* associated with the spanning forest T. The set of associated sets of vectors $\mathcal{C}_T = \{C^e | e \in A \setminus T\}$ is a basis of the circuit space $\mathfrak{C}$, see e.g. [10, Thm.3.4]. The dimension of the circuit space is therefore

$$\nu(\mathcal{G}) = |A| - |V| + k(\mathcal{G}^\circ) \tag{7.3}$$

where $k(\mathcal{G}^\circ)$ denotes the number of connected components of $\mathcal{G}^\circ$, i.e., the number of weak components of $\mathcal{G}$. Note that this is the same construction which is used in undirected graphs. Hence we can expect close relationships between the cycle space of the digraph $\mathcal{G}$ and the underlying undirected graph $\mathcal{G}^\circ$.

7.2 Elementary Circuits

From the construction of the basis it follows that $\mathfrak{C}$ has a basis consisting of vectors with coordinates -1, 0, or $+1$. Thus any vector $U \in \mathfrak{C}$ can be written in the form $U = \xi U^*$ with $\xi \in \mathbb{R}_+$ and $U^* \in \mathbb{Z}^{|A|}$; hence it suffices to consider only those elements of the cycle space with integer coordinates.

From a practical point of view those elements of $\mathfrak{C}$ that follow the directions of the arcs in $\mathcal{G}$ are of particular interest. The special role of the circuits is emphasized by the following result:

Lemma 85. *Let* $\mathbf{z}$ *be a closed walk in* $\mathcal{G}$*. Then there is a collection* $\{C_i\}$ *of elementary circuits such that* $Z = \sum_i a_i C_i$ *with* $a_i \in \mathbb{N}$.

Proof. If each vertex in $\mathbf{z}$ is visited exactly once, then $\mathbf{z}$ is an elementary circuit and there is nothing to show. Otherwise $\mathbf{z}$ contains a vertex x that is visited more than once. Let x be the initial and terminal vertex of $\mathbf{z}$. If x is visited in an intermediate step then $\mathbf{z}$ is a concatenation of two closed walks $\mathbf{z_1}$ and $\mathbf{z_2}$ starting from x, and hence $Z = Z_1 + Z_2$. We consider the two parts independently of each other. If such a part again is not simple we continue in the same way as described above. After a finite number of such decompositions x occurs only as initial/terminal vertex of each partial closed walk $\mathbf{z_i}$. Now let z be the first vertex in $\mathbf{z_i}$ that occurs more than once. We have $\mathbf{z_i} = \mathbf{z_i^1} * \mathbf{z_i^2} * \mathbf{z_i^3}$ where $\mathbf{z_i^1}$ is the walk from x to z, $\mathbf{z_i^3}$ is a the part of $\mathbf{z_i}$ after the last occurrence of z, and $\mathbf{z_i^2}$ is the closed walk between the first and last occurrence of x in $\mathbf{z}$. By construction $\mathbf{z_i^1} * \mathbf{z_i^3}$ is a circuit and $\mathbf{z_i^2}$ is a closed walk. This leads to a decomposition of $\mathbf{z}$ into a concatenation of (not necessarily distinct) circuits. Thus the vector Z associated with $\mathbf{z}$ is a sum of circuits with positive integer coefficients. □

It is natural to consider the *non-negative cone* of the circuit space $\mathbb{K} = \{x \in \mathfrak{C} | x_k \geq 0\}$. A vector $u \in \mathbb{K}$ is *extremal* if

$$u = \sum_e \lambda_e v^e \quad \text{with} \quad v^e \in \mathbb{K} \quad \text{and} \quad \lambda_e > 0 \quad \text{implies} \quad v^e = \xi_e u \quad \text{and} \quad \xi_e > 0\,, \tag{7.4}$$

i.e., if u cannot be represented as a positive linear combination of other vectors from the cone $\mathbb{K}$. Denote the *support* of a vector u by $\mathrm{supp}(u) = \{e \in A | u(e) \neq 0\}$.

Lemma 86. *If u is extremal in $\mathbb{K}$ and $v \in \mathbb{K}$ such that $v \neq 0$ and $supp(v) \subseteq supp(u)$; then $v = \mu u$ for some $\mu > 0$.*

Proof. Let $\mu = \min \{u(e)/v(e) \mid e \in \mathrm{supp}(v)\} > 0$. Then $\mu v(e) \leq u(e)$ with equality for at least one $e \in \mathrm{supp}(v)$. Consider $w = u - \mu v$. We have $w \in \mathbb{K}$ since $w(e) \geq 0$ for all $e \in A$. Thus we can write $u = w + \mu v$ as a positive linear combination of vectors in $\mathbb{K}$, contradicting the extremality of u. Thus $w = 0$ and hence $u = \mu v$. □

The following proposition is well known, see e.g. [118].

Proposition 87. *The elementary circuits of $\mathcal{G}$ are exactly the extremal vectors of the cone $\mathbb{K}$.*

Proof. It follows immediately from equ.(7.2) that the subgraph $\mathcal{G}^v$ of $\mathcal{G}$ with arc set $\mathrm{supp}(v)$ for any $v \in \mathbb{K}$ has neither a sink (vertex with out-degree 0) nor a source (vertex with in-degree 0). Therefore $\mathcal{G}^v$ contains a circuit. Consequently, no proper subset of an elementary circuit can be the support of a vector in $\mathbb{K}$, i.e., every elementary circuit is an extremal vector of $\mathbb{K}$.
To see the converse, suppose $v \in \mathbb{K}$ is extremal. Let C be an elementary circuit contained in $\mathrm{supp}(v)$, $\mu = \min_{e \in C} v(e)$, and $v' = v - \mu C$. We have $v'(e) \geq 0$ and hence $v \in \mathbb{K}$. Since v is extremal we must have $v' = 0$ and thus $\mathrm{supp}(v)$ must be an elementary circuit. □

7.3 Circuit Bases

Definition 88. *A* circuit basis *is a basis of the circuit space* $\mathfrak{C}$ *of* $\mathcal{G}(V,A)$ *consisting exclusively of elementary circuits. A* cycle basis *is a basis of the circuit space* $\mathfrak{C}$ *of* $\mathcal{G}(V,A)$ *consisting exclusively of elementary cycles.*

Lemma 85 raises the question under which conditions the circuits generate the circuit space. This question was essentially answered by Berge [10]:

Proposition 89. [10] *A strongly connected digraph* $\mathcal{G}(V,A)$ *has a circuit basis.*

The converse of Proposition 89 is easily obtained:

Theorem 90. *A digraph* $\mathcal{G}(V,A)$ *has a circuit basis if and only if each block is either strongly connected or a single arc.*

Proof. The cycle space of $\mathcal{G}(V,A)$ is the direct sum of the blocks of $\mathcal{G}$. Thus $\mathcal{G}(V,A)$ has a circuit basis if each block has a circuit basis or an empty cycle space. The only blocks with empty cycle space are isolated vertices and pairs of vertices that are connected by a single arc. □

In other words, $\mathcal{G}(V,A)$ has a circuit basis if and only if its strongly connected components are linked together in a tree-like fashion by individual arcs or sequences of individual arcs. Because of this simple structure we shall restrict ourselves to 2-connected digraphs from here on.

Double edges, i.e., circuits of length 2, play a special role, since they are a major difference between graphs and digraphs. For instance, the cyclomatic number of the underlying undirected graph is

$$\nu(\mathcal{G}^\circ) = |A^\circ| - |V| + \mathrm{c}(\mathcal{G}^\circ) = |A| - d^*(\mathcal{G}) - |V| + \mathrm{c}(\mathcal{G}^\circ) = \nu(\mathcal{G}) - d^*(\mathcal{G}) \tag{7.5}$$

where $d^*(\mathcal{G})$ denotes the number of double edges in $\mathcal{G}$.

Lemma 91. *Let* $\mathcal{B}^\circ$ *be a cycle basis of the undirected graph* $\mathcal{G}^\circ$*, and let* $\mathcal{D}(\mathcal{G})$ *be the set of double edges of* $\mathcal{G}$*. Then* $\mathcal{B} = \mathcal{B}^\circ \cup \mathcal{D}(\mathcal{G})$ *is a cycle basis of* $\mathcal{G}$ *with length*

$$\ell(\mathcal{B}) = \ell(\mathcal{B}^\circ) + 2|\mathcal{D}| \tag{7.6}$$

Proof. The cycles in $\mathcal{B}^\circ$ are of course independent cycles of $\mathcal{G}$. At each double edge, we may choose one of the arcs to be part of the $\mathcal{B}^\circ$-cycles that contain the double edge. This shows that the double edges in $\mathcal{D}$ are indeed independent of the set of $\mathcal{B}^\circ$-cycles. Equ.(7.5) hence implies that $\mathcal{B}$ is a cycle basis of $\mathcal{G}$. Equ.(7.6) now follows immediately. □

The following theorem shows that double edges are in a sense superfluous:

Proposition 92. [147] *If $\mathcal{G}(V, A)$ is strongly 2-edge-connected then one can obtain a strongly connected graph $\mathcal{G}^*(V, A^*)$ by removing one of the two arcs of each double edge.*

The main result of this section is a variant of Berge's theorem, Proposition 89.

Theorem 93. *A strongly connected digraph $\mathcal{G}(V, A)$ has a circuit basis consisting of the $d^*(\mathcal{G})$ double edges and $\nu(\mathcal{G}^\circ)$ elementary circuits.*

Proof. We follow the construction of a cycle basis consisting of circuits described in [10, Thm.3.9] and [65] with slight modifications. Clearly the theorem is correct for $|V| \leq 2$. Suppose the assertion is correct for all $k < |V|$. Let $\mathbf{c}^* = (x_0, e_1, x_1, \ldots, x_{h-1}, e_h, x_0)$ be a shortest circuit in $\mathcal{G}$, $h \geq 2$. Such a circuit exists as a consequence of strong connectedness. Clearly, it is elementary. In particular, if $\mathcal{G}$ contains double edges, we choose one of them.

Next we construct a multi-digraph $\mathcal{G}'$ by replacing the set $W = \{x_0, \ldots, x_{h-1}\}$ of vertices of $\mathbf{c}^*$ by a single vertex x' and by replacing each arc (y, z) and (z, y), $y \neq W$, $z \in W$ by an arc from y to x'.

In particular, any double edge in $\mathcal{G}$ (except $\mathbf{c}^*$ itself if it is double edge) becomes to a double edge in $\mathcal{G}'$. This contraction step may lead to multiple parallel arcs incident with a'. The resulting multi-digraph has $|A| - h$ edges and $|V| - |W| + 1 = |V| - h + 1$ vertices, i.e., $\nu(\mathcal{G}') = \nu(\mathcal{G}') - 1$.

Instead of iterating this construction immediately as in the original proofs of Prop. 89 [10, 65] we first take care of the multiple arcs in $\mathcal{G}'$. To this end we select one of the multiple arcs, say g; if one of them is part of a double edge of $\mathcal{G}$ it gets selected first. Let C_g be a shortest circuit through g in $\mathcal{G}'$; Note that if g was part of a double edge in $\mathcal{G}$, then C_g is just this double edge. We store C_g in $\mathcal{C}^*$ and delete the arc g from $\mathcal{G}'$, obtaining a multi-digraph $\mathcal{G}''$. We repeat this procedure until, after removing q arcs, there are no further parallel arcs and we are left with a di-graph $\mathcal{G}^*$. All double edges that have become part of a multiple arcs in $\mathcal{G}'$ are now contained in $\mathcal{C}^*$, all other double edge are passed on as double edges to $\mathcal{G}^*$.

Thus $\mathcal{G}^*$ has $|V| - h + 1$ vertices and $|A| - h - q$ edges, i.e., its cyclomatic number is $\nu(\mathcal{G}^*) = \nu(\mathcal{G}) - 1 - q$. Clearly the circuits in $\mathcal{C}^*$, which are elementary by construction, are independent since each uniquely contains one of the q removed parallel arcs. Consequently, the union $\mathcal{C}^{**}$ of $\mathcal{C}^*$ with any cycle basis of $\mathcal{G}^*$ consists of $\nu(\mathcal{G}) - 1$ independent cycles and hence is a basis of the circuit space of the multi-digraph $\mathcal{G}'$. The induction hypothesis assumes that there is a circuit basis of $\mathcal{G}^*$, hence $\mathcal{C}^{**}$ can be chosen such that it is circuit basis of $\mathcal{G}'$.

Now recall that each edge incident to x' in $\mathcal{G}'$ corresponds to an edge incident with a particular vertex $x_k \in W$. Thus each circuit $\mathbf{c} \in \mathcal{C}^{**}$ is either an elementary circuit in $\mathcal{G}$ if it does not contain x' or it can be lifted to a unique circuit $\hat{\mathbf{c}}$ in $\mathcal{G}$ by replacing x' with the vertices at which $\mathbf{c}$ "enters" and "leaves" $\mathbf{c}^*$ and the unique path within $\mathbf{c}^*$ that connects these two vertices. The set

$$\mathcal{C} = \{\hat{\mathbf{c}}|\mathbf{c} \in \mathcal{C}^{**}\} \cup \{\mathbf{c}^*\} \tag{7.7}$$

contains $\nu(\mathcal{G}') + 1 = \nu(\mathcal{G})$ elementary circuit, among which are all $d^*(\mathcal{G})$ double edges. Finally, consider the equation

$$\sum_{\mathbf{c} \in \mathcal{C}^{**}} a_{\mathbf{c}} C_{\hat{\mathbf{c}}} + a^* C_{\mathbf{c}^*} = 0\,. \tag{7.8}$$

First we note that $C_{\mathbf{c}^*}(e) = 0$ for all $e \in A \setminus \mathbf{c}^*$. Thus, restricting equ.(7.8) to the arcs in $A \setminus \mathbf{c}^*$ and using that the arcs $\mathbf{c} \in \mathcal{C}^{**}$ are linearly independent, we obtain $a_{\mathbf{c}} = 0$ for all $\mathbf{c} \in \mathcal{C}^{**}$. Therefore $a^* C_{\mathbf{c}^*} = 0$, and $\mathcal{C}$ is indeed a set of $\nu(\mathcal{G})$ independent circuits of $\mathcal{G}$. □

For a vector $Z \in \mathfrak{C}$ with integer coordinates we set

$$|Z| = \sum_{e \in A} |Z(e)| \tag{7.9}$$

It follows from lemma 85 that for any closed path $\mathbf{z}$ we have $|Z| = \sum_i a_i |C_i|$ with $a_i \in \mathbb{N}$. Furthermore, we have $|Z| \geq |\mathrm{supp}(Z)|$ with equality if and only if $Z(e) \in \{+1, 0, -1\}$, i.e., iff Z is an edge-disjoint union of cycles. In particular, the elementary circuits are the minimal integer-valued elements of $\mathbb{K}$. Bases of the circuit space with minimum total length

$$\ell(\mathcal{B}) = \sum_{C \in \mathcal{B}} |C| \tag{7.10}$$

consisting of integer-valued vectors, $C(e) \in \mathbb{Z}$, are of particular interest.

7.4 Minimum Circuit Bases and Relevant Circuits

Definition 94. *A* minimum cycle (circuit) basis *is a cycle (circuit) basis of* $\mathfrak{C}$ *with minimal length.*

Theorem 95. *Let* $\mathcal{G}$ *be strongly connected and let* C *be a shortest circuit through an arc* $e \in A$. *Then there is a minimum circuit basis that contains* C. *If* C *is the unique shortest circuit through* e, *then every minimal circuit basis contains* C.

Proof. Suppose $\mathcal{B}$ is a minimal circuit basis, and let $e \in A$. Set $\mathcal{B}^e = \{C \in \mathcal{B} | e \in C\}$ and $\mathcal{B}^* = \mathcal{B} \setminus \mathcal{B}^e$. Suppose C is a shortest circuit containing e, $C \notin \mathcal{B}^e$. Since $\mathcal{B}^* \cup \{C\}$ is obviously an independent set, there exists a circuit $C' \in \mathcal{B}^e$ such that $\mathcal{B}' = \mathcal{B} \cup \{C\} \setminus \{C'\}$ is a circuit basis with length $\ell(\mathcal{B}') = \ell(\mathcal{B}) + |C| - |C'| \leq \ell(\mathcal{B})$, since we have assumed $|C| \leq |C'|$. If C is the unique shortest cycle through e we have $|C| < |C'|$, and hence $\ell(\mathcal{B}') < \ell(\mathcal{B})$, contradicting the minimality of $\mathcal{B}$. Thus $C \in \mathcal{B}$ for every minimal circuit basis. □

The argument used the proof of theorem 95 is the same as in the case of minimum cycle bases of undirected graphs [129].

Corollary 96. *Every minimal circuit basis $\mathcal{B}$ of a strongly connected digraph contains the set $\mathcal{D}$ of double edges.*

Proof. If $e \in A$ is part of a double edge, then the double edge $D = \{e, e'\}$ is the unique shortest circuit containing e. By theorem 95 D is an element of every minimal circuit basis. □

It is sometimes useful to consider undirected graphs as symmetric digraphs, i.e., as digraphs in which $(x, y) \in A$ implies (y, x) in A. The following result shows that minimum cycle bases of undirected graphs and minimum circuit bases of symmetric digraphs are essentially the same.

Theorem 97. *Let $\mathcal{G}$ be a symmetric digraph. Then every minimum circuit basis consists of $\mathcal{D}$ and a set $\mathcal{B}$ of circuits such that $\mathcal{B}^\circ = \{C^\circ | C \in \mathcal{B}\}$ is a minimum cycle basis of the undirected graph $\mathcal{G}^\circ$.*

Proof. It follows from equ.(7.5) that a minimum circuit basis of $\mathcal{G}$ cannot be shorter than $2|\mathcal{D}| + L$, where L is the length of a minimum cycle basis of $\mathcal{G}^\circ$. Conversely, if $\mathcal{B}$ is a minimum circuit basis, then $\mathcal{B} \setminus \mathcal{D}$ is a set of $\nu(\mathcal{G}^\circ)$ independent proper cycles and corresponds to a cycle basis of $\mathcal{G}^\circ$ with the same length.

Now assume that $\mathcal{G}$ is symmetric, i.e., $2|\mathcal{D}| = |A|$. We will show that every minimum cycle basis $\mathcal{B}^\circ$ of $\mathcal{G}^\circ$ can be lifted and extended to a circuit basis of $\mathcal{G}$ with length $L+|A|$, which, as a consequence of the previous paragraph must then be a minimum circuit basis. To this end we identify each edge e of $\mathcal{G}^\circ$ with one of the two arcs of $\mathcal{G}$ forming with the corresponding the double edge. This amounts to lift $\mathcal{B}^\circ$ to the digraph $\mathcal{G}$. Clearly, $\mathcal{B}^* = \mathcal{B}^\circ \cup \mathcal{D}$ is a basis of the circuit space with the minimum possible length. However, the cycles $C \in \mathcal{B}^\circ$ will in general not be circuits.

For each "negative" edge e, $C(e) = -1$, of a basis cycle C, there is a double edge $D = \{e, e'\} \in \mathcal{D}$ such that either $C' = C + D$ or $C' = C - D$ is a cycle that coincides with C except for e, which is replaced by the positive edge e', $C'(e') = +1$. Clearly, C

and C' have the same length and belong to the same cycle C° of the undirected graph $\mathcal{G}^\circ$. Since D is contained in $\mathcal{B}^*$, $\mathcal{B}^{**} = \mathcal{B}^* \cup \{C'\} \setminus \{C\}$ is a basis of the circuit space with the same length. Repeating this argument for all negative edges in C replaces the basis cycle C with a basis circuit $C^>$ of the same length. Note that C and $C^>$ by construction belong to the the same cycle C° of $\mathcal{G}^\circ$. We finally obtain a circuit basis of $\mathcal{G}$ with length $L + |A|$. □

The set of circuits of $\mathcal{G}(V, A)$ forms of course a matroid. A basis of the cycle space with minimum weight can therefore be obtained by means of the greedy algorithm [89] from the set of all circuits.

Definition 98. *Let $(\mathcal{Q}, \mathfrak{J})$ be a matroid and let $|\,.\,| : \mathcal{Q} \to \mathbb{R}^+$ be a non-negative weight function on $\mathcal{Q}$. Then $A \in \mathcal{Q}$ is $|\,.\,|$-relevant if there is a minimum weight basis $\mathcal{B}$ of $(\mathcal{Q}, \mathfrak{J})$ containing A.*

An analogous definition for maximum weight bases of course is also meaningful. Definition 98 is the obvious generalization of the relevant cycles discussed in section 3.4. The set $\mathcal{R}_{|\,.\,|}$ of $|\,.\,|$-relevant circuits can be extracted from $\mathcal{Q}$ by means of the modified greedy algorithm 8.

Algorithm 8 is of course essentially the same as algorithm 3 for the relevant cycles in a graph. We repeat it here for arbitrary weighted matroids.

Algorithm 8 R-Greedy [149]

Input: $(\mathcal{Q}, \mathfrak{J})$, $|\,.\,|$
Output: $\mathcal{R}$ /* Set of $|\,.\,|$-relevant elements. */

Sort $\mathcal{Q}$ by weight: $\{A_1, A_2, \ldots, A_m\}$ /* A_1 with minimal weight. */
$\mathcal{B}_< \leftarrow \emptyset$; $\mathcal{B}_= \leftarrow \emptyset$; $\mathcal{R}_= \leftarrow \emptyset$; $\mathcal{R} \leftarrow \emptyset$;
for $k = 1$ to m **do**
  if $|A_k| > |A_{k-1}|$ **then**
    $\mathcal{R} \leftarrow \mathcal{R} \cup \mathcal{R}_=$; $\mathcal{B}_< \leftarrow \mathcal{B}_< \cup \mathcal{B}_=$;
    $\mathcal{R}_= = \mathcal{B}_= = \{A_k\}$;
  else
    if $\{A_k\} \cup \mathcal{B}_< \in \mathfrak{J}$ **then**
      $\mathcal{R}_= \leftarrow \mathcal{R}_= \cup \{A+_k\}$;
    if $\{A_k\} \cup \mathcal{B}_< \cup \mathcal{B}_= \in \mathfrak{J}$ **then**
      $\mathcal{B}_= \leftarrow \mathcal{B}_= \cup \{A+_k\}$;
$\mathcal{R} \leftarrow \mathcal{R} \cup \mathcal{R}_=$;

Lemma 99. *Algorithm 8,* R-Greedy*, works.*

Proof. Let us write $\mathcal{Z}_{<w} = \{A \in \mathcal{Z} \big| |A| < w\}$ and analogously $\mathcal{Z}_{=w}$, and let $\mathcal{B}$ be a minimum weight basis. Then A is a relevant element if and only if $\mathcal{B}_{<|A|} \cup \{A\} \in \mathfrak{J}$, since we can order $\mathcal{Q}$ such that A is the first element with weight $|A|$ in the prescribed order. □

Algorithms for listing all circuits of a digraphs are available, see e.g. [83, 97]. The number of circuits in a digraph $\mathcal{G}(V, A)$ may be very large, however. The straightforward application of the greedy algorithm or of algorithm 8 to the set of all circuits will therefore not be feasible in most cases. In the case of undirected graphs one can drastically reduce the initial set of cycles [77, 6, 149]. In the following section we consider similar constructions for circuits in digraphs. The main difference is that in the undirected graph one can work over $GF(2)$ explicitly use the vector addition of cycles. Here we have the additional problem that the sum of circuits is in general not a circuit.

7.5 Short and Isometric Circuits

Short, arc-short and isometric circuits can be defined in analogy to their undirected counterparts in section 3.2.

Definition 100. *A circuit C is* short *if for any two of its vertices x and y it contains a shortest path from x to y* or *a shortest path from y to x.*
A circuit C is strictly arc-short *if for each x in C there is an edge $e^x = (v, w)$ such that $C = P[w, x] + P[x, v] + (v, w)$ where $P[w, x]$ and $P[x, v]$ are shortest paths.*
A circuit C is arc-short *if C contains a vertex x and an arc $e = (v, w)$ such that $C = P[w, x] + P[x, v] + (v, w)$ where $P[w, x]$ and $P[x, v]$ are shortest paths.*
A circuit C is isometric *if for any two of its vertices x and y it contains a shortest path from x to y* and *a shortest path from y to x.*

Lemma 101. *Every isometric circuit is short. A circuit C is short if and only if it is strictly arc-short. Every short circuit is arc-short.*

Proof. It follows directly from the definition that an isometric circuit is short.

For two distinct vertices $x \neq y$ in C, we denote the path from x to y in C by $C[x, y]$. Furthermore we write $S[x, y]$ for a path from x to y in $\mathcal{G}$ that is shorter $C[x, y]$ provided such a di-path exists. We call $S[x, y]$ a shortcut from x to y. In this case $S[x, y] \cup C[y, x]$ is again a circuit.
Suppose C is short. First we note that in this case there cannot be a vertex in x such that there are two vertices y, y' in C and shortcuts $S[x, y]$ and $S[y', x]$. (If y lies in $C[x, y']$ then there is a shortcut $S[y, x]$, i.e., C is not short. If y' lies in $C[x, y]$ then there is shortcut $S[y, x]$, i.e., C is not short.) Hence we have to consider three cases

for each vertex x in C:
(i) There is no shortcut to or from x in C. Then we may choose any edge $e = (u, v)$ in C and see that $C[x, u]$ and $C[v, x]$ are shortest paths.
(ii) There is a shortcut from x some y in C. Then there is also shortcut $S[x, z]$ from x to every vertex w in $C[y, x]$. We can choose y such that it is maximal in the sense that there is no shortcut $S[x, w]$ for all $w \neq y$ in $C[x, y]$. Necessarily there is an edge $e = (z, y) \in C[x, y]$. Hence $C[x, z]$ is a shortest path. Since C is short $C[y, x]$ must be a shortest path and the proposition follows.
(iii) There is a shortcut from some y in C to x. This implies that there is a shortcut $S[w, x]$ for all w in $C[x, y]$. Again we choose y maximal in the sense that there is no shortcut form w to x for all $w \neq y$ in $C[y, x]$. Then there is an edge $e = (y, z) \in C[y, x]$, Fig. 7.2. If $C[x, y]$ is not a shortest path then there is a shortcut $S[x, y]$ and C is not short.

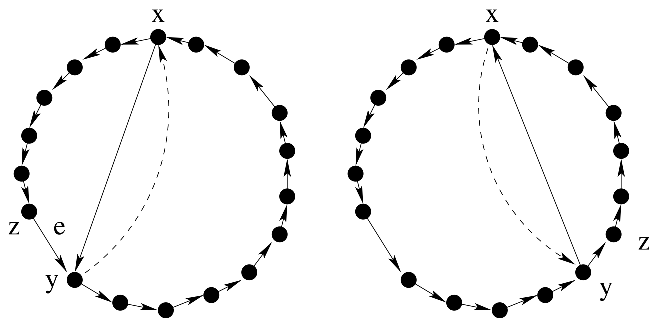


Figure 7.2. Cases ((ii) and (iii) of the proof of lemma 101. For details see text.

Conversely suppose C is not short. We show that C is not strictly arc-short. If C is not short then there are two vertices $x \neq y$ such that there are two shortcuts $S[x, y]$ and $S[y, x]$. Now suppose there is an arc $(u, v) \in C$ such that there is neither a shortcut $S[x, u]$ and $S[v, x]$. Since there is a shortcut $S[x, y]$ there are also shortcuts $S[x, y']$ for all y' in $C[y, x]$. Thus u cannot lie in $C[y, x]$. Similarly, there is a shortcut $S[y', x]$ for all y' in $C[x, y]$ and hence v cannot be in $C[x, y]$. Hence u must be in $C[x, y] \setminus \{x, y\}$ and v must be in $C[y, x] \setminus \{x, y\}$. Thus there cannot be an edge from x to y.

Finally, a strictly arc-short circuit is trivially arc-short. □

In undirected graphs, where a path from x to y is a path from y to x, a short cycle is trivially isometric. In directed graphs, a circuit is *isometric* if for all pairs of vertices $x, y \in \mathbf{c}$ the distance along the circuit equals the directed distance in $\mathcal{G}$, i.e., $d_{\mathbf{c}}(x, y) = d(x, y)$. In general, short circuits therefore are not isometric, as the example in Fig. 7.3 shows.

We observe that a double edge is obviously isometric. However, not all strongly connected digraphs graphs have cycle bases consisting of isometric circuits. The graph

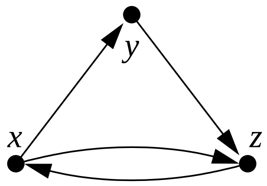


Figure 7.3. This graph has $\nu = 4-3+1 = 2$. The only circuits are the double edge $D = (x, z)$ and the triangle $T = (x, y, z)$. T is not isometric since $d_T(z, x) = 2 < d(z, x) = 1$. However, T is short, since for any two points, a shortest path in one of the two directions runs along T.

in Fig. 7.3 serves a counter-example. Arc-short circuits are useful because they can be constructed rather easily. In analogy to the undirected case only short circuits can be part of a minimum circuit basis.

Theorem 102. *If C is relevant then C is short.*

Proof. Suppose C is contained in a minimum circuit basis and it is not short. Then there are two vertices x and y such that C contains neither a shortest path P' from x to y nor a shortest path P'' from y to x. Furthermore, denote by C' and C'' the paths from y to x and from x to y along C, respectively. Note that $C' + P'$, $C'' + P''$, and $P' + P''$ are each closed paths in $\mathcal{G}$. By lemma 85, each of them can be written as a (positive) linear combination of circuits that are all not longer than $|C'| + |P'| < |C|$, $|C''| + |P''| < |C|$, or $|P'| + |P''| < |C|$. From

$$C = C' + C'' = (C' + P') + (C'' + P'') - (P' + P'')$$

we find that C itself can be written as a linear combination of circuits, all of which are strictly shorter than C itself. Since we have assumed that C is in the minimum circuit bases, at least one of these shorter circuits is not. In this case, however, we can replace C by one of these circuits in the basis, obtaining a strictly shorter basis, contradicting minimality of the circuit basis. □

Theorem 102 is a simple generalization of the analogous result for undirected graphs [77, 66]. The notion of short circuits, however, appears weaker than its undirected counterpart. Again the converse of theorem 102 is not true. As in the undirected case not all isometric circuits are relevant (Fig. 7.5), but also not all relevant circuits are isometric (Fig. 7.3). Thus it is not possible to generalize Horton's algorithm to find a minimum circuit bases.

We may, however, exploit a trick used e.g. in [69], namely perturbing the edge weights by a small amount in such a way that each subset of A has a unique weight. For instance, we assign numbers $\#e = 1, \dots, |A|$ to the arcs and set $w(e) = 1 + \varepsilon 3^{-\#e}$, $0 < \varepsilon < 1$. Now all shortest paths are unique and hence the $|A| \times |V|$ candidates for arc-short cycles must contain a minimum weight basis of the circuit space of G. This basis is independent of ε. Furthermore, the weight of two arc-sets with the same number of edges differs by less than $\epsilon/2$. Hence we obtain indeed a minimum length basis of the un-weighted problem. Of course, the weighting scheme is numerically problematic, it can, however, be replaced by a suitable lexicographic ordering of the arcs. Thus we have:

Theorem 103. *A minimum circuit basis of $\mathfrak{C}$ can be computed in polynomial time.*

But it still remains to find the appropriate prototypes of the set of relevant circuits.

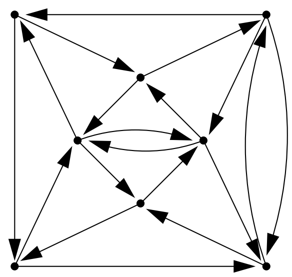


Figure 7.4. A counterexample to the converse of theorem 102. The outer quadrangle is a short circuit, but not a relevant one; by the side the quadrangle is also not isometric.

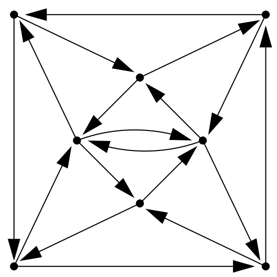


Figure 7.5. The outer quadrangle is an isometric circuit, but it is not relevant, because it is a linear combination of all the triangles.

The diagram in Fig. 7.10 shows the relationship between the different types of circuits, in directed graphs described in the preceding sections. Below the one-sided arrows, standing for one-sided implications, the number of counterexample for the inversion is given.

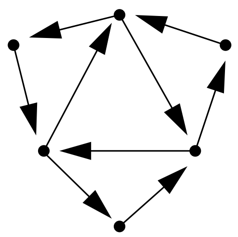


Figure 7.6. The inner triangle is a shortest circuit, but it is for none of its edge a unique shortest one.

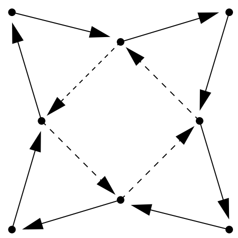


Figure 7.7. The quadrangle (dashed) is a relevant circuit, but for none of its edges the shortest one in the graph.

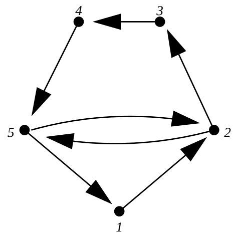


Figure 7.8. Consider the arc-short pentagon C. For the vertex 2 it is impossible to find an edge e, such that the shortest paths form both endpoints of e are contained in C.

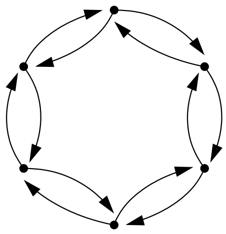


Figure 7.9. The MCB of this graph contains all double edges and one of the two hexagons.

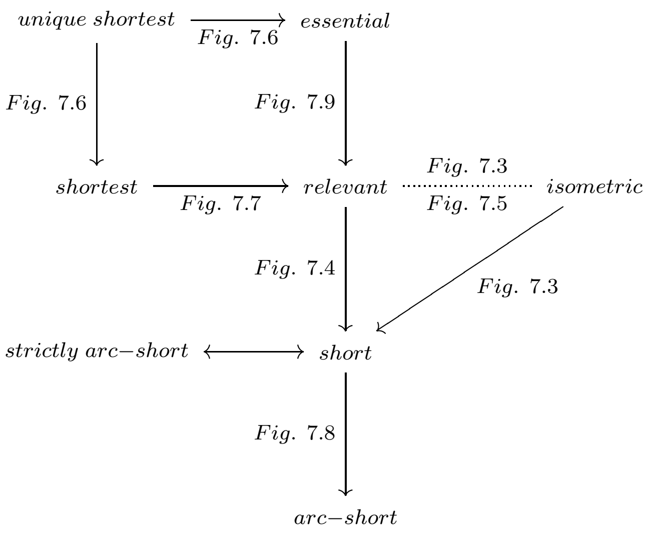


Figure 7.10. The relations between the different type of cycles. Two-sided arrows implies an iff relation. Below the one-sided arrow the number of the figure of the counterexample for the inversion is given. The dotted line stands for a “non-implication”

Chapter 8

Computations

8.1 Programs

The computational part of this work involved the development of tools for calculating the interchangeability classes of relevant cycles (introduced in chapter 6). All programs were written in ANSI C++, including STL, for attaining maximum portability and speed (`CycDeco`) and usability (`CalcGra`).

Since the GML (**G**raph **M**odelling **L**anguage) [117] is a portable file format for graphs with simple syntax, extensibility and flexibility, we decided to use the GML-file format for the input and output of our programs. A GML file consists of a hierarchical key-value lists. Graphs can be annotated with arbitrary data structures. GML is the standard file format in the `Graphlet` graph editor system [117] .

`CycDeco` calculates the relevant, shortest, unique shortest and essential cycles, as well as the interchangeability classes. If the vertex set $U \subseteq V$ is specified additionally to the graph, the U-paths are calculated additionally. First the biconnected components of the graph are calculated. Then the cycles are computed separately for each biconnected component. The following algorithm are used in `CycDeco`:

- relevant U-path prototypes: algorithm 6, see section 5.4
- MCB and relevant prototypes: algorithm 2 and 3, see section 3.4.
- relevant cycles: algorithm 4, see section 3.4
- unique shortest and shortest cycles: see section 3.2 and 3.6
- essential cycles: algorithm 5, if the interchangeability classes are not calculated, see section 3.6; else extracted from the $\leftrightarrow$-classes with $\mathrm{knar}() = 1$, using algorithm 7 and corollary 66.

- interchangeability classes: algorithm 7, see section 6.4

`CalcGra` is the graphical user interface for `CycDeco`, implemented with the Qt-library from Trolltech [143]. `CalcGra` is a simple graph editor with additional functions as the cycle decompositions. Additionally, it is possible to color each cycle separately or dependent sets of cycles.

Table 8.1 gives the CPU-time in seconds needed for cycle decompositions with `CycDeco` for random graphs with biochemical relevant degrees (3,4,5,6). For the decompositions biconnected random graphs — started with an Hamiltonian cycle — were used. One can see, that the calculation of all possible cycles and the partition does not need much more time for these graphs, as the calculation of the relevant cycles does. Even graphs with a lot vertices are decomposed very fast. The graphs of biomolecules are sub-cubic, i.e., the maximal vertex degree is 3, hence very long RNAs can be decomposed in within 15 minutes. For the small graphs ($|V| = 10$) most of the time is used for I/O of the results.

We also measured the CPU-time for more dense graphs (lower part of table 8.1). Again biconnected random graphs were used. The degree of the graph is given in percentage of the maximal possible degree. These graphs contain almost only relevant cycles of the length 3. Therefore, a lot of linear dependency test has to be done. Since, the decomposition of rather small graphs with only 100 nodes takes a long time, we did not measured bigger graphs.

8.2 RNA Structures

8.2.1 Background

The motivation for the present contribution arises from the search for a suitable energy model for RNA secondary structure computations in the presence of so-called pseudo-knots.

Definition 104. [156] *A* secondary structure *is a vertex-labeled graph $\mathcal{G}$ on n vertices with an adjacency matrix A fulfilling*

1. $a_{i,i+1} = 1$ *for* $1 \leq i < n$ *(the backbone)*
2. *For each i there is at most a single $k \neq i-1, i+1$ such that $a_{i,k} = 1$ (the base pairs)*
3. *If $a_{i,j} = a_{k,l} = 1$ and $i < k < j$ then $i < l < j$ (no overlapping base pairs).*

Table 8.1. `CycDeco`: CPU-time in sec for the calculation of the relevant cycles and the relevant, shortest, unique shortest and essential cycles as well as the interchangeability classes of a biconnected random graphs with low degrees (first and second part) and high degrees (third part) ($x = \frac{\deg}{|V|-1}$). (average values over 100 ($|V| < 100$) and 10 ($|V| \geq 100$) graphs) (using a Pentium III (Coppermine), 733 MHz, 1024 RAM)

average degree	number of nodes 10	30	50	100	300	500	1000
	relevant cycles						
3	0.0028	0.0236	0.080	0.506	13.33	66.3	677.8
4	0.0032	0.0356	0.130	0.958	29.48	241.06	2744
5	0.0046	0.0596	0.243	2.254	61.52	469.04	7372
6	0.0055	0.0934	0.410	4.716	207.6	1326.0	29032
	complete decomposition						
3	0.0028	0.0252	0.081	0.521	13.50	66.3	679.3
4	0.0040	0.0387	0.137	0.993	29.69	244.5	2749
5	0.0056	0.0649	0.254	2.299	63.18	474.7	7485
6	0.0056	0.0986	0.424	4.804	212.4	1354.0	30153

x	number of nodes 10	30	50	100	10	30	50	100
	relevant cycles				**complete decomposition**			
3	0.0027	0.53	32.24	288.5	0.0024	0.55	32.3	303.8
4	0.0034	1.55	25.47	3365	0.0039	1.59	25.7	3393
5	0.0049	0.80	28.01	5389	0.0053	0.82	29.1	5762
6	0.0053	1.93	71.06	26757	0.0060	1.99	73.9	27013
7	0.0071	4.16	154.6	49977	0.0078	4.28	159.5	50426
8	0.0103	8.08	304.5	82581	0.0103	8.43	314.7	84085

If we violate the third condition in definition 104, we produce two overlapping base pairs, the result is called a *pseudo-knot*.

The "classical" definition of the RNA secondary structures excludes pseudo-knots [156] mostly for technical reasons: The folding problem for RNA can be solved efficiently by dynamic programming [156, 171] in their absence. It is not known how to assign energies to the loops created by pseudo-knots and dynamic programming methods that compute minimum energy structures break down.

These (pseudo-knot free) secondary structures are outerplanar graphs $\mathcal{G}$. Hence they have unique minimum cycle bases $\mathcal{B}(\mathcal{G})$ [93]. Therefore any secondary structure can be uniquely decomposed into loops as shown in Fig. 8.1. The energy of an RNA secondary structure is assumed to be the sum of the energy contributions of all loops. Energy parameters for the contribution of individual loops have been determined experimentally (see e.g. [49, 80, 155]) and depend on the loop type, size and partly its sequence.

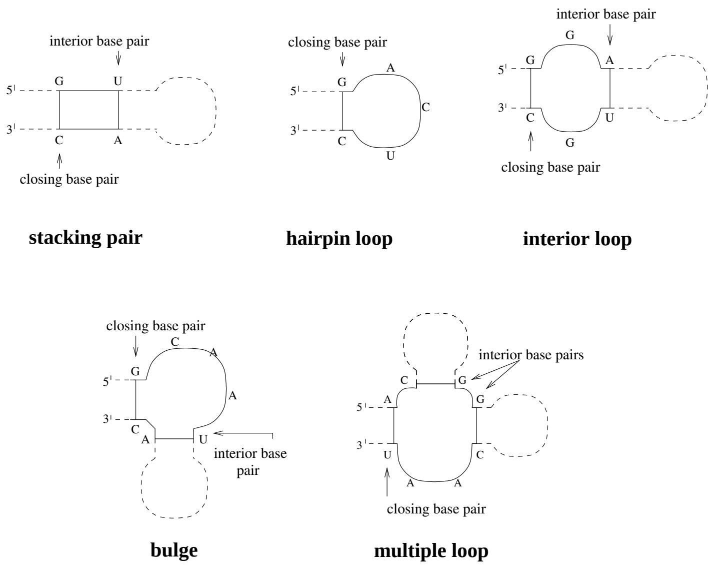


Figure 8.1. RNA secondary structure elements. Any secondary structure can be uniquely decomposed into these types of loops

These loops are exactly the relevant cycles of the graph. Assigning each relevant cycle C an energy contribution $E(C)$ leads to the standard energy model. The RNA secondary structure prediction problem can be rephrased as minimizing the energy function $E(\mathcal{G}) = \sum_{C \in \mathcal{B}(\mathcal{G})} E(C)$ over the class of secondary structure graphs ([128, 156] for details).

On the other hand, an increasing number of experimental findings, as well as results from comparative sequence analysis, suggest that pseudo-knots are important structural elements in many RNA molecules [161]. Notably, functional RNAs such as RNaseP RNA [94] (Fig. 8.9) and ribosomal RNA [88] contain pseudo-knots. The diversity of molecular biological functions performed by pseudo-knots can be subdivided into three groups. Pseudo-knots at the 5'end of mRNAs appear to adopt a role in the control of mRNA translation. For instance the expression of replicase is controlled in several viruses either by ribosomal frame shifting [16, 20, 31, 139, 146] or by in-frame read-through of stop codons [165]. Both mechanisms involve pseudo-knots. Core pseudo-knots are necessary to form the reaction center of ribozymes. Most of the enzymatic RNAs with core pseudo-knots, such as RNaseP, are involved in cleavage or self-cleavage reactions [17, 48, 62, 101]. Pseudo-knots in the tRNA-like motifs at the 3'end of the genomic RNA mediate replication control in several groups of plant viral RNA [96].

These biopolymer graphs can be decomposed into outerplanar graphs. This decomposition can be used to derive an upper bound on $\ell(\mathcal{G})$, the length of an MCB.

Definition 105. *Let $\mathcal{G} = (V, E)$ be a graph with spanning path T. Consider a partition $\{B_1, B_2, \ldots, B_\beta\}$ of $B = E \setminus T$ such that $\mathcal{G}_k = (V, B_k \cup T)$ is outerplanar. The subgraph $\mathcal{G}_k$ of $\mathcal{G}$ is called an* outerplanar constituent *and $\mathcal{G} = \mathcal{G}_1 \vee \mathcal{G}_2 \vee \cdots \vee \mathcal{G}_\beta$.*

It is clear that such a partition always exists. To see this assume that B_i contains only a single edge then $\mathcal{G}_I$ contains a single cycle and hence is outerplanar. A p-book $\mathfrak{B}$ is a set of p distinct half-planes (the *pages* of the book) that share a common boundary line ℓ, called the *spine* of the book. An embedding of a graph $\mathcal{G}$ into a book $\mathfrak{B}$ consists of an ordering of the vertices along the spine of the book together with an assignment of each edge to a page of the book, in which edges assigned to the same page do not cross. If $\mathcal{G}$ has a spanning path T and the vertices are arranged along the spine in their order of occurrence along T, we shall say for simplicity that T is the spine of the book embedding.

Note that $\mathcal{G} = \bigvee_{k=1}^{\beta} \mathcal{G}_k$ is embeddable in a β-book $\mathfrak{B}$ with spine T. The bisecondary structure graphs introduced in [128] are exactly those that have at most two outerplanar constituents. Equivalently, they are characterized as subgraph of planar Hamiltonian graphs [11].

Theorem 106. *Let $\mathcal{G} = \bigvee_{k=1}^{\beta} \mathcal{G}_k$. Then:*

$$\ell(\mathcal{G}) \leq \sum_{k=1}^{\beta} \ell(\mathcal{G}_k) \tag{8.1}$$

Proof. First we observe that $\mathcal{G}_k$ is connected for $1 \leq k \leq \beta$, hence $\nu(\mathcal{G}_k) = |T| + |B_k| - |V| + 1 = |B_k|$, while $\nu(\mathcal{G}) = |B| + |T| - |V| + 1 = |B| = \sum_k |B_k|$, i.e., $\nu(\mathcal{G}) = \sum_{k=1}^{n} \nu(\mathcal{G}_k)$.

The minimum cycle bases $\mathcal{M}_k$ of the outerplanar components are $\mathcal{G}_k$ are easily constructed: they are given by the faces of the outerplanar embeddings [93]. Each of these cycles contains at least one edge in B_k and none of the edges in B_l, $l \neq k$, whence $\mathcal{M} = \bigcup_{k=1}^{\beta} \mathcal{M}_k$ is a set of independent cycles of $\mathcal{G}$ containing $\sum_k |\mathcal{M}_k| = \sum_k \nu(\mathcal{G}_k) = \nu(\mathcal{G})$ cycles. In other words, $\mathcal{M}$ is a cycle basis of $\mathcal{G}$. Equation (8.1) now follows from $\ell(\mathcal{M}) = \sum_k \ell(\mathcal{M}_k) = \ell(\mathcal{G}_k)$. $\square$

Indeed, most known RNA structures with pseudo-knots are bi-secondary structures (which do not involve nested pseudo-knots) (for details see [128]). Bi-secondary structures correspond to planar graphs while secondary structures form the sub-class of outerplanar graphs. The virtue of bi-secondary structures is that they capture a wide variety of RNA pseudo-knots, while at the same time they exclude true knots such as the structure in Fig. 8.2.

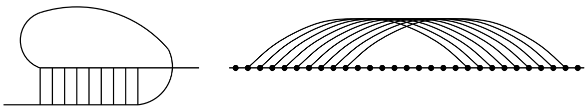


Figure 8.2. This pseudo-knot do not belong to the class of bi-secondary structures (from [128]).

The *book-thickness* (sometimes also called the page-number) of a graph is the minimal number p of pages of a book into which the graph can be embedded. Thus ordinary secondary structure graphs need $p = 1$ pages, a pseudo-knot at least $p = 2$ pages. An upper limit for the book-thickness consequently constricts the pseudo-knot complexity. Structures with $p = 2$ pages correspond to the class of planar graphs. Almost all known structures fall into this class (one exception: α-mRNA - Fig. 8.3, see section 8.2.2).

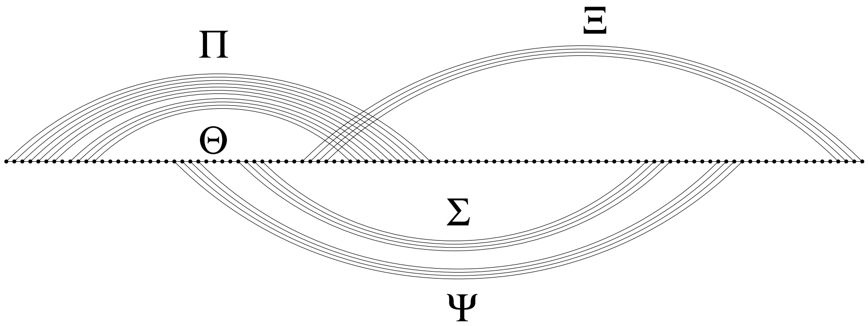


Figure 8.3. Diagram of the contact structure of *E. coli* α-mRNA (Pseudoknot of the regulatory region of the alpha ribosomal protein operon, EMBL number: XO2543). The structure contains 5 stems, labeled by uppercas Greek letters (from [128]).

Pseudo-knots violate outerplanarity and often they leads to graphs with a non-unique minimum cycle basis. The set $\mathcal{R}$ of relevant cycles seems to be a good candidate for extending the energy model. However, as the Fig. 8.7 of the pk_2 of *E. coli* tmRNA shows, sometimes there is a large class of relevant cycles associated with what bio-physically is a single structural element. These are exactly the interchangeable cycles described in chapter 6.

The folding of an RNA molecule is largely determined by the formation of base pairs, leading to short, double-stranded, stem regions connected by single-stranded loop regions like hairpins, bulge, internal and multi-branched loops (Fig. 8.1), each having its own characteristic folding pattern, dependent on the particular base sequence present [108]. Pseudo-knots result from base pairing of nucleotides of a single-stranded loop region with complementary sequence outside this loop [112, 130]. The simplest form of a pseudo-knot is the so-called H(airpin)-type [114, 161], which characterized by the presence of two stems and two loops (see Fig. 8.4).

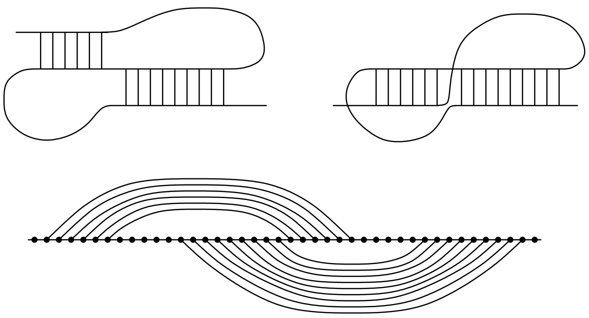


Figure 8.4. The h(airpin)-type pseudo-knot results from base pairing of nucleotides of a single-stranded loop region with complementary sequence outside this loop [113]

Each of these two loops correlate to one relevant cycle or to one interchangeability class. The free energy of a secondary structure is the sum of independent energies for each loop in the structure. It seems natural therefore to average over the contributions of interchangeable cycles or to define the energy parameters in such a way that all interchangeable cycles contribute the same energy.

Hence we suggest that

$$E(G) = \sum_{\mathcal{W}\in\mathfrak{P}} \frac{\operatorname{knar}(\mathcal{W})}{|\mathcal{W}|} \sum_{C\in\mathcal{W}} E(C) \tag{8.2}$$

serves as a suitable generalization of the standard energy model for nucleic acid structures.

To be able to differ between the graph of the RNA secondary structure without and with pseudo-knots, we denote the graphs $\mathcal{G}$ and $\mathcal{G}^+$.

8.2.2 *Escherichia coli* α-operon mRNA

The *Escherichia coli* α-operon mRNA folds into a structure that is required for allosteric control of translational initiation [137]. Compensatory mutations have defined an unusual pseudo-knotted structure [136], the thermodynamics of which were subsequently investigated in detail [57]. The structure cannot be drawn without intersections, see Fig. 8.5. To our knowledge it is the only known RNA structure that cannot be embedded in a 2-page book.

Nevertheless, the graph of the α-mRNA has a unique MCB and therefore each $\leftrightarrow$-class contains only one relevant cycle. Table 8.2 gives the cycle length distribution of the unique MCB. The three stams Σ, Ξ and Ψ causes 9 additional quadrangles and three longer relevant cycles (red, blue and yellow colored in Fig. 8.5).

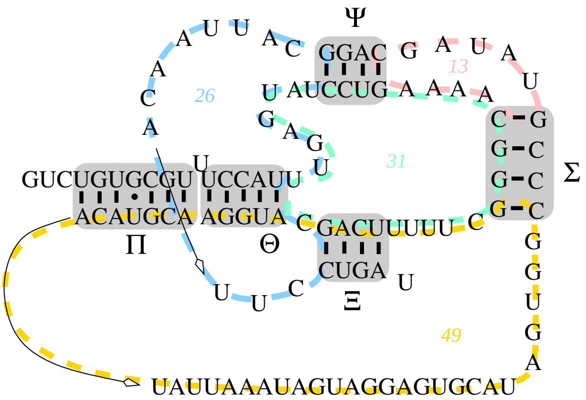


Figure 8.5. Diagram of the contact structure of *E. coli* α-mRNA (Pseudo-knot of the regulatory region of the alpha ribosomal protein operon, EMBL number: XO2543). The structure contains 5 stems, labeled by uppercase Greek letters (from [128]).

Table 8.2. Length distribution of the unique MCB $\mathcal{M}$ of *Escherichia coli* α-operon mRNA with $(\mathcal{G}^+)$ and without $(\mathcal{G})$ pseudo-knots.

length	4	5	13	26	31	49
$\mathcal{M}(\mathcal{G}^+)$	19	1	1	1	1	1
$\mathcal{M}(\mathcal{G}$	10	1	-	-	1	-

8.2.3 tmRNA

The bacterial tmRNA (also known as 10Sa RNA or SsrA) got its name because of its dual tRNA-like and mRNA-like nature [86]. It is employed in a remarkable trans-translation process to add a C-terminal peptide tag to the incomplete protein product of a broken mRNA. The tmRNA-directed tag targets the abnormal protein for proteolysis.

Four pseudo-knots are usually present in each tmRNA, so the database is rich with information on pseudo-knot variability. The tmRNA Website* currently contains 114 non-duplicate tmRNA sequences from 99 species.

Fig. 8.6 shows the secondary structure of the tmRNA of *E. coli* also as a linked graph, where the sequence is placed on a line and the connecting bases are represented through arc. The tmRNA of *E. coli* has four so called *h-type* pseudo-knots [113], so the book-thickness is $p = 2$ and the graph $\mathcal{G}^+$ is still planar.

Only pk_2 violates the uniqueness of the MCB. This pseudo-knot causes a large class of relevant cycles of length 30, which are associated with what biophysically is a single structural element. In Fig. 8.7 one of these relevant cycle is indicated by the thick blue line, the other $7 \times (3 + 4)$ possible paths are shown by dashed blue lines. All these 49 cycles belong to the same $\leftrightarrow$-class $\mathcal{W}$ with knar $(\mathcal{W})\ = 1$.

*URL:`http://www.indiana.edu/~tmrna/`

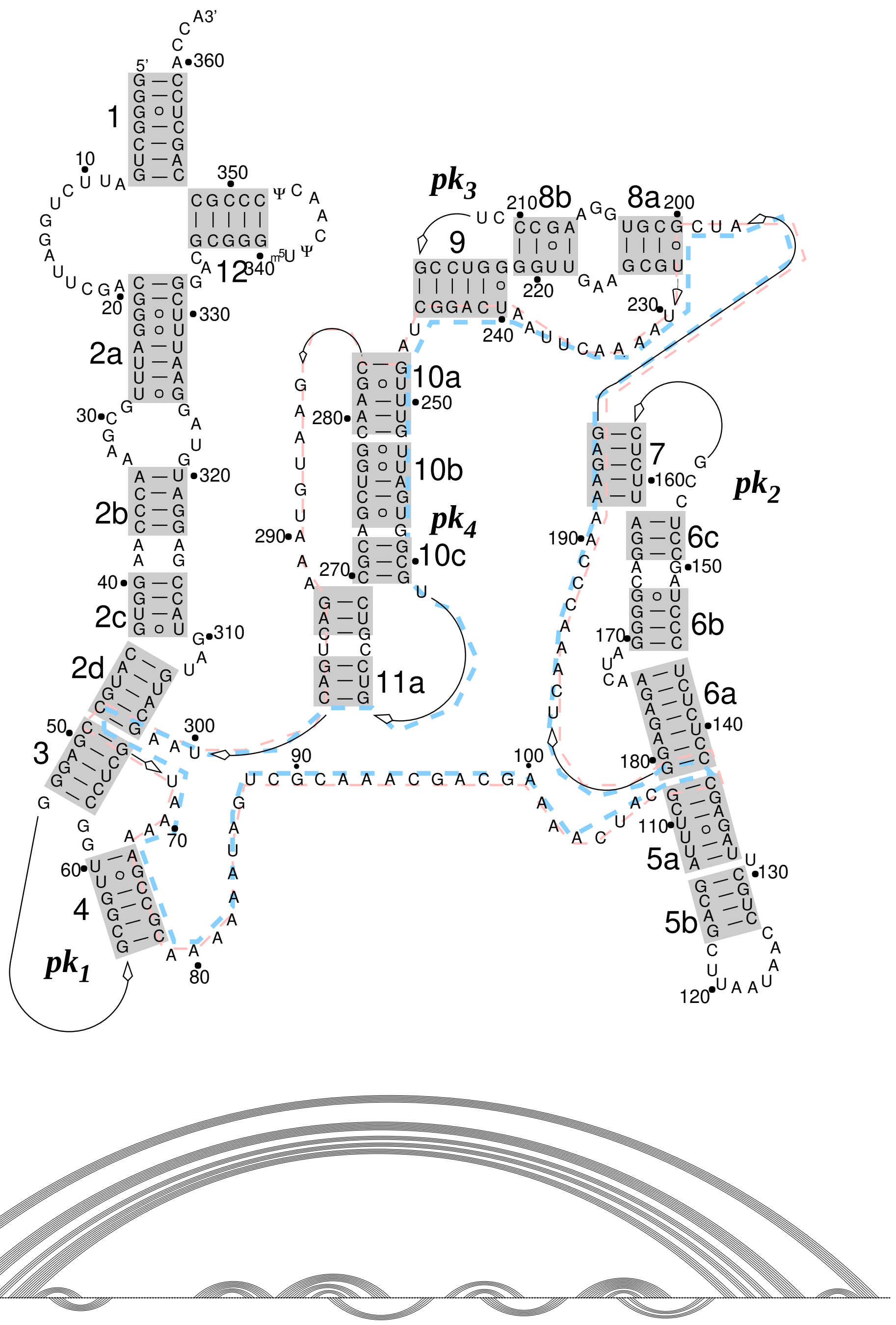


Figure 8.6. *Escherichia coli* tmRNA

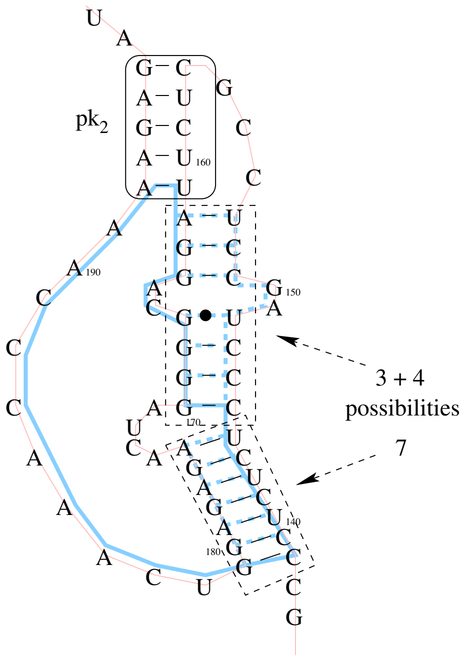


Figure 8.7. Pseudo-knot pk_2 of *E. coli* tm-RNA, the box indicates the stem which causes the pseudo-knot. The backbone of the RNA is represented by the red line.
The thick blue line shows one of the 21 interchangeable relevant cycles. The others can be found by an $\oplus$-sum with the blue dashed stacks.

Table 8.3 shows the cycle length distribution of the set of relevant cycles and an MCB of the secondary structure without ($\mathcal{G}$) and with ($\mathcal{G}^+$) pseudo-knots. Immediately, the additional quadrangles leaps to the eye. They come from the pseudo-knot stams: 5 from pk_1, 4 from pk_2, 5 from pk_3 and $2+2$ from pk_4. The additional hexagon comes from internal loop of the two stams pk_4. pk_1 and stam number 3 forms the 14-edges cycle, pk_3 and stams $8a$ and $8b$ leads to the 23-edges cycle and pk_4 and stams $10a$, $10b$ and $10c$ encloses the 25-edges cycle.

Table 8.3. Length distribution of the relevant cycles $\mathcal{R}$ and cycles of an MCB $\mathcal{M}$ of tmRNA *Escherichia coli* secondary structure graph with ($\mathcal{G}^+$) and without ($\mathcal{G}$) pseudo-knots.

length	4	5	6	7	8	9	10	11	13	14	22	23	25	30	98	107
$\mathcal{M}(\mathcal{G}^+)$	85	1	2	1	3	1	4	2	1	1	1	1	1	1	1	-
$\mathcal{M}(\mathcal{G})$	67	1	1	1	3	1	4	2	1	-	1	-	-	-	-	1
$\mathcal{R}(\mathcal{G}^+)$	85	1	2	1	3	1	4	2	1	1	1	1	1	49	1	-
$\mathcal{R}(\mathcal{G})$	67	1	1	1	3	1	4	2	1	-	1	-	-	-	-	1

In Fig. 8.6 the two longest relevant cycles of $\mathcal{G}$ (red line) and $\mathcal{G}$ (blue dashed line) are shown. Again a shorter path, arose by pk_4, leads to the shorter longest relevant cycle.

The tmRNA from *Cyanophora paradoxa cyanelle* (Fig. 8.8) contains only one h-type pseudo-knot (pk_1), which forms a bi-secondary structure [128] $\mathcal{G}^+$ with a unique MCB. Therefore $\mathcal{G}^+$ is embeddable in a book with 2 pages.

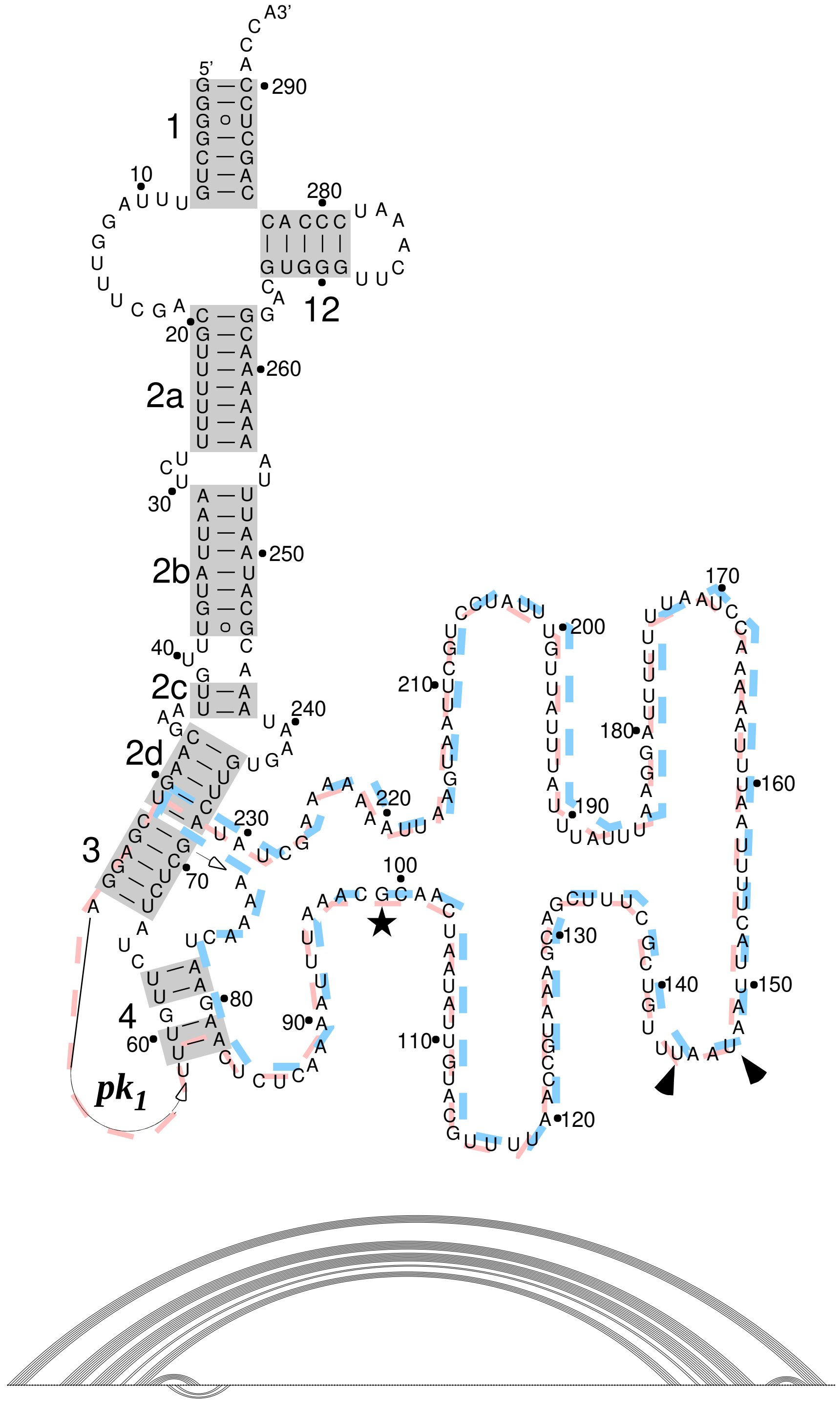


Figure 8.8. *Cyanophora paradoxa cyanelle* tmRNA

Of course, there are some differences in the length distribution of the relevant cycles of $\mathcal{R}(\mathcal{G})$ and $\mathcal{R}(\mathcal{G}^+)$, see Table 8.4.

Table 8.4. Length distribution of the relevant cycles $\mathcal{R}$ of tmRNA *Cyanophora paradoxa cyanelle* secondary structure graph with ($\mathcal{G}^+$) and without ($\mathcal{G}$) pseudo-knots.

length	4	6	9	12	16	21	160	164
$\mathcal{R}(\mathcal{G}^+)$	35	1	3	2	1	1	1	-
$\mathcal{R}(\mathcal{G})$	33	-	3	2	-	1	-	1

The two stems of the pseudo-knot leads to the two additionally quadrangles and the hexagon, which arises from the internal loop between the two stams. The big hairpin loop of the pseudo-knot free secondary structure with length 164, indicated by the blue dashed line in Fig. 8.8, is in the bi-secondary structure a little bit shorter (the red line in Fig. 8.8). It consists of a path containing a "pseudo-knot-edge".

8.2.4 Ribonuclease P

Ribonuclease P (RNaseP) RNA is a well studied molecule which is found in all cells that carry out tRNA synthesis. It is a processing endonuclease that specifically cleaves precursors of tRNA. RNaseP generates the mature 5' end of tRNAs by removing 5' leader sequences from pre-tRNAs. In bacteria (and some Archaea) the RNA subunit alone is catalytically active in vitro, i.e. it is a ribozyme. In vivo it is associated with a small protein but is clearly the catalyst. It acts as a true enzyme, in the sense that it reacts with multiple substrates. Unlike most ribozymes, RNase P recognizes its substrate through tertiary RNA-RNA interactions, rather than through extensive Watson-Crick base-pairing.

The secondary structure from *Pseudomonas fluorescens* RNaseP[†] given in Fig. 8.9, contains two core pseudo-knots pk_1 and pk_2. These core pseudo-knots are necessary to form the reaction center of the ribozyme. Most of the enzymatic RNAs with core pseudo-knots are involved in cleavage or self-cleavage reactions.

Fig. 8.10 gives the relevant cycles occurred by both pseudo-knots. The pseudo-knot pk_1 (l.h.s. of Fig. 8.10) leads to two additional stams, represented through 4 and 2 quadrangles, separated through an internal loop of length 5. The remaining 3 additional quadrangles arise from pk_2 (r.h.s. of Fig. 8.10).

pk_1 leads only to relevant cycles, which are also essential. pk_2 on the other hand violates the uniqueness of the MCB. As shown in table 8.5 a 16 34-edged cycles occurs in the contact graph of the RNaseP with the pseudo-knots. These long relevant cycles, indicated by the blue lines, belong to one $\leftrightarrow$-class $\mathcal{W}$ with $\operatorname{knar}(\mathcal{W}_i) = 1$. One of

[†] from "The RNase P Database", URL:`http://www.mbio.ncsu.edu/RNaseP/home.html`

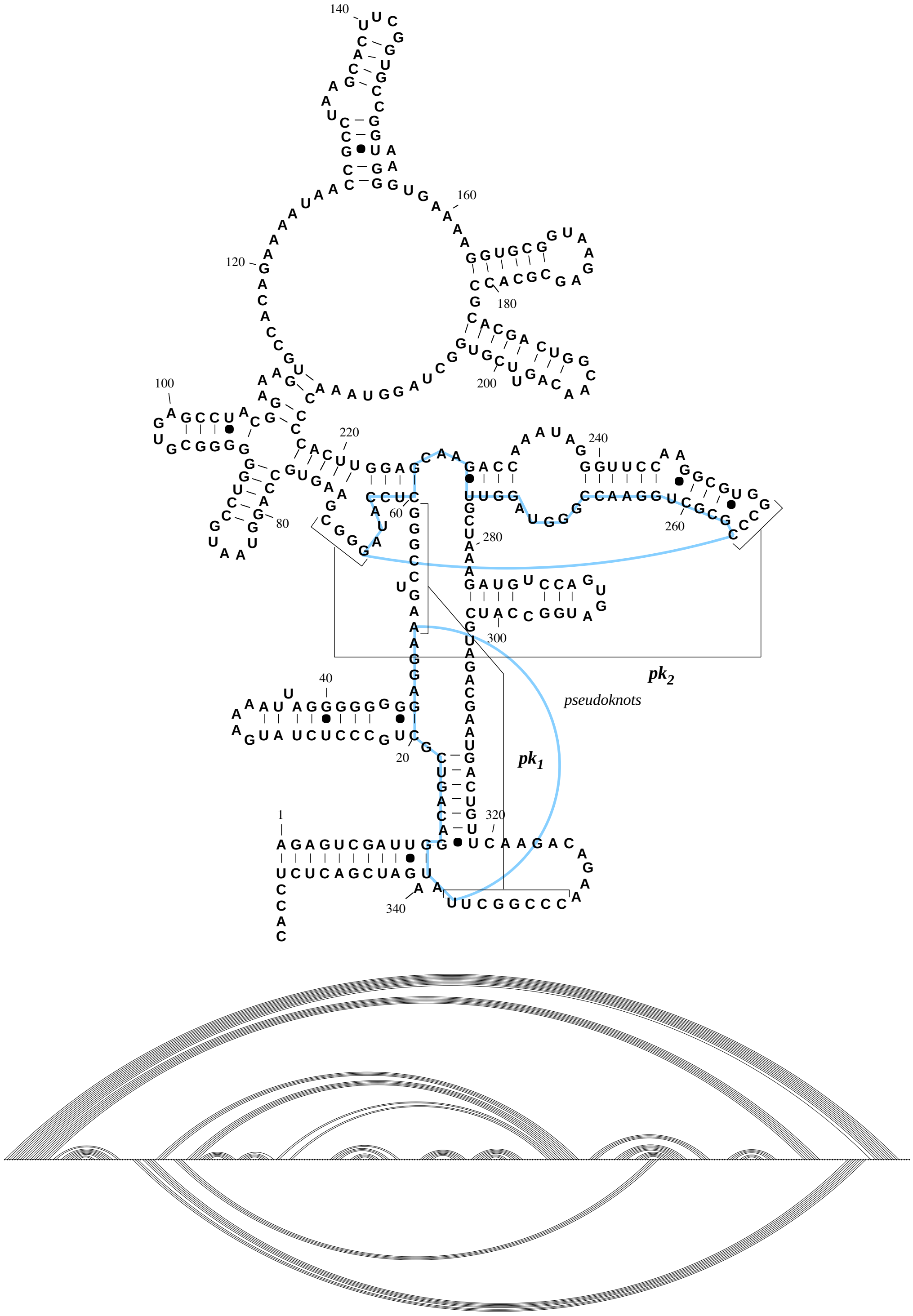


Figure 8.9. *Pseudomonas fluorescens* RNase P

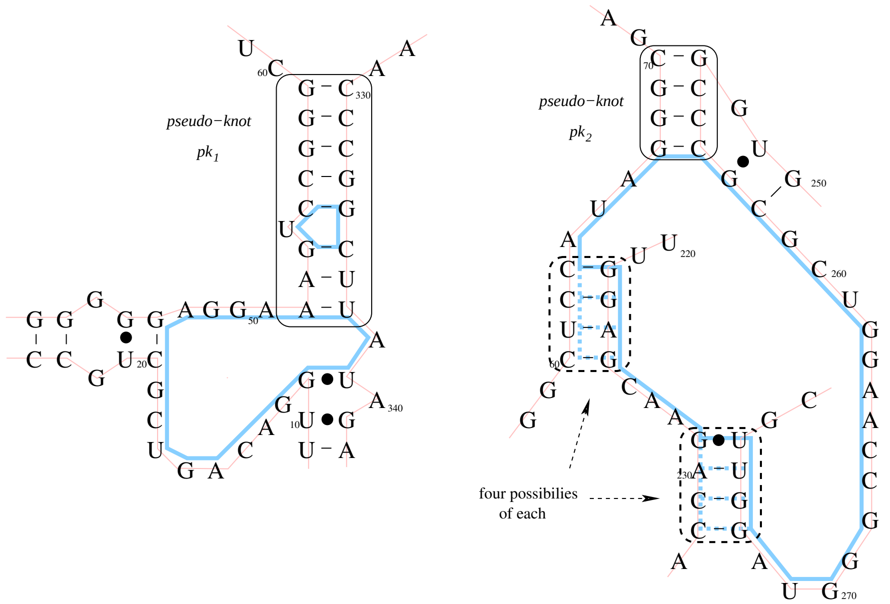


Figure 8.10. Pseudo-knots pk_1 (l.h.s.) and pk_2 (r.h.s.) of *Pseudomonas fluorescens* RNaseP, indicated by the boxes. The backbone of the RNA sequence is shown by the red lines. The relevant cycles, caused by pseudo-knots are represented by blue lines (for details see text)

them is shown in right picture of Fig. 8.10 as a solid blue line, the alternative paths are indicated by dashed blue lines.

In table 8.5 the cycle length distribution of the set of relevant cycles and an MCB is shown. Immediately two additional group of long relevant cycles leap to the eye: the 19-edges one and 34-edges ones. Both are shown in Fig. 8.9 by thick blue lines.

Table 8.5. Length distribution of the relevant cycles $\mathcal{R}$ and cycles of an MCB $\mathcal{M}$ of tmRNA *Pseudomonas fluorescens* secondary structure graph with ($\mathcal{G}^+$) and without ($\mathcal{G}$) pseudo-knots.

length	4	5	6	7	8	9	12	15	19	23	34	39	43
$\mathcal{M}(\mathcal{G}^+)$	81	5	9	3	2	1	1	1	1	1	1	1	1
$\mathcal{M}(\mathcal{G})$	72	4	9	3	2	1	1	1	-	1	-	1	1
$\mathcal{R}(\mathcal{G}^+)$	81	5	9	3	2	1	1	1	1	1	16	1	1
$\mathcal{R}(\mathcal{G})$	72	4	9	3	2	1	1	1	-	1	-	1	1

8.3 Chemical Networks

8.3.1 Reaction Networks

Recent surveys, in particular [82, 154, 46], have revealed that metabolic reaction networks belong to the class of small world networks in the wider sense: they have a diameter that is much smaller than what one would expect for an uncorrelated random graph with the same number of vertices and edges.

Small world networks have received considerable attention since the seminal paper by Watts and Strogatz [158]. In a recent paper [2], Amaral *et al.* present evidence that there are (at least) three structurally different classes of networks that are distinguished by the distribution $P(d)$ of the vertex degrees d:

(a) *Single Scale Networks* with a sharp distribution of vertex degrees exhibiting exponential or Gaussian tails. This class includes also the Erdős-Rényi model of uncorrelated random graphs [43, 13].

(b) *Scale Free Networks* with a power law distribution $P(d) \sim d^{-\gamma}$. A simple model for this type of networks was introduced recently by Barabási *et al.* [7, 8]. Metabolic networks [154, 82] and food-webs [103] belong to this class.

(c) *Broad Scale Networks* for which $P(d)$ has a power-law regime followed by a sharp cut-off, e.g. exponential or Gaussian decay of the tail. An example is the movie-actor network described in [157]

The most common model of graph evolution, introduced by Erdős and Rényi [43], assumes a fixed number $n = |V|$ of vertices and assigns edges independently with a certain probability p [13]. In many cases ER random graphs turn out the be quite different from a network of interest. The Watts-Strogatz [158] model of small world networks starts with a deterministic graph, usually a circular arrangement of vertices in which each vertex is connected to k nearest neighbours on each side. Then edges are “rewired” (in the original version) or added [106, 105] with probability p. We shall consider the latter model for $k = 1$, denoted SW1 below, which corresponds to adding random edges to a Hamiltonian cycle. Both ER and SW1 graphs exhibit an approximately Gaussian degree distribution.

The other extreme is scale-free BA model [7, 8] with a degree distribution of the form $P(d) \sim d^{-3}$: Starting from a small core graph, at each time step a vertex is added together with m edges that are connected to each previously present vertex k with probability

$$\Pi(k) = d(k) / \sum_j d(j)\,, \tag{8.3}$$

where $d(j)$ is the degree of vertex j. A recent extension of the model allows the tuning of the scaling exponent γ in the range $2 \le \gamma \le 3$ [1].

Much of the literature discusses small world networks in terms of the average path length between two vertices [105] or of the network's clustering coefficient [74, 9] which measures how close the neighbourhood of a each vertex comes on average to being a complete subgraph (clique) [158]. In this contribution we consider the small cycle of small world networks in detail. This approach is motivated by the following two observations:

Recent work on the spread of epidemics on a small world network [109] emphasizes the importance of "far-reaching" edges. The idea is that clipping a far edge will force a (relatively) long detour in the network. Hence it is these edges that are responsible for the small diameter of the graph $\mathcal{G}$. In section 3.2 we have seen that detours are intimately related to the cycles in the graph. In particular, we describe the connection between cycles in directed and undirected models and argue that the collection of *relevant cycles* is the appropriate mathematical object for our purposes. In section 1.2.3 a brief outline of the relationship between the cycle structure of a reaction network and *Chemical Flux Analysis* was given. In the following sections the distribution of triangles and longer relevant cycles is discussed for uncorrelated random graphs as well as for small world models.

Triangles in Reaction Networks

It is clear that all triangles in a graph are relevant, since a triangle is necessarily a shortest cycle through each of its edges. Hence $|\mathcal{R}(G)| \ge \Delta$, where Δ denotes the number of triangles in G. We expect $\langle\Delta\rangle_{\mathrm{ER}} = \binom{n}{3}p^3$ triangles in an ER random graph with edge-drawing probability p. For the SW1 graphs we obtain a similar expression:

$$\langle\Delta\rangle_{\mathrm{SW1}} = np + n(n-4)p^2 + \frac{1}{6}n(n^2-9n+20)p^3 \,. \tag{8.4}$$

The MCB will therefore consist almost exclusively of triangles if $\Delta \gg \nu(G)$. The average vertex degree is $d = 2|E|/n = p(n-1)$ for ER and $d = 2 + p(n-3)$ for SW1, resp. Assuming that n is large we expect to find only triangles in $\mathcal{R}(G)$ for $d \gg \sqrt{3n}$. Numerical simulations show that this is indeed the case, see Fig 8.12 in the following section. In this regime, we have $|\mathcal{R}(G)| \sim d^3/6$, and the graph contains no far edges. Not surprisingly, there is little difference between SW1 and ER random graphs for large n.

Since the BA model is constructed such that it yield a fixed *average* vertex degree d, it should be compared to random graph models with the same vertex degree d instead of random graphs with a fixed edge drawing probabilities p. We have an asymptotically constant number of triangles for both ER and SW1: $\Delta_{\mathrm{ER}} \to d^3/6$ and $\Delta_{\mathrm{SW1}} \to d^3/6 - d + 2/3$, resp. Note that as a consequence the clustering coefficient vanishes asymptotically. In SW networks with *a priori* connectivity $k > 1$ we find of course a number of triangles that grows at least linearly with n, since the initial ($p = 0$) networks already contains $(k-1)n$ triangles. The clustering coefficient stays finite for large n in this case [157].

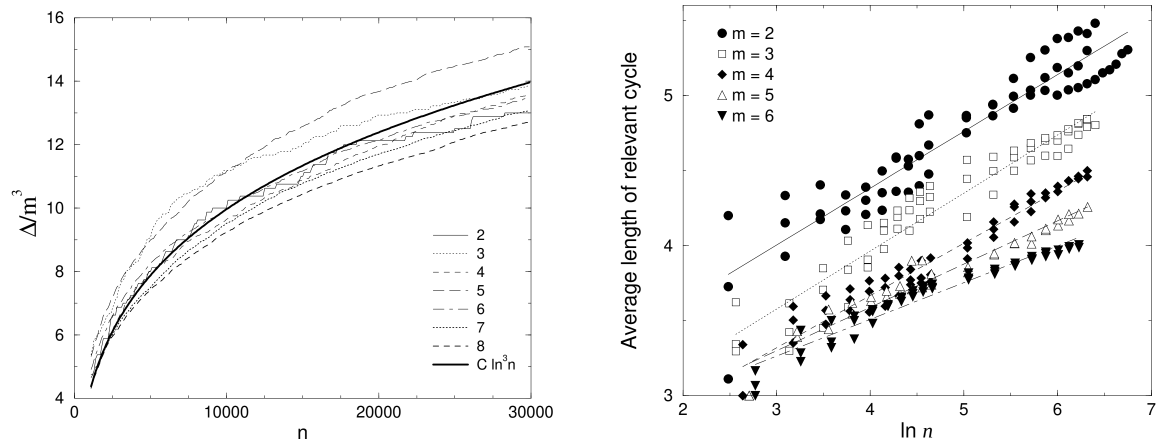


Figure 8.11. Cycles in the BA model.
L.h.s.: triangles in BA models with different values of m.
R.h.s.: mean length of a relevant cycle in BA networks.

The large vertex degree of the "early" vertices in the BA model suggests that there should be many more triangles than in ER or SW1 models. The expected degree of vertex s at "time" t is known [34]: $d(s|t) = m[\sqrt{t/s} - 1]$. The probability of an edge between s and t, $t > s$, is therefore $p_{st} = md(s|t-1)/2(t-1)m$, where $2(t-1)m$ is the sum of the vertex degrees at "time" $t - 1$. We have therefore

$$\begin{aligned} \langle \Delta \rangle &= \sum_{r<s<t} p_{rs} p_{st} p_{rt} \\ &\approx \frac{m^3}{8} \int_{1<r<s<t}^{n} (1/st^2) \left(\sqrt{\frac{s}{r}} - 1\right)\left(\sqrt{\frac{t}{r}} - 1\right)\left(\sqrt{\frac{t}{s}} - 1\right) \\ &\sim Cm^3 \ln^3 n + \mathcal{O}(\ln^2 n) \end{aligned} \tag{8.5}$$

The l.h.s. panel in Fig. 8.11 shows Δ for typical BA-random graphs with $m = 2, \dots, 8$ as a function of "time". The behavior of Δ in a individual growing network is well represented by equ.(8.5).

An extension of the BA model generates graph with $2 < \gamma \leq 3$. In addition to the growth of the network, the model includes to rewiring operations: (i) addition of m new edges such that the initial points of the edges are chosen randomly while the terminal points are selected according to equ.(8.3), and (ii) rewiring of m randomly selected edges by leaving on endpoint fixed and re-attaching the other endpoint according to equ.(8.3). Since the scaling exponents depend on the relative frequency of the two rewiring operations, a quantitative comparison of chemical reaction networks with the extended BA model does not seem to meaningful at this point.

There is, however, a universal scaling relation between $P(d) \sim d^{-\gamma}$ and the degree $d(s|t) \sim (t/s)^{\beta}$ of vertex s and time t [34], namely $\beta = 1/(\gamma - 1)$ and $2 \leq \gamma \leq 3$, i.e., $\frac{1}{2} \leq \beta \ll 1$. Using the same reasoning as above the number of triangles should scale as $\langle\Delta\rangle \sim C(\beta)n^{2\beta-1}\ln n$ for $2 < \gamma < 3$. Thus we again expect the fraction of triangles to vertices to approach zero for large systems. The number of triangles in graphs with the same number of edges in vertices, on the other hand, increases with decreasing values of γ.

Longer Cycles in Reaction Networks

Much less can be said in general about longer relevant cycles. Computationally we find that the number $L = |\mathcal{R}| - \Delta$ of non-trivial relevant cycles has its maximum around $|E| \approx 0.74n^{3/2}$ independent of the random graph model, Fig 8.12. The scaling of is consistent with $L \sim Cn^{5/2}$, where the constant $C \approx 0.036$ is the same for ER and SW1 random graphs and $C \approx 0.016$ for the BA models. For small vertex degrees, $d \ll |V|^{1/2}$ we find $\mathcal{R}(G) \approx \nu(G)$, i.e., the MCB is (almost) unique.

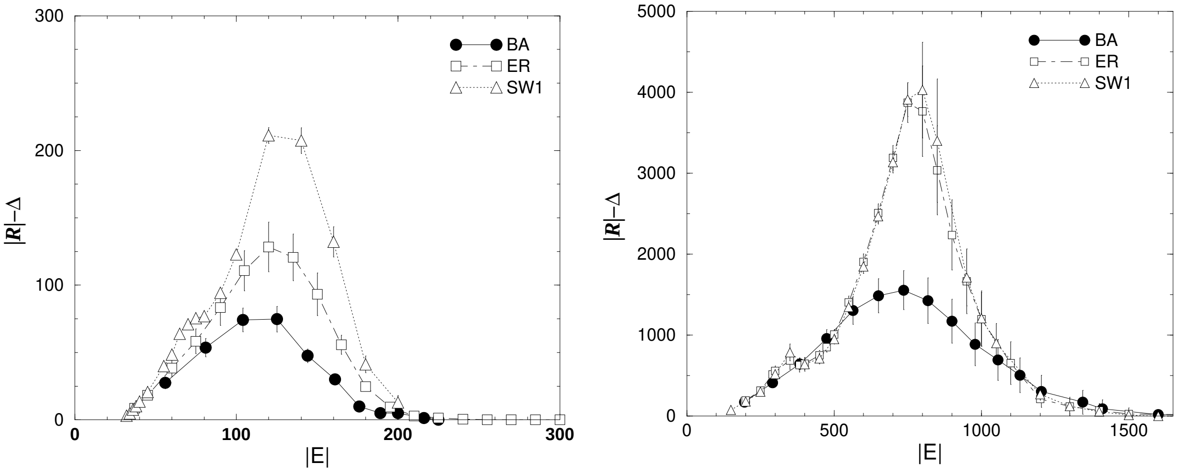


Figure 8.12. Relevant non-triangles in ER (□), SW1 (△), and BA (•) random graphs with $n = 30$ (l.h.s) and $n = 100$ (r.h.s).

The cyclomatic number of a BA random graph is $\nu(G) \sim (m/2 - 1)n$; Hence, asymptotically, almost all relevant cycles must be long for $\beta < 1$, i.e., $\gamma > 2$. The l.h.s.

of Fig. 8.11 shows that the average length of a relevant cycle grows logarithmically with n in the BA model. Not surprisingly, the slopes decrease with m.

8.3.2 A Metabolic Network

Metabolic networks form a particular class of chemical reaction networks which is distinguished by the fact that all reactions are associated with specific enzymes that catalyze the reaction.

For our analysis of metabolic graphs, we use the substrate graph of the `Ecoli1` core metabolism, a set of chemical reactions representing the central routes of energy metabolism and small-molecule building block synthesis. Similar to [154], we omit the following substrates from the graph: CO_2, NH_3, SO_4, AMP, ADP, and ATP, their deoxy-derivatives, both the oxidized and reduced form of thioredoxine, organic phosphate and pyrophosphate. The resulting graph has $n = |V| = 272$ vertices and $|E| = 652$ edges. Its analysis is summarized in Table 8.6.

Recent results by Barabasi et al. [82] show that the degree distribution of a variety of metabolic networks follows a power law with scaling exponent $\gamma \approx 2.2$. Note that these author did not use the substrate graph Σ. Instead, they used the digraph representation of the reaction network, discussed in Section 1.2.3 and Figure 1.6, whose vertices are the substrates, the reactions, and the enzymes catalyzing the reaction. The numerical values of γ are not necessarily comparable between different graphical representations of reaction network.

The extended BA model [1], which is based on both growth and partial re-wiring of the networks can explain scaling exponents γ between 2 and 3. The discussion in [46, 154] shows that a sequentially growing metabolic network is consistent with data because the evolutionary oldest metabolites have the largest vertex degrees.

The longest relevant cycles in a metabolic network are of particular interest since they reflect parts of the network that cannot easily be replaced by alternative routes. In Fig. 8.13 we show the largest such cycle in `Ecoli1`. We emphasize that the cycles in our analysis represent routes for transmission of perturbations, but not necessarily of mass, as it is commonly considered in MFA. This is apparent from Fig.8.13, which does not correspond to a pathway from a biochemical chart, but links several pathways together. Note that the notion of a pathway requires from the outset the distinction between "substrates" or products and intermediates; for our purposes such a distinction is not necessary.

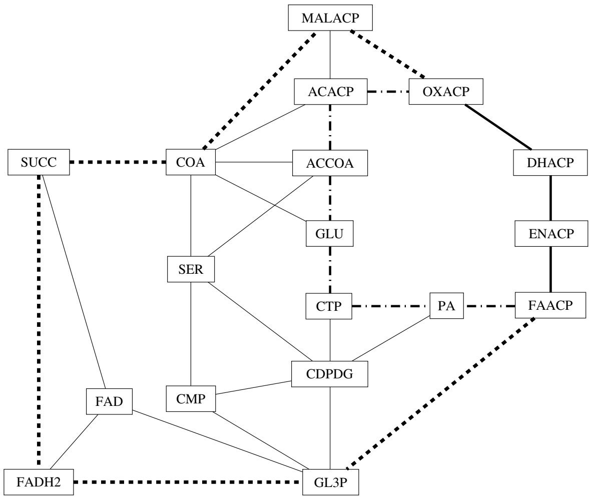


Figure 8.13. The subgraph of *E. coli* spanned by the relevant cycles of length 9. Two of these long cycles are highlighted. The edges shown in bold are part of each of the 16 relevant 9-cycles.

8.3.3 Planetary Atmospheres

It seems interesting to compare metabolic networks to reaction networks that are not governed by the enzymatic reactions. A class of large and well-understood models are the chemical networks of planetary atmospheres. The data reported here are taken from the book [167]. For details on these reaction networks we refer to [167] and the references therein.

The largest network included in this study is a model of Earth's atmosphere which contains a large number of reactions involving halogen species including the CFCs implicated in global warming.

The atmospheres of the Jovian planets Jupiter, Saturn, Uranus, and Neptune are dominantly reducing. The thermodynamically stable form of carbon in the giant planets is methane CH_4. The photolysis of CH_4 leads to the production of higher hydrocarbons, some of which have been detected Earth-based or space-craft observations. The network of the most important reactions inter-converting carbon species is denoted HC in Table 8.6 below.

Smaller networks model the atmospheres of the planets Mars and Venus, the Jovian satellite Io and the Saturn satellite Titan. The bulk of the atmospheres of both Mars and Venus is CO_2. While a pure CO_2 atmosphere should contain sizeable amounts of CO and O_2 small amounts of H_2O stabilize CO_2 through a network of reactions involving $\cdot OH$ radicals. In addition, both atmospheres contain N_2 and exhibit the associated chemistry of nitrogen oxides. Venus furthermore exhibits an interesting sulfur chemistry. Io's thin atmosphere is dominated by the photo-chemistry of SO_2. Titan possesses a mildly reducing atmosphere exhibiting a rich hydrocarbon and nitrogen chemistry with HCN as a core species.

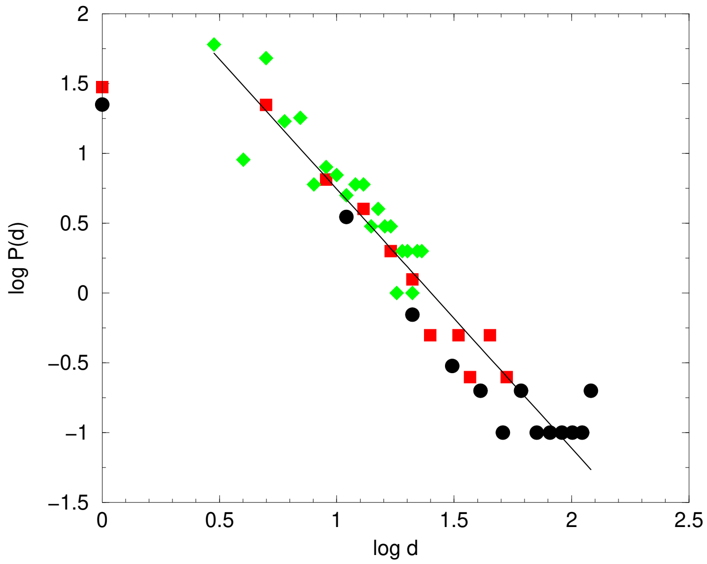


Figure 8.14. Degree distribution of the atmospheric reaction network of Earth. The symbols correspond to three different bin widths of the histogram for $P(d)$. The data are consistent with a power law with $\gamma = 1.86 \pm 0.09$ (full line).

Fig. 8.14 shows that the atmosphere chemistry models also appear to have a scale free degree distribution with a scaling exponent $\gamma \approx 1.9$. This is surprising since these reaction networks could not have arisen by a stepwise mechanism. A possible explanation is a strong bias in the choice of chemical species and reaction pathways: the network models have been constructed to describe inter-conversion of a relative small number of dominating (or interesting) species, which naturally favors a "hub and spine" arrangement.

8.3.4 Comparison of the Chemical Networks

Table 8.6 shows that the three random models BA, SW1, and ER agree at least qualitatively with each other. The BA random graphs exhibit a much broader distribution of cycle sizes (not shown) than the ER and SW1 models. As a consequence, the average cycle numbers for ER and SW1 have statistical uncertainty of about 2%, while the uncertainty of the BA values is 5 to 10 times higher. Note that ER and SW1 have a similar number of relevant cycles, but the cycles are slightly longer in SW1.

The substrate graphs of the planetary atmosphere models have a much larger average vertex degree. This accounts for the increased number of triangles and the lack of long relevant cycles.

Two features distinguish the metabolic network `Ecoli1` from all three random network models:

(1) The number Δ of triangles is almost 10 times larger than expected. This can be explained by two effects. In part this might be an artifact of the substrate graph representation. The ratio $282/379 \approx 0.744$ indicates that almost all triangles are contained in 4-cliques, since in each 4-clique we have three triangles that belong to a particular MCB, while the fourth face of the tetrahedron is their $\oplus$-sum [54]. More importantly, however, the discussion in section 8.3.1 leads us to expect an increased number of triangles in scale free networks with small scaling exponent $\gamma < 3$, as is the case in metabolic networks [154, 82]. A quantitative comparison between metabolic networks and the extended scale-free model [1] does not appear to be useful since the rewiring mechanism of the extended BA model is too artificial to apply to metabolic networks.

(2) There is a much smaller number of relevant pentagons and hexagons, which results in an overall somewhat reduced number of relevant cycles: 723 compared to about 1060 (BA), 904 (ER), and 805 (SW1). This is most likely again a consequence of the small value of the scaling exponent γ.

The atmosphere chemistry networks have a significantly larger average vertex degree. This explains the fact that almost all relevant cycles are triangles.

The vertices with the largest degree d in the raw data of many of the above networks are in some cases exceptional. In metabolic networks, for instance, ATP is involved as "universal energy currency". Many of the reactions in planetary atmosphere involve a background gas atom as a means to removing excess energy from a reaction or photons $\mathsf{h}\nu$. Following [154] we argue that one should consider the network topology without these "special purpose" vertices. Almost all relevant cycles involving these exceptional species are triangles. We remark that their inclusion does not lead to qualitative changes of either the degree distribution or the distribution of relevant cycles apart from the obvious increase in the total number of cycles.

Table 8.6. Cycle Structure of Networks.

Model	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert C\rvert$	3	4	5	6	7	8	9	$\sum$
BA	272	625	MCB	78	158	124	20	0.4	0.01	0	380
			$\mathcal{R}$	81	285	527	161	5.5	0.4	0	1060
			$\mathcal{S}$	81	273	414	144	5.5	0.4	0	918
ER	272	625	MCB	18	58	163	131	11	0.4	0	381
			$\mathcal{R}$	18	61	212	528	82	3.2	0	904
			$\mathcal{S}$	18	61	205	311	68	3.2	0	666
SW1	272	625	MCB	15	46	131	167	21	1.1	0.03	381
			$\mathcal{R}$	15	48	157	427	151	7.1	0.2	805
			$\mathcal{S}$	15	48	155	301	108	6.5	0.2	634
`Ecoli1`	272	652	MCB	282	51	19	20	3	5	1	381
			$\mathcal{R}$	379	114	90	83	5	36	16	723
			$\mathcal{S}$	379	56	24	42	2	14	16	533
Earth	250	780	MCB	431	98	2	0	0	0	0	531
			$\mathcal{R}$	918	332	6	0	0	0	0	1256
			$\mathcal{S}$	918	303	6	0	0	0	0	1227
Titan	56	305	MCB	243	2	0	0	0	0	0	245
			$\mathcal{R}$	646	52	0	0	0	0	0	697
			$\mathcal{S}$	646	0	0	0	0	0	0	646
Venus	43	207	MCB	159	3	0	0	0	0	0	162
			$\mathcal{R}$	438	10	0	0	0	0	0	448
			$\mathcal{S}$	438	10	0	0	0	0	0	448
Mars	32	167	MCB	130	0	0	0	0	0	0	130
			$\mathcal{R}$	342	0	0	0	0	0	0	342
			$\mathcal{S}$	342	0	0	0	0	0	0	342
HC	40	299	MCB	260	0	0	0	0	0	0	260
			$\mathcal{R}$	1017	0	0	0	0	0	0	1017
			$\mathcal{S}$	1017	0	0	0	0	0	0	1017

Chapter 9

Conclusion and Outlook

The perception of cyclic structures is a crucial step in the analysis of graphs. The smallest canonical set of cycles which describes the cyclic structure of a graph is the *union of all the minimum cycle bases*, called the *set of relevant cycles* $\mathcal{R}$. A cycle is if it is not the sum of shorter cycles. The set of relevant cycles is of particular importance for "ring perception" in computational chemistry. In chapter 4 we have reviewed this application, clarifying a number of inconsistencies in the literature. The concept of relevant cycles is then extended to a new vector space $\mathfrak{U}^*$, generated by the incidence vectors of the uv-paths — paths connecting two vertices u and $v \in U$, where U is a subset of the vertex set. This construction is of interest in the context of chemical networks, where a subset U of all chemical species V is fed into the system from the outside or is harvested from the system. The uv-paths hence correspond to productive pathways. We gave a polynomial time algorithm to compute *the set of relevant cycles and U-paths*. A partitioning of $\mathcal{R}$ has been described such that each cycle in a class $\mathcal{W}$ can be expressed as a sum of other cycles in $\mathcal{W}$ and shorter cycles. It is shown that each minimum cycle basis contains the same number of representatives of a given class $\mathcal{W}$. We extended this partitioning to the U-space and gave a polynomial-time algorithm to compute this partition.

A well known theorem in network flow theory states that any strongly connected digraph has a directed circuit basis, i.e., a basis of the cycle space consisting of circuits. We generalized the idea of relevant cycles to relevant circuits and show that a minimum circuit basis can be computed in polynomial time. Even though all relevant circuits are short — a generalization of the analogous result for undirected graphs — it is not possible to generalize either Horton's algorithm for to minimum circuit bases or Vismara's prototypes of relevant cycles to the directed case. Both procedures work in the undirected case since one can show that, during building a initial set of cycles, *any* fixed choice of a — usually not unique — shortest path can be used. The ideas behind this proof, however, do not work with directed paths. The efficient calculation of relevant circuits thus remains an open problem. Although we found a not very elegant way to calculate a minimum circuit basis, we have not implemented this algorithm.

We showed that some definitions commonly used for the chemical problem of ring perception (ESSR and ESER) yield neither a superset nor a subset of the minimum cycle basis. This problem of defining extended ring-sets will more appealing lies beyond the scope of this work, however.

Since most of the chemical graphs are planar, it seem to be a good idea to implement the faster Hartvigsen's algorithm [67] for planar graphs. This would speed up the time complexity from $\mathcal{O}(|V|^4)$ to $\mathcal{O}(|V|^2 \log |V|)$ (Hartvigsen). Furthermore, there exists faster algorithms for other graph classes with special structure.

The partitioning of the union of all minimum cycles bases into interchangeability classes appears to be a suitable starting point for the RNA prediction problem of secondary structures in the presence of pseudo-knots. Each structural element is represented through one class. Not being restricted to a special type of pseudo-knots, we can define for each single structural element an energy parameter, by taking the average over the contributions of each representatives of the class associated with this element. We found a suitable framework for a generalization of the standard energy model for nucleic acid structures. The actual parameters could be estimated from known structures, since datas from thermodynamical experiments are not available at present (with very few exceptions [61].

The direct comparison of minimum U-bases with biological relevant metabolic pathways as obtained e.g. from MFA is not reasonable, since in chemical reaction mechanisms are in the latter case described by hypergraph [169].

The cycles are one part of the perception of the cyclic structure of a graph. The mutual arrangement of the cycles is also of great interest. There are three different definitions of cycle graphs:

Definition 107. *A cycle graph $\mathcal{C}_{\mathcal{E}}(\mathcal{G}), \mathcal{C}_{\mathcal{S}}(\mathcal{G}), \mathcal{C}_{\mathcal{R}}(\mathcal{G})$ of a graph $\mathcal{G}$ is the graph whose vertex set is the set*

1. *of all elementary cycles of $\mathcal{G}$*
2. *of all chordless cycles of $\mathcal{G}$*
3. *of all relevant cycles of $\mathcal{G}$*

and two vertices in $\mathcal{C}_*(\mathcal{G})$ are adjacent whenever the corresponding cycles have at least one edge in common.

Of source (3) of definition 107 is a subset of (2) which is again a subset of (1). The particular interest lies on the cycle graph of type (3) and its correlation with the interchangeability classes of relevant cycles. We know that the relevant cycles of the same Vismara cycle family belongs to a connected subgraph of the cycle graph $\mathcal{C}_{\mathcal{R}}(\mathcal{G})$, since from their definition refdef:vismara they have at least one edge in common. This does not hold true for the cycles of same interchangeability class (see Fig. 9.1), also not for homotopic equal-length cycles and strong-equivalent cycles.

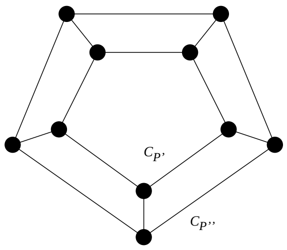


Figure 9.1. The two pentagons $C_{P'}$ and $C_{P''}$ of the upper graph belongs to the same $\leftrightarrow$-class, but do not have any edges in common. Thus the cycle graph of this $\leftrightarrow$-class consists of two isolated vertices.

Our whole mathematical construct is based on the cycle space, but certainly it can be applied on other vector spaces of the graph, e.g., the cut space. Maybe the minimum bases and relevant elements can not be calculated in polynomial time. In the case of the vector space, there are restrictive conditions for the initial set, from which the relevant elements are extracted by the greed algorithm. One have also to find such constrictive initial conditions.

Bibliography

[1] R. Albert and A.-L. Barabási. Topology of evolving networks: local events and universality. *Phys. Rev. Lett.*, 25:5234–5237, 2000.

[2] L. A. N. Amaral, A. Scala, M. Barthélémy, , and H. E. Stanley. Classes of small world networks. *Proc. Natl. Acad. Sci. USA*, 97:11149–11152, 2000.

[3] A. T. Balaban, D. Fărcaşiu, and R. Bănică. Graphs of multiple 1,2-shifts in carbonium ions and related systems. *Rev. Roum. Chem.*, 11:1205–1227, 1966.

[4] A. T. Balaban, P. Filip, and T. S. Balaban. Computer program for finding all possible cycles in graphs. *J. Comput. Chem.*, 6:316–329, 1985.

[5] R. Balducci. *Symbolic Structural Modeling*. PhD thesis, University of Texas at Austin, USA, Aug 1992.

[6] R. Balducci and R. S. Pearlman. Efficient exact solution of the ring perception problem. *J. Chem. Inf. Comput. Sci.*, 34:822–831, 1994.

[7] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[8] A.-L. Barabási, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272:173–187, 1999.

[9] A. Barrat and M. Weigt. On the properties of small-world network models. *Europ. Phys. J. B*, 13:547–566, 2000.

[10] C. Berge. *Graphs*. North-Holland, Amsterdam, NL, 1985.

[11] F. Bernhart and P. C. Kainen. The book thickeness of a graph. *J. Comb. Theor. B*, 27:320–331, 1979.

[12] N. L. Biggs, E. K. Lloyd, and R. J. Wilson. *Graph Theory 1736 - 1936*. Clarendon Press, Oxford, UK, 1976.

[13] B. Bollobás. *Random Graphs*. Academic Press, London UK, 1985.

[14] B. Bollobás. *Modern Graph Theory*. Springer, New York, 1998.

[15] J. A. Bondy. Trigraphs. *Discr. Math.*, 75:69–79, 1989.

[16] I. Brierley, N. J. Rolley, A. J. Jenner, and S. C. Inglis. Mutational analysis of the RNA pseudoknot component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol*, 229:889–902, 1991.

[17] J. W. Brown. Structure and evolution of ribonuclease P RNA. *Biochemie*, 73:689–697, 1991.

[18] J. Cai. Counting embeddings of planar graphs using DFS trees. *SIAM J. Discrete Math.*, 6:335–352, 1993.

[19] A. Cayley. On the analytic forms called trees, with applications to the theory of chemical combinations. *Rept. Brit. Assoc. Adv. Sci.*, 45:257–305, 1875.

[20] M. Chamorro, N. Parkin, and H. E. Varmus. An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. *Proc. Natl. Acad. Sci. USA*, 89:713–717, 1992.

[21] C. Champetier. On the null-homotopy of graphs. *Discr. Math.*, 64:97–98, 1987.

[22] C.-T. Chen, P. Gantzel, J. S. Siegel, K. K. Baldrigde, R. B. English, and D. M. Ho. Synthese und Struktur eines nanodimensionierten Multicyclophans: das "Kuratowski-Cyclophan" - ein Beispel für ein achirales, nichtplanares $k_{3,3}$-Stereoelement. *Angew. Chem.*, 107:2870–2873, 1995.

[23] N. Chiba, T. Nishizeki, A. Abe, and T. Ozawa. A linear algorithm for embedding planar graphs using *pq*-trees. *J. Comput. Sys. Sci.*, 30:54–76, 1985.

[24] D. M. Chickering, D. Geiger, and D. Heckerman. On finding a cycle basis with a shortest maximal cycle. *Inform. Processing Let.*, 54:55–58, 1994.

[25] L. O. Chua and L. Chen. On optimally sparse cycle and coboundary basis for a linear graph. *IEEE Trans. Circuit Theory*, 20:54–76, 1973.

[26] B. L. Clarke. Stoichiometric network analysis. *Cell Biophys.*, 12:237–253, 1988.

[27] E. J. Corey and G. A. Petersson. An algorithm for machine perception of synthetically significant rings in complex cyclic organic structures. *J. Am. Chem. Soc.*, 94:460–465, 1972.

[28] E. J. Corey and W. T. Wipke. Computer-assisted disgn of complex organic syntheses. *Science*, 166:178–192, 1969.

[29] N. Deo. *Graph Theory with Applications to Engineering and Computer Science*, pages 1–148. Prentice-Hall Series in Automatic Computation. Prentice-Hall, Inc., Englewood Cliffs, New York, 1974.

[30] N. Deo, G. M. Prabhu, and M. S. Krishnamoorty. Algorithms for generating fundamental cycles in a graph. *ACM Trans. Math. Software*, 8:26–42, 1982.

[31] J. D. Dinman, T. Icho, and R. B. Wickner. A-1 ribosomal frameshifting in a double-stranded RNA virus of yeast forms a gag-pol fusion protein. *Proc. Natl. Acad. Sci. USA*, 88:174–178, 1991.

[32] G. A. Dirac. On rigid circuit graphs. *Abh. Math. Sem. Hamburg*, 25:71–76, 1961.

[33] E. T. Dixon and S. E. Goodman. An algorithm for the longest cycle problem. *Networks*, 6:139–149, 1976.

[34] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, 85:4633–4636, 2000.

[35] G. M. Down. Ring perception. In P. von R.S̃chlezer, N. L. Allinger, J. G. T.C̃lark, P. A. Kollman, H. F. Schaefer, and P. R. Schreiner, editors, *Encyclopedia of Computational Chemistry*, volume 4, pages 2509–2515. John Wilez & Sons Ltd, Chichester,UK, 1998.

[36] G. M. Downs, V. J. Gillet, J. D. Holliday, and M. F. Lynch. Computer storage and retrieval of generic chemical structures in patents. 9. an algorithm to find the extended set of smallest rings in structurally explicit generics. *J. Chem. Inf. Comput. Sci.*, 29:207–214, 1989.

[37] G. M. Downs, V. J. Gillet, J. D. Holliday, and M. F. Lynch. Review of ring perception algorithms for chemical graphs. *J. Chem. Inf. Comput. Sci.*, 29:172–187, 1989.

[38] G. M. Downs, V. J. Gillet, J. D. Holliday, and M. F. Lynch. Theoretical aspects of ring perception and development of the extended set of smallest rings concept. *J. Chem. Inf. Comput. Sci.*, 29:187–206, 1989.

[39] P. Duchet, M. Las Vergnas, and H. Meyniel. Connected cutsets of a graph and a triangle basis of the cycle space. *Discr. Math.*, 62:145–154, 1986.

[40] J. Edmonds. Matroids and the greedy algorithm. *Math. Progam.*, 1:127–136, 1971.

[41] J. D. Edwards and B. Ø. Palsson. The *escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA*, 97:5528–5533, 2000.

[42] S. B. Elk. Derivation of the principle of smallest set of smallest rings from Euler's polyhedron equation and a simplified technique for finding this set. *J. Chem. Inf. Comput. Sci.*, 24, 1984.

[43] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci., Ser. A*, 5:17–61, 1960.

[44] L. Euler. Solutio problematis ad geometriam situs pertinentis. *Comm. Acad. Sci. Imp. Petropol.*, 8:128–140, 1736. (Latin).

[45] D. A. Fell. *Understanding the Control of Metabolism.* Portland Press, London, 1997.

[46] D. A. Fell and A. Wagner. Structural properties of metabolic networks: implications for evolution and modelling of metabolism. In J.-H. S. Hofmeyr, J. M. Rohwer, and J. L. Snoep, editors, *Animating the cellular map*, pages 79–85, Stellenbosch, ZA, 2000. Stellenbosch University Press.

[47] J. Figueras. Ring perception using breadth-first search. *J. Chem. Inf. Comput. Sci.*, 36:986–991, 1996.

[48] A. C. Forster and S. Altman. Similar cage-shaped structures for the RNA component of all ribonuclease P and ribonuclease MRP enzymes. *Cell*, 62:407–409, 1990.

[49] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. (USA)*, 83:9373–9377, 1986.

[50] S. Fujita. Logical perception of ring-opening, ring-closure, and rearrangement reactions based on imaginary transition structures. selection of the essential set of essential rings (ESER). *J. Chem. Inf. Comput. Sci*, 28:1–9, 198.

[51] S. Fujita. A new algorithm for selection of synthetically important rings. the essential set of essential rings for organic structures. *J. Chem. Inf. Comput. Sci*, 28:78–82, 198.

[52] J. Gasteiger and C. Jochum. An algorithm for the perception of synthetically important rings. *J. Chem. Inf. Comput. Sci*, 19:43–48, 1979.

[53] N. E. Gibbs. A cycle generation algorithm for finite undirected linear graphs. *J. Assoc. Comput. Mach.*, 16:564–568, 1969.

[54] P. M. Gleiss, J. Leydold, and P. F. Stadler. Interchangeability of relevant cycles in graphs. *Elec. J. Comb.*, 7:R16 [16pages], 2000. See `http://www.combinatorics.org`.

[55] P. M. Gleiss and P. F. Stadler. Relevant cycles in biopolymers and random graphs. Presented at the Fourth Slovene International Conference in Graph Theory, Bled, Slovenia (Santa Fe Institute Preprint 99-07-042, `http://www.santafe.edu`), 1999.

[56] P. M. Gleiss, P. F. Stadler, A. Wagner, and D. A. Fell. Relevant cycles in chemical reaction network. *Adv. Cmplx. Syst.*, 4:0–0, 2001. accepted.

[57] T. C. Gluick and D. E. Draper. Thermodynamics of folding a pseudoknotted mRNA fragment. *J. Mol. Biol.*, 241:246–262, 1994.

[58] R. Gould. The application of graph theory to the synthesis of contact networks. In *Proceedings of an International Symposium on the Theory of Switching (April 2-5, 1957)*, volume 29 of *The Annals of the Computation Laboratory of Harvard University*, pages 244–292, Cambridge, UK, 1957. Harvard University Press.

[59] R. Gould. Graphs and vector spaces. *J. Math. and Phys.*, 37:193–214, 1958.

[60] N. A. B. Gray. *Computer-Assisted Structure Elucidation*, chapter IX.E, pages 312–324. John Wiley & Sons, New York, 1986.

[61] A. P. Gultyaev, F. H. van Batenburg, and C. W. Pleij. An approximation of loop free energy values of RNA H-pseudoknots. *RNA*, 5:609–617, 1999.

[62] E. S. Haas, D. P. Morse, J. W. Brown, J. F. Schmidt, and N. R. Pace. Long-range structure in ribonuclease P RNA. *Science*, 254:853–856, 1991.

[63] P. R. Halmos. *Finite-Dimensional Vector Spaces*. Van Nostrand Reinhold, New York, 1958.

[64] T. Hanser, P. Jauffret, and G. Kaufmann. A new algorithm for exhaustive ring perception in a molecular graph. *J. Chem. Inf. Comput. Sci.*, 36:1146–1152, 1996.

[65] M. Hartmann, H. Schneider, and M. H. Schneider. Integral bases and p-twisted digraphs. *Europ. J. Combinatorics*, 16:357–369, 1995.

[66] D. Hartvigsen. Minimum path bases. *J. Algorithms*, 15:125–142, 1993.

[67] D. Hartvigsen and R. Mardon. Cycle bases from orderings and coverings. *Discr. Math.*, 94:81–94, 1991.

[68] D. Hartvigsen and R. Mardon. The prism-free planar graphs and their cycle bases. *J. Graph Theory*, 15:431–441, 1991.

[69] D. Hartvigsen and R. Mardon. When do short cycles generate the cycle space. *J. Comb. Theory, Ser. B*, 57:88–99, 1993.

[70] D. Hartvigsen and R. Mardon. The all-pair min-cut problem and the minimum cycle basis problem on planar graphs. *SIAM J. Discr. Math.*, 7:403–418, 1994.

[71] D. Hartvigsen and E. Zemel. Is every cycle basis fundamental? *J. Graph Theory*, 13:117–137, 1989.

[72] R. Heinrich and S. Schuster. *The Regulation of Cellular Systems*. Chapman & Hall, New York, 1996.

[73] J. B. Hendrickson, D. L. Grier, and A. G. Toczko. Condensed structure identification and ring perception. *J. Chem. Inf. Comput. Sci.*, 24:195–203, 1984.

[74] H. Herzel. How to quantify "small world networks"? *Fractals*, 6:301–303, 1998.

[75] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:165–188, 1994.

[76] F. E. Hohn. *Elementary Matrix Algebra*. MacMillan, New York, 1958.

[77] J. D. Horton. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM J. Comput.*, 16:359–366, 1987.

[78] E. Hubicka and M. M. Sysło. Minimal bases of cycles of a graph. In M. Fiedler, editor, *Recent Advances in Graph Theory, Proc. 2nd Czechoslovak Symp. on Graph Theory*, pages 283–293. Academia, 1975.

[79] T. Ito and M. Kizawa. The matrix rearrangement procedure for graph-theoretical algorithms and its application to the generation of fundamental cycles. *ACM Trans. Math. Software*, 3:227–231, 1977.

[80] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA*, 86:7706–7710, 1989.

[81] R. E. Jamison. On the null-homotopy of bridged graphs. *Europ. J. Comb.*, 8:421–428, 1987.

[82] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.

[83] D. B. Johnson. Finding all the elementary circuits of a directed graph. *SIAM J. Comput.*, 4:77–84, 1975.

[84] S. Kammermeier, H. Neumann, F. Hampel, and R. Herges. Diels-Alder reactions of tetradehydrodianthracene with electron-rich dienes. *Liebigs Ann.*, 1996:1795–1800, 1996.

[85] I. Kant. *Metaphysische Anfangsgründe der Naturwissenschaft*. Hartknock Verlag, Riga, 1786.

[86] K. C. Keiler, P. R. Waller, and R. T. Sauer. Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science*, 271:990–993, 1996.

[87] G. Kirchhoff. Über die Auflösung der Gleichungen, auf welche man bei der Untersuchungen der linearen Verteilung galvanischer Ströme geführt wird. *Poggendorf Ann. Phys. Chem.*, 72:497–508, 1847.

[88] D. A. M. Konings and R. R. Gutell. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA*, 1:559–574, 1995.

[89] J. B. Kruskal. On the shortest spanning subgraph of a graph and the travelling salesman problem. *Proc. Amer. Math. Soc.*, 7:48–49, 1956.

[90] D. Kuck and A. Schuster. Centrohexaindan, der erste Kohlenwasserstoff mit topologisch nicht-planarer Molekülstruktur. *Angew. Chem.*, 100:1222–1224, 1988.

[91] C. Kuratowski. Sur le problème des courbes gauches en topolgie. *Fund. Math.*, 15:271–283, 1930.

[92] A. Lehman. A solution of the shannon switching game. *J. Soc. Indust. Appl. Math.*, 12:687–725, 1964.

[93] J. Leydold and P. F. Stadler. Minimal cycle basis of outerplanar graphs. *Elec. J. Comb.*, 5:R16, 1998. See `http://www.combinatorics.org`.

[94] A. Loria and T. Pan. Domain structure of the ribozyme from eubacterial ribonuclease P. *RNA*, 2:551–563, 1996.

[95] S. MacLane and G. Birkhoff. *Algebra.* MacMillan, New York, 1967.

[96] R. Mans, C. Pleij, and L. Bosch. Transfer RNA-like structures: Structure, function and evolutionary significance. *Eur. J. Biochem.*, 201:303–324, 1991.

[97] D. Marcu. On finding the elementary paths and circuits of a digraph. *Polytech. Univ. Bucharest Sci. Bull. Ser. D Mech. Engrg.*, 55:29–33, 1993.

[98] P. Mateti and N. Deo. On algorithms for enumerating all circuits of a graph. *SIAM J. Comput.*, 5:90–99, 1976.

[99] D. Mathews, J. Sabina, M. Zucker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[100] T. A. McKee. Induced cycle structure and outerplanarity. *Discr. Math.*, 223:387–392, 2000.

[101] F. Michel and E.Westhof. Modelling of the three-dimmensional architecture of group I catalytic introns based on comparative sequence anaysis. *J. Mol. Biol.*, 216:585–610, 1990.

[102] G. J. Minty. On the axiomatic foundations of the theories of directed linear graphs, electrical networks and network programming. *Journ. Math. Mech.*, 15:485–520, 1966.

[103] J. M. Montoya and R. V. Solé. Small world patterns in food webs. Technical Report 00-10-059, Santa Fe Institute, 2000.

[104] P. Mutzel and R. Weiskircher. Optimizing over all combinatorial embeddings of a planar graph. In G. Cornuejols, R. E. Burkard, and G. J. Wöginger, editors, *Integer Programming and Combinatorial Optimization — 7th International IPCO Conference, Graz, Austria, June 9-11*, volume 1610 of *Lect. Notes Comput. Sci.*, pages 361–376, Berlin, 1999. Springer.

[105] M. E. J. Newman, C. Moore, and D. J. Watts. Mean-field solution of the small-world network model. *Phys. Rev. Lett.*, 84:3201–3204, 2000.

[106] M. E. J. Newman and D. J. Watts. Renormalization group analysis of the small-world network model. *Phys. Lett. A*, 263:341–346, 1999.

[107] H. Nickelsen. Ringbegriffe in der Chemie-Dokumentation. *Nachr. Dok.*, 3:121–123, 1971. (and associated microfiche).

[108] J. Nowkowski and J. I. Tinoco. RNA structure and stability. *Seminars in Virology*, 8:153–165, 1997.

[109] S. A. Pandit and R. E. Amritkar. Characterization and control of small-world networks. *Phys. Rev. E*, 60:R1119–R1122, 1999. chao-dyn/9901017.

[110] A. M. Patterson, L. T. Capell, and D. F. Walker. *The Ring Index.* American Chemical Society, Washington, D. C., 2 edition, 1960.

[111] M. Petitjean, B. T. Fan, A. Panaye, and J.-P. Doucet. Ring perception: Proof of a formula calculating the number of the smallest rings in connected graphs. *J. Chem. Inf. Comput. Sci.*, 40:1015–1017, 2000.

[112] C. W. A. Pleij. Pseudoknots: A new motif in the RNA game. *Trends Biochem. Sci.*, 15:143–147, 1990.

[113] C. W. A. Pleij. *Appendix 2:RNA Pseudoknots in the RNA World.* Cold Spring Harbor Laboratory Press, 1993.

[114] C. W. A. Pleij and L. Bosch. RNA pseudoknots: structure, detection and prediction. *Methods Enzymol.*, 180a:289–303, 1989.

[115] M. Plotkin. Mathematical basis of ring-finding algorithms in CIDS. *J. Chem. Doc.*, 11:60–63, 1971.

[116] R. Rado. Note on independence functions. *Proc. Lond. Math. Soc.*, 7:300–320, 1957.

[117] M. Raitner. Gml file format. `http://www.infosun.fmi.uni-passau.de/Graphlet/GML/`.

[118] R. T. Rockafellar. *Convex Analysis.* Princeton Univ. Press, Princeton NJ, 1970.

[119] B. L. Roos-Kozel and W. L. Jorgensen. Computer-assisted mechanistic evaluation of organic reactions. 2. perception of rings, aromaticity and tautomers. *J. Chem. Inf. Comput. Sci.*, 21:101–111, 1981.

[120] C. H. Schilling, D. Letscher, and B. Ø. Palsson. Theory for the systematic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, 203:229–248, 2000.

[121] B. Schmidt and J. A. Fleischhauer. A Fortran IV program for finding the smallest set of smallest rings of a graph. *J. Chem. Inf. Comput. Sci.*, 18:204–206, 1978.

[122] S. Schuster, D. A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnol.*, 18:326–332, 2000.

[123] A. K. Sen. Quantitative analysis of metabolic regulation. A graph-theoretic approach using spanning trees. *Biochem. J.*, 275:253–258, 1991.

[124] S. Seshu and M. B. Reed. *Linear Graphs and Electrical Networks.* Addison-Wesley, Reading,Mass., 1961.

[125] W.-K. Shiha and W.-L. Hsub. A new planarity test. *Theor. Comp. Sci.*, 223:179–191, 1999.

[126] E. G. Smith. *The Wiswessner Line-Formula Chemical Notation.* McGraw-Hill, New York, 1968.

[127] P. F. Stadler. Minimal cycle bases of Halin graphs. *J. Graph Theory*, 2000. submitted, see `http://www.tbi.univie.ac.at`.

[128] P. F. Stadler and C. Haslinger. RNA structures with pseudo-knots: Graph-theoretical and combinatorial properties. *Bull. Math. Biol.*, 61:437–467, 1999.

[129] G. F. Stepanec. Basis systems of vector cycles with extremal properties in graphs. *Uspekhi Mat. Nauk. 2*, 19:171–175, 1964. (Russian).

[130] G. M. Studnicka, G. M. Rahn, I. W. Cummings, and W. A. Salser. Computer method for predicting the secondary structure of single-stranded RNA. *Nucleic Acid Res.*, 5:3365–3387, 1978.

[131] J. J. Sylvester. On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics, with three appendices. *Amer. J. Math.*, 1:64–128, 1878.

[132] M. M. Sysło. Characterizations of outerplanar graphs. *Discrete Math.*, 26:47–53, 1979.

[133] M. M. Sysło. On cycle bases of a graph. *Networks*, 9:123–132, 1979.

[134] M. M. Sysło. An efficient cycle vector space algorithm for listing all cycles of a planar graph. *SIAM J. Comput.*, 10:797–808, 1981.

[135] Y. Takahashi. Automatic extraction of ring substructures from a chemical structure. *J. Chem. Inf. Comput. Sci.*, 34:167–170, 1994.

[136] C. K. Tang and D. E. Draper. An unusual mRNA pseudoknot structure is recognized by a protein translation repressor. *Cell*, 57:531–536, 1989.

[137] C. K. Tang and D. E. Draper. Evidence for allosteric coupling between the ribosome and repressor binding sites of a translationally regulated mRNA. *Biochemistry*, 29:4434–4439, 1990.

[138] O. N. Temkin, A. V. Zeigarnik, and D. Bonchev. *Chemical Reaction Networks: A Graph-Theoretical Approach.* CRC Press, Bocal Raton, FL, 1996.

[139] E. B. Ten Dam, C. W. A. Pleij, and L. Bosch. RNA pseudoknots and translational frameshifting on retroviral, coronaviral and luteoviral RNAs. *Virus Genes*, 4:121–136, 1990.

[140] C. Thomassen. Embeddings and minors. In R. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics*, pages 301–349. North-Holland, Amsterdam, 1995.

[141] K. Thulasiraman and M. N. S. Swamy. *Graphs: Theory and Algorithms.* J. Wiley & Sons, New York, 1992.

[142] J. C. Tiernan. An efficient search algorithm to find the elementary circuits of a graph. *Commun. Assoc. Comput. Mach.*, 13:722–726, 1970.

[143] Trolltech. Qt free edition - version 2.3.1. `http://www.trolltech.com/`. (under GPL).

[144] W. T. Tutte. A homotopy theorem for matroids I and II. *Trans. Amer. Math. Soc.*, 88:144–174, 1958.

[145] W. T. Tutte. Lectures on matroids. *J. Res. Nat. Bur. Stand.*, 69B:1–48, 1965.

[146] T. H. Tzeng, C. L. Tu, and J. A. Bruenn. Ribosomal frameshifting requires a pseudoknot in the saccharomyces cerevisiae double-stranded RNA virus. *J. Virology*, 66:999–1006, 1992.

[147] M. L. Vergnas. Sur le nombre de circuits dans un graphe fortement connexe. *Cahiers C.E.R.O*, 17:261–265, 1975.

[148] P. Visamara. *Reconnaissance et représentation d'éléments stucturaux pour la description d'objects complexes. Application à l'élaboration de stratégies de synthèse en chimie organique*. PhD thesis, Université Montpellier II, France, 1995. 95-MON-2-253.

[149] P. Vismara. Union of all the minimum cycle bases of a graph. *Electr. J. Comb.*, 4:73–87, 1997. Paper No. #R9 (15 pages), see `http://www.combinatorics.org`.

[150] M. V. Vol'kenshtein. *Physics of Enzymes*. Sciences, Moscow, 1965.

[151] M. V. Vol'kenshtein and B. N. Gol'dshtein. *Dokl. Akad. Nauk SSSR*, 170:963, 1965.

[152] M. V. Vol'kenshtein and B. N. Gol'dshtein. Models of allosteric enzymes and their analysis using the graph theory method. *Biokhimiia*, 4:679–686, 1966. russian.

[153] M. V. Vol'kenshtein and B. N. Gol'dshtein. A new approach to the problems of stationary kinetics of enzymic reactions. *Biokhimiia*, 3:541–547, 1966. russian.

[154] A. Wagner and D. A. Fell. The small world inside large metabolic networks. Technical Report 00-07-041, Santa Fe Institute, 2000. See `http://www.santafe.edu`.

[155] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Müller, D. H. Mathews, and M. Zuker. Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predicions of RNA folding. *Proc. Natl. Acad. Sci. USA*, 91:9218–9222, 1994.

[156] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Adv. Math. Suppl. Studies*, 1:167–212, 1978.

[157] D. J. Watts. *Small Worlds*. Princeton University Press, Princeton NJ, 1999.

[158] D. J. Watts and H. S. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.

[159] J. T. A. Welch. A mechanical analysis of the cyclic structure of undirected linear graphs. *J. Assoc. Comput. Mach.*, 13:205–210, 1966.

[160] D. J. A. Welsh. Kruskal's theorem for matroids. *Proc. Camb. Phil. Soc.*, 64:3–4, 1968.

[161] E. Westhof and L. Jaeger. RNA pseudoknots. *Current Opinion Struct. Biol.*, 2:327–333, 1992.

[162] H. Whitney. Congruent graphs and the connectivity of graphs. *Amer. J. Math.*, 54:150–168, 1932.

[163] H. Whitney. On the abstract properties of linear dependence. *Am. J. Math.*, 57:509–533, 1935.

[164] E. P. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. *Comm. Pure Appl. Math.*, 13:1–14, 1960.

[165] N. Wills, R. F. Gesteland, and J. F. Atkins. Evidence that a downstream pseudoknot is required for translational readthrough of the moloney murine leukemia virus gag stop codon. *Proc. Natl. Acad. Sci. USA*, 88:6991–69995, 1991.

[166] W. T. Wipke and T. M. Dyott. Use of ring assemblies in a ring perception algorithm. *J. Chem. Inf. Comput. Sci.*, 15:140–147, 1974.

[167] Y. L. Yung and W. B. DeMore. *Photochemistry of Planetary Athmospheres.* Oxford University Press, New York, 1999.

[168] A. Zamora. An algorithm for finding the smallest set of smallest rings. *J. Chem. Inf. Comput. Sci.*, 16:40–43, 1976.

[169] A. V. Zeigarnik. On hypercycles and hypercircuits in hypergraphs. In P. Hansen, P. W. Fowler, and M. Zheng, editors, *Discrete Mathematical Chemistry*, volume 51 of *DIMACS series in discrete mathematics and theoretical computer science*, pages 377–383, Providence, RI, 2000. American Mathematical Society.

[170] M. Zuker. `mfold-2.0`. `pub/mfold.tar.Z (nrcbsa.bio.nrc.ca)`. (Public Domain Software).

[171] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46(4):591–621, 1984.

[172] A. A. Zykov. *Theory of Finite Graphs.* Nauka, Novosibirsk, USSR, 1969. (Russian).

Curriculum Vitae

Petra Manuela Gleiss
* 9. April 1972 in St. Pölten, Niederösterreich

09/1978 – 07/1982	Volksschule der Engl. Frl., St. Pölten
09/1982 – 06/1990	WkRg der Engl. Frl., St. Pölten
06/1990	Matura
10/1990	Beginn Studium der Biologie an der Universität Wien
10/1991 – 10/1998	Studium der Biochemie an der Universität Wien
10/1995 – 01/1997	Diplomarbeit am Institut für Biochemie und Molekulare Zellbiologie der Universität Wien, bei Prof. Dr. Helmut Ruis
10/1998	2. Diplomprüfung, Sponsion zum Mag. rer. nat.
10/1998 – 9/2001	Dissertation am Institut für Theoretische Chemie und Molekulare Strukturbiolgie der Universität Wien, bei Prof. Dr. Peter F. Stadler